BUPBM Unit 1 V3 (1).docx

Building useful Predictive Business models_BBA_3

Building useful Predictive Business models_BBA_3

ATLAS SkillTech University

Document Details

Submission ID

trn:oid:::3618:127668267

Submission Date

Feb 6, 2026, 3:15 PM GMT+5:30

Download Date

Feb 6, 2026, 3:20 PM GMT+5:30

File Name

BUPBM Unit 1 V3 (1).docx

File Size

222.7 KB

21 Pages

4,133 Words

25,843 Characters

*% detected as AI

AI detection includes the possibility of false positives. Although some text in this submission is likely AI generated, scores below the 20% threshold are not surfaced because they have a higher likelihood of false positives.

Caution: Review required.

It is essential to understand the limitations of AI detection before making decisions about a student's work. We encourage you to learn more about Turnitin's AI detection capabilities before using the tool.

Frequently Asked Questions

How should I interpret Turnitin's AI writing percentage and false positives?
The percentage shown in the AI writing report is the amount of qualifying text within the submission that Turnitin's AI writing detection model determines was either likely AI-generated text from a large-language model or likely AI-generated text that was likely revised using an AI paraphrase tool or word spinner.

False positives (incorrectly flagging human-written text as AI-generated) are a possibility in AI models.

AI detection scores under 20%, which we do not surface in new reports, have a higher likelihood of false positives. To reduce the likelihood of misinterpretation, no score or highlights are attributed and are indicated with an asterisk in the report (*%).

The AI writing percentage should not be the sole basis to determine whether misconduct has occurred. The reviewer/instructor should use the percentage as a means to start a formative conversation with their student and/or use it to examine the submitted assignment in accordance with their school's policies.

What does 'qualifying text' mean?
Our model only processes qualifying text in the form of long-form writing. Long-form writing means individual sentences contained in paragraphs that make up a longer piece of written work, such as an essay, a dissertation, or an article, etc. Qualifying text that has been determined to be likely AI-generated will be highlighted in cyan in the submission, and likely AI-generated and then likely AI-paraphrased will be highlighted purple.

Non-qualifying text, such as bullet points, annotated bibliographies, etc., will not be processed and can create disparity between the submission highlights and the percentage shown.

Unit 1 – Introduction to Data Mining (Basics)

Learning Objectives

1. Define data mining and explain its role in knowledge discovery and decision-making.

2. Differentiate between data, information, and knowledge, and understand their relevance in the data mining process.

3. Identify the key steps involved in the data mining process, including data collection, cleaning, integration, and transformation.

4. Recognize major data mining techniques such as classification, clustering, association, and regression.

5. Describe common applications of data mining across industries like business, healthcare, finance, and e-commerce.

6. Discuss challenges and issues in data mining, including data quality, scalability, privacy, and ethical considerations.

7. Understand the role of tools and technologies (such as databases and machine learning) in supporting data mining activities.

8. Develop a foundational perspective on how data mining supports business intelligence and strategic decision-making.

Content

1.0 Introductory Caselet

“Unlocking Insights at ShopEase Online”

ShopEase is a medium sized online store hosting thousands of products and has seen extremely rapid growth over the last five years. The company, with many thousands of daily visitors, accumulates reams of data — customer demographics; browsing, search and purchase histories; even customer reviews.

At first, ShopEase used this data only for rudimentary reporting like monthly sales numbers and customer satisfaction scores. But management recognised the data could be a veritable gold mine for honing business strategies. Using data mining, ShopEase was starting to find patterns such as:

• Those who bought smartphones typically purchased protective cases within two weeks.

• Some product searches (such as “budget laptop”) were very seasonal, surging during school admissions windows.

• Frequent high-value customers generally responded well to personal recommendations as opposed to generalist promotional emails.

With this knowledge in hand, ShopEase was able to refine marketing strategies, enhance inventory management and boost customer loyalty. For instance, offering accessory suggestions on the fly post a high-value purchase resulted in a substantial upsurge in sales.

The ShopEase case serves as an example of how raw data, when sufficiently 'digested' can become knowledge able to influence decisions and give competitive advantage.

Critical Thinking Question

If ShopEase would like to predict customer churn (i.e., customers who will probably stop buying from the platform) by using data mining, what type of data mining technique should it use and how can accurate and fair predictions be difficult?

1.1 Introduction to Data Mining

1.1.1 Definition and Meaning of Data Mining

Data mining is the practice of searching through large datasets to identify patterns, relationships and insights. Data mining is not just simple retrieval of the data, instead it combines advanced analytical techniques, including statistics, machine learning algorithms and artificial intelligence, with a model for analyzing patterns in large databases.

The phrase "data mining" suggests the extraction of knowledge that is not obvious from looking at the data. For instance, a sales report might say how many of an item have been

sold, data mining can tell you that people who buy laptops often buy insurance for them within a week and that bundled offer could be promoted.

Key aspects of its meaning:

• It is knowledge discovery (not data storage) driven.

• Its thrust is to discover new or concealed facts.

• It is a process (preparation, exploration, analysis, interpretation) and an infrastructure (facilities and tools).

Example Similarity: Regarding the healthcare industry, there is data mining for predicting early onset of diseases by studying patient records and identify statistical correlations between demographic group information (lifestyle and genetic factors) and a health status.

1.1.2 Evolution of Data Mining as a Discipline

The evolution of data mining has not happened abruptly, it has grown with computing, storage and analytical capabilities. Its development can be followed through several phases of history:

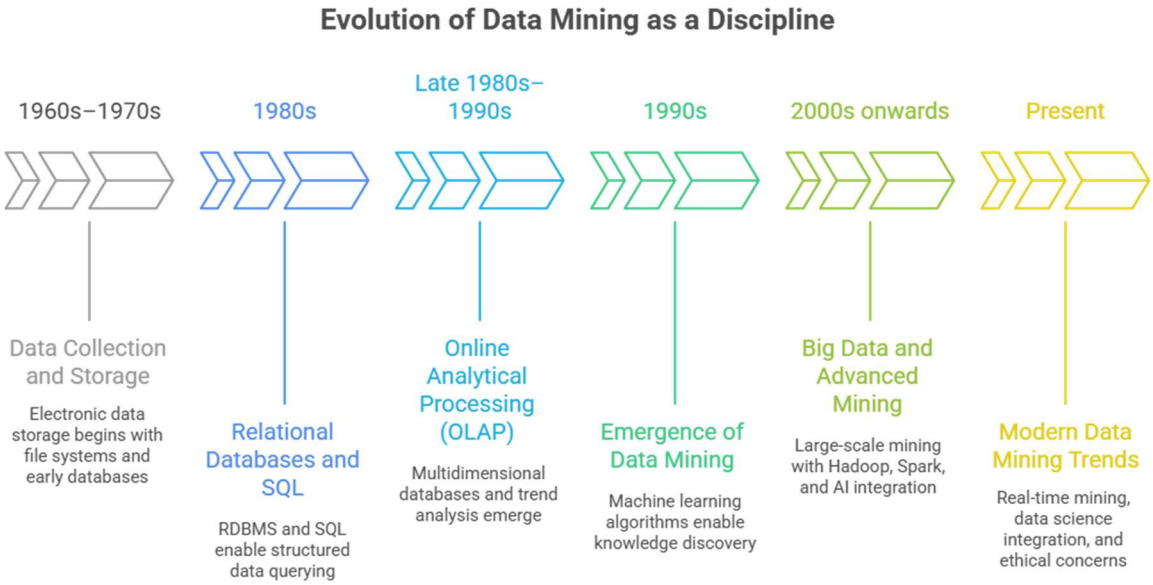

Figure: Evolution of Data Mining as a Discipline

Data Collection and Storage (1960s–1970s)

o Design: Electronic data storage.

o Technology- filesystems, hierarchical and network databases.

o Limitation: Data could be stored, but not analyzed efficiently.

Relational databases/query languages (1980's) For those familiar with and accustomed to a relational database query language, knowledge of SQL can be reused.

o Introduction of RDBMS.

o SQL made it easy to query the system.

o Example: A company would be able to type in a question like "show me all customers from Mumbai who have made a purchase during the past month."

o Limitation: Yes but only descriptive, not predictive/pattern analysis.

Online Analytical Processing (OLAP) (Late 1980s – early 1990s)

o Introduction of multidimensional databases.

o Allowed trend analysis, drill down reports and forecasting.

o Example – Business can study monthly sales trend per regions.

Emergence of Data Mining (1990s)

o ML algorithms like decision trees, clustering, association rules.

o Move from a focus on data storage' to the goal of "knowledge discovery."

o Example: Identifying products frequently purchased together in retail (market basket analysis).

Big Data and Advanced Mining (since the 2000s)

o Massive increases in internet, social media and sensor-originated data.

o Huge scale mining was made possible by tools such as Hadoop, Spark and cloud computing.

o Incorporation of artificial intelligence and deep learning in image, video and speech analysis.

Modern Data Mining Trends (Present)

o Data mining in real time from streaming data (e.g. fraud detection, banking).

o Integrating with data science and predictive analytics.

o Ethical and privacy issues, for example regarding GDPR-compliance in Europe.

Example: Today, Netflix and Amazon mine real-time data to recommend movies or products in which I might be interested based on my viewing history, my ratings and what I have purchased.

1.1.3 Key Aspects of Data Mining

Data mining integrates a number of fundamentals to allow for effective knowledge discovery, and application:

Data Cleaning and Preparation

o Temporary access can be given to some subset of the raw data.

o Preprocessing ensures high-quality results.

o Techniques include:

♣ Dealing with missing values (e.g. imputation of mean).

♣ Removing duplicates.

♣ Standardizing data (e.g., converting all currency to a common unit).

o Example: Sanitizing customer phone numbers before the analysis of contact history.

Pattern Discovery

o Main task: Discovering (non-spurious) structure in the data.

o Techniques include:

♣ Clustering: Partition into clusters similar data points (e.g., partitioning customer in buying behaviour).

♣ Classification: Labeling (e.g., labeling spam emails as such or not).

♣ Association Rules: A search for co-occurrence relationships (e.g., “If A is bought, then B is bought”).

Prediction and Forecasting

o Predicting the future via historical data.

o Examples:

♣ Predicting stock prices.

♣ Forecasting electricity demand.

♣ Predicting churn of customers in the telecom industry.

Knowledge Representation

o Representing the discovered patterns in an interpretable form.

o Tools are the dashboards, decision trees, visual graphs and summary rules.

o Example: A tree of decisions for loan approval.

Scalability and Efficiency

o Algorithms have to be scalable, as data size keeps increasing.

o New systems are based on distributed computing and parallel processing.

o Example: Google search engine for distributed mining to process billions of web pages in seconds.

Integration with Other Disciplines

o Data mining incorporates methods from:

♣ Statistics for hypothesis testing.

♣ Machine learning for adaptive models.

♣ Artificial intelligence for reasoning.

♣ Storage and retrieval database systems.

Applications in Real Life

o Banking: Fraud prevention through detection of abnormal spending behavior.

o Retail: Recommendations and market basket analyses.

o Healthcare: Disease or treatment outcome forecasting.

o Education: Early recognition of children dropping out.

o Telecommunications: Enhancing consumer satisfaction by utilising usage data for analysis.

1.2 Concepts of Data Mining

1.2.1 Data Preprocessing

Preprocessing is the key for any successful data mining. Since raw data is usually incomplete and noisy, one needs to pre-process it before being able to analyze it.

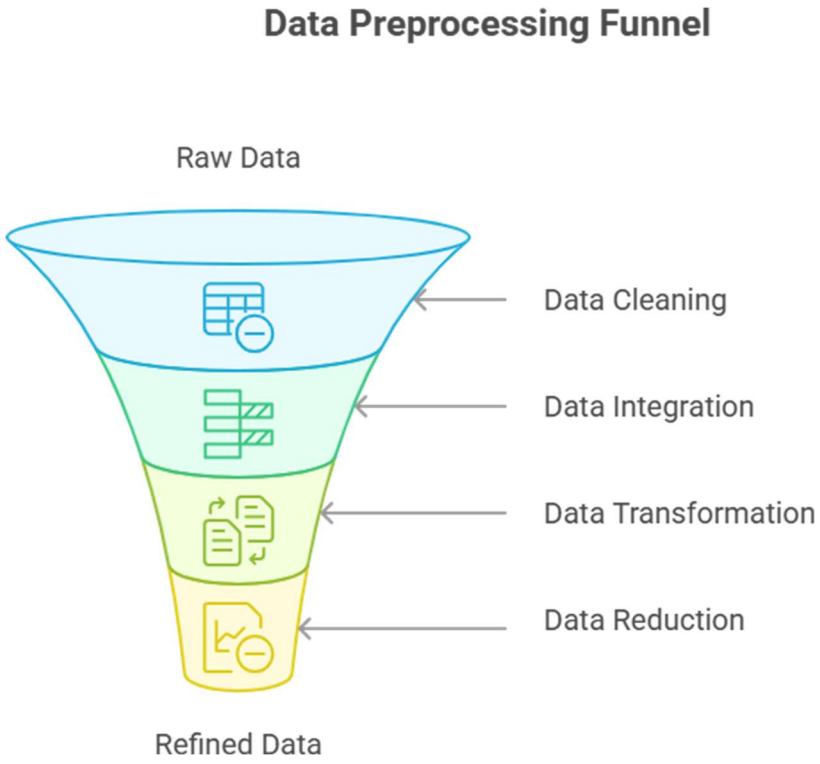

Figure: Data Preprocessing Funnel

Stages of Data Preprocessing

Data Cleaning

o Removal of duplicate records.

o Dealing with missing values (e.g. impute resolution - replace with mean/median, predictive imputation).

o Removing outliers or correcting inconsistencies.

• Example: a dataset has "Age = -5" this data has to be corrected or removed.

Data Integration

o Federating various data into integrated, unified dataset.

o Example: A bank may combine customer transaction details, loan data and credit card usage patterns.

Data Transformation

o Normalization: rescaling data to a common scale (0 to 1).

o Aggregation: Preparing data (e.g. daily sales → monthly sales).

o Converting categorical variables to binary (for ex., “Male = 0, Female = 1”).

Data Reduction

o Particularly in audio information, decreasing the volume but maintaining important content.

o Statistical feature selection with dimensionality reduction for example using Principal Component Analysis.

o Example: Reduce 100 features instead of taking all and coming out with the 10 key attributes that explain most of your variance.

Importance

- Provides assurance that results are accurate and reliable.
- Computational cost is saved by eliminating irrelevant features.
- It can be used to smooth the operation of machine learning algorithms.

Did You Know?

“Nearly 80% of the time in a data mining project is spent on data preprocessing rather than actual model building. Without proper cleaning, integration, and transformation, even the most advanced algorithms produce misleading results. High-quality preprocessing is the hidden key behind accurate predictions and reliable insights.”

1.2.2 Data Exploration

Exploratory Data Analysis (EDA), is the process through which data modelers begin to understand a dataset prior to running File-selection processes.

Techniques Used in Exploration

Descriptive Statistics

o Measures like mean, median, mode, variance, stand dev.

o Helps in summarizing data distribution.

Visualization Tools

o Histograms for frequency distribution.

o Box plots to detect outliers.

o Scatter plots for identifying correlations.

Correlation and Dependency Analysis

o Measures relationships between variables.

o Example: Evidence of a high correlation between advertisement spending and sales revenue.

Example

An e-commerce company is digging through some sales data:

• From histogram we conclude that most customers are at the age of (25–35).

• The scatter plot shows a positive relationship between time spent on website and predicted purchase probability.

Purpose

• To understand data characteristics.

• To lead selection of features which should be considered in more detailed analysis.

• Which mining technique to use (cluster ing vs. classification), to choose appropriate.

1.2.3 Association Rule Mining

Association rule mining is a technique used to find patterns of relationship or association among set of items in large databases.

Core Concepts

• Rules format: X → Y (If antecedent X, consequent Y should occur).

• Measures of Rule Quality:

Support: How frequently X and Y appear together.

Example : If 10 of the 100 transactions contain {bread, butter}, Support = 10%.

Confidence: The probability of Y given X.

♣ Example: 80% of people who purchase bread also buy butter → Confidence = 80%.

Lift: Value of observed confidence compared to that expected if X and Y were independent.

♣ Lift > 1 means the rule is meaningful.

Algorithms

• Apriori Algorithm: creates frequent item sets and association rules.

• FP-Growth Algorithm – Employ tree structures to make discovery faster.

Applications

• Retail: Basket analysis (e.g., people who buy diapers also tend to buy baby wipes).

• E-commerce: Recommendation systems.

• Healthcare: Discovering symptom-disease relationships.

1.2.4 Classification

It is the process of categorizing information into predetermined groups though supervised learning.

Process

Building a Model – Based on the training data with known labels.

Testing the Model – To make predictions using the model.

Prediction: Labelling new inputs.

Algorithms

• Decision Trees (e.g., ID3, C4. 5, CART)

• Naïve Bayes Classifier

• Support Vector Machines (SVM)

• Neural Networks Example

• Classification is employed by a bank to make “Low Risk” or “High Risk” assessment of applicant applying for loan based on the customer’s income, employment history and credit score.

• Email servers label the incoming email as “Spam” or “Not Spam”.

Applications

• Fraud detection.

• Sentiment analysis.

• Medical diagnosis.

1.2.5 Clustering

Flying as a UI to cluster (view here): Clustering is an unsupervised learning approach that partitions objects into groups such that the objects within the same group are most similar and the distance between groups items are maximally different without any prior knowledge of group membership.

Methods of Clustering

Partitioning Methods – K-Means algorithm.

Hierachical Methods – Constructs cluster tree structures.

Density-Based Methods- In DBSCAN, clusters are formed when the points in a cluster are close enough and their surroundings are not as dense or crowded.

Example

• Within retail organizations, customer segments are often defined: repeat purchasers, seasonal purchasers and one-time purchasers.

• In biology, clustering of gene expression data is useful for grouping together genes which have similar functionality.

Applications

• Customer segmentation in marketing.

• Image compression.

• Social network analysis.

1.2.6 Regression Analysis

It discovers the relation between a dependent variable (target) and independent variables (predictors).

Types

Linear Regression – It is used to make predictions for continuous values based on straight line relationship.

o Example: Estimating the price of house, depending on area and location.

Logistic Regression - Lets you predict outcomes that are categorical (for example yes/no, true/false).

o Example: The forecasting of whether or not a student will fail an exam based on study hours.

Multiple Regression – Uses multiple predictors.

figure: Types

Importance

• Used for forecasting and prediction.

• Assists in measuring the effect of predictor variables.

Applications

• Predicting demand in supply chains.

• Sales forecasting.

• Healthcare predictions (disease risk).

1.2.7 Anomaly Detection

Anomaly detection (sometimes called outlier detection) is the identification of items, events or observations which do not conform to an expected pattern.

Techniques

Methods: Statistical: Through the use of standard deviation, z-scores and probability distributions.

Machine Learning Techniques – Isolation forests, autoencoders.

Methods Based on Distance – Outliers are distant to the means of the clusters.

Example

• In bot detection, a rapid sequence of queries is regarded as an anomaly. • For credit card fraud, an exorbitant transaction in a different nation is similarly tagged by the algorithm.

• In cybersecurity, abnormal login attempts signal a potential hack.

• In health care, anomalous values in patient vitals could indicate a medical emergency.

Applications

• Fraud detection in finance.

• Fault detection in manufacturing.

• Network intrusion detection.

“Activity: Exploring Data Mining Concepts in Action”

Students will be divided into groups and assigned one concept from data mining (preprocessing, exploration, association, classification, clustering, regression, anomaly detection). Each group must prepare a real-world example, illustrate it with data or a simple chart, and present findings. This encourages collaborative learning and practical application of concepts.

1.3 Technologies Used in Data Mining Process

1.3.1 Database and Data Warehousing Technologies

A. Database Technologies

Databases are the foundation of data-driven organizations. They offer an organized way of handling data storage/retrieval operations.

Databases for Data Mining

Relational Databases (RDBMS):

o Record information in rows and columns (table).

o Perform Data Queries and Manipulations using SQL (Structured Query Language).

o Example: Oracle, MySQL, PostgreSQL.

Object-Oriented Databases:

o Expand relational systems with storing of objects such as images, audios or videos.

o Applicable in multimedia mining, medical imaging etc.

NoSQL Databases:

o Handle unstructured or semi-structured data.

o Examples: MongoDB, Cassandra.

o Commonly used from the analysis of web logs, social media and sensor data.

Role in Data Mining

• Act as a single-source of truth for data mining.

• Permit query-needs filtering prior to mining algorithms execution.

• Maintain data consistency and accuracy.

• Provide massive storage and retrieval for high volume data.

Example: A supermarket provides analysts with the ability to obtain frequent purchase patter variation (bread and butter) for market basket analysis.

B. Data Warehousing Technologies

Data Warehouse is a collection of data designed to support management decision making.

Features of a Data Warehouse

• Subject-Oriented: Structured on business concept (sale, finance, customers).

• Inclusive: Integrates information from various places.

• Time-variant: Retains historical information for long term trend analysis.

• Prologue: After being written, data is not updated often so that the result is stable.

Key Components

ETL Process (Extract, Transform, Load):

o Extracts data from multiple sources.

o Transforms it into consistent formats.

o Loads it into the warehouse.

OLAP (Online Analytical Processing):

o Enables multidimensional analysis of data.

o To slice the data means to look at one dimension of the cube; to dice it is to view multiple dimensions; roll-up means to summarise, drill-down shows a more detailed level.

Schemas for Organization:

o Star Schema – A central fact table and related dimension tables.

o Snowflake Schema: A more normalized form of the star schema.

Role in Data Mining

• Delivers clean, fully-integrated and historical data.

• Enables faster query performance.

• Delivers business intelligence applications such as dashboards and reporting.

Example: A telecom provider analyzes their customer call records using a data warehouse to help predict churn and create targeted promotions.

1.3.2 Machine Learning and Statistical Techniques

A. Machine Learning Techniques

Machine Learning (ML)—The subset of AI that actually trains systems to recognize patterns in data.

without being explicitly programmed.

Types of Machine Learning techniques in Data Mining

Supervised Learning

o Labeled data is used to train models.

• Algorithms: Decision Trees, Random Forests and Support Vector Machines.

o For example, decide if a loan applicant is high or low risk.

Unsupervised Learning

o Exploring hidden patterns in the unlabeled data.

o Algorithms: Type.cursor- highlighting, LDA Topic Modelling, K-Means clustering.

o Example: Identification of customer segments for focused high conversion rate marketing.

Semi-Supervised Learning

o Is able to utilize both labelled and unlabelled data.

o Useful if labeling data is expensive or time-consuming.

Reinforcement Learning

o Feedback driven (rewards and punishments).

o Example: Online recommendation systems learning from user clicks.

Role in Data Mining

• Automates pattern recognition.

• Enhances predictive modeling.

• Is compatible with adaptive systems that get better with use.

B. Statistical Techniques

Statistics serves as a body of theory for data mining and it provides testing, validation, and measurement tools from that theory.

Common Statistical Techniques

Regression Analysis

o Represents the relationships between dependent and independent variables.

o Example: Predicting sales from the advertising expense.

Probability Models

o Estimate the likelihood of events.

o Example: Predict risks of disease in healthcare data.

Hypothesis Testing

o Indicates whether or not observed patterns were statistically significant.

o Example: to determine whether a new marketing program led to significantly higher sales.

Bayesian Methods

o Increase probabilities as more information is received.

o Used extensively for spam filtering and fraud detection.

Role in Data Mining

• Guarantees mathematical soundness to verify identified patterns.

• Guarantees credibility and transferability of findings.

• Assists estimate confidence and uncertainty in predictions.

Example: In credit card fraud prevention, statistical models predict the likelihood of a transaction being fraudulent and machine learning models further refine this prediction by using historical behavior.

Knowledge Check 1

Choose the correct option:

1. Which of the following is the primary role of a data warehouse in data mining?

a) Transaction processing

b) Historical data analysis

c) File storage

d) Real-time messaging

2. Which schema is commonly used in data warehousing?

a) Ring schema

b) Star schema

c) Flow schema

d) Chain schema

3. Which machine learning type is used when labels are not available?

a) Supervised learning

b) Reinforcement learning

c) Unsupervised learning

d) Semi-supervised learning

4. Logistic regression is mainly used for predicting:

a) Continuous values

b) Binary outcomes

c) Clustering groups

d) Time series trends

1.4 Summary

- Data mining refers to the extraction of useful knowledge from large datasets.
- It is not just a query and reporting tool it applies some additional methodologies like classification, clustering, regression, association rule mining etc.
- Data mining as a field has grown from mere databases in 1960s to the big data analytics and machine learning together with AI etc.
- Raw data cleaning, integration, transformation and reduced for proper analysis are the four initial steps of Data Preprocessing.
- Data exploration Through explorative analysis, analysts can use descriptive statistics, chart visualizations and correlation analysis to understand data features.

- ❖ Association rule mining discovers the relationship between items, most often utilize in retail and recommendation system.
- ❖ Classification categorized data in predefined groups, which distinguishes clustering that only clusters together similar data examples.
- ❖ Regression analysis describes the relationships between variables and enables to forecast and predict.
- ❖ Anomaly detection recognizes rare or abnormal patterns, crucial in applications such as fraud detection, cyber security and health care.
- ❖ Database, and Data Warehouse constitute a back-end repository where data are stored, organized and processed for mining.
- ❖ Machine learning and statistical methods constitute the analytic heart of data mining: predictive modelling, pattern discovery (data dredging), and validation.

1.5 Key Terms

1. Data Mining – It is the process of finding hidden patterns and useful knowledge from large databases.
2. Data Preprocessing–Refinement and transformation of the raw data to remove outliers, irrelevant or redundant information's and integration of useful information for analysis.
3. Association Rule Mining – It is method to discover relationship between items in large transactional database.
4. Classification – A type of supervised learning which involves placing data into preexisting categories.
5. Clustering Unsupervised method which puts similar points together in clusters.
6. Regression Analysis – A statistical process of estimating relationships between a dependent variable and one or more independent variables.
7. Anomaly Detection – Detection of unusual or rare patterns in a dataset.
8. Data Warehouse – A collection of integrated, historical data for decision support.
9. Machine Learning – Branch of AI that trains systems to recognize patterns and make predictions with data.

1.6 Descriptive Questions

1. Define data mining. Define it and explain its significance to contemporary firms.
2. Describe the development of data mining as a field. What has it evolved from databases before to big data analytics today?
3. Discuss the significance of data preprocessing in the Knowledge discovery process. Do you think this is crucial and it should be performed before analysis?

4. Explain the concepts of association rule mining, classification and clustering with appropriate examples.
5. What is regression analysis? How is it used in prediction/f414_fp40_0292.html and forecasting?
6. Explain anomaly detection. Provide brief description of its application in fraud and cybersecurity.
7. Explain the contribution of database and data warehousing technologies for data mining.
8. How can data mining patterns be discovered with the aid of machine learning and statistical techniques?

1.7 References

1. Han, J., Kamber, M., & Pei, J. (2012). Data Mining: Concepts and Techniques (3rd ed.). Morgan Kaufmann.

2. Tan, P. N., Steinbach, M., & Kumar, V. (2018). Introduction to Data Mining (2nd ed.). Pearson.

3. Witten, I. H., Frank, E., Hall, M. A., & Pal, C. J. (2016). Data Mining: Practical Machine Learning Tools and Techniques. Morgan Kaufmann.

4. Aggarwal, C. C. (2015). Data Mining: The Textbook. Springer.

5. Mitra, S., & Acharya, T. (2003). Data Mining: Multimedia, Soft Computing, and Bioinformatics. Wiley.

6. Larose, D. T., & Larose, C. D. (2014). Discovering Knowledge in Data: An Introduction to Data Mining. Wiley.

7. Kantardzic, M. (2020). Data Mining: Concepts, Models, Methods, and Algorithms (3rd ed.). Wiley.

8. Berkhin, P. (2006). A Survey of Clustering Data Mining Techniques. Springer.

9. Online Resource: The Data Mining Group (DMG) – http://www.dmg.org

Answers to Knowledge Check

Knowledge Check 1

1. b) Historical data analysis

2. b) Star schema

3. c) Unsupervised learning

4. b) Binary outcomes

1.8 Case Study

Data mining to find patterns and trends in ShopEase

Introduction

In the digital age, companies collect and retain large amounts of data from the transactions they execute, interactions with their customers, and online platforms. But without any sort of data consumer, that raw data is essentially useless. Data mining has an important role in discovering unknown patterns, predicting the results and making decision.

ShopEase is an expanding online sales platform and it could see that, even though the company had tons of customer and sales data, most of it was being employed in order to generate some pretty standard reports. Using methods like data mining, ShopEase started to notice buying trends, customer likes as well new trends. This allowed the company to create efficient marketing plans, optimize stocks, and enrich customer experiences.

Background

ShopEase collected structured data (transaction, product categories, demographics) and unstructured evidence (reviews, browsing history). At the outset, there were data integrity issues in terms of records being incomplete or duplicates, and data was spread out all over. The data preprocessing assisted in the cleaning of this data and integrating it so that it can be used for deeper analysis.

Analysts doing data exploration discovered seasonal peaks in searches for products like “budget laptops” and “school bags.” Using association rule mining, they found that patrons who bought a smartphone would frequently buy protective cases within two weeks. With classification the company was able to estimate which customers were likely to respond to personalized offers, and clustering helped in dividing buyers into categories such as premium customers and discount-seekers. The organization used regression analysis to predict monthly sales, enabling anomaly detection for abnormal purchasing patterns and thereby nipping malpractice in the bud.

Issue 1 - Dealing with the Data Quality.

ShopEase faced the challenge of incomplete and inconsistent data which were making it hard to analyze accurately. Solution: The use of strong pre-processing methods such as cleaning missing values, fusion of data sources and normalization of records provided the mining tasks with solid and quality data.

Problem Statement 2: Customer Purchase Patterns As a business owner, it is crucial for you to develop a recommender system which recommends your products based on the purchase history of your customers.

The company had to know what other products people were buying together. Solution – Association rule mining of transaction data enable ShopEase to implement effective product bundling and recommendations tailored for individual users.

Problem 3: Customer churn forecast It is also possible to predict customer behaviour.

ShopEase hoped to decrease the number of customers leaving by pinpointing which were likely to leave. Solution: It developed classification models to classify customers as “likely to stay” and “likely to churn” for proactive retention campaigns using historical data.

MCQ Example

Q: Which type of data mining algorithm would assist ShopEase in finding which products are often bought together?

a) Regression

b) Classification

c) Association rule mining

d) Anomaly detection

Answer: c) Association rule mining

Conclusion

The instance of ShopEase is a clear example of how you can apply advanced analytics to raw data to get real business insights. ShopEase was able to derive competitive edge by taking on challenges of data quality, purchasing patterns, customer churn prediction and fraud prevention. Data mining is an invaluable asset for businesses that want to utilize data in order to help make key decisions.

BUPBM Unit 2 V3.docx

Building useful Predictive Business models_BBA_3

Building useful Predictive Business models_BBA_3

ATLAS SkillTech University

Document Details

Submission ID

trn:oid:::3618:127670887

Submission Date

Feb 6, 2026, 3:39 PM GMT+5:30

Download Date

Feb 6, 2026, 3:42 PM GMT+5:30

File Name

BUPBM Unit 2 V3.docx

File Size

144.7 KB

19 Pages

3,869 Words

23,499 Characters

0% detected as AI

The percentage indicates the combined amount of likely AI-generated text as well as likely AI-generated text that was also likely AI-paraphrased.

Caution: Review required.

It is essential to understand the limitations of AI detection before making decisions about a student's work. We encourage you to learn more about Turnitin's AI detection capabilities before using the tool.

Detection Groups

0 AI-generated only 0%
Likely AI-generated text from a large-language model.

0 AI-generated text that was AI-paraphrased 0%
Likely AI-generated text that was likely revised using an AI-paraphrase tool or word spinner.

Frequently Asked Questions

How should I interpret Turnitin's AI writing percentage and false positives?
The percentage shown in the AI writing report is the amount of qualifying text within the submission that Turnitin's AI writing detection model determines was either likely AI-generated text from a large-language model or likely AI-generated text that was likely revised using an AI paraphrase tool or word spinner.

False positives (incorrectly flagging human-written text as AI-generated) are a possibility in AI models.

AI detection scores under 20%, which we do not surface in new reports, have a higher likelihood of false positives. To reduce the likelihood of misinterpretation, no score or highlights are attributed and are indicated with an asterisk in the report (*%).

The AI writing percentage should not be the sole basis to determine whether misconduct has occurred. The reviewer/instructor should use the percentage as a means to start a formative conversation with their student and/or use it to examine the submitted assignment in accordance with their school's policies.

What does 'qualifying text' mean?
Our model only processes qualifying text in the form of long-form writing. Long-form writing means individual sentences contained in paragraphs that make up a longer piece of written work, such as an essay, a dissertation, or an article, etc. Qualifying text that has been determined to be likely AI-generated will be highlighted in cyan in the submission, and likely AI-generated and then likely AI-paraphrased will be highlighted purple.

Non-qualifying text, such as bullet points, annotated bibliographies, etc., will not be processed and can create disparity between the submission highlights and the percentage shown.

Unit 2 – "Data Mining: Types, Applications, and Challenges"

Learning Objectives

1. Explain different types of data mining techniques such as predictive, descriptive, text mining, and web mining.

2. Differentiate between classification, clustering, association, and regression techniques with respect to their functions and applications.

3. Identify real-world applications of data mining across domains such as business, healthcare, finance, retail, and education.

4. Analyze case-specific uses of data mining for tasks like fraud detection, customer segmentation, recommendation systems, and risk prediction.

5. Recognize challenges in data mining including data quality issues, privacy concerns, ethical implications, and algorithmic limitations.

6. Understand the role of big data and emerging technologies in enhancing the scope and capabilities of data mining.

7. Evaluate the advantages and limitations of applying data mining in decision-making and strategic planning.

8. Develop a critical perspective on balancing opportunities and risks while implementing data mining solutions in organizations.

Content

2.0 Introductory Caselet:

"Data Mining at MedicoHealth Solutions"

MedicoHealth Solutions is a healthcare analytics company that partners with hospitals to provide data-driven actionable insights aimed at improving patient care. The company pulls in a vast amount of data from patient records, lab tests, prescriptions and even wearable health devices.

Originally this data is only utilized for record and to be compliant. Yet, thanks to Mh's use of data mining methods, the latter got a second chance. They generated models with classification to predict the probability of a patient suffering from chronic diseases (e.g., diabetes). Using clustering, they classified patients by their lifestyle habits and created targeted wellness programs. Association rule mining facilitated the identification of association between symptoms in relation to potential health risks, and regression analysis was useful for prediction of patients who would come to emergency rooms.

Despite these benefits, challenges remain. Privacy of patient data is a big concern because healthcare data are very sensitive. There are challenges in quality of data between various hospitals, and biases may be present in algorithms leading to recommendations of unfair treatment.

The MedicoHealth use-case exemplifies the power of big data in healthcare initiatives, and it describes some of the practical uses, benefits, and challenges that companies should approach responsibly.

Critical Thinking Question

If MedicoHealth's algorithms appear to be producing biassed results that harm particular patient groups, what should the company do in order to ensure they are using data mining fairly, accurately and ethically in respect of health care decisions?

2.1 Mining on Various Kinds of Data

2.1.1 Mining Structured Data (Relational Databases, Warehouses)

Structured data is the original and most common format of organization. It has a clear structure of rows and columns, which is easier to store, query and analyze.

• Sources:

o Database system concepts, including SQL based database RDBMS eg. MySQL, Oracle or SQL Server.

o The data warehouse consolidates various databases into centralized storehouses.

• Characteristics:

o Ready to use schema (tables with attributes and links).

o One can easily extract and organize structure data that could be queried from SQL queries.

o Normally either a continuous or discrete (numerical or categorical) variable.

• Techniques Used:

o Association Rule Mining: Finding purchase trends in sales data.

o Classification and Prediction – Could use labeled data about an event to aid in categorizing the likelihood that an event is of a particular type (fraud detection, loan application).

o Clustering: Segregating customers or products into clusters by some attributes.

o OLAP (Online Analytical Processing): Facilitates slicing, dicing, roll-up of data in warehouse.

• Applications:

o Banking: Identifying anomalous transactions from relational data.

o Retail: Market basket analysis with warehouse sales data.

o Health care: Patient diagnosis streams from hospital databases.

Structured data mining has been well received due to its reliability and the maturity of DB systems.

2.1.2 Mining Semi-Structured Data (XML, JSON, Web Data)

Semi-structured data is not merely a flat file and does not adhere strictly to a relational format, but we still have relieved tags or attribute or meta-data markers that help guide us through our documents.

• Sources:

o XML (Extensible Markup Language) files.

o Javascript Object Notation (JSON) utilized in web apis.

o HTML documents on websites.

o Log files from web servers.

• Characteristics:

o Flexible, self-describing structure.

o No schema, but has hierarchical or tag-based structure.

o Can store complex nested data.

• Techniques Used:

o Parsing and Transformation: Parsing XML/JSON into understandable formats.

o Web Mining: Exploring the content, structure, and usage of web resources.

o Clickstream: Analysis, i.e., examining a user's browsing steps in order to guess behaviour.

• Applications:

o E-commerce: Extracting structured product information from JSON APIs.

o Web Analytics- Identify most frequent paths of navigation by users.

o Social Platforms: Extracting profile and user interaction from semiformal log.

Semi-structured mining is crucial in the era of internet as a large amount of information exchanging on the web is under XML or JSON form.

Did You Know?

"Nearly 70% of data exchanged on the web is semi-structured, often in formats like XML and JSON. These flexible formats power APIs, e-commerce platforms, and social networks. Mining semi-structured data enables businesses to analyze customer behavior, personalize recommendations, and uncover trends hidden in web logs and online interactions."

2.1.3 Mining Unstructured Data (Text, Images, Video, Social Media)

Unstructured data is not organized in a pre-defined manner and is therefore more difficult to analyze. This category represents almost about 80–90% of all data existing worldwide right now.

• Sources:

o Text: Emails, news print, blogs posts, tweets.

o Images: Photos, X-rays, satellite imagery.

o Video: Surveillance footage, YouTube content.

o Social: Posts, likes, shares, comments.

• Characteristics:

o No rigid schema or format.

o Content and form of personalized information may vary greatly.

o Needs NLP and computer vision tools.

• Techniques Used:

o Text Mining & NLP (Natural Language Processing):

♣ Sentiment analysis (positive/negative opinions).

♣ Topic modelling (to discover what is discussed in documents).

♣ NER (Named Entity Recognition - identifying names, places, events).

o Image & Video Mining:

♣ Object recognition, facial recognition.

♣ Content classification using deep learning.

o Social Media Mining:

♣ Trending, user-influence and opinion patterns detection.

• Applications:

o Marketing: Monitoring reception of the brand on social media.

o Security: Identify suspects from the facial features.

o Healthcare: Disease diagnosis with medical image analysis.

Unstructured mining is a complex task, yet it offers rich information on human behavior and trend.

2.1.4 Mining Time-Series, Spatial, and Spatio-Temporal Data

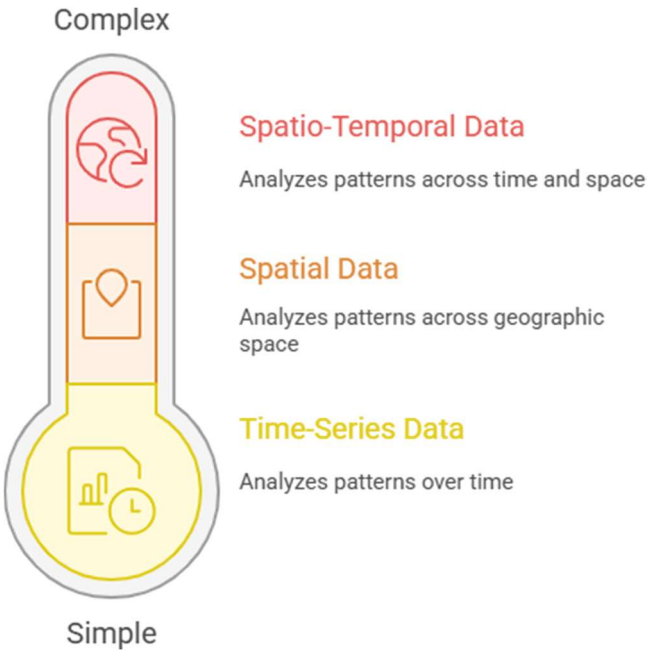

Figure: Mining Time-Series, Spatial, and Spatio-Temporal Data

In addition to traditional data features, special feature types describe patterns over time, space or time and space.

Fig.: Mining of Time-Series, Spatial and Spatio-Temporal Data

A. Time-Series Data

• Definition: Information collected over the course of time.

• Examples: Stock prices, daily sales, temperature measures, Internet of Things (IoT) sensor readings.

• Techniques:

o Statistical models like ARIMA.

o Deep learning models like LSTMs.

o Trend, seasonal, and cyclical analysis.

• Applications:

o Forecasting electricity demand.

o Predicting stock market movements.

o Monitoring patient vitals in healthcare.

B. Spatial Data

•Definition: Information associated to physical or geographic space.

• Examples: GPS coordinates, maps, satellite images, urban planning data.

• Techniques:

o Spatial clustering (e.g., DBSCAN).

o Proximity analysis (nearest neighbour).

• Applications:

o Geographic Information Systems (GIS).

o Location Base Services (ride sharing application, food delivery).

o Environmental monitoring.

C. Spatio-Temporal Data

• Time and space: A definition of data as a combined time and location-wise entity.

• Examples: Dispersal of epidemics, traffic flow, weather.

• Techniques:

o Trajectory mining.

o Space-time pattern recognition.

• Applications:

o Monitoring the spread of disease over time and geographic areas.

o Investigate traffic jam in the smart cities.

o Climate modelling and prediction.

2.2 Applications of Data Mining

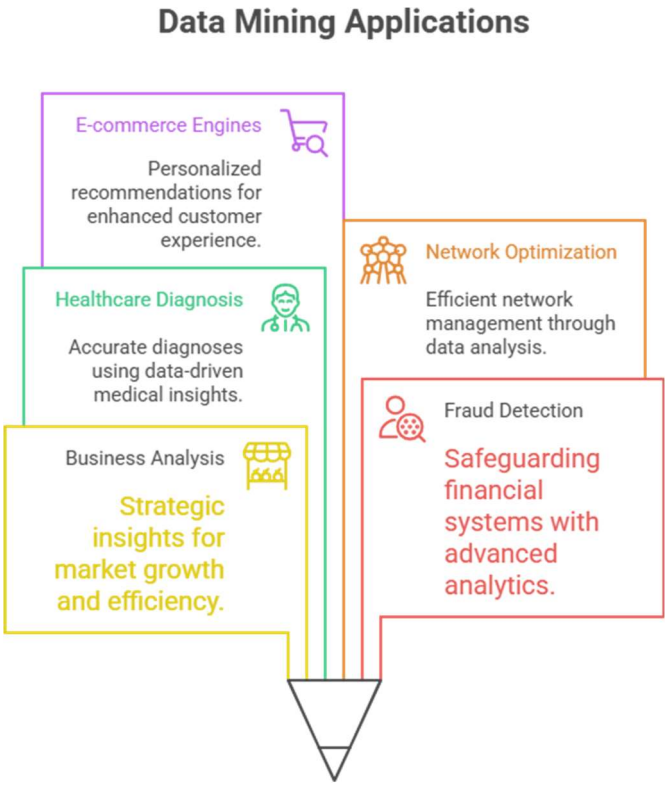

figure: Data Mining

2.2.1 Business and Market Analysis

Companies are always seeking to look through the markets to grasp the consumer behavior, preference and competition." The raw transactional and demographic data must be mined to convert it into strategic knowledge.

• Key Applications:

o Customer Segmentation: Clustering can help you highlight high-value customers, one-time buyers, and customers who are very price sensitive.

O Market Basket Analysis: Association rule mining is used to identify the items that commonly get sold together (e.g “customers who buy bread also tend to buy butter”).

o Forecasting: Regression and time-series models are used to predict sales and seasonal behaviour in the future.

o Customer Lifetime Value Prediction: Classification helps firms to predict customer profitability in the long run.

• Example: A grocer applies data mining to loyalty card records. By separating out buyers, it homes in on promotions that convert, lifting sales and retention.

2.2.2 Fraud Detection in Banking and Finance

Fraud costs the finance industry billions of dollars annually. Data mining offers proactive means to predict and avert it.

• Techniques:

o Anomaly Detection: Detects when transaction are not normal (e.g. large purchase in another country at once).

o Classification Models: Will help determine whether a transaction is “fraudulent” or “genuine.”

o Pattern Recognition: Identifies repeat offenses by associating suspicious activities.

o Real Time Monitoring: Real time streaming analytics notifies fraudulent behaviour.

• Example: A credit card company uses anomaly detection models to detect the fact that a customer purchased luxury goods in Paris on the same day he or she made a small grocery purchase in Mumbai. The system sets off an alarm and puts the card on a temporary lockdown.

2.2.3 Healthcare and Medical Diagnosis

There is plenty of data in healthcare: from a patient’s medical history to their test results, images or even wearables, health organizations have access to large-scale data. Data analysis or data mining contributes in converting it into useful medical knowledge.

• Key Applications:

o Predictive Diagnosis: We use classification models to predict the potential risk of diseases such as diabetes, heart attack or cancer.

o Patient Clustering: Groups similar symptoms or collective treatment responses of patients for precision medicine.

o Drug Discovery: Text mining on medical research papers to discover promising compounds.

o Medical Image Mining: Applying computer vision and deep learning to identify tumours in X- rays and MRIs.

• Example: Hospitals use regression models to estimate patient admissions during a seasonal flu, better allocating both beds and staff.

2.2.4 Telecommunications and Network Optimization

The telecom sector also uses data mining extensively, since it generates call detail records, internet logs," and network logs.

• Key Applications:

o Churn Prediction: Companies can predict which customers are more likely to switch providers through classification models and intervene with offers.

o Pattern Clustering Analysis: Customers can be clustered into high, medium or low user categories for tailor-made plans.

o Network optimization: Data mining with time-series and geographical information to comprehend traffic congestion behaviour to benefit service quality.

o Fraud Detection: Discovery of SIM cloning, fake calls behaviour, unusual usage.

• For example, a telco realizes from clustering that one segment of its users always exhausts its data allowance mid-month. It adds a new “top-up” plan, increasing revenue and customer satisfaction.

2.2.5 E-commerce and Recommendation Engines

Personalization is bread-and-butter for e-commerce, and data mining powers the recommendations that ultimately draw customers in.

• Techniques:

Association Rule Mining: Finds all the products that are purchased together often (e.g., phone + case + charger).

o Collaborative Filtering: Recommends items by comparing user similarities and purchase histories.

o Content Based Filtering: Provides suggestions based on attributes of the product like previous purchase.

o Sentiment Analysis: Extracting product reputation from customers’ reviews.

• Example: Amazon’s recommender system processes purchasing history and browsing patterns to propose products that are complementary or act as an alternative. Netflix does something similar when it suggests movies to watch based on what a user has been watching.

Knowledge Check 1

Choose the correct option:

1. Which data mining technique is commonly used in market basket analysis?

a) Clustering

b) Regression

c) Association rules

d) Anomaly detection

2. In banking, data mining is mainly applied for:

a) Customer loyalty

b) Fraud detection

c) Ad placement

d) Image processing

3. Which method helps group patients with similar symptoms in healthcare?

a) Clustering

b) Classification

c) Regression

d) Anomaly detection

4. Recommendation engines in e-commerce primarily use:

a) Sentiment analysis

b) Collaborative filtering

c) Regression analysis

d) Time-series models

2.3 Challenges of Data Mining

2.3.1 Data Quality and Preprocessing Issues

It is absolutely true that garbage in, garbage out does not just refer to the perception of a model. Weak data leads to weak insights that can lead organizations astray.

A. Common Data Quality Problems

Missing Data: Values that are unavailable in the dataset because of collection errors, inadvertently skipped questions or technical malfunctions.

o Example: In a retail data set, customer addresses or phone numbers may be missing.

Noisy Data: Incorrect or irregular outliers that corrupt the performance of an algorithm.

o Example: Malfunctioning resulted in the reading of a 500 degree C temperature.

1.2 Inconsistent Data: Data is conflicting in different sources or systems

o Example: A member's age was 35 from one database and 37 from another.

Duplicate / Redundant Data: More than one record of the same thing.

o Example: The patient who is entered into the database twice by a hospital.

B. Challenges in Preprocessing

• Data Cleaning: Identifying and correcting mistakes, filling in missing data, and removing noise.

• Integrate: combine dissimilar data types (such as relational databases, flat files, cloud-based storage).

• Data Formatting: Standardizing (dates, currencies, units of measure.

• Feature Selection Using dimension reduction (for instance PCA) to be used to large scale features.

C. Example

Medical records from different hospitals often using are not compatible with one another. These records need to be cleaned, combined and normalized or our predictions will be worthless before we can mine disease patterns.

Did You Know?

“More than 60% of a data scientist’s time is spent on data cleaning and preprocessing, not on model building. Issues like missing values, duplicates, and inconsistencies are the biggest hurdles in mining. Without proper preprocessing, even advanced algorithms produce unreliable or misleading insights, making data quality the top priority.”

2.3.2 Scalability, Privacy, and Security Concerns

Data in the modern world is often big, fast and diverse,' which poses challenges unique to this situation.

A. Scalability

• Problem: Classical mining algorithms cannot process very large data sets (i.e., billions of transactions a day). They can be inefficient or too memory-intensive.

• Solutions:

o Distributed Computing: Experience with Hadoop, Spark or in a cloud environment to process data on multiple machines.

o Parallel Algorithms: Assignment of tasks to processors in order to speed up the computation.

o Incremental Mining (updating models as new data comes in, rather than retraining from scratch).

• Example : Google, Amazon handle (in petabytes) of click stream daily stored on disks and mined in a distributed way.

B. Privacy Concerns

• Problem: Mining may involve sensitive personal data (health records, financial transactions, browsing history). And it can cross individual privacy lines if misapplied.

• Risks:

o Identifying individuals from anonymized datasets.

o Applying knowledge to surveillance or discrimination.

• Solutions:

o Anonymized Data: Potentially identifiable data that are de-identified before analysis is conducted.

o Differential Privacy – Adding noise to the results so that people cannot be identified.

o Federated Learning: Method of training algorithms across devices (decoupled learning from raw data).

• Example: In health care, if you mine patient data to predict risk of disease, it must adhere to regulation such as HIPAA in the U.S. and GDPR in Europe.

C. Security Concerns

• Problem: Not only other sources of data, but also huge datasets themselves and the models used to mine them can be attacked. Hacker types often try just a couple of things: 1) Steal

data (i.e., the recent Anthem break in), 2) Modify it, and/or then 3) Futz with it such as later deleting/corrupting, or even worse.

• Risks:

o More data breaches that result in the loss of personal or financial information.

o Poisoning, where adversarial agents inject poisonous data to corrupt models.

• Solutions:

o Encryption: Securing data in motion and at rest.

o Data Security: Who can see what data o Access Control: Restricting access to people who do not need to see it or change it.

o Review and Audit: Ongoing observation of system behavior for abnormal indicators.

• Example: A bank's fraud detection model may be vulnerable to hackers who access and alter the training data in a way that causes it to overlook signs of fraud.

"Activity: Identifying Data Mining Challenges"

Students will be divided into groups, each assigned one challenge of data mining (data quality, preprocessing, scalability, privacy, or security). They will research a real-world case where this challenge occurred, present the issue, and suggest possible solutions. This promotes critical thinking and practical understanding of mining challenges.

2.4 Summary

- Market dynamics depict demand and supply forces on the equilibrating price and quantity in a market.
- An excess exists when supply exceeds demand at a certain price point, creating unsold stock and depressing prices.
- Typically, excess supplies are generated after prices have been set too high or when over-production follows from overly optimistic expectations.
- A scarcity occurs when the quantity demanded exceeds the quantity supplied at a particular price; this causes buyers to compete for whatever is available and lends upward pressure on the price.
- Scarcity often arises when prices are fixed below what is termed the equilibrium price level, in times of demand surges or because supply constraints have limited availability.

- The price system mechanism—the process of surplus or shortage operation as a result of higher and lower prices (higher: eliminate surplus by increasing demand; lower: eliminate shortages by attracting supply).
- In free market, this autoregulation mechanism means that the balance between demand and supply is gradually achieved without interference.
- Examples of cases, including various price adjustments for necessary goods under situations of crises, illustrate can give insight into market behavior driven by surpluses and shortages.
- Estimating Surpluses, Shortages, and Shifts Effectively analyzing real markets –houses in Scottsdale, TV sets Use dodgeball to introduce Elasticity Understanding Richard Simmons to teach Elasticity 3 2 Supply & Demand Determine prices of goods & services Create a price mechanism (flexible like a rubber band) for how resources are used.

2.5 Key Terms

1. Market Dynamics – The manner in which pricing and quantity is set between demand and supply in a market.
2. Excess of supply – the condition that exists when the quantity supplied exceeds quantity demanded at a certain price.
3. Scarcity – a condition and concession of the fact that quantity demanded is greater than quantity supplied at the prevailing price level.
4. Price Equilibrium – The price level at which the quantity of a good demanded equals the quantity supplied in a market.
5. Curve – A curve that shows the relationship between the price of a good and the quantity demanded by consumers.
6. Supply Curve – A graph of the relationship between Price and Quantity Supplied.
7. Adjustment Mechanism – The mechanism by which prices move surplus or shortage into equilibrium.
8. Price Distortions - Prices in the market as a result of changes in demand and supply conditions.

2.6 Descriptive Questions

1. Define market dynamics. How do supply and demand interact to construct equilibrium in a market?
2. Explain the concept of surplus. What drives surpluses and its implication on the price in the market?
3. What is a shortage in economics? Explain the reasons and results of real market shortages.

4. Use a diagram to show how excess supply pressures prices down towards equilibrium.
5. Show graphically how shortage makes prices rise toward the equilibrium.
6. Explain the mechanism that operates to restore equilibrium when markets are in surplus or shortage.
7. Offer an example and describe how the market reacted to excess (surplus).
8. Give an example of a shortage, cite the real world situation, and demonstrate how it was resolved.
9. Explain why knowing about surpluses and shortages is important to policymakers and business decisions.

2.7 References

1. Mankiw, N. G. (2021). Principles of Economics (9th ed.). Cengage Learning.

2. Samuelson, P. A., & Nordhaus, W. D. (2010). Economics (19th ed.). McGraw-Hill Education.

3. Case, K. E., Fair, R. C., & Oster, S. M. (2017). Principles of Economics (12th ed.). Pearson.

4. Krugman, P., & Wells, R. (2020). Microeconomics (6th ed.). Worth Publishers.

5. Lipsey, R. G., Chrystal, K. A., & Lipsey, R. G. (2015). Economics (13th ed.). Oxford University Press.

6. Parkin, M. (2019). Microeconomics (13th ed.). Pearson.

7. Nicholson, W., & Snyder, C. (2019). Microeconomic Theory: Basic Principles and Extensions (12th ed.). Cengage Learning.

8. Baumol, W. J., & Blinder, A. S. (2015). Microeconomics: Principles and Policy (13th ed.). Cengage Learning.

9. Varian, H. R. (2014). Intermediate Microeconomics: A Modern Approach (9th ed.). W. W. Norton & Company.

Answers to Knowledge Check

Knowledge Check 1

1. c) Association rules

2. b) Fraud detection

3. a) Clustering

4. b) Collaborative filtering

2.8 Case Study

“Fraud Detection in Banking ”

Introduction

Banking financial fraud is a significant problem in global banking sector. As digital banking grows at an unprecedented pace, offenders take advantage of the loopholes and emerge to commit credit card fraud, identity theft, money laundering, phishing scams etc. Legacy rule based systems fall short of the mark since fraudsters continually adapt their tactics. Data mining and anomaly detection is a very strong solution by finding abnormal transaction patterns that don't follow normal customer behavior. This provides idea to bank that they can work intelligently by enable them to implement fraud prevention techniques, minimizing financial loss as well acquiring customers’ trust.

Background

Banking fraud detection is challenging since the fraudulent transactions often look very similar to legitimate ones. An example of this is when a thief makes series of little purchases with a stolen credit card before the big one. If a customer typically shops in their home country and starts making numerous international purchases, models alert to a potential problem.

Data mining processes – including clustering, classification, decision trees and neural networks—allow to spot these anomalies by sifting through a large amount of historical and real-time data. Banks use these techniques to:

• Real-time detection of suspect activities

• The number of falsely accused will decrease and less actions taken against innocent customers.

• Machine learning to adapt to new fraud patterns

Problem Statement 1: Fraud Detection in real time

Problem: Many fraudulent transactions are identified after it is too late for banks to do anything about them.

Answer: Deploying anomaly detection algos on transaction streams for real-time alerts.

MCQ:

From a detection perspective, which of them is the best for detecting fraud as it occurs?

a) Manual verification

b) Customer complaints

c) Real-time anomaly detection in data-stream with data mining

d) Ignoring small deviations

Answer: c) Data mining for real time anomaly detection

Problem Statement 2: The High False Positives: Our second observation is that many newcomers introduced to the community are lost.

Problem: Existing fraud detection systems are known to capture actual fraud activities, but often also reject legitimate transactions which can lead to customer dissatisfaction.

Solution: Advanced machine learning models that learn a customer user behavior over time, making sure not to create many false alarms while catching real fraud.

MCQ:

What is the disadvantage of having high false positives in fraud detection?

a) Fraud remains undetected

b) Customers are made to experience inconvenience when transaction is legitimate

c) Banks earn more revenue

d) All suspicious transactions get ignored

Answer: b) Inconvenience to customers is caused even in cases where the transaction is legitimate.

Issue 3: Adapting to Changes in Fraud Methodology

Problem: Scammers are constantly updating their tactics, rendering static detection processes obsolete. Solution: Adaptive data mining models that refreshes these models on a regular basis with new information so that they can identify fraud pattern early.

MCQ:

Why do we need to update fraud detection models?

a) To reduce operational costs

b) Being able to adjust with new fraud methods

c) To increase manual investigation

d) To limit transaction volumes

Answer: b) To keep up with evolving methods of fraud

Conclusion

Detecting fraud in the banking sector is important for protecting bank assets and customer confidence. Using data mining tools in conjunction with anomaly detection, banks are able to discover fraudulent activity as it occurs, dramatically diminishing the number of false positive alerts and allowing them to stay one step ahead of rapidly changing fraud tactics. Banks arm themselves with strong tools for fighting financial crime through machine learning and big data analytics. Such a proactive system is secure and efficient for banking in the modern day.

BUPBM Unit 3 V3.docx

Building useful Predictive Business models_BBA_3

Building useful Predictive Business models_BBA_3

ATLAS SkillTech University

Document Details

Submission ID

trn:oid:::3618:128374763

Submission Date

Feb 16, 2026, 12:13 PM GMT+5:30

Download Date

Feb 16, 2026, 12:16 PM GMT+5:30

File Name

BUPBM Unit 3 V3.docx

File Size

165.1 KB

21 Pages

4,676 Words

27,058 Characters

0% detected as AI

The percentage indicates the combined amount of likely AI-generated text as well as likely AI-generated text that was also likely AI-paraphrased.

Caution: Review required.

It is essential to understand the limitations of AI detection before making decisions about a student's work. We encourage you to learn more about Turnitin's AI detection capabilities before using the tool.

Detection Groups

1 AI-generated only 0%
Likely AI-generated text from a large-language model.

0 AI-generated text that was AI-paraphrased 0%
Likely AI-generated text that was likely revised using an AI-paraphrase tool or word spinner.

Frequently Asked Questions

How should I interpret Turnitin's AI writing percentage and false positives?
The percentage shown in the AI writing report is the amount of qualifying text within the submission that Turnitin's AI writing detection model determines was either likely AI-generated text from a large-language model or likely AI-generated text that was likely revised using an AI paraphrase tool or word spinner.

False positives (incorrectly flagging human-written text as AI-generated) are a possibility in AI models.

AI detection scores under 20%, which we do not surface in new reports, have a higher likelihood of false positives. To reduce the likelihood of misinterpretation, no score or highlights are attributed and are indicated with an asterisk in the report (*%).

The AI writing percentage should not be the sole basis to determine whether misconduct has occurred. The reviewer/instructor should use the percentage as a means to start a formative conversation with their student and/or use it to examine the submitted assignment in accordance with their school's policies.

What does 'qualifying text' mean?
Our model only processes qualifying text in the form of long-form writing. Long-form writing means individual sentences contained in paragraphs that make up a longer piece of written work, such as an essay, a dissertation, or an article, etc. Qualifying text that has been determined to be likely AI-generated will be highlighted in cyan in the submission, and likely AI-generated and then likely AI-paraphrased will be highlighted purple.

Non-qualifying text, such as bullet points, annotated bibliographies, etc., will not be processed and can create disparity between the submission highlights and the percentage shown.

Unit 3: "SPSS Basics: Environment, Data Attributes, and Descriptive Statistics"

Learning Objectives

1. Familiarize with the SPSS environment by identifying key windows, menus, and tools for data analysis.

2. Understand different types of data attributes (nominal, ordinal, interval, ratio) and their representation in SPSS.

3. Learn how to enter, edit, and manage datasets within SPSS, including defining variables and assigning labels.

4. Differentiate between variable view and data view and explain their roles in handling datasets.

5. Apply descriptive statistical techniques such as mean, median, mode, variance, and standard deviation using SPSS.

6. Generate frequency distributions and cross-tabulations to summarize and interpret categorical data.

7. Use SPSS to create basic graphical representations such as bar charts, histograms, and pie charts for exploratory analysis.

8. Interpret SPSS output tables and charts to draw meaningful insights and prepare data for advanced analysis.

Content

3.0 Introductory Caselet

"Using SPSS to Analyze Students' Performance Data"

Dr. Meera, a professor in a university, was interested in predicting academic success of her students across multiple courses she was teaching. She collected statistics from the students like name, gender, age, course registration information and mid-term scores as well as trait scores. At first, she worked with the data in Excel, but had difficulty running meaningful statistical summaries and producing informative visualizations.

To resolve this, she switched to SPSS. In the Variable View, she described among other things student id (nominal), gender (nominal), age (scale), and exam score(s) scale. In Data View, she inputted student-level data. She determined mean, median and standard deviation using SPSS's descriptive statistics functions and constructed frequency tables with respect to gender distribution. She also prepared histograms to display the distribution of scores from the exam.

These outputs enabled her to discern, not just general trends in performance overall, but also differences between groups. For example, she observed that although the mean mark of the final exam had been good, in a few cases students always had a final exam score lower than one standard deviation which would reflect an academic risk.

This case demonstrates how SPSS offers an organized and efficient platform for managing data attributes, performing descriptive statistics, and analyzing results as opposed to manipulating them by hand.

Critical Thinking Question

If Dr. Meera wants to test whether men and women are different in their overall performance in the final exams by using SPSS, what statistical procedures should she take and why would a description of statistics for the performance not be enough?

3.1 Introduction to SPSS Environment

3.1.1 Overview of SPSS Interface – Data View, Variable View, Menus, and Toolbars

Raw data entry is separated from variable creation to enhance clarity and reduce errors.

A. Data View

- Works as a spreadsheet (Excel-style) similarity.
- Rows = cases/observations (such as a student, patient, or customer).
- Columns = variables (e.g. age, gender, marks, income).

• Example: In a student data set, row 1 may be “Student A” and column as his age, gender or score.

B. Variable View

• Indicates metadata or descriptions for each variable.

• Key attributes:

o Name: Name of variable (no spaces i.e., “Age” or “Exam_Score”).

o Type: Number, string, date or currency.

o Label: Description (e.g., “Student Age in Years”).

o Values: Coding of nominal scale (e.g., 1 = Men, 2 = Women).

o Missing: A missing value. Values that you want to treat as missing (for example, “999”).

o Scale: Nominal (categories), Ordinal (ranked), Scale (interval/ratio).

• Example: A “Gender” variable could be set as numeric, with 1 = Male, 2 = Female.

C. Menus

• Is situated at the top of the SPSS window and offers an organized way to access operation.

• Common menus:

o File: New, open, save and export data sets.% The Library is there to use but beyond it is for developer only-Name: How you would like to be referred -any Name o wget c git clone d hg a svn 5 RANCH Y Burns Unit TH forbid each compete postural drainage exercises billing machines.

o Data: Filtering, Merging and Joins on data sets.

o Transformation: Create new variables, and recode ‘old’ ones.

o Analyze: Get advanced statistical tools in addition to all of the features available in QCADesigner.

o Graphs: Build bar graphs, histograms, scatterplots, and pie charts.

o Utilities: Look at properties of a variable and dataset information’s.

D. Toolbars

• Include shortcut buttons for common actions like open file, save file, undo, redo, print and run analysis.

• Zoom in/out options and quick-access icons save time when menu scrolling is on.

3.1.2 File Management in SPSS – Creating, Opening, Saving, and Importing Datasets

Use SPSS to manage a variety of data sources.

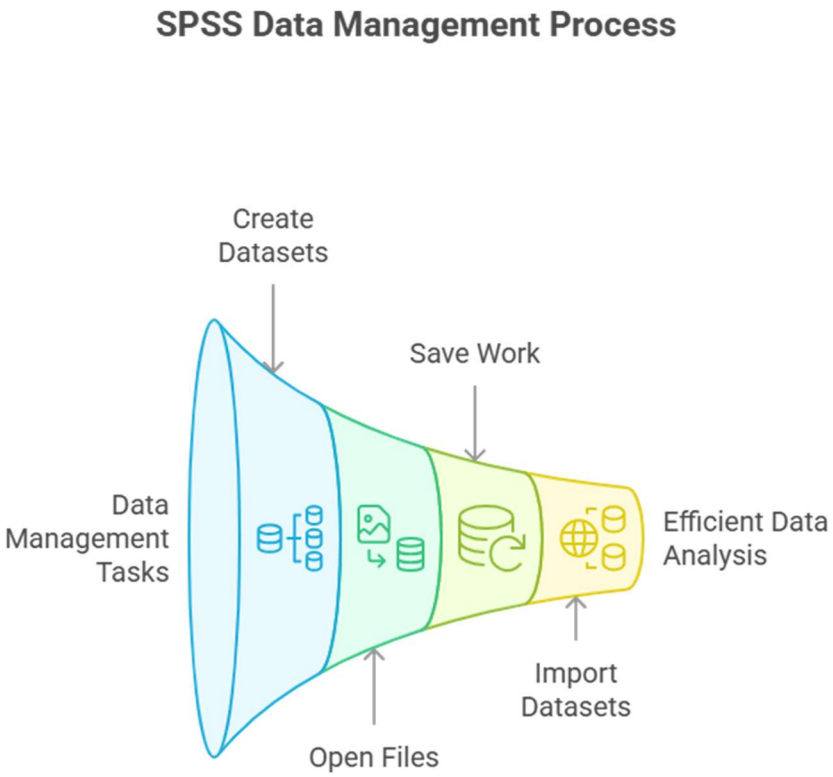

figure: File Management in SPSS

A. Creating Datasets

•Users have the ability to start a dataset from nothing.

• Variables are defined in Variable View and values entered in Data View.

• For example, a professor enters students' IDs, genders, and exam scores into SPSS.

B. Opening Datasets

• SPSS reads dataset files in..sav format in the following way: – Data > Sort data by ascending to open data; then navigating to the appropriate folder and picking up that file. sav format.

• Supports other formats like:

o Excel files (.xls,.xlsx).

o CSV (comma-separated values).

o Text files (.txt).

o Databases (e.g., SQL, Access).

C. Saving Datasets

• Data can be saved in. sav format while keeping all labels, types and metadata.

• Output can be saved independently as. spv files.

• Results can also be saved on Word, Excel and PDF for reporting.

D. Importing Datasets

• Data can be imported from Excel spreadsheets, web surveys or database links.

• Delimiters, variable names and formats are defined during import.

• Example: An organization imports its monthly sales data from Excel to SPSS for analysis.

Through organized transfer of the files, you'll have more reliable, easier sharing and better continuing of your research projects.

3.1.3 Navigation and Basic Operations – Entering Data, Defining Variables, and Using Dialog Boxes for Simple Analyses

SPSS is menu driven and does not require programming for statistical analysis.

A. Entering Data

• Data is manually entered into Data View (much like Excel).

• One row = one observation (eg, respondent on a survey).

• Each column = a different variable (e.g., age, gender, income).

• Example: A survey is completed by a researcher, creating one new row at a time.

B. Defining Variables

• To ensure proper analysis, variables must be defined in Variable View.

• Steps:

o Give the variable a name (for example, 'Age').

o Select type (numeric, string, etc.).

o Type a label (e.g., “Age in Years”).

o Create value labels (i.e., 1 Yes, 2 No).

o Record missing values if appropriate (e.g., “999=Missing”).

o Choose type of measurement (nominal, ordinal, scale).

• Example: For a 'Satisfaction' variable, specify levels as 1 = Very Dissatisfied to 5 = Very Satisfied.

C. Application to Simple Analyses by the Use of Dialog Box

• Dialouge boxes in SPSS for tests of significance, descriptive statistics and graphs.

• Example Operations:

o Frequencies: Convert categorical responses into numbers (e.g., male vs. female).

o Descriptives: Mean, median, mode, variance and standard deviation.

o Charts: Create bar charts, pie charts, and histograms.

• Output is presented in the Output Viewer with tables and graphs for easy interpretation.

Dialog boxes eliminate the need for programming, so SPSS is fun or beginner-friendly but allows you to do advanced analysis as an expert.

Did You Know?

"In SPSS, correctly defining variables in Variable View is just as important as entering data in Data View. Without proper labels, value codes, and measurement levels, the software may misinterpret variables, leading to inaccurate results. Dialog boxes simplify analyses, letting even beginners perform complex statistics with just a few clicks."

3.2 Installation of SPSS Software

3.2.1 Installation of SPSS Software – Student / Trial Version Download from IBM SPSS (Elaborated)

Described further a set of detailed step-by-step information that includes how IBM SPSS student/academic access typically functions, what to check, and how you should proceed. And places where screenshots are useful are indicated; you'd pop in pictures there.

• Certify whether your institution is having a site-license or campus license Many universities/colleges buy SPSS with the help of campus edition or site license for IBM SPSS. …, students usually get to access for free (or it's subsumed into their tuition) or through local university labs and computers.

• Know what "Student / GradPack / Academic Version" means IBM has "GradPack" things are just less expensive if you are a student.

The feature editions have Base, Standard and Premium versions with an increasing level of module based support. Base version also comes with basic stats.

• Be aware if there is a free trial offering

IBM provides 30 days of free use (as a desktop version) so students get the opportunity to experiment with full-featured software for a month.

Step-by-Step Instructions (If You Want to be Free / Academic / Trial Access)

This is how other people does it install the SPSS student / trial version if you got either institutional access or trial access without paying lots of money. (If it's on order via your school or employer, or if IBM is billing you for the software, then you will take similar steps, but with payment in play.)

Procedure 1) Login into the Institution or Academic Portal/Internal IBM site

• Go to your university's academic software portal (like OnTheHub, campus software store, or the academics page of SPSS GradPack vendor) and log in as a student.

• Or Go to the SPSS IBM website → search for "GradPack" or Suche für "Student / Academic".

Step 2: Confirm Eligibility

• Upload necessary verification: student ID, registration, or login.

• Verify whether the version is free (through your institution) or a free trial. If trials: indicate the number of days (usually 30).

Step 3: Choose the Edition and Click Install Choose the edition of Windows (Windows 7/8/10) that you want to install, then click Next.

• If you have options (Base / Standard / Premium), select the edition of SPSS that meets your requirements. And the Base version often is enough for many course needs.

• Select platform: Windows or Mac. Also check if available, or 32-bit.

• Download the installer from IBM, or academic vendor portal. ¬ Screenshots here: The download page with version options.

Step 4: System Requirements

• Make sure your computer has the minimum hardware requirement: enough RAM (4-8GB often), disk space (a few GB), compatible operating system version (Windows 10/11, recent MacOS).

• If you have admin rights (Windows) you should close other programs.

Step 5: Run Installation

• Launch the installer. Accept the license agreement. Pick your install directory (Defaults are fine for most).

• Select components/modules you need. Yes, in Student/Trial Version some of the higher end module will not be available.

• Follow through steps; installer copying files, perhaps configuring license wizard. ¬ Add screen shot: Installation wizard first page.

Step 6: Activate / Authorise License.

• If it is a trial following installation you may only need to accept a trial license or login with your IBM id.

• If you have been given an authorization code / license key (by your institution or a software vendor), click on "Authenticate by License key" (or equivalent) and enter the code.

• You will need to be online so that SPSS can validate the license with IBM's servers.

Step 6: License Activation / Authorization License.

• If it is a trial version, following the installation you may want to simply accept a trial license or use IBM id sign in.

• For authorized user licenses (license key / authorization code from institution or vendor) select "Authorized user license" (or similar) and enter the code.

• Internet connection to be able to authorize the license with IBM's servers.

Step 7: Launch and Verify

• Now launch your SPSS application(in the start Menu for windows, Applications on mac).

• Navigate to Help → About SPSS Statistics. It shows you the Version, License (Trial is academic), and expiration date if trial. ¬ Enter-screen capture: 'About SPSS' dialogue with license status.

Step 8: Post-installation Setup

• Install IBM Updates or Patches if necessary.

• If missing modules: see if they are included in your edition. Some sections may need to be turned on individually.

• For time-limited licenses, keep the expiration date in mind; when it arrives, renew through your institution if possible or switch to another deal.

Important Notes & Caveats

• There are also feature restrictions, even in the academic versions. Some high-end statistical modules (complex modelling, big data connectors etc.) could not be included in the student edition.

• Free trial = limited duration. Software is deactivated and no longer active at the end of the trial period if you have no current license.

• Often schools will give out share lab or remote VM access to fully licensed copy of SPSS and students do not install their own.

• Always check the policy of your university / department: sometimes there is no charge to students, some a nominal one, some full but much cheaper than from software organisations.

"Activity: Example Use Case for Students"

A student downloads and installs the SPSS student version to analyze a survey dataset. After defining variables in Variable View and entering responses in Data View, they use the Analyze → Descriptive Statistics menu to calculate means, frequencies, and generate histograms. This illustrates the importance of correct installation — without a properly licensed copy, the software cannot process datasets or produce results.

3.3 Data Objects and Attribute Types

3.3.1 Definition of Data Objects

The simplest join is the cartesian join or cross-join, which contains all permutations between two data objects. Every object represents a temporal entity and holds values for a series of attributes.

• Structure of a Dataset:

o Rows → Data Objects (what are being studied).

Columns → Properties (features about these entities).

• Examples:

o In a collection of student performances: Each performance is a data object. Features can be Student ID, Gender, Age, GPA and Exam Scores.

o In A hospital dataset: Each patient is an object. Features can be Patient ID, Age, Diagnosis, Blood Pressure or Treatment.

o In a retail data: A each sale is one data object. An example for attributes is the transaction identifier, the product, quantity and price.

Accordingly, data objects are what or whom we analyze, whereas attributes refer to the characteristics of interest.

3.3.2 Types of Attributes – Nominal, Ordinal, Interval, Ratio

Key characteristics can be categorized on the basis of levels of measurement. This level of hierarchy serves to establish what kind of numerical operations and formal statistics are sensually applicable.

Nominal Attributes

• Definition: List categories that have no intrinsic order.

• Nature: Classificatory as labels, qualitative nature.

• Permitted: Mode; Counting; Frequency Analysis; Cross tabulation\Security.

• Example:

o Gender: Male, Female, Other.

o Marital Status: Single, Married, Divorced.

o Blood Group: A, B, AB, O.

Ordinal Attributes

• Description: They symbolize categories with a meaningful order, where the distances between them are unknown.

• Nature: Rank-based qualitative data.

• Permitted Operations: Median, percentiles, non-parametric testing (Mann Whitney U, Kruskal-W allis).

• Example:

o Customer Satisfaction: 1 = Low, 2= Medium, 3 = High, and 4 = Very High.

o Socio-economic Class: Low, Middle, High.

o Qualifications: 0-Primary School, 1-High School Graduate, 2-College Grad.,3-Grad./Professional.

Interval Attributes

• Definition: Numeric values on which differences are computed and on which a zero point is not natural.

• Nature: Quantitative, but ratios not meaningful.

• Operations Available: Mean, S.D., correlation, regression.

• Example:

o Temperature (Celsius, Fahrenheit).

• Dates in a calendar (years apart is significant but the "zero year" is an arbitrary choice).

Ratio Attributes

"Ratio level of measurement ,A type of measure that possesses all the characteristics of an interval measure together with a zero point - an absolute reference point.

• Nature: Full mathematical models were conducted no quantitative restrictions were imposed.

• Eligible Operations: Any statistical operations, including the geometric mean and coefficient of variation.

• Example:

o Higher (170 cm is two times higher than 85 cm).

o Weight (60 kg is double that of 30 kg).

o Income, Age, Distance.

Hierarchy of Measurement:

Nominal < Ordinal < Interval < Ratio Note: The interval order here seems wrong, I'll check that.

The higher levels include the properties of the lower levels, with additional features.

3.3.3 Discrete vs Continuous Attributes

We may also classify attributes according to the type of values such as countable or measurable.

Discrete Attributes

• Definition: Features which can have finite or countable number of values.

• Nature: Often integers or categories.

• Examples:

o Number of children in a home (0, 1, 2 …).

o Own no car / 1 car / 2 cars (0, 1, 2…).

o Shoe Size (which is numerical, but it is a fixed category).

• Analysis: Frequency, bar charts; chi-square.

Continuous Attributes

• Attributes that can assume an unlimited number of values within a given range to arbitrarily fine levels.

• Out in the Natural world: Quantifiable, can easily be written as decimal numbers.

• Examples:

o Height (cm,, etc.).

o Weight (62.5 kg).

o Temperature (36.6°C).

• Statistical Software: Mean, SD, Pearson correlation, multiple regression analysis and t-tests.

Key Difference:

• Discrete = Values that can be counted (finite categories, gaps between values).

• Continuous = You can have an infinite number of values within a range (no gaps, and it is measured in fine detail).

Knowledge Check 1

Choose the correct option:

1. A data object in SPSS is usually represented as:

a) Column

b) Row

c) Cell

d) Variable

2. Blood group (A, B, AB, O) is an example of which attribute?

a) Nominal

b) Ordinal

c) Interval

d) Ratio

3. Temperature in Celsius is classified as:

a) Nominal

b) Ordinal

c) Interval

d) Ratio

4. Height of a person is an example of:

a) Discrete attribute

b) Continuous attribute

c) Nominal attribute

d) Ordinal attribute

3.4 Basic Statistical Descriptions of Data

3.4.1 Measures of Central Tendency (Mean, Median, Mode)

Measures of central tendency tell us a value that represents the middle or centre of our data.

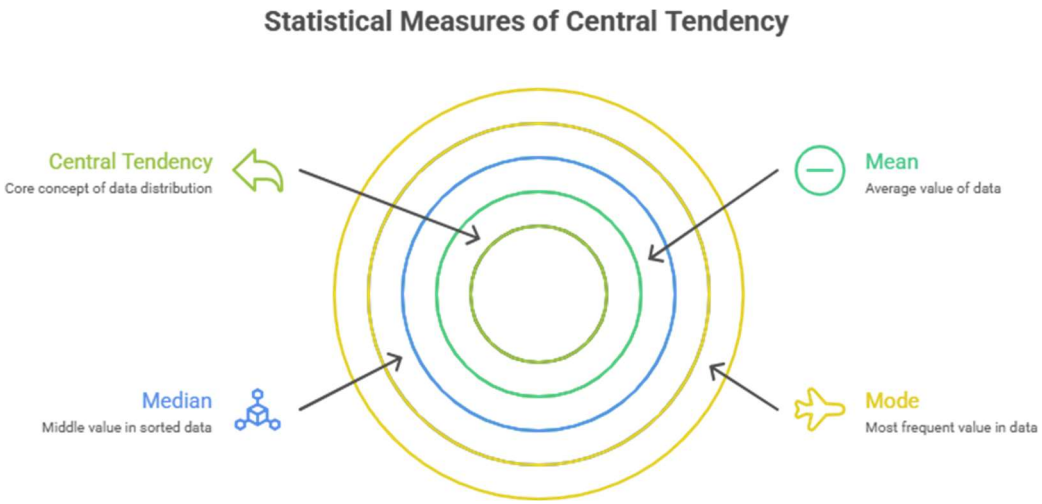

Figure: Measures of Central Tendency

A. Mean (Arithmetic Average)

• Mean: Average of all values.

• Pros: Includes all data values; the statistic is frequently used in inferential statistics.

• Disadvantages: sensitive to extreme values (outliers).

Example: Exam scores (40, 50, 60, 95). Average (40+50+60+95)/4 = 61.25.

B. Median (Middle Value)

• Definition: A middle number in a set of data when the objects are listed in order from least to greatest. If there are an even number of values, it is the average of the two middle numbers.

• Pros: No influence of outliers.

• Drawbacks: No overall scale for remaining data points.

• For example: Exam scores = {40, 50, 60, 95}. Median =(50+60)/2 = 55.

C. Mode (Most Frequent Value)

• Definition: A value that occurs most frequently in a data set.

• Pros: Appropriate for categorical/nominal data.

• Demerits: A set of data can lack a mode, have one mode (unimodal), or more than one mode (multimodal).

• For example: The blood groups of a class = {A, O, O, B, AB,O}. Mode = O.

In SPSS:

• Path: Analyze, Descriptive Statistics , Frequencies/Descriptives.

• SPSS creates table that reports mean, median, and mode for variables specified by the user.

3.4.2 Measures of Dispersion (Range, Variance, Standard Deviation)

Measures of dispersion quantify spread from the center. Two sets of data could have the same average and yet vary widely.

A. Range

• Definition: The maximum minus the minimum.

• Formula: Range = Max – Min.

• Advantages: Easy to calculate.

• Disadvantages: Affected heavily by outliers.

• Example: Ages = {18, 19, 20, 22, and 25}. Range = 25 – 18 = 7.

B. Variance

• Definition: The mean of the squared differences from the mean.

• Advantages: Considers all data points.

• Cons: Not directly interpretable, measurements are in squared units.

• For example, if exam scores were {50, 60, 70}, then variance = 66.67.

C. Standard Deviation (SD)

• Definition: The square root of variance; indicates the average deviation between values and the mean.

•Advantages: Presented in the same units of measurement as the data; commonly used in research.

• For example: If exam scores = {50,60,70}, SD = 8.16.

• Interpretation: A low standard deviation indicates that the data are close to the mean, a high standard deviation means the data are widely spread.

In SPSS:

• Path: Analyze → Descriptive Statistics → Descriptives → Choose “Std. Deviation.”

Did You Know?

“The standard deviation is the most widely used measure of dispersion in research. Unlike the range, which only considers extremes, or variance, which uses squared units, standard deviation expresses variability in the same units as the data. This makes it easier to interpret how spread-out values truly are.”

3.4.3 Data Distribution: Skewness and Kurtosis

Apart from centrality and dispersion, the form of the distribution assists in knowing if data is normal, skewed or peaked.

A. Skewness

• Definition: Quantifies the asymmetry of a distribution.

• Types:

o Positive Skew (Right skewed): Extended right tail. Mean > Median. Example: Income distribution.

o Negative Skew (Left-skewed): Tail of the data points off to the left. Mean 3): High peak, thick tails. More extreme values. Example: Stock returns.

o Platykurtic (K<3): The distribution is flat with lighter tails. Example: Uniform-like distributions.

o Mesokurtic (K = 3): The standard distribution which we observe.

• SPSS interpretation: Kurtosis near to 0(normalized) is normal data.

In SPSS:

• Path: Analyze → Descriptive Statistics → Descriptives → Click “Skewness” and “Kurtosis.”

• Provides output of skewness and ka aggregation with their standard errors.

3.5 Summary

- ❖ SPSS is an excellent statistical package that offers user-friendly environment for manipulation, visualization and analysis of data.
- ❖ The SPSS system working interface contains Data View for data input of cases, Variable View for defining features and menu/toolbars type of analysis.
- ❖ ϖ You can organize your work through file management in SPSS; you have the options to create, saver, open and import datasets for example:.sav;.axles. sav, Excel, and CSV.
- ❖ Guides the user with dialog boxes, to do an analysis such as a frequency or a descriptives and graphs without coding needed.
- ❖ Entities are represented by data objects, and their properties are described in terms of nominal, ordinal, interval or ratio attributes.
- ❖ Attributes can be discrete (value is countable) and continuous (the values are measurable), its type will depend on which statistical tests that might be used.
- ❖ Measures of central tendency (mean, Median, Mode): These expresses the typical value of a data set.
- ❖ Descriptive statistics (range, variance, standard deviation) describe how spread out values are from the average.
- ❖ Skewness and kurtosis are parameters of the shape of the distribution (they define how much it is asymmetric and peaked, respectively compared to a normal distribution).

3.6 Key Terms

1. SPSS - A Program or Software for performing statistical analysis as well as data management and graph creation.
2. Data View - A SPSS window that is used to enter and view raw data (cases & variables).
3. Variable View – The SPSS window that allows the creation of variables (name, type, label, measure).
4. Data Object – A data record such as row of a given dataset which represents an entity or an observation.
5. Nominal Attribute – A variable for which there is no inherent order (eg., sex, blood-type).
6. Ordinal Attribute: An attribute which has ordered categories, but the difference between the ranks is not known (like ranked data -1st rank vs 2nd rank)eg: levels of satisfaction.
7. Interval Attribute – Numerical variables with meaningfully ordered differences but no true zero (e.g., temperature in celsius).
8. Conditional Ratio Attribute – A numeric attribute with a true zero and equal intervals in which ratios can be meaningfully computed (e.g., weight, income).
9. Standard Deviation – A measure of dispersion that illustrates the typical distance between the values in a distribution and its mean.

3.7 Descriptive Questions

1. Describe the Elements of the SPSS Interface. Segments often lead to more refined equations that are accurate or at least usable based on your research question. What are the functions of Data View and Variable View?
2. Explain the process for building, saving, and importing datasets in SPSS.
3. How do we define variables in SPSS? Explain using examples of nominal, ordinal, interval and ratio scale type attributes.
4. Distinguish between discrete and continuous attributes. Provide examples from real-world datasets.
5. Define measures of central tendency. What are the differences between mean, median and mode in interpretation and application?
6. Describe measures of dispersion and provide examples, with particular reference to range, variance and standard deviation. why does so many analyses use the standard deviation?
7. Explain skewness. "Inference About the Median Under Dependent Alternation of Positively and Negatively Skewed Distributions. And how positive and negative skewness influence our interpretation of data?
8. Define kurtosis. Distinguish between leptokurtic, mesokurtic, and platykurtic distributions Give examples.

9. Explain why get moving is case sensitive. What is the difference between ascending and descending order? 7) What happens when you use AutoSum in a sensitive document? Explain how would SPSS obtain ideas to write descriptive statistics? What actions can you take when descriptions do not make sense? What are the typical menu paths to get the mean, median, mode, range, variance, standard deviation and skewness & kurtosis?

3.8 References

1. Pallant, J. (2020). SPSS Survival Manual: A Step by Step Guide to Data Analysis Using IBM SPSS (7th ed.). McGraw-Hill Education.

2. Field, A. (2018). Discovering Statistics Using IBM SPSS Statistics (5th ed.). SAGE Publications.

3. George, D., & Mallery, P. (2022). IBM SPSS Statistics 28 Step by Step: A Simple Guide and Reference (18th ed.). Routledge.

4. Brace, N., Kemp, R., & Snelgar, R. (2016). SPSS for Psychologists (6th ed.). Routledge.

5. Landau, S., & Everitt, B. S. (2004). A Handbook of Statistical Analyses using SPSS. Chapman & Hall/CRC.

6. Verma, J. P. (2015). Data Analysis in Management with SPSS Software. Springer.

7. Sarstedt, M., & Mooi, E. (2019). A Concise Guide to Market Research: The Process, Data, and Methods Using SPSS Statistics (3rd ed.). Springer.

Answers to Knowledge Check

Knowledge Check 1

1. b) Row

2. a) Nominal

3. c) Interval

4. b) Continuous attribute

3.9 Case Study

"Analyzing Customer Profiles Using SPSS"

Introduction

Ring has engaged in many business practices, such as extracting demographic information regarding customers and retrieving the personal information of those who knew a Ring customer, to analyze and market to its "customer base. Demographic data sets, though, typically contain more than one attribute – e.g., age, gender, income and purchase preference – that must be systematically treated. 2-5 SPSS is a friendly software for declaration of variable types, presenting descriptive time to predict the dynamic communication, (author) extremum and plotting customer manoeuvre cases marking marketer's life style.

Background

An online retailer polled 500 customers to better understand what drives their demographic and buying decisions. The dataset included:

• Customer ID (Nominal)

• Gender (Nominal)

• Age Group (Ordinal)

• Monthly Income (Ratio)

• Satisfaction Rating (Interval)

• Annual Spending (Ratio)

The raw data was entered into a Microsoft Excel database and then analyzed using SPSS. In Variable View, the investigator characterized attributes by labeling, coding gender (1 = Male, 2 = Female), and scale levels. The row-by-row responses of customers were organized in Data View. Mean income, median spending, mode of age groups and standard deviations were computed by the firm with Descriptive Statistics on SPSS for continuous variables. Frequency tables and histograms were used to determine level of satisfaction for their customers and income data.

Problem 1: Segmentation approach by customer demographic variables Our first problem is about segmentation analysis in demographics.

The original dataset was problematic for analysis in that variable naming codes were inconsistent and categorical variables like sex were not coded.

Solution: The calculation was made with SPSS Variable View defining the characteristics, inserting labels and coding values (1 = Male, 2= Female). This made the customer attributes more clearly displayed in output tables.

MCQ:

What SPSS window do you use to specify the properties of a variable (e.g. name/variable type/labels?

a) Data View

b) Output Viewer

c) Variable View

d) Chart Editor

Answer: c) Variable View

PROBLEM#2: Condensing Key Customer Testimonials Effectively FROM THE TRENCHES The following day, I leveraged short testimonial videos from thrilled customers on Facebook.

Who would have the time to manually work out averages, and spend patterns across 500 people.

Resolution: By using the analyze → Descriptive Statistics, in a few seconds summaries of the mean income and median spending were made, mode of age group can also be reported as well as variance and standard deviation. This allowed rapid classification of averages among the different customers.

MCQ:

Where in SPSS can you get desciptive statistics such as mean and standard deviation?

a) Graphs

b) Analyze

c) Transform

d) File

Answer: b) Analyze

Problem 3: Understanding Distribution of Customer Expenditure and Satisfaction

The customer-spending data was spotty because a few high spenders caused outliers. It was impossible to understand without a visual.

Solution: The author produced histograms and skewness for values in SPSS. The result was a spending distribution that clearly broke to the right, meaning that most patrons spent relatively little on their visit and yet some high spenders accounted for a far larger part of revenue.

MCQ:

If a data set exhibits a long tail to the right, then what type of skewness does the data have?

a) Negative skew

b) Positive skew

c) Normal distribution

d) Platykurtic distribution

Answer: b) Positive skew

Conclusion

This example will show you how SPSS turns raw customer demographics into actionable marketing strategy. Through the proper description of attributes, application of descriptive statistics, and spending pattern visualization, corporations can target appropriate customer groups, segment the market, as well as optimize marketing activities. SPSS saves the manual work and allow us for accurate, efficient decision which leads to high quality of customer analytics capabilities.

BUPBM Unit 4 V3.docx

Building useful Predictive Business models_BBA_3

Building useful Predictive Business models_BBA_3

ATLAS SkillTech University

Document Details

Submission ID

trn:oid:::3618:128374762

Submission Date

Feb 16, 2026, 12:13 PM GMT+5:30

Download Date

Feb 16, 2026, 12:17 PM GMT+5:30

File Name

BUPBM Unit 4 V3.docx

File Size

146.7 KB

15 Pages

3,302 Words

18,997 Characters

*% detected as AI

AI detection includes the possibility of false positives. Although some text in this submission is likely AI generated, scores below the 20% threshold are not surfaced because they have a higher likelihood of false positives.

Caution: Review required.

It is essential to understand the limitations of AI detection before making decisions about a student's work. We encourage you to learn more about Turnitin's AI detection capabilities before using the tool.

Frequently Asked Questions

How should I interpret Turnitin's AI writing percentage and false positives?
The percentage shown in the AI writing report is the amount of qualifying text within the submission that Turnitin's AI writing detection model determines was either likely AI-generated text from a large-language model or likely AI-generated text that was likely revised using an AI paraphrase tool or word spinner.

False positives (incorrectly flagging human-written text as AI-generated) are a possibility in AI models.

AI detection scores under 20%, which we do not surface in new reports, have a higher likelihood of false positives. To reduce the likelihood of misinterpretation, no score or highlights are attributed and are indicated with an asterisk in the report (*%).

The AI writing percentage should not be the sole basis to determine whether misconduct has occurred. The reviewer/instructor should use the percentage as a means to start a formative conversation with their student and/or use it to examine the submitted assignment in accordance with their school's policies.

What does 'qualifying text' mean?
Our model only processes qualifying text in the form of long-form writing. Long-form writing means individual sentences contained in paragraphs that make up a longer piece of written work, such as an essay, a dissertation, or an article, etc. Qualifying text that has been determined to be likely AI-generated will be highlighted in cyan in the submission, and likely AI-generated and then likely AI-paraphrased will be highlighted purple.

Non-qualifying text, such as bullet points, annotated bibliographies, etc., will not be processed and can create disparity between the submission highlights and the percentage shown.

Unit 4: Data handling

Learning Objectives

1. Understand the concept of data handling – Explain what data handling means and why it is essential in organizing and interpreting information.

2. Identify different types of data – Distinguish between qualitative and quantitative data, primary and secondary data, and raw and processed data.

3. Collect and organize data effectively – Learn various methods of collecting data (e.g., surveys, observations, experiments) and present it systematically using tables.

4. Represent data using diagrams and charts – Demonstrate the ability to construct and interpret bar graphs, pie charts, histograms, and line graphs.

5. Apply measures of central tendency – Understand and calculate mean, median, and mode as tools to summarize data.

6. Interpret and analyze graphical data – Develop skills to read, compare, and draw conclusions from different data representations.

7. Develop problem-solving skills through data handling – Apply data handling techniques to solve real-life problems and case-based questions.

8. Recognize the importance of accuracy and reliability – Understand errors in data collection, bias in surveys, and the need for accurate data representation.

9. Relate data handling to practical contexts – Connect concepts of data handling to everyday life situations, business decision-making, and academic research.

Content

4.8 Case Study

4.0 Introductory Caselet

"Decoding School Attendance Data"

Green Valley School wanted to analyze the attendance rate of students from various classes. The principal observed that even though attendance was a concern in general it looked all fine but they found some classes having absentees often on particular days of the week. To get a better sense of this, the school gathered attendance records for one month.

Daily observations were collected and summarized in table. It was also noted for instance that attendance in Class VIII on Mondays stood at 95 percent, while it plummeted to 78 percent on Fridays. Class IX had an almost stagnant average throughout the week around 88%, while Class X presented a wider spread between days - ranging from 96% on exam days to only 70% on the day after a school event.

In order to make sense of this, the data was presented in bar graph and line chart for teachers to recognise patterns. The principal also employed the averages (mean and mode) to measure attendance across various classes and days.

This exercise enabled the school to make data-informed decisions:

•Incentivizing students with incentives for sustained attendance.

• Planning high impact events around lower traffic days.

• Providing counselling for students who have erratic attendance.

With the right data capture, organization and presentation, the school transformed raw numbers into useful information on how to improve scholastic participation.

Critical Thinking Question:

Imagine you were one of the members of the school management team: What other data would you like to have about these students that would provide greater insight into the problem and how would it be represented in order to make a good, informed decision about influencing (changing) student attendance behavior?

4.1 Data Visualization

4.1.1 Importance of Visualization in Data Mining

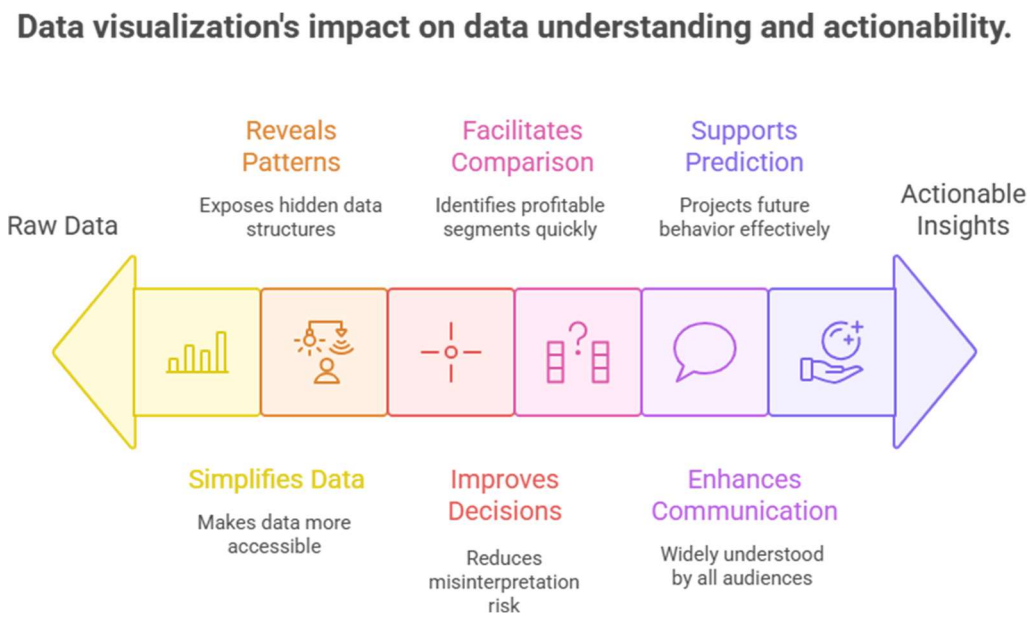

Figure: Data Visualization

Simplifies Complex Data

Raw data sets tend to be huge, unorganized and very hard to understand. There are visualisation tools like graphs and plots which simplify these data sets, and make them user-friendly.

o For instance, 10000 records of sales data in a spreadsheet is unmanageable—but a line graphic that depicts monthly sales trends makes it meaningful in seconds.

Reveals Hidden Patterns

o Data mining searches for valuable relationships, patterns that are too complex to be detected in tabular data, or anomalies.

o Visualization makes these latent structures be observable. For instance, clustering in a scatter plot could uncover groups of customers exhibiting similar buying patterns.

Improves Decision-Making

o Decision-makers need clear, actionable insights. Graphics are useful for the quick comprehension of data and to avoid misunderstanding.

o For example, a business making decisions on where to invest and where not can use a heatmap to spot high performing and low performing areas.

Facilitates Comparison

o Visualize, easily compare across groups, by categories or time frames.

o Example: A bar chart comparing profits for different product categories reveals the best and worst performing segments at a glance.

Enhances Communication

o Visualizations are broadly accessible; they make sense to non-experts also.

o This makes them great for reporting, sharing and cross-functional cooperation.

Supports Predictive Analysis

o Insightful visualization.,Äù ==> Visualization is good at revealing trends from the past, and giving insight to predict future.

o Example: A time-series plot of stock exchange prices aids planners in predicting future price changes.

4.1.2 Types of Visualizations

a) Histograms

• Definition: A histogram is a bar graph that displays the frequency of which data values fall into certain ranges (sometimes called bins).

• Structure:

o The intervals of the data, i.e., score ranges, are shown on the x-axis.

o The frequency (how many times the values fall in that interval) is shown on the y-axis.

• Purpose:

o To see what the data distribution looks like: normal, skewed or uniform.

o A sense of concentration of points of data and variability.

• Example If we take an exam and have 50 students, then record the marks of 50 students which are given in ranges like (0-10), (11-20), (21-30) etc a histogram will tell us how many students scored in each range.

• Educational Insight: Histograms are most useful in an area such as statistics and quality control where shape of distribution (bell curve, skewed curve) is important for analysis.

b) Scatter Plots

• Definition: A scatter plot is a graph that shows the values of two numerical variables; the values are represented on a set of axes using dots.

• Structure:

o The x-axis represents one variable.

o The y-axis represents another variable.

o Each point in the graph represents a value pair.

• Purpose:

o To find out the association (positive, negative and no correlation) amongst two variables.

o Inspection of outliers or any odd values.

• For example each point in the plot may represent the number of “hours studied” (x axis) and corresponding “marks obtained" (y axis) for a group of students, in which case you may observe positive correlation: More you study better the marks.

• Educational Insight: “Scatter plots are an important tool in academic research, economics and business to focus on what is driving a change -for example is our advertising spend leading to higher revenue."

c) Boxplots (or Whisker Plots)

• Definition: A boxplot is a visual representation of the distribution of a dataset or individual categorical group using five values summaries: Minimum, First quartile (Q1), Median, Third quartile(Q3), Maximum.

• Structure:

o The middle box shows the interquartile range, IQR (the middle 50% of scores).

o The median (middle number) is marked with the line in the box.

o The “whiskers” range up to the lowest and down to the highest non-outlying value.

o Outliers are presented as individual points outside the whiskers.

• Purpose:

o To comprehend variation aberration and spread of data.

o To determine which ones are outliers or compare how distributions do differ across groups.

• For example, a boxplot of monthly salaries among three departments will quickly reveal which department has the most spread in monetary pay levels, where the central value estimate is for each group and if hiring exists above or below typical gross earnings.

• Educational Insight : Boxplots are common in many business analytics, test environment and research where outliers should be given extra attention.

“Activity: Exploring Data with Visuals”

Collect marks of 20 students in Mathematics and record them in a table. Create a histogram to show the distribution of marks, a scatter plot to compare study hours with marks, and a boxplot to identify outliers. Discuss patterns and relationships visible from each visualization.

4.2 Data Distribution

4.2.1 Understanding Frequency Distributions

• Definition:

A frequency distribution is a particular tally that shows how often certain values, or ranges of values, occur in a collection. Commonly included as a table, histogram or bar chart.

• Structure:

Class Intervals (Bins): Continuous intervals in which data values are grouped (e.g., 0–10, 11–20).

Frequency: How many of the data values fall into each interval.

Refer: Ratio of number of data values in each interval to total.

Dragging Scatter Plot: Plots the Scatter chart and on dragging, it displays the trend of the plotted ILI values with respect to time. QRSTUV Cumulative Frequency: The running total of frequencies in a class for each interval.

• Example:

That is, you have all the test scores of 40 students. By breaking up into intervals of 10 (0–10, 11–20, etc.), the frequency distribution tells you how many students are in each range. This quickly shows whether most students scored above average, high or low.

• Importance:

Through frequency distributions, massive amounts of data can be reduced and organized in a simple form, to reveal main trends, variability and exceptions.

4.2.2 Measures of Central Tendency (Mean, Median, Mode)

Central Tendency Measure of central tendency describes the “center” or general value within a dataset.

Mean (Arithmetic Average):

o Formula:

o Strength: Utilizes every data value.

o Weakness: Not robust to outliers.

o Example: Mean of {5, 10, 15} = (5+10+15) ÷ 3 =30 ÷ 3=10.

Median (Middle Value):

o Puts data in order and finds the middle.

o If the data is even, then find median = average of middle two values.

o Advantage: Not influenced by outliers.

o Example: For {5, 7, 9, 15, 18}, median is equal to 9.

Mode (Most Frequent Value):

o The number that appears most frequently in the set of numbers.

o Possibly more than one (bimodal/multimodal).

o Example : {2,4,4,6,8,8,8} Mode = 8.

• Application:

o Mean: Ideal for data that is not heavily tailed.

o Median: Ideal for skewed distributions (e.g., income).

o Mode: Appropriate for categorical data (e.g., most preferred product).

Did You Know?

“Did you know that the mean, median, and mode are not always equal? In a perfectly symmetrical distribution, they overlap at the same point. However, in skewed data like income levels, the mean is pulled by extreme values, while the median and mode remain closer to typical observations.”

4.2.3 Measures of Dispersion (Range, Variance, Standard Deviation)

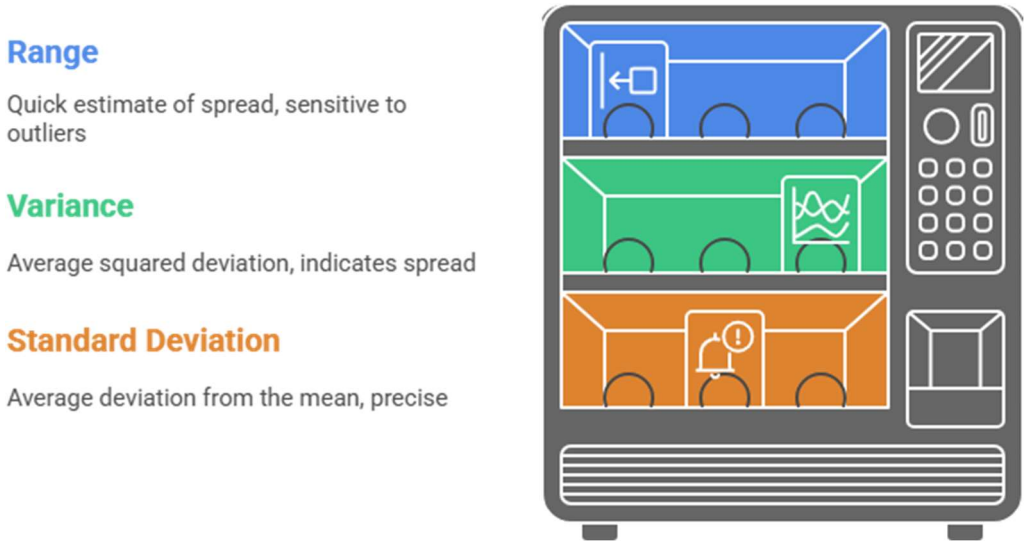

Figure: Measures of Dispersion (Range, Variance, Standard Deviation)

"Central tendency gives you a middle point, whereas a measure of dispersion explains how spread out your data is.

Range:

o Formula: Highest value – Lowest value.

o Gives a rough measure of spread, but is very sensitive to outliers.

o For Example: Let us take range of {4,7,10,20} = 20 – 4 = 16.

Variance:

o Estimates mean squared (average squared) deviation for each value.

o Formula:

o Larger variance = greater spread.

Standard Deviation (SD):

o Root of variance, with the same unit as data.

o Means deviation from the mean.

o Example: If students scores have little spread about the mean, the SD is small. This way, the greater the differences amongst scores, the higher SD becomes.

• Application:

o Range: Quick estimate of spread.

o Variance / SD: Recommended for accurate statistical analysis (especially in scientific research, business and quality control)

4.3 Relationship among Variables

4.3.1 Covariance

• Definition:

A measurement of how two variables change together UI Alert Covariance. It tells you whether when one variable is higher, the other is normally higher or lower.

• Formula:

• where the xi's and yi's are observations, are their sample means.

• Interpretation:

o Positive Covariance: One variable tends to increase as the other increases.

o Negative COVARIANCE – the covariance between the other one&s, inversely increased as the first increases.

o Zero covariance: There is not a consistent relationship of change between the two variables.

• Example:

• If X and Y are positively cosed, it means that as hours of study increase, test scores tend to increase.

o If price covaries negatively with demand, then the higher the price, the lower is the quantity demanded.

• Limitation:

Covariance only reveals the direction of the relationship, and no such strength of the connection. It is possible for two factors to have identical covariance but very different grades of association.

4.3.2 Correlation: Positive, Negative, and Zero

• Definition:

Correlation is a standard measurement of how two variables (direction, strength, and variability) are related to each other. Correlation is a unit less measure and it lies between -1 to 1 unlike covariance.

• Formula (Pearson's r):

• Types of Correlation:

Direct Correlation: The two variables move in the same direction.

♣ For instance: Height and weight—it is frequently the case that taller people weigh more.

An Increased negative relationship: A rise in one, leads to a decrease in the other and vice versa.

♣ Example: Car speed and travel time-higher speeds leads to lower travel time.

Zero relationship; No correlation between 1,2 and 3.

♣ For example: Shoe size and exam performance – unrelated items.

• Advantages:

Correlation does not just tell us direction, but it also tells us how strong of a relationship there is. For example, r=0.85r = 0.85r=0.85 represents a strong positive relationship and r=0.20r = 0.20r=0.20 shows a weaker one.

4.3.3 Coefficient of Determination

• Definition:

The R2 (coefficient of determination) measures the proportion of the variance in dependent variable that is predictable from independent variables.

• Formula:

for a one predictor model. In the case of multiple regression, the computer calculates R2 by means of sums of squares.

• Interpretation:

Application:

• In regression models to test fit.

• The higher the R2 value, the more explanatory power it has, but very high values may signify overfitting if too many variables are included.

Knowledge Check 1

1. Covariance between two variables primarily indicates:

a) Strength of relationship

b) Direction of relationship

c) Standard deviation

d) Variance

2. Correlation coefficient (r) always lies between:

a) –10 to +10

b) 0 to 1

c) –1 to +1

d) 0 to 100

3. If correlation between X and Y is 0, it means:

a) Perfect positive relation

b) Perfect negative relation

c) No linear relation

d) Strong relation

4. Coefficient of determination (R^2) is calculated as:

a) r ÷ 2

b) r^2

c) 2r

d) √r

4.4 Summary

- Data distribution – tells spread of the values in a dataset and is used to find out trends, cluster and pattern.
- Frequency distribution helps in summarizing a given raw data by classifying it into intervals, corresponding frequencies and also cumulative frequencies.
- Central tendency measures (mean, median, mode) give information about the location of data in a dataset.
- Mean is the average value, median is central and mode of the most observation.
- The statistics of dispersion (range, variance and standard deviation) inform us the extent to which data points spread from its center.

- ❖ Range indicates how much spread there are between high and low values, variance measures the average of the squares of the readings from their means and standard deviation provides a measure for distribution in actual units.
- ❖ Skewness provides the degree of asymmetry in a distribution whether it is negative, positive and symmetric.
- ❖ For our bivariate normal case, kurtosis reflects the "peakiness" of the distribution relative to a normal one.
- ❖ Covariance reveals whether two variables are moving in similar or contrary directions.
- ❖ Correlation and coefficient of determination measure the strength and percentage of variation described in association between the variables.

4.5 Key Terms

1. Frequency Distribution: A table that lists how many times a value or range of values occurs in a dataset.
2. Mean: The sum of all values in a dataset divided by how many there are (the arithmetic average).
3. Median: The value in the middle when data are arranged by size (either smallest to largest or largest to smallest).
4. Mode: Observation that appears most often in a dataset.
5. The range: The range of the highest and lowest value in a set.
6. Standard Deviation: A measure of the spread of data values around the mean (on average).
7. Skewness: how asymmetrical the distribution of the data is.
8. Correlation: A statistic revealing how strongly, and in what direction, two variables are related.

4.6 Descriptive Questions

1. Discuss the role of data visualization in data mining with examples.
2. Explain the differences between histograms, scatter plots and boxplots using examples.
3. What is a frequency distribution? How does it help in analyzing of big data sets?
4. Explain measures of central tendency (mean, median, mode) with examples. Were each measure the most useful in which settings?
5. Explain what measures of dispersion are, and discuss their purpose. What is the difference between variance and standard deviation?
6. Explain the terms skewness and kurtosis? Illustrate with suitable diagrams.
7. Explain the concept of covariance. How does it differ from correlation?
8. What are the classes of correlation? Elaborate both types with the help of examples.

9. Define coefficient of determination. How is it deployed in a regression analysis?

4.7 References

1. Anderson, D. R., Sweeney, D. J., & Williams, T. A. (2019). Statistics for Business and Economics. Cengage Learning.

2. Gupta, S. C. (2017). Fundamentals of Statistics. Himalaya Publishing House.

3. Levin, R. I., & Rubin, D. S. (2017). Statistics for Management. Pearson Education.

4. Moore, D. S., McCabe, G. P., & Craig, B. A. (2021). Introduction to the Practice of Statistics. W.H. Freeman.

5. Spiegel, M. R., Schiller, J., & Srinivasan, R. A. (2018). Schaum's Outline of Probability and Statistics. McGraw-Hill.

6. Field, A. (2018). Discovering Statistics Using IBM SPSS Statistics. Sage Publications

7. Tukey, J. W. (1977). Exploratory Data Analysis. Addison-Wesley.

8. Freedman, D., Pisani, R., & Purves, R. (2007). Statistics. W.W. Norton & Company.

9. Online Resource: Khan Academy. (n.d.). Statistics and Probability. Retrieved from https://www.khanacademy.org/math/statistics-probability

Answers to Knowledge Check

Knowledge check 1

1. b) Direction of relationship

2. c) –1 to +1

3. c) No linear relation

4. b) r^2

4.8 Case Study

"Healthcare Data Preparation for Predictive Analysis Visualization and Distribution"

Introduction

A large hospital system had the need to better understand how patient demographics, lifestyle, and clinical characteristics play a role in developing chronic diseases like diabetes and hypertension. To do this, the hospital administrators gathered patient data from electronic health records (EHRs) and opted to perform data visualization, frequency distribution and measures of central tendency and dispersion in order to ensure that they could prepare their dataset for predictive modelling.

Background

The raw data set was comprised of 1,000 patients and the variables such as:

• Patient ID (Nominal)

• Age (Ratio)

• Gender (Nominal)

• BMI (Body Mass Index, Ratio)

• Blood Pressure (Interval)

• Cholesterol Level (Interval)

• Diagnosis of Diabetes (Nominal: Yes/No)

It was a complicated and rather inscrutable dataset, to begin with. This is a series of code to prepare it for predictions.

• Histograms were used to check for the shapes of the continuous variables like BMI and cholesterol.

• Boxplots indicated outliers, for instance extremely high blood pressure values.

• Frequency tables were used to describe categorical data such as sex and diagnosis.

• Scatter plots were employed to find associations and potential predictors, for example, BMI versus blood pressure.

• Measures of central tendency (mean, median and mode) and dispersion (range, variance and standard deviation) we used to identify the patterns of related to patient health indicators.

Problem 1: Enrollment of Reliable Averages in Patient Data

The mean BMI values were influenced by particularly high readings that did not represent the majority of the patient population.

Solution: A comparison of mean and median BMI among researchers led to the conclusion that median provided a more representative measure of typical patient BMI, whereas the mean offered context for understanding such values at extreme levels.

Issue 2: Addressing Variability in Health Status and Outcomes

Blood pressure measurement" readings were very different in other patients – with consistent patterns of response for some groups and extreme variability in others.

Solution: The hospital use the standard deviation and variance of blood pressure to find groups with higher variability. Patients such as these were identified for closer monitoring, as higher variability is associated with increased risk of cardiovascular disease.

Problem 3: Investigating the Relationship between Risk Factors and Outcomes

Scatter plots demonstrated a positive strong association of BMI and BP, and cross-tabulations showed higher prevalence of diabetes in subjects with elevated BMI. A regression model showed that BMI and cholesterol levels accounted for 72% of the variance (R2 = 0.72) in diabetes diagnosis.

Solution: Based on this knowledge the hospital focused on programs for BMI and cholesterol management, expecting that by doing so the diabetes risk would be markedly reduced in these patient population.

Conclusion

Using visualisation tools, descriptive statistics and correlation analysis, they were able to turn raw patient records into structured knowledge. This groundwork allowed for an increased comprehension of patient health distributions and also served as a stepping stone to more advanced predictive analytics in clinic (e.g., clinical operative prognostics) that supported them in anticipation of and designing targeted preventive healthcare activities.

BUPBM Unit 5 V3.docx

Building useful Predictive Business models_BBA_3

Building useful Predictive Business models_BBA_3

ATLAS SkillTech University

Document Details

Submission ID

trn:oid:::3618:128376768

Submission Date

Feb 16, 2026, 12:32 PM GMT+5:30

Download Date

Feb 16, 2026, 12:37 PM GMT+5:30

File Name

BUPBM Unit 5 V3.docx

File Size

83.4 KB

24 Pages

7,056 Words

34,893 Characters

*% detected as AI

AI detection includes the possibility of false positives. Although some text in this submission is likely AI generated, scores below the 20% threshold are not surfaced because they have a higher likelihood of false positives.

Caution: Review required.

It is essential to understand the limitations of AI detection before making decisions about a student's work. We encourage you to learn more about Turnitin's AI detection capabilities before using the tool.

Frequently Asked Questions

How should I interpret Turnitin's AI writing percentage and false positives?
The percentage shown in the AI writing report is the amount of qualifying text within the submission that Turnitin's AI writing detection model determines was either likely AI-generated text from a large-language model or likely AI-generated text that was likely revised using an AI paraphrase tool or word spinner.

False positives (incorrectly flagging human-written text as AI-generated) are a possibility in AI models.

AI detection scores under 20%, which we do not surface in new reports, have a higher likelihood of false positives. To reduce the likelihood of misinterpretation, no score or highlights are attributed and are indicated with an asterisk in the report (*%).

The AI writing percentage should not be the sole basis to determine whether misconduct has occurred. The reviewer/instructor should use the percentage as a means to start a formative conversation with their student and/or use it to examine the submitted assignment in accordance with their school's policies.

What does 'qualifying text' mean?
Our model only processes qualifying text in the form of long-form writing. Long-form writing means individual sentences contained in paragraphs that make up a longer piece of written work, such as an essay, a dissertation, or an article, etc. Qualifying text that has been determined to be likely AI-generated will be highlighted in cyan in the submission, and likely AI-generated and then likely AI-paraphrased will be highlighted purple.

Non-qualifying text, such as bullet points, annotated bibliographies, etc., will not be processed and can create disparity between the submission highlights and the percentage shown.

Unit 5: "Data Cleaning and Preparation for Analysis"

Learning Objectives

1. Understand the concept of data cleaning – Explain what data cleaning means and why it is an essential step before performing analysis.

2. Identify common data quality issues – Recognize problems such as missing values, duplicates, inconsistencies, and outliers in raw datasets.

3. Apply techniques for handling missing data – Learn methods such as deletion, imputation, and substitution to manage incomplete records effectively.

4. Detect and remove duplicates – Develop skills to identify redundant records and ensure accuracy in datasets.

5. Standardize and transform data – Apply formatting, normalization, and scaling techniques to prepare data for statistical and machine learning models.

6. Handle outliers effectively – Understand methods to detect, analyze, and treat outliers without losing valuable insights.

7. Integrate data from multiple sources – Learn how to combine, merge, and reconcile datasets while ensuring consistency.

8. Ensure data readiness for analysis – Gain the ability to prepare a clean, structured, and reliable dataset that enhances the quality of insights and decisions.

Content

5.0 Introductory Caselet

“Organize Your Customer Data the Way You Think”

ABC Retail, a retailer, needed to launch its loyalty program with a targeted push via marketing message. The marketing team had managed to gather customer information from a range of touchpoints – website sign-ups, in-store sales, and social media engagements. Yet, as analysts took a closer look at the information, some concerns arose:

Some customer names were recorded in the reviewing process inconsistently (for example John Smith, J Smith, Jon Smith) making unique customer identification difficult. There was a lot of missing or mis keyed email addresses, so even the mailing list is not entirely accurate. Duplicates screwed up customer count and inconsistent formats (i.e., some date of births as 12/05/1990, others as 05-12-90) made for messy analysis.

The data needed much cleaning before we could do anything useful with it. Duplicates were excluded, spelling errors corrected, formats standardized and missing data filled in via the use of imputations. The marketing team were able to segment customers and run a campaign with the clean dataset which resulted in 25% increased customer interaction.

This case demonstrates that large datasets without data cleaning and preparation can lead to misleading insights, wastage of resources and failed strategies.

Critical Thinking Question:

If you were a data analyst at ABC Retail, what else would you do (besides deleting duplicates and standardizing formats) to create a clean dataset that is solid for predictive marketing analysis?

5.1 Introduction to Colab

5.1.1 Overview of Google Colab Environment

Features of Google Colab

• Cloud-Based Platform:

Colab runs in the cloud entirely, so no setup is required. All the users need is a Google account.

• Support for Python and Libraries:

Colab is also preinstalled with popular Python libraries, like TensorFlow, PyTorch, and Matplotlib, among others.

• Hardware Acceleration:

A user can use GPUs (Graphics Processing Units) and TPUs (Tensor Processing Units), which carry out complex and large computations productively.

• Integration with Google Drive:

Each notebook is in Google Drive, so you can replicate in your drive and edit as needed for time-saving.! This also takes the hassle out of sharing/working together.

• Collaboration Features:

Like Google Docs, several individuals can work on the same notebook at once. Contributors can leave comments, suggest edits and make changes in real time.

• Rich Media Support:

Notebooks may also contain formatted text (Markdown), mathematical and scientific equations (LaTeX), images, and even interactive data visualization output.

Interface of Google Colab

• Notebook Layout:

The layout resembles Jupyter Notebook. It comprises two principal cell types:

o Code cells where you write and execute Python code.

o Text cells for documentation, explanations or mathematical formulas written with Markdown.

• Toolbar and Menus:

The menu bar at the top has functionalities to create new notebooks, upload existing ones, add cells, modify runtime settings and manage file connections.

• File Explorer Panel:

On the side, theres a panel for connected files and file uploads downloads that is also used for project related file management.

• Runtime Menu:

It allows users to choose between CPU, GPU and TPU runtimes. The menu also includes options to restart or reset the runtime.

Advantages over Traditional IDEs

No installation/setup, no PyCharm or Anaconda Well-known IDEs like PyCharm/Anaconda need to be installed and configured. Colab lives directly in the browser.

Cross-platform Accessibility: Continuation of work from any computer online, without the need to install new software.

•Resource Availability: The ease of access to GPUs and TPUs drastically cuts down costs on running large computations on when running from personal machines.

• Collaborate & Share Notebooks can be shared with a link and have custom permissions for private or public sharing (view, comment or edit).

• Resources for learning: Colab is a great tool to learn from (programming and data analysis for beginners, ML engineers wanting to work with new libraries like Hugging Face or Uber's Ludwig) as well as a powerful tool for professionals.

5.1.2 Basic Operations in Colab

Creating Notebooks

• A new notebook can be made by clicking Google Colab and “New Notebook.”

• Or start from Google Drive > New > More > Google Colaboratory to create a notebook.

• Every notebook is saved as a. ipynb extension, just like in Jupyter Notebooks.

Running Code Cells

• Code goes into code cells. These cells are performed individually.

• To run a cell hit Shift + Enter or click the Run button that appears next to it.

• The results are shown just below the cell and they can be numerical values, text or images.

• Cells are re-runnable, editable, and rearrangeable enabling agile experimentation with code.

Importing Libraries

• Colab already contains most of the popular Python libraries. You'll be able to import them just as we have for the standard commands - e.g.

• import numpy as np

• import pandas as pd

• import matplotlib. pyplot as plt

• Anyone who needs more libraries can install them in the notebook with pip commands. Example:

• !pip install seaborn

• This adds a great deal of system flexibility and enables users to use large numbers of packages without leaving the Colab environment.

Saving Work

• Colab notebooks are stored in the Colab Notebooks folder on Google Drive.

• Users can also download their notebooks in various formats like . ipynb (Jupyter Notebook),. py (Python script), or. pdf for documentation purposes.

• Keeps track of versions, so it's safe to go back.

• Work can easily be shared with others via a shareable Google Drive link and you have the ability to control who can view, or edit.

"Activity: Getting Started with Google Colab"

Open Google Colab and create a new notebook. Add a code cell to print "Hello, World." Next, import the NumPy library and create an array of numbers. Save the notebook in Google Drive, then download it as a .ipynb file to understand storage and sharing options.

5.2 Importing the Files to Colab

5.2.1 Uploading Files from Local System

Definition and Use Case:

This way people can send files from their computer to the Colab environment.

The Colab environment is really helpful when you are working with datasets or when you are testing things with just a little bit of data.

You can use the Colab environment to try out things with the Colab environment. It is very useful, for the Colab environment.

Steps to Upload Files:

1. First you need to import the file upload module. The file upload module is very important, for this process so make sure you import the file upload module correctly.

2. from google.colab import files

3. uploaded = files.upload()

4. Now a box will pop up that lets you pick files. You can choose one file or a bunch of files from your computer.

5. When you have uploaded the files you can find them using their filenames. For example:

6. import pandas as pd

7. df = pd.read_csv('sample_data.csv')

Important Notes:

- When you upload files they go to the Colab session for a while. The files will be erased when the Colab session is over or when the computer stops talking to the runtime. This means you will not be able to see the files anymore after the Colab session ends. The Colab session is what stores the files so when it is done the uploaded files are gone too.
- When you want to use the files, in future sessions you have to upload the files again. This way you can reuse the files in sessions.
- Best suited for small, temporary tasks such as testing scripts or running practice exercises.

Advantages:

- Quick and simple to use.
- No need to manage external connections.
- Useful for beginners working with small CSV or text files.

Limitations:

- The files are not kept forever the files are only stored for a while because the files are not permanently stored.
- Uploading large files repeatedly can be inefficient.

Did You Know?

“Did you know that when you upload files from your local system to Google Colab, they are stored only for the current session? Once the runtime disconnects or resets, those files are automatically deleted, which means you need to re-upload them every time you restart.”

5.2.2 Importing Files from External Sources (Google Drive)

Definition and Use Case:

When you are working with a lot of data or you have a project that is still going on or you are working with people, the best way to get your files is to import them from Google Drive. Google Colab works well with Google Drive so you can be sure that your files will be there even when you are not working on them. This way Google Drive makes sure that your files are always available even when you start a session, with Google Colab.

Steps to Import Files from Google Drive:

1. Mount Google Drive within Colab:

2. from google.colab import drive

3. drive.mount('/content/drive')

4. Now you will see a message asking you to sign in with your Google account. When you do this Colab will make a connection to Google Drive. This connection will be made at the path /content/drive, on Google Drive. You will use Google Drive through this link that Colab creates.

5. Navigate through folders to access files. Example:

6. file_path = '/content/drive/My Drive/Colab Notebooks/data.csv'

7. df = pd.read_csv(file_path)

Features and Benefits:

• Persistent Storage: The files you upload to Google Drive stay there forever. This is really different, from when you upload something to your computer because those things are gone when you are done using the computer. Google Drive keeps your files safe in one place so you can use them whenever you want.

• When we talk about collaboration teams can do something useful. They can share datasets in Google Drive. This means that many people can look at the files in Google Colab. So collaboration with datasets in Drive makes it easy for teams to work together on the files, in Colab.

• Supports Larger Datasets: Efficient for machine learning projects and big data tasks.

• Organization: You can put your files into folders in Drive. This helps you manage your projects in a way. It is like keeping your files in order so you can easily find what you need for your project. Using Drive for organization is an idea because it helps you keep your files neat and tidy. Drive is very useful for managing your files and folders which's important, for organization.

Additional Tips:

• Data that you store in Drive can be in lots of formats. For example it can be in CSV format. It can be an Excel file. You can also store JSON files, images and even zipped archives in Drive. The Drive can hold all these types of data, like CSV, Excel, JSON, images and zipped archives.

• You can also use sources like GitHub, with Colab. To do this you can clone GitHub repositories into Colab using Git commands. This way GitHub and Colab work together. GitHub repositories are cloned into Colab using these Git commands.

Advantages:

• Convenient for long-term projects.

• Seamless integration with Google’s ecosystem.

• This thing really helps because it means you do not have to upload things over and again. Uploading the thing multiple times can be a real hassle so this feature of uploading is very useful because it reduces the need, for repeated uploading.

Limitations:

• This thing needs you to be connected to the internet. It also needs you to sign in with your Google account.

• Access permissions must be managed properly if files are shared.

5.3 Data Preprocessing

5.3.1 DataFrame Attributes and Methods for Data Exploration

A DataFrame is a table that has rows and columns. It is from the pandas library in Python. People use DataFrames a lot to organize and look at their data. DataFrames are really helpful for this because they are two-dimensional and have labels. This makes it easy to work with the data, in the DataFrame.

Common Attributes:

- df.shape → Returns the number of rows and columns (e.g., (100, 5)).
- df.columns → Lists all column names.
- df.dtypes → Displays data types of each column.
- df.index → Shows row index values.

Common Methods:

• When you use df.head(n) it will show you the first n rows of the dataset. This is really helpful when you want to see what the dataset looks like. The dataset is what df.head(n) is working with so it only shows you the first n rows of the dataset.

- df.tail(n) → Displays the last n rows of the dataset.
- df.info() → Provides summary including data types, non-null counts, and memory usage.

• When you use df.describe() it gives you a bunch of measures for your data. These measures include things like how many itemsre in your data the average value, how spread out the values are, the smallest value, the biggest value and where the middle values are. The statistical measures you get from df.describe() are really useful, for understanding your data. You get to see the count the mean, the deviation, the minimum, the maximum and the quartiles, which are all important things to know about your data when you are working with df.describe().

• The df.value_counts function is really useful. It shows us how times each unique value appears in a column. This means we get to see the frequency counts of these values. The df.value_counts function is very helpful for understanding what is, in our data. We use df.value_counts to get these frequency counts.

The main reason for these methods and attributes is that they provide an overview of the dataset structure. They also show us the distribution of the dataset. Help us find potential issues with the dataset, such as missing values or inconsistent values in the dataset. These issues with the dataset can be things like inconsistent values, in the dataset.

5.3.2 Handling Missing Values (Drop, Impute, Fill)

Missing values happen when some of the information in the dataset is empty. We just do not have it. The thing is, if we do not take care of missing values they can cause a lot of problems. We might get errors or our models might not be fair. Missing values can also make our models less accurate. When we are working with missing values we have to be careful because missing values can really affect our results. Missing values are a deal because they can make a big difference, in how well our models work.

Techniques:

1. Dropping:

o Remove rows with missing values: df.dropna()

o Remove columns with missing values: df.dropna(axis=1)

The best time to use this method is when the missing data is small and it does not matter much. This method works well when the missing data is not very important.

2. Imputation (Replacing with Statistics):

o Replace missing numeric values with mean, median, or mode.

o Example:

o df['Age'].fillna(df['Age'].mean(), inplace=True)

o Preserves dataset size while reducing bias.

3. Filling with Constants:

o Replace missing values with fixed substitutes such as 0, “Unknown,” or another placeholder.

o Example: Filling missing city names with “Not Provided.”

The thing that is really important to remember is that the method you choose is based on what kind of data you're working with how many values are missing and how important that variable is. The method chosen for handling missing values depends on the data type the proportion of missing values and the importance of the variable.

5.3.3 Encoding Categorical Variables (Label Encoding, One-Hot Encoding)

Categorical variables have things, like names or labels in them. These machine learning algorithms cannot understand them as they are so we need to change them into numbers that the algorithms can work with. We do this because machine learning algorithms need numbers to do their job and categorical variables are not numbers.

1. Label Encoding:

- Assigns each category a unique integer.
- Example:

Gender → Male = 0, Female = 1

- Pros: Simple and compact.
- Cons: The categories may not actually be in any order even though it looks like 0 comes before 1. This can be confusing because it makes the categories seem like they are ranked when really they are not. The problem is that the categories are not really in order they are the categories of the thing we are talking about like the categories of food and food does not have an order it is just food. The categories are the same, as the categories of the thing. The thing is just the thing and the categories are just the categories of the thing and that is all.

2. One-Hot Encoding:

• This makes columns that are either 0 or 1, for each category of the categories. It does this by taking each category and creating binary columns for the categories.

- Example:

Color → Red = (1,0,0), Blue = (0,1,0), Green = (0,0,1)

- Pros: Avoids artificial ordering of categories.
- Downsides: This will make the dataset bigger if there are a lot of categories, in the dataset.

5.3.4 Normalization and Standardization Techniques

When we look at datasets we see that they have variables that are measured in different ways. For example age is measured in years. Salary is measured in dollars. This can be a problem because some algorithms use distance or weight to make decisions. If we do not adjust the scale of the data the results can be misleading. Normalization and standardization are techniques that help fix this issue by making sure all the features in the dataset are on the scale so they all contribute equally to the results of the algorithms. This way the features like age and salary which're datasets are treated fairly. Datasets with features are more useful, for algorithms that rely on datasets.

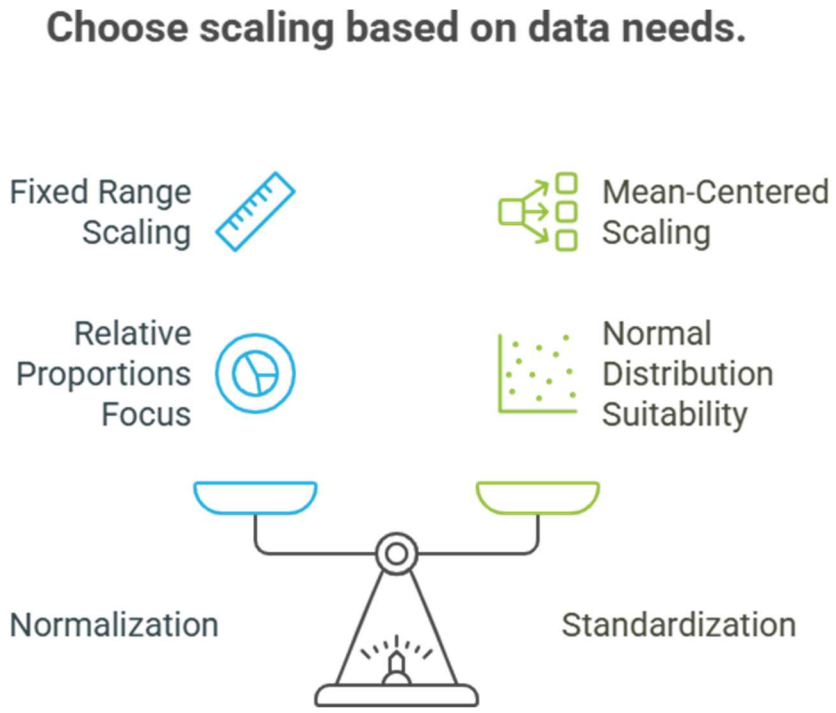

figure: Normalization and Standardization Techniques

1. Normalization (Min–Max Scaling):

- Rescales values into a fixed range, usually 0 to 1.
- Formula:

$x' = (x - \min(x)) \div (\max(x) - \min(x))$

• Example: If a column ranges from 50, to 200 the column value 125 becomes a value. To find this value we subtract 50 from the column value 125. So that is 125. 50 = 75. Then we divide this result by the range of the column. The total range of the column is 200. 50 = 150. So we divide 75 by 150. This gives us 0.5. The new value of the column value 125 is 0.5.

- Suitable when relative proportions are more important.

2. Standardization (Z-Score Scaling):

- Centers data around mean 0 with standard deviation 1.
- Formula:

$z = (x - \mu) \div \sigma$

where μ = mean of the variable, σ = standard deviation.

- Example: If $\mu = 100$, $\sigma = 20$, and $x = 120$: $z = (120 - 100) \div 20 = 20 \div 20 = 1$
- Suitable for algorithms assuming normally distributed data (e.g., regression, PCA).

Knowledge Check 1

Choose the correct option:

1. Which method shows the number of rows and columns in a DataFrame?

a) df.info()

b) df.describe()

c) df.shape

d) df.head()

2. Replacing missing values with mean, median, or mode is called:

a) Dropping

b) Imputation

c) Normalization

d) Encoding

3. One-hot encoding is mainly used to:

a) Normalize values

b) Remove duplicates

c) Convert categories into binary columns

d) Drop missing values

4. The formula $z = (x - \mu) \div \sigma$ represents:

a) Range

b) Normalization

c) Standardization

d) Encoding

5.4 Summary

* Data preprocessing is an important step. It is about taking the data and turning it into a clean and structured format. This format is suitable for analysis and modeling of the data. The data preprocessing step is necessary to make sure the data is ready, for use. Data preprocessing helps to get the data into a shape so that it can be used for analysis and modeling of the data.

* DataFrames in pandas have some things like shape, columns and dtypes that tell us about the data.

DataFrames in pandas also have some methods.

These methods include head, info and describe.

We can use these methods like head() info() and describe() to look at our data

This is really good for exploring the data, in our DataFrames in pandas.

* When we have missing values in our data we can deal with them in a ways. We can remove the records that have missing values.. We can fill in the missing values with the average value, the middle value or the value that appears most often in the data. We can also fill in the missing values, with a value. Missing values are a problem that we need to solve. We have a few options to handle missing values.

* We need to change variables into numbers.

We do this by using techniques like Label Encoding and One-Hot Encoding.

These techniques help us work with variables in a different way.

Categorical variables are not numbers so we have to use Label Encoding and One-Hot Encoding to make them into numbers.

This is important because computers can only understand numbers, not variables, like words or names.

So we use Label Encoding and One-Hot Encoding to convert variables into a numeric form that computers can understand.

* Label Encoding is a way to give numbers to categories but it can also make it seem like there is an order when there really is not. Label Encoding does this by assigning integer values to the categories. This can be a problem because Label Encoding can make people think that the categories have some kind of order or ranking when that's not what was intended.

* One-Hot Encoding is a way to make columns for each category in the data. This helps because it does not assume that the categories have any order. One-Hot Encoding is useful for things, like colors or types of food, where the categoriesre not related to each other in any specific way. One-Hot Encoding creates these columns so that the computer can understand the categories better.

* Normalization changes the data so it fits into a range like from 0 to 1 which makes it easier to compare things fairly. This way Normalization helps us see how different sets of data are related to each other by using Normalization. Normalization is really useful for things, like Normalization of data.

* Standardization is a process that changes data so it has a mean of 0 and a standard deviation of 1. This is really useful when we are dealing with Standardization of distributed data. The goal of Standardization is to make sure that all the data is on the scale, which is

important for normally distributed data. When we apply Standardization to distributed data it helps with the analysis. Standardization is about getting the data ready, for analysis especially for normally distributed data.

* Preprocessing is really important because it makes sure that all the features are treated equally. This helps the algorithm to work and it also makes the results of the analysis more accurate. Preprocessing does a job of improving the accuracy of the analytical results and this is very useful.

5.5 Key Terms

1. Data Preprocessing: The process of cleaning and transforming raw data into a usable format for analysis.

2. DataFrame: A two-dimensional labeled data structure in pandas for organizing datasets.

3. Attribute: A property of a DataFrame, such as shape, columns, or dtypes.

4. Method: This is a function that we use on a DataFrame. We use it to do things like look at the data or change it in some way. We can use a method to explore the data in the DataFrame or to transform the data, in the DataFrame. This is really helpful when we are working with a DataFrame and we want to do something with the data.

5. Missing Values: Sometimes you have data points that're not there or do not have a value in a dataset. These are called Values. Missing Values are a problem because they can affect the results of your analysis. Missing Values happen when data points are absent or undefined, in a dataset.

6. Imputation: Replacing missing values with statistical measures such as mean, median, or mode.

7. Dropping is when we get rid of rows or columns that have missing values, in our data. We do this by removing the rows or columns that have these missing values. Dropping these rows or columns helps us to clean up our data and make it easier to work with. When we are Dropping we are basically removing the parts of our data that're not complete. This way we can focus on the parts of our data that're complete and have all the information we need.

8. Encoding: The process of converting categorical variables into numeric form.

9. Label Encoding: Assigning unique integers to categorical values.

10. One-Hot Encoding: Creating separate binary columns for each category in a variable.

11. Normalization: Rescaling values into a fixed range, usually 0 to 1.

12. Standardization: Transforming values to have mean 0 and standard deviation 1.

13. Outlier: An outlier is a data point that's really different from the other data points. This can make it hard to understand the data because the outlier is so different from the rest of the data points. The outlier can also change the results of the analysis so it is important to look at outliers when you are working with data points, like these outliers.

5.6 Descriptive Questions

1. So you want to know about data preprocessing and how it helps with data analysis. Data preprocessing is really important because it makes sure the data is good and clean before we try to make sense of it. We have to do this because if the data is bad or messy it will be hard to get any information from it. This is why data preprocessing is a step before we apply any models to the data. The data preprocessing step is like getting everything ready before we start our analysis. We do things like handle missing data and make sure everything is in the format. This helps us get results when we apply the models to the data. Data preprocessing is essential, for data analysis because it helps us get results and make good decisions.

2. When we are working with pandas DataFrame there are some things we need to know about. A pandas DataFrame has some attributes and methods that help us explore our datasets.

The pandas DataFrame has attributes like columns, which gives us the column labels of the pandas DataFrame. We also have index, which gives us the row labels of the pandas DataFrame.

We use methods like head and tail to look at the first and last few rows of the pandas DataFrame. The head method shows us the five rows of the pandas DataFrame by default. We can also specify the number of rows we want to see.

For example if we have a pandas DataFrame called data we can use the head method like this: data.head(). This will show us the first five rows of the data pandas DataFrame.

We also have methods like info. Describe to get more information about the pandas DataFrame. The info method gives us a summary of the pandas DataFrame. The describe method gives us some summary statistics of the pandas DataFrame like the standard deviation.

So to explore our datasets we can use these attributes and methods of the pandas DataFrame. For instance we can use the columns attribute to see the column labels of the pandas DataFrame and the describe method to get some summary statistics of the pandas DataFrame.

Let us say we have a pandas DataFrame called data that has information about students, including their names, ages and grades. We can use the head method to look at the few rows of the data pandas DataFrame. We can use the info method to get a summary of the data pandas DataFrame. We can use the describe method to get some summary statistics of the data pandas DataFrame, like the age and standard deviation of grades.

By using these attributes and methods of the pandas DataFrame we can explore our datasets and get the information we need. We can use the pandas DataFrame to work with our datasets and get insights, from them.

3. What are missing values? Missing values are values that're not available in a data set. They can be numbers or words that are supposed to be but are not.

The thing about missing values is that they can cause a lot of problems when we try to analyze the data.

So we need to handle missing values.

There are techniques for handling missing values.

For example we can use a technique called deletion.

This is when we delete the row of data if there is a missing value.

Missing values can be handled in this way because it is simple to do.

Another technique for handling missing values is imputation.

This is when we replace the missing value with the value of that column.

Missing values are handled in this way because it gives us an idea of what the value might be.

We can also use a technique called regression imputation.

This is when we use a regression model to predict the missing value.

Missing values are handled in this way because it is very accurate.

For instance if we have a data set with missing values we can use regression imputation to fill in the missing values.

The data set will then be complete. We can analyze it.

We can also use a technique called interpolation.

This is when we use the values around the missing value to estimate what the missing value might be.

Missing values are handled in this way because it gives us an idea of what the value might be.

Different techniques, for handling missing values are useful because they help us to make sense of the data.

Missing values are a problem. Handling them is very important.

4. Differentiate between Label Encoding and One-Hot Encoding. In what situations is each method more appropriate?

5. So what is normalization and standardization? Normalization is a process where we make sure that the data is on the scale. This is really important because it helps the models to treat all the data points. Normalization of data is usually done by subtracting the minimum value and then dividing by the range of the data.

Standardization of data is a bit different. Standardization of data is done by subtracting the mean and then dividing by the deviation. This is really helpful when we are dealing with data that has a range of values.

The main difference between normalization and standardization is the approach they use. Normalization is good when we know the maximum values of the data. Standardization is good when we do not know the maximum values of the data.

In terms of application normalization is often used in image and signal processing. Standardization is often used in machine learning models. So the choice, between normalization and standardization really depends on the problem we are trying to solve and the type of data we are working with. We use normalization and standardization to make sure that the data is consistent and this helps the models to work.

6. So why do we need to change variables into a special code before we use them in machine learning algorithms? We need to do this because machine learning algorithms work well with numbers. Categorical variables are often just words or names. For example if we are trying to predict something based on the color of an object the machine learning algorithm will not know what to do with the words " "blue" or "green".. If we encode these colors into numbers like 1, 2 or 3 the algorithm can understand them. This is why encoding variables, like colors or names is such an important step, before using them in machine learning algorithms like decision trees or neural networks. We have to make sure the machine learning algorithm can understand the variables and encoding them into numbers is the way to do it. This way the machine learning algorithm can use the variables, like colors or names to make predictions or classify things. Categorical variables are a part of many machine learning projects so encoding them is a crucial step.

7. So let us talk about the imputation of missing values in a dataset. The imputation of missing values is a way to improve data quality. For example the imputation of missing

values can help us to keep the dataset size the same. We do the imputation of missing values by using the imputation of missing values methods.

The imputation of missing values methods are many. We can use the mean of the dataset to do the imputation of missing values. We can also use the median of the dataset to do the imputation of missing values. The imputation of missing values can also be done by using the mode of the dataset.

For instance let us say we have a dataset of student scores. The dataset of student scores has some missing values. We can use the mean of the student scores to do the imputation of missing values. This way the imputation of missing values will not reduce the dataset size. The dataset size will remain the same. The imputation of missing values will improve the data quality.

The imputation of missing values is very important, for data analysis. The imputation of missing values can help us to get results. The imputation of missing values can also help us to make decisions. So the imputation of missing values is a way to improve data quality without reducing dataset size. The imputation of missing values is an useful method.

8. Let us talk about the limitations of dropping missing values when we are getting our data ready. Dropping missing values is something we do to prepare our data.. When is it not a good idea to do this? We should think about when dropping values is not the best choice. Dropping values has some limitations. For example if we have a lot of missing values, in our data dropping them might not be the idea. This is because we will be losing a lot of information. The data we have might not be enough to make decisions. So we need to be careful when we are dropping missing values. We need to think about what will happen to our data if we drop the missing values. Dropping missing values should be avoided when we have information that is missing. If we drop this information our results might not be accurate. We need to make sure we are making the choice when we are dealing with missing values. Missing values are a problem. Dropping them is not always the solution.

9. Illustrate a step-by-step preprocessing workflow on a sample dataset covering exploration, handling missing values, encoding, and scaling.

5.7 References

1. Gupta, S. C. (2017). Fundamentals of Statistics. Himalaya Publishing House.

2. Han, J., Kamber, M., & Pei, J. (2011). Data Mining: Concepts and Techniques. Morgan Kaufmann.

3. Aggarwal, C. C. (2015). Data Mining: The Textbook. Springer.

4. Géron, A. (2019). Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow. O'Reilly Media.

5. McKinney, W. (2017). Python for Data Analysis. O'Reilly Media.

6. Kuhn, M., & Johnson, K. (2013). Applied Predictive Modeling. Springer.

7. Field, A. (2018). Discovering Statistics Using IBM SPSS Statistics. Sage Publications.

8. James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An Introduction to Statistical Learning. Springer.

9. Online Resource: Towards Data Science. (n.d.). Data Preprocessing Techniques in Machine Learning. Retrieved from https://towardsdatascience.com

Answers to Knowledge Check

Knowledge check 1

1. c) df.shape
2. b) Imputation
3. c) Convert categories into binary columns
4. c) Standardization

5.8 Case Study

Preparing Customer Purchase Data for Predictive

The people, at XYZ Supermarket wanted to start using a system that could tell them what products to suggest to customers. This system would look at what customers bought and make good guesses. They got lots of information from XYZ Supermarket stores over six months.. Before they could use this information to make predictions they had to make sure the XYZ Supermarket transaction data was correct and good to use. The XYZ Supermarket dataset needed a lot of work to get it ready.

Background

The raw dataset had some problems. There were a things wrong, with the raw dataset. The raw dataset was not perfect. It had several issues. The main thing was that the raw dataset contained issues.

- Missing Values: A lot of the information was incomplete because it did not include the customer age or the method of payment that the customers used. The customer age was missing in entries and so was the payment method that the customers chose to use for the customers.

- Duplicates: The company found that some transactions were written down than one time, which made the sales figures look bigger than they really were. The sales figures were not correct because of these transactions. The company had transactions that made their sales look better than they actually were.

- The dates were over the place because they were stored in different ways. For example some dates were written like this: 12/05/2023 and others were written like this: 2023-05-12. The dates were stored in formats, like 12/05/2023 and 2023-05-12.

- Categorical Variables: The payment method, which includes Cash, Card and UPI needs to be converted into a form because the payment method is made up of categorical variables, like Cash, Card and UPI.

• Unequal Scales: The amount people spent was really high in the thousands but the customer age was in years so we had to make sure the numbers were fair to compare this is what Unequal Scales is about it is like comparing apples and oranges we need to make the Unequal Scales equal so the purchase amount in thousands and the customer age, in years can be compared properly this is done by scaling the Unequal Scales.

If we use the dataset without doing some work on it the dataset will give us wrong answers and that will lead to the company making bad choices with the dataset. The dataset is really important. We need to make sure it is correct before we use the dataset to make big decisions with the dataset.

Problem Statement 1: Handling Missing Values

The dataset had 15% entries for the "Customer Age" column. This was a problem because the "Customer Age" column had a lot of missing information. If we only used the values for the "Customer Age" column it could give us the wrong idea, about the "Customer Age" column.

The people who looked at the numbers found a way to deal with the missing ages. They used the age of the customers to replace the missing ones. This way they made sure everything was balanced. It did not change the way the ages were spread out. The analysts did this to the ages of the customers.

Problem Statement 2: Managing Duplicates

I found that there are some transaction IDs that're the same in a few different rows. The transaction IDs are repeated in these rows. This is a problem because the transaction IDs should be unique. The same transaction IDs are showing up more, than once.

The solution, to the problem was to get rid of records. This was done by looking at the transaction IDs. If two records had the transaction ID, one of them was removed. This way the sales reports are accurate. The sales reports show what is really going on with the sales. This is because the deduplication process removed the duplicate records with transaction IDs.

Problem Statement 3: Encoding Categorical Data

The payment method and the product categories were kept as text. The algorithms were not able to work with the payment method and the product categories because they were just text.

The team used a method called One-Hot Encoding for the payment methods. This helped to turn the payment methods into numbers. They also used Label Encoding for the product categories. This was done to change the product categories into numbers too. The team did this so that the payment methods and product categories would still make sense as numbers. They wanted to keep the meaning of the payment methods and product

categories which's why they used One-Hot Encoding, for payment methods and Label Encoding for product categories.

Problem Statement 4: Scaling Features

The dataset had some features that were really different from each other. For example the "Purchase Amount" was on a different scale than the "Customer Age". This is because "Purchase Amount" and "Customer Age" are two features that are measured in different ways. The "Purchase Amount" feature and the "Customer Age" feature had a difference, in terms of their scales.

The team used a method called Standardization to make sure all the features were, on the level. They did this by using a formula: Standardization is calculated as z equals the value of x minus the mean divided by the deviation. This means z = the value of x minus the mean divided by the deviation. The team wanted all features to have a mean of 0 and a standard deviation of 1. This way all the features would contribute equally to the model. The team used Standardization to achieve this.

By applying data cleaning and preprocessing techniques—handling missing values, removing duplicates, encoding categorical variables, and scaling numerical features—the supermarket transformed a messy dataset into a reliable foundation for predictive analytics. This process not only improved model accuracy but also provided managers with trustworthy insights to design better marketing strategies.

BUPBM Unit 6 (1).docx

Building useful Predictive Business models_BBA_3

Building useful Predictive Business models_BBA_3

ATLAS SkillTech University

Document Details

Submission ID

trn:oid:::3618:128391978

Submission Date

Feb 16, 2026, 3:46 PM GMT+5:30

Download Date

Feb 16, 2026, 3:56 PM GMT+5:30

File Name

BUPBM Unit 6 (1).docx

File Size

166.5 KB

20 Pages

4,621 Words

28,567 Characters

0% detected as AI

The percentage indicates the combined amount of likely AI-generated text as well as likely AI-generated text that was also likely AI-paraphrased.

Caution: Review required.

It is essential to understand the limitations of AI detection before making decisions about a student's work. We encourage you to learn more about Turnitin's AI detection capabilities before using the tool.

Detection Groups

1 AI-generated only 0%
Likely AI-generated text from a large-language model.

0 AI-generated text that was AI-paraphrased 0%
Likely AI-generated text that was likely revised using an AI-paraphrase tool or word spinner.

Frequently Asked Questions

How should I interpret Turnitin's AI writing percentage and false positives?
The percentage shown in the AI writing report is the amount of qualifying text within the submission that Turnitin's AI writing detection model determines was either likely AI-generated text from a large-language model or likely AI-generated text that was likely revised using an AI paraphrase tool or word spinner.

False positives (incorrectly flagging human-written text as AI-generated) are a possibility in AI models.

AI detection scores under 20%, which we do not surface in new reports, have a higher likelihood of false positives. To reduce the likelihood of misinterpretation, no score or highlights are attributed and are indicated with an asterisk in the report (*%).

The AI writing percentage should not be the sole basis to determine whether misconduct has occurred. The reviewer/instructor should use the percentage as a means to start a formative conversation with their student and/or use it to examine the submitted assignment in accordance with their school's policies.

What does 'qualifying text' mean?
Our model only processes qualifying text in the form of long-form writing. Long-form writing means individual sentences contained in paragraphs that make up a longer piece of written work, such as an essay, a dissertation, or an article, etc. Qualifying text that has been determined to be likely AI-generated will be highlighted in cyan in the submission, and likely AI-generated and then likely AI-paraphrased will be highlighted purple.

Non-qualifying text, such as bullet points, annotated bibliographies, etc., will not be processed and can create disparity between the submission highlights and the percentage shown.

Unit 6: Prediction models

Learning Objectives

1. Understand the fundamental concepts of prediction and forecasting in business and finance.
2. Explain the role of statistical and machine learning models in prediction.
3. Differentiate between qualitative and quantitative prediction techniques.
4. Apply regression models to analyze and predict future trends.
5. Evaluate the accuracy of prediction models using appropriate error metrics.
6. Interpret time series models for short-term and long-term forecasting.
7. Compare traditional prediction methods with modern AI-based approaches.
8. Develop critical thinking for selecting the most suitable prediction model for decision-making.

Content

6.0 Introductory Caselet

“Predicting Sales for SmartMart”

HyperMart, a mid sized general retail chain, is experiencing intermittent monthly sales variations as a result of promos, seasonality and competitors. For the past 2 years, the company has a data about sales revenue, adspend, festive season and customer footfall.

This management now would like to leverage this data to better predict future sales. They intend to employ prediction models in order to:

- Project sales for the next holiday season.
- Determine the appropriate advertising budget for maximum income.
- Adjust supply based on known needs to prevent over or under stocking.

The analysts recommend to create predictive insights via the use of regression models and time series forecasting techniques. But some senior executives say they would rather trust experience and intuition than mathematical models, noting that predictive tools might not take into account sudden shifts in markets.

SmartMart has to determine whether it should follow a data based predictive modeling approach or rely on the manager judgment (traditional way).

Critical Thinking Question

As a member of SmartMart's decision committee, how do you integrate models' predictions with your gut for making business decisions?

6.1 Introduction to Prediction Models

6.1.1 Introduction to Predictive Modeling

Predictive modeling--one of the mechanic's most important data-analysis arms--is all about using past information to project what might happen in the future or under circumstances we don't yet know. Predictive modeling differs from descriptive analysis in that it focuses on looking forward, providing insight into what is likely to happen in the future – information an organization can use to take informed action. It is the essential part of today's data drive strategies, especially when it comes to finance, healthcare, retail and education domain as anticipating future trends can have huge operational and economic consequences.

At the heart, predictive modeling entails recognizing patterns in historical data and utilizing these patterns to predict future similar events. This consists of number of steps that outline the process — from clearly formulating the problem, to gathering and preparing data, selecting appropriate modeling techniques, training and testing models, following through with deployment in a real-world application. All of these stages are important to ensure that the model's output is reliable, valid and useful.

It's an iterative process when building predictive models; sometimes it takes more than 1 pass to get there. As better data emerges, models can be retrained and updated so that they become more accurate and usable. In reality, predictive modeling is facilitated by a variety of statistical algorithms and computational tools which it employs to make it scalable and more stable.

The steps below are the standard form of working in predictive modeling:

Problem Definition

The very first and most important stage in predictive modelling is properly identifying the problem to be solved. Without a clear sense of what is being predicted, the whole modeling exercise can turn out to be misdirected and yield results with no meaning or value. A properly defined problem defines the scope, states the target variable and make clear how the modeling exercise helps to achieve certain business or research objectives.

For instance, a university might be interested in forecasting which students are likely to drop out based on past academic and behavioral information. An e-commerce company, for example, might want to predict customer churn in order to tailor retention programs effectively. Each one will require a different modeling strategy and an articulated definition of what success means.

Key considerations include:

• Determining the prediction target: Formulate outcome variable (e.g., purchase, attrition risk, loan default).

o It gives a model clear objective and assist in selecting the appropriate algorithm.

• Contextualize the business or operational landscape: Know why we want our prediction and how it will be used.

o This drives data curation, stakeholder involvement and implementation on the ground.

• Determine success metrics: Determine how model performance will be measured (e.g., accuracy, error rate).

o Early recognition of successful criteria leads to validation and understanding.

A badly shaped problem is too often a waste of resources, and provides worthless predictions. Thus, it is required in the problem definition stage for domain experts, data scientists and stakeholders to come together to match technical solutions with practical requirements.

Data Collection

For now, I'll move forward assuming that the problem is well-formed; get to the data you need after you've determined what your predictive model will look like. Data is the foundation of predictive modeling; its quality, completeness and relevancy has a clear bearing on how the model will behave. The source of data can be very different according to which domain, and problem are in place. It could either internal (e.g., transaction logs) or external data sources like public datasets.

Broad and deep data collection should be the goal. This is to say that we want data that comprehensively represents the problem at hand, but with granular level of detail that would allow for accurate modelling. This could include fusing data from more than one source - both structured (like spreadsheets and databases) and unstructured (such as text, images).

Typical data sources include:

• Operational files and transaction records: Sales logs, inventories, billing statements.

o These are tangible consequences of past actions or events pertinent to predictions.

• Survey and feedback information: Customer satisfaction ratings, employee engagement surveys.

o These provide subjective contexts that can enhance the context of the model.

• External data sets: Market trends, economic indicators, public health records.

o These are additional variables which might help predicting accuracy in case of such sources.

• Sensor or IoT data: Machine logs, environmental monitors, wearables.

o These are becoming popular in fields such as manufacturing and healthcare where real-time prediction is required.

It is necessary to examine the reliability and validity of the collected data before advancing to stage two. Biases or noisy signals in these can be directly input to the amodal object detector resulting in biased predictions.

Data Preparation

And once data is gathered, it usually comes in unstructured and raw format. Thus, data preparation is a crucial step in which the raw data is prepared into a usable form. At this stage, there are several sub-tasks such as taking care of missing data, handling outliers, discarding irrelevant variables and also encoding categorical variables followed by normalization or standardization. Good data quality will make the model learn well and generalize well as a result.

The aim is to produce a clean data set which can be directly used for modeling without noise and bias. This frequently requires statistical methods as well as domain knowledge to decide which data to retain, alter or discard.

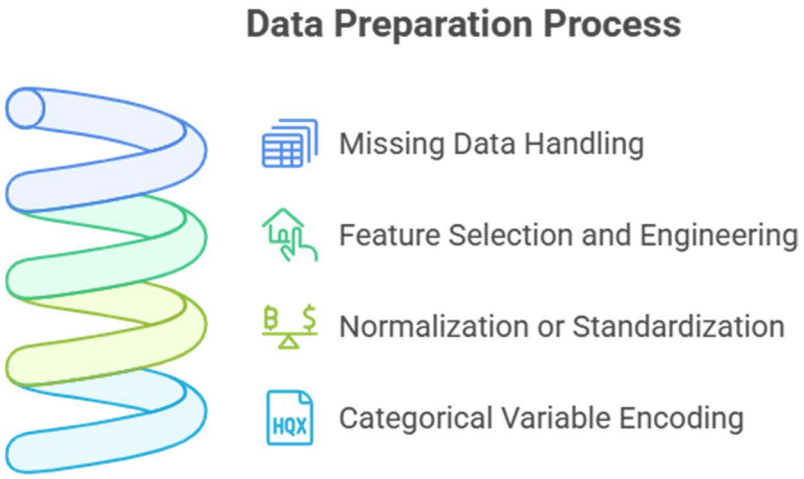

figure: Key steps in data preparation

Some important data preparation steps are:

Missing data processing: Methods such as Imputation (mean, median or model-based) are performed for missing value imputation.

o It can lead to biased patterns and poor model performance if missing values are not handled.

• Feature selection and engineering: Discover and develop variables that have a substantial impact to model accuracy.

o This helps the model to capture more important patterns in your data.

• Normalisation or standardisation: Scale the features to a range (or distribution) of values.

o Most of the machine learning algorithms work well with scaled data.

• Categorical encoding: Mapping categories to some form of numerical representation (e.g., one-hot encoding).

o Critical for models unable to handle categorical input values.

Data preparation is part art and part science: knowing what transformations will support or detract from the modelling process. Better quality in the data makes better models and less headache during preprocessing.

Model Selection

Model selection is a critical step in predictive modelling, as it chooses the algorithm or statistical technique that will be applied to discover patterns within data. This choice is a function of the prediction task at hand, the data structure and what type of output is expected. Whether the PointNet architecture predicts a numeric value or discretizes it into

their respective classes as well as predicting time series, each modeling approach has its own set of strengths and weaknesses that must be factored in.

In general, there are several categories in which predictive modeling techniques can be classified. We usually use regression models for empirical prediction of a continuous numerical variable. For classification problems (where the output is in one of a few predefined categories), you would typically use classification algorithms such as decision trees, logistic regression, or support-vector machines. On the other hand, for data that changes with time such as stock prices or sales numbers, you may do better using a time series model such ARIMA or an LSTM neural network.

Common predictive modeling techniques include:

• Linear and logistic regression: Great for uncomplicated variable relationships and forecasting for numeric or binary outcomes (respectively).

o Reggression models are interpretable and can be used as a baseline in many prediction problems.

•Decision trees,ensemble methods (Random Forest,XGBoost) -For powerful to model complex interactions and non-linear relationships.

o They can be used for both categorical and numerical data.

• SVM: Useful for classification of high-dimensional feature spaces.

o SVM are not so susceptible to overfitting, especially when working with smaller sample sizes.

• Neural Networks: Appropriate for big data and complicated patterns, particularly image, text or speech data.

o They are very flexible but computationally intensive.

• Time series models (ARIMA, Prophet): Intended for time-based data with trends and seasonality.

o The time-based models take temporal dependencies in observations into consideration.

Selecting the appropriate model is routinely decided through informal experimentation, in light of both theoretical works and practical applicability. Often times, several models are run at once to see which model gives the best accuracy-speed-interpretability balance.

Model Training

Model training refers to feeding a certain machine learning algorithm selected as the model with historical data for it to learn or find patterns and correlations between variable/s. This is where you pass the dataset to the model, such that it can learn its internal parameters

according to the mapping from inputs to outputs. The goal is for the model to generalize well, performing accurately on new, unseen data by having learned from previously observed ones.

During training, the model leverages an optimization algorithm (such as gradient descent) to reduce a loss function representing the prediction errors. The quality of the training: how much and how representative data are available, how is the algorithm configured (as in hyperparameters), and what is the complexity of the target relationships.

Key considerations with respect to model training are:

• Training set: A sub-set of the data that is used for learning during training only.

o This part is typically 60-80% of all data and allows the model to “learn”.

• Loss function: A mathematical rule used to measure prediction errors.

o Popular loss functions are Mean Squared Error for regression, and Cross-Entropy for classification.

• Overfitting and underfitting: Overfitting happens when the model learns the noise in the data, rather than a general trend.

while underfitting indicates it hasn’t captured enough of the pattern in the data.

o Methods like regularization, pruning and dropout layers are used to prevent such problems.

• Hyperparameter tuning: Changing settings, such as the learning rate, tree depth or hidden layers.

o Grid search or randomized search helps you to find what configuration works best for you.

It is during training that predictive modeling can become computationally expensive, particularly with large or powerful models. Nowadays cutting-edge tools like Scikit-learn, TensorFlow, and PyTorch are frequently used to simplify the process and handle the training flow.

Model Validation and Testing

Once the model is trained, you will want to validate and test it to make sure that it performs well and can perform accurate predictions on new data. Here, you apply the model to some portion of the data that wasn't part of the training set—usually previously unseen or "test" data. The ultimate aim is to test how well the model can generalize over the data that it has not been exposed on.

Validation helps to detect common problems like overfitting which occurs if a model is good for the training data but bad for new data. Testing also means there is a way to contrast different models or configurations and see which one performs best. A number of performance metrics can be used to quantify predictive performance, depending on the type of task (regression or classification).

Common evaluation metrics include:

• Mean Squared Error (MSE): Average of the squared differences between predicted and actual values.

o Smaller the MSE better the regression performance and less is the prediction error.

• R-squared (R^2): Represents the amount of variance in the dependent variable that is predictable from the independent variable in regression.

o Higher R^2 values indicate better fit of the explanatory variables.

• Precision: True positive rate in classification.

o Quick/simple and intuitive though not always robust to imbalances in data.

• Precision, Recall and F1-Score: Evaluate the goodness of classification especially with class imbalance.

o These give a deeper perspective than the accuracy alone.

• Confusion Matrix: A table used to show the performance of a classification model (or "classifier") on a set of test data for which the true values are known.

o Assist in knowing certain points where model is having misclassification.

Validation and testing not only test the predictive power of the model but also offer direction for final modifications prior to deployment. Methods such as k-fold cross validation can be used to make the evaluation more robust and not depend on a single test split.

6.1.2 Understanding Dependent and Independent Variables

How does a prediction model work? Prediction models function by discovering associations between various features. Building these relationships will require to be familiar with dependent and independent variables.

Dependent Variable (Target Variable)

• The response variable is what you are trying to predict, the Y in f(X) = Y.

• It is commonly known as the target variable, response variable, or predictor.

• The dependent variable varies as a function of the independent variables. Its value is believed to depend on other factors.

• Examples:

o The dependent variable for the forecast of sales is the sales amount.

o The dependent variable for predicting student exam performance is the score on the exam.

o Dependent variable for Loan Default prediction: In predicting loan default, the dependent variable is whether or not the borrower defaults (Yes/No).

Independent Variables (Predictor Variables)

• Independent variables that are responsible for, or that account for the differences in, the dependent variable.

• They are also referred predictors, explanatory variables or features.

• These are input to the model that serve to estimate the outcome.

• Examples:

o In sales forecast, independent variables would be ad spend, price strategy and seasonal influence.

oThe independent variables in modeling the performance of students may include study hours, attendance and previous academic record.

o For the loan default, a few explanatory variables can be income, credit history or due amount.

4 The Relationship Between Dependent and Independent Variables

• Predictive models attempt to provide a measure of the relationship between independent and dependent factors.

• Regression analysis could, for instance, indicate how much a boost in advertising spend (independent variable) boosts sales (dependent variable).

• The ability to understand and measure it allows organizations to make better strategic decisions -- how much budget to spend on marketing, which pricing strategy is best and so on.

Illustrative Example

For example, pretend a university is attempting to determine the final exam scores of students.

• Dependent Variable: Final Exam result(the variable to be predicted).

• Predictors: Study hours, attendance percentage, previous GPA, participating in casting/tutorials. The model will examine the impact of independent variables on exam results and then use that understanding to predict student future performance.

“Activity: Identifying Variables in Real-Life Scenarios”

Students will be divided into small groups and asked to choose a real-life situation such as predicting exam results, sales performance, or weather. Each group must identify one dependent variable and at least three independent variables influencing it, then present their reasoning briefly.

6.2.1 Regression Models and Classification

Regression Models

The purpose of regression models is to try to determine a relationship via a mathematical formula between a dependent variable and some independent variables. Second, these values not only allow for predictions into the future but also describe the strength and direction of relationships.

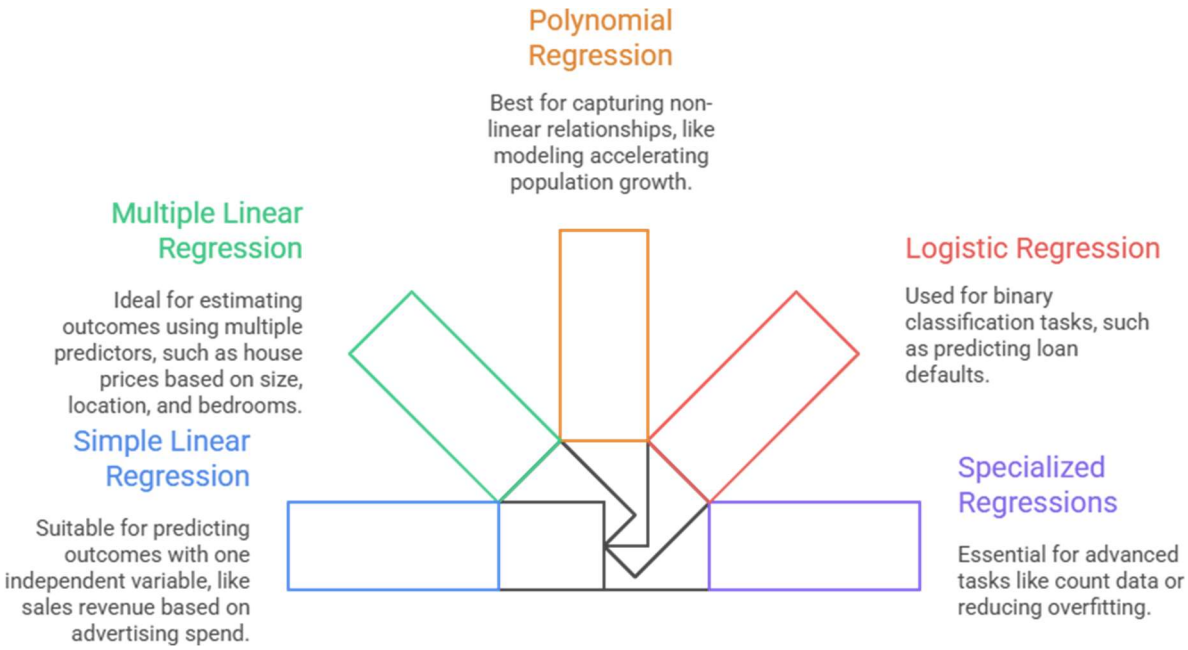

figure: Types of Regression Models

Types of Regression Models:

Simple Linear Regression

o Makes prediction with one predictor variable.

o Example: Predicting sales revenue from advertising spend only.

Multiple Linear Regression

o Uses more than one predictor to predict the dependent variable.

o Example: Predicting home values based on square footage, neighborhood and number of bedrooms.

Polynomial Regression

o Represents the non-linear relationship by adding powers of input variables.

o Example: When you do population modelling where growth increases exponentially over time.

Logistic Regression

o Technically, it is known as regression, but we use this for binary classification (Yes/No).

o Example: Predicting if a customer will be defaulter on loan.

Other “Fancy” Regressions (Poisson, Ridge, Lasso, etc.)

o For advanced prediction tasks such as count data or mitigating overfitting in large datasets.

Classification Models

Classification is assigning observations to classes or categories based on input variables. The model is not predicting a quantifiable outcome; it's predicting a category.

Types of Classification Models:

Decision Trees

o Decompose decisions into simple rules.

o Example: Diagnosing a study subject as having or not having the disease based on symptoms.

Random Forests

o Ensemble several decision trees to enhance the accuracy.

o For example, categorizing customers as buyers or non-buyers.

Support Vector Machines (SVMs)

o Determine the best boundary (hyperplane) which will divide classes.

o Example: Image identification (Cat Vs Dog classification).

Naïve Bayes Classifier

o By means of probability and Bayes’ theorem.

o Example: Email Spam and filtering for non spam.

Neural Networks

o Brain-inspired models that classify difficult data using simulations of the human brain.

o Example: Recognizing faces or identifying voices.

6.2.2 Selecting Appropriate Regression Models

The selection of a particular regression model is based on the characteristics of the data and type of dependent variable, and not study aims.

Nature of Dependent Variable

o Variable-continuous (e.g., income, sales) → Linear or multiple regression.

o Binary variable (e.g., default/non-default) → Logistic regression.

o Counts (number of accidents, etc.) → Poisson regression.

Nature of Relationship

o Linear relation → Linear regression is appropriate.

o Non-linear relationship → Polynomial or non-linear fit needs to be considered.

Number of Predictors

o One predictor → Simple regression.

o Multiple predictors → Multiple regression.

Assumptions Check

o Regress has assumptions of linear, independent, homoscadastic, normality in the erros. When assumptions are not met, other methods such as Ridge, Lasso or Non-parametric regression are applied.

Example:

A firm that anticipates sales may begin with the help of linear regression, if its relationship by advertising should be linear. If seasonality impacts sales, they might include polynomial terms or turn to time-series regression.

6.2.3 Identification of Regression and Classification Models for Various Problem Statements

• Predicting households' monthly electricity use → Multiple Linear Regression.

• Stock price prediction → Polynomial regression or ML (advance) for regression.

• Quantifying the effect of training hours spent on employee output → Simple Linear Regression.

Classification Applications

• Predicting churn of Wearhouse Customer → Regression or Decision Trees.

• Spam Email/Not Spam Email → Naïve Bayes Classifier.

• Whether a patient has diabetes → Random Forest or Support Vector Machine.

• Whether a loan is risky or not → Logistic Regression.

6.2.4 Role of Prediction in Business and Investigation

Prediction is a common practice in business and academia.

• In Business

o Sales Forecast: Facilitates in inventory and production planning.

o Risk Management: Forecasting credit default or finding fraud.

o Customer Behaviour Analysis: Predicting the behaviour of customers to plan marketing tactics.

o Resource Allocation: Predicting need to balance supply chains.

• In Research

o Test of Hypothesis: Models are being created to test the relationship between the variables.

o Policy Analysis: Economists make forecasts about the effects of fiscal policy.

o Health Care Research: Assay prediction in treatment efficacy.

o Social Sciences: Predicting the voter preferences, population growth or education outcomes.

Predictive models help in minimizing the uncertainty and thereby aiding organizations, and decision makers to take more rational decisions.

Did You Know?

“Did you know that businesses using predictive models improve decision-making accuracy by up to 30%? In research, predictive analytics allows scientists to simulate real-world scenarios before testing them. From forecasting stock markets to predicting disease outbreaks,

prediction models significantly reduce uncertainty in both business and academic research contexts."

6.2.5 Applications of Prediction, Regression, and Classification

Prediction Applications

o Retail: Predicting demand for holiday sales.

o Banking: Will the borrower default or not.

o Health care: Predicting an outbreak of disease, or a patient's time to recovery.

o Education: Anticipating dropout rates.

Regression Applications

o Real Estate: Converting residential details and the place into housing costs.

o Economics: Simulating the impact of inflation on consumer expenditures.

o Marketing: Quantification impact of advertising on sales growth.

o Weather Forecasting – To predict rain fall or temperature.

Classification Applications

o Telecom: Forecasting if a customer will churn (leave and stay).

o Medicine: Identifying tumors cells in a cancerous vs benign state.

o Cyber security: Differentiating the network traffic as being normal, or suspicious.

o Finance: Fraud detection for credit card transactions.

Knowledge Check 1

Choose the correct option:

1. Which model is best suited for predicting continuous numeric values?

a) Logistic Regression

b) Linear Regression

c) Decision Tree

d) Naïve Bayes

2. Classification models are mainly used when the outcome variable is:

a) Continuous

b) Categorical

c) Random

d) Independent

3. Predicting house prices using size, location, and amenities is an example of:

a) Logistic Regression

b) Decision Tree

c) Multiple Regression

d) Naïve Bayes

4. Spam vs. Non-spam email detection is an example of:

a) Regression

b) Classification

c) Time Series

d) Correlation

6.3 Summary

- Prediction models They are used to forecast future events using historical data, and statistical algorithms.
- Regretion model use when we have a continuous(dependent)variable.
- When the dependent variable is categorical (groups/labels), classification models are used.
- The plain ones could be linear, non-linear (polynomial), probability ((simple) logistic), ordinal (as in the ordering of the levels of an ordinal factor; this could use a binary or more generally thresholder response…or sar/vgam if we're willing to assume that being between levels has no effect — and r is merely computing odds ratios which given it's not really doing anything but transforming).
- Decision trees, random forests, support vector machines and Naïve Bayes are some of the popular classification algorithms.

- ❖ Choice of regression models is based on type of data, nature of variable relationships and objectives of the study.
- ❖ Regression forecasts such as sales revenues, prices and consumption patterns.
- ❖ Tasks: Classification is utilized in problems such as fraud discovery, diagnosis of diseases and prediction of customer attrition.
- ❖ Predictive models help reduce uncertainty in business as well as research for making evidence based decisions.
- ❖ Applications are in the areas of retail, finance, healthcare, telecom and education.

6.4 Key Terms

1. Prediction Model – Statistical or machine learning technique used to predict future results from historical data.
2. Regression – A method used to forecast continuous number values by determining relationships between variables.
3. Categorization – A technique where data is categorized into certain classes or groups."
4. Dependent Variable – The target or result variable which a model aims to predict.
5. Up to Index Independent Variable – the primary variable that explains or causes changes in a dependent variable.
6. Logistic Regression: Logistic regression is another statistical model we can use to predict binary outcomes, such as Yes/No or True/False.
7. Decision Tree – A classification model that uses decision rules to divide data into branches to produce decisions.

6.5 Descriptive Questions

1. Define regression models. Describe various kinds of regression with appropriate examples.
2. What are classification models? Explain their significance in the context of predictive analytics.
3. What are real life examples of regression and classification models?
4. Discuss the considerations in choosing a suitable model of regression.
5. Explain how prediction has been important in business policy with appropriate examples.
6. How do they use regression/classification models in research work?
7. Describe the uses of prediction models in retail, banking, and healthcare.
8. Discuss with examples how the dependent and independent variables are introduced in regression models.
9. Provide examples on how classification models can be used in fraud detection and prediction of customer churn, using case studies.

6.6 References

1. Montgomery, D. C., Peck, E. A., & Vining, G. G. (2012). Introduction to Linear Regression Analysis. Wiley.

2. James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). An Introduction to Statistical Learning with Applications in R. Springer.

3. Hosmer, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). Applied Logistic Regression. Wiley.

4. Bishop, C. M. (2006). Pattern Recognition and Machine Learning. Springer.

5. Hair, J. F., Black, W. C., Babin, B. J., & Anderson, R. E. (2018). Multivariate Data Analysis. Cengage.

6. Han, J., Kamber, M., & Pei, J. (2012). Data Mining: Concepts and Techniques. Morgan Kaufmann.

7. IBM. (2023). Regression vs Classification in Machine Learning. Retrieved from https://www.ibm.com/topics/regression-vs-classification

8. Towards Data Science. (2022). A Beginner's Guide to Regression and Classification. Retrieved from

https://towardsdatascience.com

9. ResearchGate Articles on Regression and Classification Models (various academic sources).

Answers to Knowledge Check

Knowledge check 1

1. b) Linear Regression

2. b) Categorical

3. c) Multiple Regression

4. b) Classification

6.7 Case Study

Regression and Classification in Privacy-control Work Benedek et al.

Introduction

The retail industry is very competitive today and customer needs keep evolving. If forecasts of sales and customer demand can be predicted precisely, so too are purchasing patterns. Prediction models, e.g., regression and classification, give a business powerful tools for analyzing historical data for predicting the future. Such models also serves the key purposes to identify growth opportunities as well as mitigate risk of oversupply, stockouts and customer churn.

This study discusses the use of regression and classification in retail decisions, presents implementation challenges, and gives ideal solutions to increase predictive accuracy.

Background

ShopSmart, a retail store spanning across metropolitan cities, has been facing unpredictable sales and erratic inventory handling. The company accumulates a lot of data on the sales, promotions, customer characteristics, and seasonality but does not have an ordered prediction method.

• Sales managers themselves use a lot of intuition, which results in incorrect demand forecasts.

• Marketing organizations struggle to cater the right customers without predictive understandings.

• It has been observed that there is growing customer attrition (churn), though no model exists to detect vulnerable customers.

ShopSmart opts to use regression models for sales prediction and classification models

to support the adjustment toward more data-based marketing regarding customer behavior.

Problem Statement 1: Sales Forecast Not Accurate

With bad sales forecasts, ShopSmart routinely overstocks or stockouts.

Solution: Use Multiple regression analysis with price, promotional and seasonal demand to predict sales will happen more accurately.

MCQ:

Which model would be the best to predict sales revenue for ShopSmart?

a) Logistic Regression

b) Multiple Regression

c) Decision Tree

d) Naïve Bayes

Answer: b) Multiple Regression

So Multiple Regression is able to deal with multiple independent variables (such as price, promotions and seasonality) and predict sales revenue well.

Problem 2: Predicting Customer Churn Let's say you're a telecommunications company and you want to predict if some of your customers will move to another competitor.

Predicting which customers are likely to stop shopping with them is also hard for Israel-based company.

Solution : Develop classifiers like logistic regression or decision trees to pick out at-risk customers using purchase frequency, complaints and spending behavior.

MCQ:

What is the best model to predict customers who "churn" or do not churn"?

a) Linear Regression

b) Logistic Regression

c) Polynomial Regression

d) Time Series

Answer: b) Logistic Regression

Notes and references Logistic regression is appropriate, if churn vs. retain classiying problem is binary.

Problem Statement 3: "Application of Regression and Classification Models within "

Poor customer segmentation is one of the key reasons for marketing campaign failure.

Solution: Use a decision tree and random forest classification algorithms to group customers as high value, medium value and low-value. This also makes it easier to send targeted promotions.

MCQ:

Which one could segment customers into multiple groups for specific marketing?

a) Decision Tree

b) Simple Linear Regression

c) Logistic Regression

d) Poisson Regression

Answer: a) Decision Tree

ANSWER: Explanation: Decision trees have the capability for dividing customers into several segments, targeting marketing campaigns better and increasing sales.

Conclusion

Utilizing regression and classification models, ShopSmart will be able to increase accurate sales predictions, recognize the customers that need your respect and how often they need it! These predictive models turn guesswork into decisions with empirical validation for efficiency, client satisfaction and long-term business growth.

BUPBM Unit 7.docx

Building useful Predictive Business models_BBA_3

Building useful Predictive Business models_BBA_3

ATLAS SkillTech University

Document Details

Submission ID

trn:oid:::3618:128389900

Submission Date

Feb 16, 2026, 3:18 PM GMT+5:30

Download Date

Feb 16, 2026, 3:23 PM GMT+5:30

File Name

BUPBM Unit 7.docx

File Size

70.8 KB

25 Pages

5,648 Words

34,461 Characters

*% detected as AI

AI detection includes the possibility of false positives. Although some text in this submission is likely AI generated, scores below the 20% threshold are not surfaced because they have a higher likelihood of false positives.

Caution: Review required.

It is essential to understand the limitations of AI detection before making decisions about a student's work. We encourage you to learn more about Turnitin's AI detection capabilities before using the tool.

Frequently Asked Questions

How should I interpret Turnitin's AI writing percentage and false positives?
The percentage shown in the AI writing report is the amount of qualifying text within the submission that Turnitin's AI writing detection model determines was either likely AI-generated text from a large-language model or likely AI-generated text that was likely revised using an AI paraphrase tool or word spinner.

False positives (incorrectly flagging human-written text as AI-generated) are a possibility in AI models.

AI detection scores under 20%, which we do not surface in new reports, have a higher likelihood of false positives. To reduce the likelihood of misinterpretation, no score or highlights are attributed and are indicated with an asterisk in the report (*%).

The AI writing percentage should not be the sole basis to determine whether misconduct has occurred. The reviewer/instructor should use the percentage as a means to start a formative conversation with their student and/or use it to examine the submitted assignment in accordance with their school's policies.

What does 'qualifying text' mean?
Our model only processes qualifying text in the form of long-form writing. Long-form writing means individual sentences contained in paragraphs that make up a longer piece of written work, such as an essay, a dissertation, or an article, etc. Qualifying text that has been determined to be likely AI-generated will be highlighted in cyan in the submission, and likely AI-generated and then likely AI-paraphrased will be highlighted purple.

Non-qualifying text, such as bullet points, annotated bibliographies, etc., will not be processed and can create disparity between the submission highlights and the percentage shown.

Unit 7: Model Development (Regression Models)

Learning Objectives

1. Explain the fundamental concepts and assumptions underlying regression models.
2. Differentiate between simple and multiple regression models and their applications.
3. Construct regression models using appropriate variables and datasets.
4. Interpret regression coefficients, intercepts, and error terms in practical contexts.
5. Evaluate model performance using statistical metrics such as R^2, adjusted R^2, and RMSE.
6. Diagnose multicollinearity, heteroscedasticity, and other regression model issues.
7. Apply regression models to real-world business and financial decision-making scenarios.
8. Use software tools (e.g., Excel, R, Python) to build, analyze, and validate regression models.

Content

7.0 Introductory Caselet

"Regression Analysis for Strategic Decision-Making: ShopEase Case Study"

ShopEase is an e-commerce company which is of mid-size and the business has got some potential to grow, ShopEase would like to know what are the factors that contributes towards sales revenue in each month. The management has recorded two years worth of data including advertising spend (online and offline), traffic on the website, average product rating, offer to customers as per their discount, and competitor price index.

Early analysis reveals strong positive correlation between advertising spend and sales, though not necessarily linear. The same way, the more traffic you have on your website, it doesn't always mean that you can make more money through them because there conversion rate is different. Management particularly wants to better understand which factors have the greatest effect on sales and how much of the variance in sales they explain.

Your team suggests that they would like to create a multiple regression model with monthly revenue as the dependent variable and looking at the independent variables. They will also statistically test if all the independent variables significantly contribute, whether multicollinearity between advertising and discount offers (that could similarly influence customer purchases) is detected.

And over the time, goal of this model is to be able to predict sales revenue in near future and make strategic decision on budget spending for next month advertising spend, discount program spend and customer engagement based on it.

Critical Thinking Question

R^2 is high in the regression model but some of the independent variables are not statistically significant should management still keep those variables in the model? Explain your response with respect to the balance between model accuracy, interpretability and action.

7.1 Model Identification

7.1.1 Introduction to Simple Linear Regression

This makes simple linear regression particularly well-suited to the problem of trying to quantify and explain the gain between two things: spam related words and whether spam occurs.

- Single dependent variable Y which we are trying to predict or explain.
- One predictor variable (X) that explains or predicts Y.

The Simple Regression Equation:

$Y = \beta_o + \beta_1 X + \varepsilon$

• β_o (Intercept): This is the expected value of Y when X = 0. For example, when we are trying to predict the monthly sales (Y) on the basis of advertisement spend (X), then β_o is the level of sales when advertising spend is zero.

• β_1 (Slope or Beta Coefficient): The amount of change in Y for a one-unit change in X. If β_1 = 50, that means that for every $1 increase in advertising, we expect sales to increase by $50.

• ε (Error Term): Represents the fluctuation of Y around X.

Example:

A clothing store would like to understand the impact of advertising on sales. By pooling 12 months of data, they report:

Sales = 20,000 + 8.5 * Advertising Spend

Here sales increase by 8.5 units for each unit invested in advertising and the number of sales without any advertising being 20,000.

Applications:

• Estimating test scores from study hours.

• Predicting crop yield from rainfall.

• Predicting sales as a function of promotional costs.

7.1.2 Introduction to Multiple Linear Regression

Solving Multiple Regression Multiple regression is simply an extension of the concept of simple regression but applying it for two or more independent variables at a time. This makes possible a more natural and realistic description of the complex situations under consideration in which the future is not usually determined by just one factor.

The general model is:

$Y = \beta_o + \beta_1X_1 + \beta_2X_2 + \ldots + \beta nXn + \varepsilon$

• β_o (Intercept): Y's value when all independent variables are zero.

• β_1, β_2, …, βn (Coefficients): They quantify the magnitude of change depending on an independent variable when everything else is kept constant.

Example:

An airline is seeking to forecast ticket sales (Y) from advertising spend (X_1), ticket price (X_2), and customer loyalty score (X_3). The regression model might appear as:

Sales = 50,000 + 6.5(Advertising) – 120(Ticket Price) + 200(Loyalty Score)

Interpretation:

• Elasticity: For every $1 increase in advertising, sales go up by 6.5 units (ignoring change in the price and loyalty)

• The number sold falls 120 for every $1 hike in price.

• Sales increase by 200 units for every additional point of loyalty.

Applications:

• Forecasting the valuation of homes based on square footage, number of bedrooms, and location.

• Making assumptions about how much a company earns from its marketing and employees and overall economic conditions.

• Predicting demand for a product based on price, competitor prices, and season.

7.1.3 Selecting Appropriate Regression Models

Choosing the right regression model is a key factor in achieving accuracy, precision and interpretability in predictive modelling. Regression analysis aims to describe the relation between a single dependent variable and one or more independent variables. Selection of the most appropriate regression method is influenced by a number of factors, such as the number of predictors, data types, performance criteria, simplicity and theoretical context.

And how the findings may be interpreted and used should they hold up to all or most of these fucking-awful accuracy, completeness regulations. A poor theoretical model and fit can be misleading even if the data are well-fit to the statistical model. On the other hand, a model that is too complex might fit the training dataset very well but not generalize to new observations.

Below are developed first steps that help guide the choice of regression models. There is example Python for each subsection to cut-and-paste when screen sharing.

Number of Predictors

Factors that determine whether a researchers uses simple or multiple regression include the number of independent variables in the data.

• Simple linear regression: If there is single independent variable.

• Multiple Linear Regression: This is used when more than one independent variables influence the dependent variable.

In reality, almost all real-world issues are multifaceted such that multiple regression is more frequent.

Python - Simple Regression Example: Here is a simple example of regression.

Simple Linear Regression import pandas as pd from

import statsmodels.api as sm

data = pd. DataFrame({ 'Experience': [1, 2, 3, 4, 5],

'Salary': [30000, 35000, 40000, 45000, 50000]}

})

X = sm. add_constant(data['Experience']) y = data['Salary']

model = sm. OLS(y, X). fit() print(model. summary())

Multiple Linear Regression data = pd. DataFrame({

‘Experience’:[1,2,3,4,5]

‘Education_Level’ : [10, 12, 14, 16, 18],

'Salary':[30000, 35000, 40000, 45000, 50000]

})

X = sm. add_constant(data[['Experience', 'Education_Level']]) y = data['Salary']

model = sm. OLS(y, X). fit() print(model. summary())

Type of Data

Most regression models work with a continuous dependent variable. The predictor variables may be continuous or categorical. If the features are categorical, they need to be converted into numerical form-dummy variable encoding is a common approach.

• Dependendent variable Numeric and Continuous: The essential feature of regression analysis is that the dependent or response variable is a continuous and numeric value.

• Categorical Independent Variables: These need to be encoded i.e., It needs to be converted from the string format to numeric. This can be performed using different techniques like one-hot encoding or Label Encoder.

Python Example – Categorical Variables Encoding:

Importing the libraries import numpy as np import matplotlib. DataFrame({

'Gender': ['Male', 'Female', 'Female', 'Male'], 'Experience': [1, 3, 5, 7],

'Salary': [30000, 40000, 50000, 60000]

})

Convert 'Gender' to dummy variables df = pd.concat([df,pd.get_dummies(df['Gender'])],axis = 1)

data_encoded = pd. get_dummies(data, columns=['Gender'], drop_first=True)

X = sm. add_constant(data_encoded[['Experience', 'Gender_Male']]) y = data_encoded['Salary']

model = sm. OLS(y, X). fit() print(model. summary())

Goodness of Fit Measures

Assessing the fit of a regression model is critical. A number of statistical measures help to differentiate between these two possibilities:

• R^2(Coefficient of Determination): Represents the percent proportion of variance in the dependent variable that is predictable from independent variables.

• Adjusted R^2: Adjusts R^2 according to the number of predictors present, so as to account for overfitting.

• AIC and BIC: Akaike and Bayesian Information Criteria, model selection criteria that adjust for complexity. Lower values are preferred.

Python Example – Model Evaluation Metrics Here is the Python code of kid to get all the model evaluation metrics: #kid modèle logit et metriques de quality le son

print("R-squared:", model. rsquared) print("Adjusted R-squared:", model. rsquared_adj) print("AIC:", model. aic)

print("BIC:", model.bic)

These measures are used to compare multiple models and select one that strikes a balance between fit (performance) and simplicity.

Parsimony

Parsimony is based on the belief that, other things being equal, simpler models are better. “The model should only consist of the predictors that significantly add to the description of the dependent variable.”

• Simple models are easier to interpret and generalize better.

• Excessive predictors may well overfit the data, particularly with fewer observations.

Feature selection approaches can be utilized to select the most informative variables.

Python Example – Feature Selection with RFE: among from import_python_venn as pv from sklearn. linear_model import LinearRegression from sklearn. feature_selection import RFE

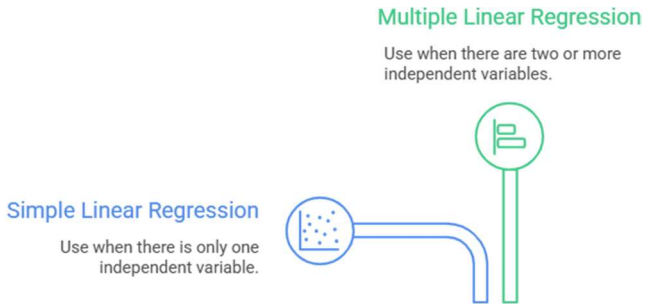

figure: Number of Predictors

data = pd. DataFrame({ 'Experience': [1, 2, 3, 4, 5],

'Education': [10, 12, 14, 16, 18],

'Gender_Male': [1, 0, 0, 1, 1],

'Salary': [30000, 35000, 40000, 45000, 5000]

})

X = data[['Experience','Education','Gender_Male']] y = data['Salary']

model = LinearRegression()

selector = RFE(model, n_features_to_select=2) selector = selector. fit(X, y)

print("Selected Features:", X.columns[selector. support_])

Theoretical Justification

We not only want the model to be statistically strong, but it should have a solid theoretical basis as well. It's misleading to include a variable just because it correlates, statistically, with the dependent variable without logic explanation there is connection.

• There is no causation between statistical co-relation.

• Variables ought to be theoretically relevant or found in the literature.

For example, for a model that predicts loan defaults, we may add credit score and income. But boating up an irrelevant variable like "number of pets owned" (even if statistically significant) could be senseless from a domain standpoint.

Python Example – Dropping irrelevant Predictors:

data = pd.DataFrame({

'Income':[30, 50, 70, 90, 110],

'Credit_Score': [600, 650, 700, 750, 800],

'Pets_Owned': [2, 1, 4, 3, 2],

'Loan_Default': [0, 0, 1, 1, 0]

})

Exclude 'Pets_Owned' from model

X = sm. add_constant(data[['Income', 'Credit_Score']]) y = data['Loan_Default']

model = sm. OLS(y, X). fit() print(model. summary())

The removal of irrelevant predictors is useful to keep the model interpretable and theoretically accretive.

7.1.4 Assumptions in Regression Analysis

Conclusions Analysis of regression is correct when: some conditions are met. These ones allow for valid parameter estimates, statistical inferences and forecasts.

Linearity:

Proportionality between dependent and independent variables the relationship of which is defined by linearity must exist. Preventive dentistry Perception for child 5. If the association is not linear, transformations, or polynomial regression are suspected.

Example: Sales versus advertising spend results often make use of log transformation for diminishing returns line.

Independence of Errors:

The residuals (errors) should be independent of each other. This is particularly significant for time- series data. Autocorrelation leads to misleading estimates.

Example – Most stock price predictions break this rule due to temporal dependence.

Homoscedasticity:

Homoscedasticity The residual variance should be consistent for all levels of the independent variables. Heteroscedasticity is when variance goes up or down in an organized way.

For instance: Variability in spending between higher and lower-income groups may be larger for income than expenditures.

Normality of Residuals:

Mistakes are supposed to be distributed normally. Such is imperative for statistical hypothesis testing and the confidence estimates.

Illustrative example: If residuals are non-normal, tests of the significance of the parameters can be suspect.

No Multicollinearity:

The independent variables should not be too closely related to one another, since such relations can provide misleading coefficients. We calculate the VIF (Variance Inflation Factor) to check for multicollinarity.

Example: Adding both "years of experience" and "age" in the prediction of salary could cause problems like high correlation.

No Autocorrelation:

Residuals should be uncorrelated over time. If the errors in one period affect the errors in another, it breaches this condition. The Durbin-Watson is a test for auto-correlation.

Example: In a time-series data such as monthly sales, autocorrelation is common unless seasonality is modeled.

Consequences of Violations:

• Estimates obtained from the regression analysis may be biased or inefficient when assumptions are not satisfied.

• Estimates may not be accurate and statistical tests can have misleading conclusions.

• Some of the remedies in these situations are transforming variables, introducing interaction terms, implementing robust regression or fitting other model types (e.g., generalized linear models).

“Activity: Exploring Regression through Real-World Data Patterns"

Students will collect a small dataset (e.g., study hours vs. exam scores or advertising spend vs. sales) and apply both simple and multiple linear regression. They will test assumptions such as linearity, independence, and homoscedasticity using residual plots, interpret coefficients, and decide which regression model best explains the data.

7.2 Model Development

7.2.1 Steps in Building Regression Models

It's a methodical dance to build a regression model that is accurate and meaningful.

Problem Definition and Objective Setting

And the starting point is an unambiguous statement of what you want the regression model to do.

• Are goals those of prediction (predicting future) or explanation (understanding what is related to what)?

• The precision of the problem, which means all variables relevants to be entered while irrelevants excluded. Example: A hospital might be interested in predicting recovery time (dependent variable) as a function of age, type of treatment, and number of hospital visits(independent variables).

Data Collection

Good data is the foundation of a valuable regression model.

• Sources may be surveys, data from companies, government data sources; sensors or observations recorded by secondary sources.

• Both dependent and independent must be compiled. For a sales model, for example, gather monthly sales data, advertising spend, social media impressions, competitor prices and seasonal dummy variables.

Data Preparation and Cleaning

Generally speaking, the raw data need preprocessing before any data analysis.

• Missing Values : Methods to impute missing values such as mean, median or any other variable.

• Slam shut on and treat as outliers that may distort.

• Do log or square-root transformations of skewed variables.

• Create dummies for categorical variables (e.g. 0 = Male, 1 = Female). Example: If a dataset is missing some customers' income, replacing them with the median values can be used to omit bias.

Exploratory Data Analysis (EDA)

EDA act as a tool for discovering trends in the data.

• Explore variable relationships using correlation analysis.

• Create scatterplots to examine linearity between predictors and response variable.

• Check multicollinearity (presence of strong intercorrelations among the predictors). Example: If "advertising spend" and "discount offers" are highly correlated, adding both could bloat error terms.

Model Specification

The latter step is the issue of selecting the appropriate type of regression equation.

• Choose between simple linear regression (one predictor) and multiple regression (multiple predictors).

• Exclude irrelevant variables to avoid overfitting. Example: Include square footage and location to model housing prices but not unrelated variables such as the house paint color.

Parameter Estimation

Regression coefficients are estimated by fitting to the model statistical methods such as Ordinary Least Squares (OLS).

• And each coefficient is measuring the impact of a predictor on the dependent variable.

• Statistical software (such as Excel, R, Python and SPSS) takes care of this automatically. Example: A coefficient of 2.5 for "advertising spend" implies that for each dollar increase in advertising, sales go up by $2.50 (all else being equal).

Model Diagnostics and Assumption Testing

Validity of regression Validity of a regression analysis is conditional to the satisfaction of some assumptions, such as necessity for linearity, independence, homoscedasticity or not.

• Use residual plots to verify that homoscedasticity is held.

• Detect multicollinearity using Variance Inflation Factor (VIF).

• Apply Durbin-Watson test to check autocorrelation in time-series data.

For example, the linear model is probably not appropriate if residuals exhibit a clear pattern.

Model Evaluation

Assess the model using varied metrics:

•R^2 y R^2 Ajustado: Cuanta varianza en Y explica el modelo.

• RMSE (Root MeanSquare Error): is calculated as the square root of the average of the squares of error.

• MAE (Mean Absolute Error): The average absolute difference between the predicted and actual values.

For example: If $R^2 = 0.85$, then 85% of the variance in the dependent variable is explained.

Model Validation

Validation makes sure your model does not overfit training data.

• Common methods: hold-out validation (train-test split) and cross-validation (k-fold). For instance, if we train on 70% of the data then our model should behave properly for the remaining 30% to be reliable.

Model Deployment and Interpretation

3.6 Apply the model to practical problems.

• Translate coefficients into actionable insights.

• Give clear, concise interpretations that are directly relevant to managers/policymakers. Example: After realizing that online advertising has a strong positive impact on sales, a retailer might opt to not just maintain but actually increase it.

7.2.2 Training and Testing Data Split

It is the key most important thing in model building to split the data set into training and test sets, while working with regression models.

Concept of Data Split

• The data is split into training (used to make the model) and testing sets (used to assess how it performs).

• 70%, 30% or 80:20 (where dataset size is large).

Training Data

• Larger portion of the dataset.

• The model “learns” relationships between variables here by estimating coefficients.

Overfitting: If the model gets too 'smart' and is train on data that's too specific .

Testing Data

• Smaller portion of the dataset.

• Unseen by the model during training.

• Unbiased assessment of model performance on new, unseen data.

• Makes sure that the model is not only recalling but generalizing.

Importance of Splitting

• Reduces overfitting, i.e. the model overfits and fits noise in data (which is not supported on new data).

• Recommendations to ensure your evaluation metrics match real-world predictive performance.

Methods of Splitting

Random Splitting: The dataset is randomly splitted into the training and testing sets.

Stratified Split: maintained the proportion of categories in the dependent variable in each split.

Cross-Validation (e.g., k-fold): The dataset is divided into k parts; the model is trained and tested k times, each time using a distinct part for testing. This is better in that it is more reliable, particularly with smaller training sets.

Example

Imagine that a company has 1,000 customers and it wants to forecast their spending using the customer's income, age and loyalty score.

• Training sample (800 customers): For regression equation building purposes.

• Test set (200 customers): Applied to measure predictive accuracy.

If you have very high R^2 on the train data, but very low on the test data (R^2 = 0.92 and 0.55 respectively), it means

overfitting. The model overfits the training dataset and does not generalize.

Practical Application

• For financial prediction, the training phase could be based on performance that has occurred in the past for a company and testing is predicting next quarter results.

• In health care, patient data is divided to train a model to predict the risk of disease and then tested on new patients to see if it makes accurate predictions.

Did You Know?

"Did you know that splitting data into training and testing sets helps prevent overfitting, ensuring models perform well on new data? A typical 70:30 or 80:20 split mimics real-world scenarios, where models must predict outcomes for unseen cases, making validation essential for reliability and accuracy."

7.3 Model Evaluation

7.3.1 Model Perfomance Evaluation (R^2, RMSE, MAPE, MSE, MAE)

Various metrics account for different information on model accuracy. Depending on one measure might be mistaken so certain metrics are considered together.

R^2 (Coefficient of Determination)

• Represents the proportion of variance in the dependent variable accounted for by independent variables.

• Formula:

$R^2 = 1 – (SSR / SST)$

o SSR = Sum of Squared Residuals (unexplained variation).

o SST = Sum of Squares Total (variance between individuals).

• Values range from 0 to 1. High R2 value represents strong explanatory capability.

Example:

If $R^2 = 0.85$ on a sales forecast, it means that advertising, price and customer loyalty help explain 85% of the variability in sales. The other 15% comes from things that can't be modeled.

Limitations:

• High value of R^2 does not necessarily mean the model is good.

• No matter how irrelevant it is, adding more predictors will always increase R^2.

• For this reason, Adjusted R^2 is commonly used in practice because it adjusts for the number of predictors.

RMSE (Root Mean Squared Error)

• The square root of the mean (per image) squared error.

• Formula:

$RMSE = \sqrt{(\Sigma (Y_i – \hat{Y}_i)^2 / n)}$

• Poorly suited to large errors in data, since it squares error values before taking the mean.

Example:

For example, if in a demand forecasting model the RMSE = 5, this means that per average predictions miss actuals by about 5 units.

Strengths: It's very punishing of large errors, which may be helpful if a big mistake is particularly costly (for instance, in medical predictions).

MSE (Mean Squared Error)

• The average, before taking the square root, of the squared errors (how wrong) or difference between predicted and actual values.

• Formula:

$MSE = \Sigma (Y_i - \hat{Y}_i)^2 / n$

Lower values are always better, and the eigenvalue must be non-negative.

Example:

MSE = 25 means squared prediction errors are on average 25 units². It is less interpretable in real units, but it is often used as the optimization criterion during training the model.

MAE (Mean Absolute Error)

• Average of absolute prediction errors.

• Formula:

$MAE = \Sigma |Y_i - \hat{Y}_i| / n$

• More intuitive than MSE/RMSE as same units as the dependent variable.

Example:

If MAE = 3 when predicting exam scores, we are off by 3 points on average.

Pros: Less sensitive to outliers compared with RMSE.

MAPE (Mean Absolute Percentage Error)

• Expresses errors as percentages.

• Formula:

$(\Sigma (|Y_i - \hat{Y}_i)/Y_i | \times 100) / n$ MAPE: Where, Y_i = Actual value $\hat{Y}_i$=predicted value n=number of measurements.

• Good for business forecasting because it is relative error.

Example:

MAPE = 8% tells you that predicting the demand of product, your model' s prediction would be wrong, on an average by 8%.

Limitations:

• Not applicable when the real values (Y_i) are zeros.

• Insensitive to the very small value of Y.

Summary of Use-Cases:

• R^2/Adjusted R^2 → Explanatory power.

• RMSE / MSE → Error size, penalize large error.

• MAE → Interpretive error in actual units.

• MAPE → Business-friendly percentage error.

Did You Know?

"Did you know that a model with extremely high training accuracy can still fail in real-world predictions due to overfitting? Conversely, an underfitted model misses key patterns, performing poorly everywhere. Model validation techniques like cross-validation help strike the right balance, ensuring reliable and generalizable predictive performance."

7.3.2 Overfitting, Underfitting, and Model Validation

A model's performance should not be judged solely on its fit to training data, but also on its ability to generalize to new data. Here is where the ideas of overfitting, underfitting, and validation comes into play.

Overfitting

• Occurs when model memorizes the noise in training data instead of generalizing the underlying pattern.

• Characteristics:

o Very high R^2 on training data/ very poor performance on test data.

o Too many variables or/and complex transformations.

• Cause: The model is too wobbly.

Example:

50 Gmail spam features can work well in predicting spam (in classification) with only 2.5% test error, but using these same 50 predictors to predict the size of a house gives a ludicrous model that correctly predicts zero new houses!

Results: False predictions in practical application.

Underfitting

• Happens if the model is simplistic to represent data trends.

• Characteristics:

o Poorly performing on both train and test data.

o Important predictors are ignored.

• Cause: The model lacks complexity.

Example:

Predicting house value with addresses and bedrooms as features leaving out the neighbourhood and square footage.

Result: The model fails in any situation.

Model Validation

Validation mechanisms are employed to avoid of underfitting and overfitting.

Hold-Out Validation (Train-Test Split):

o Data is divided into training (e.g., 80%) and testing (20 %) subsets.

o Simple and commonly used but outcome varies according to data splitting.

k-Fold Cross-Validation:

o Data is divided into k approximately equal partitions 82 Image by: redeviveri/kaggle-competition/ Page of 105 86 (folds).

o Model is fitted on k-1 folds and tested on 1 fold.

o Repeated until the test set was all used.

o Provides more reliable performance estimates.

Bootstrapping:

o Random sub-samples (with replacement) from the dataset are repeatedly drawn.

o Models are tested on these samples to test for robustness.

Example of Validation:

If you build a churn model for customers that has 95% accuracy on the training data and just 60% accuracy in test, it's having an overfit. Cross-validation may also suggest that some redundant predictors need to be removed or the model should be regularized.

Practical Application

• For credit risk predictions in finance, precise model assessment is important.

• As shown in marketing, error metrics Have the ability to guide budget allocation decision making by providing campaign returns.

• In health care, preventing overfitting is important for ensuring that predictive models of diseases can extend well to other patient groups.

7.4 Multilinear Regression

7.4.1 Concept of Multiple Linear Regression

Definition

o Relation of a single dependant variable with several independent variables is formulated by it.

o We assume that the impact of each predictor is linear and independent.

Purpose

o For more accurate prediction of the response by including all factors that could potentially influence the response.

o To test the significance of each predictor.

Example

o Forecasting house prices (Y) from predictors as square footage (X_1), number of bedrooms (X_2), location index (X_3), and the age of the property (X_4).

Equation:

Price = 30,000 + 100(Area) + 8,000(Bedrooms) +15,000(Location factor) −500(Age in years.

Interpretation: Each of the factors uniquely adds to the ability of predicting house price.

7.4.2 Interpreting Coefficients in Multilinear Regression

Regression coefficients should be interpreted according to both their magnitude and direction:

Intercept (β_o)

o The mean E(Y) when all the predictors are zero.

o Has marginal value in reality, but required for the formula.

Estimate (β_1, β_2, …, βn) of the Slope Coefficients

o Standing for the change in DV that accompanies a one-unit increase in the predictor,

holding all other variables constant.

o Positive coefficient→ predictor – Y gets larger.

o Negative coefficient → predictor predicts Y decreases.

Standardized Coefficients (Beta values)

o Facilitates comparisons of the relative significance/importance of predictors when the variables are measured on different scales.

Statistical Significance (p-values)

o A coefficient is significant if $p < 0.05$, this means that the predictor affects Y to a meaningful extent.

Example:

In a salary prediction model:

Salary = 25,000 + 2,000(Years of Experience) + 5,000(Master's Degree Dummy)

• So the coefficient of Years experience is: "with an increase in 1 year, you can expect your salary to increase by $2,000".

• The dummy for Master's Degree means having the degree makes one earns $5,000 more than no having it.

7.4.3 Business and Research Uses of Multilinear Regression

Multilinear is applied in all fields where the dependent variable can be influenced by more than one factor.

Business Applications

o Marketing: The effect of advertising expenditure, social media engagement and competitive pricing on sales.

o Finance: Predicting stock returns with interest rates, GDP growth and inflation.

o Operations: Estimate the delivery time based on distance, traffic and number of shipments.

o Staff: Measuring individual performance in terms of job experience, training hours and scores for job satisfaction.

Research Applications

o Economics: Investigating the impact of education, experience and region on earnings.

o Medical: How long will it take for a patient to recover depending on their age, treatment type and lifestyle?

o Social Sciences: Analyzing the effects of factors on academic performance like hours of study, attendance and social background.

o Environmental Studies: Prediction of pollution based on population density, level of industrial activity and number of vehicles.

Advantages

o Allows a complete view of intricate relationships.

o Enable adjustment for confounding by breaking down the effect of each predictor.

o Provides better prediction accuracy than simple regression.

Limitations

o Prone to multicollinearity: when predictors are correlated.

o They rely on linearity and additivity which does not hold in all practical cases.

o Interpretation is difficult in presence of interactions between predictors.

7.5 Summary

- Regression analysis models that can be used to determine and quantify the relationship between dependent and independent variables.
- Simple linear regression has one predictor, and multiple linear regression has two or more predictors.
- Model building proceeds through structured stages: definition of the problem, preparation of the data, specification of the model, estimation, appraisal and validation.
- Dividing the data to training and testing sets can overcome overfitting problem as well as improve generalization.
- Model performance has been evaluated based on R^2, RMSE, MSE, MAE and MAPE for accuracy and reliability.
- Overfitting is when we have a model that is too complex, underfitting occurs when a model is not complex enough.
- Validation procedures such as cross-validation tradeoff between model complexity and predictive strength.
- In multi-logistic regression (form of) the coefficients are telling you what is the effect of one other predictor variable than either =1 and large than or not.
- Some business and research uses of multiple regression are: Marketing Finance Health care Economics Social science
- A good regression model should be statistically and also practical useful in the decision- making process and forecasting.

7.6 Key Terms

1. Regression Model: A statistical analysis which describes the association between dependent and independent varia bles.
2. Response (Y axis) – The variable whose values are being predicted or explained.
3. Independent Variable (X): The variable that is responsible for the change in the dependent variable.
4. Intercept (β_0) – The value of the dependent variable when all independent variables are also equal to zero.
5. Coefficient (β) – The amount of change in the dependent variable per unit change of one predictor variable, while controlling for other variables.
6. R^2 (Coefficient of Determination) – Percentage of variation in the dependent variable that is predictable from the independent variables.
7. RMSE (Root Mean Squared Error) – The root mean of the squared differences between the predicted and actual values, punishing a large error more.
8. Overfitting The model becomes too much trained into data, that it performs pathetically well on new data.
9. Multicollinearity – when the independent variables are too much correlated for them to be used as separate factors.

7.7 Descriptive Questions

1. Differentiate between simple linear regression and multiple linear regression with appropriate examples.
2. Explain the high-level process to go from collecting data to deploying a model and all steps in between.
3. What are the major assumptions that needs to hold for a regression analysis and why it is important to check them?
4. Explain the importance of training and test data in validation of a regression model.
5. Describe the various types of model evaluation metrics (R^2, RMSE, MAE, MAPE,MSE) that are used to evaluate regression performance.
6. What is over-fitting and under-fitting in regression? How could the validation of models factor in on this issue?
7. How do you explain the coefficients from a multiple linear regression? Provide a business- related example.
8. Describe three or more business applications in which multiple regression is appropriately used.

7.8 References

1. Draper, N. R., & Smith, H. (1998). Applied Regression Analysis (3rd ed.). Wiley-Interscience.

2. Kutner, M. H., Nachtsheim, C. J., & Neter, J. (2004). Applied Linear Regression Models (4th ed.). McGraw-Hill/Irwin.

3. Montgomery, D. C., Peck, E. A., & Vining, G. G. (2012). Introduction to Linear Regression Analysis

(5th ed.). Wiley.

4. Wooldridge, J. M. (2016). Introductory Econometrics: A Modern Approach (6th ed.). Cengage Learning.

5. Gujarati, D. N., & Porter, D. C. (2009). Basic Econometrics (5th ed.). McGraw-Hill.

6. James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An Introduction to Statistical Learning: With Applications in R. Springer.

7. Field, A. (2017). Discovering Statistics Using IBM SPSS Statistics (5th ed.). Sage Publications.

8. Freedman, D. A. (2009). Statistical Models: Theory and Practice (Rev. ed.). Cambridge University Press.

Answers to Knowledge Check

Knowledge check 1

1. c) One dependent, multiple independents

2. b) Change in Y per unit change in X

3. c) Predictor increases Y

4. b) Estimating sales from advertising and price

7.9 Case Study

Forecasting Retail Sales with Regression Models

Introduction

Retail businesses have to make uncertain future sales forecasting everyday considering various factors for example advertising, pricing strategy the customer response and what their competitor is doing. Such a structured method applied through regression models help managers to understand the main factors affecting sales and make data based decisions. This case describes how one retail chain used regression analysis to increase the accuracy of its sales forecasts and improve strategic planning.

Background

A mid-tier retail chain that worked in five cities need to know, why and which factors drove the revenue for their monthly sales. Management gathered information about advertising costs, customer discounts, website visitors and competitive prices during a two-year period. Descriptive analysis only showed high-level trends, and the business wanted more – they were looking for a model that could predict future sales and optimise resources.

Key challenges included:

• Multiple factors simultaneously influencing sales.

• Challenging to determine the contribution of predictors separately.

• Overfitting when utilizing too many predictors.

To tackle this problem, the analytics community instead opted to create multiple linear regression models.

Problem 1: The important factors that are affecting the sales.

The retailer didn't know which of these independent variables was contributing the most to sales. Adding all the variables would simply overcomplicate the model.

Solution:

Stepwise regression and Adjusted R^2 were used by the team to identify the predictive factors.

Results indicated that advertising and consumer promotions impacted sales, while competitor price was the least important.

MCQ:

Which strategy to determine the most relevant predictors in regression?

a) Random selection

b) Stepwise regression

c) Ignoring insignificant predictors

d) Using all available variables

Answer: b) Stepwise regression

2) The Model May be Overfitting Problem # 2: There is a risk that the model may have overfit.

Our initial attempt to build a regression model achieved a very high R^2; however, it performed extremely poorly on new data. This suggested overfitting.

Solution:

The dataset was divided to the team into 70% for training and 30% for testing. However, they also employed cross- validation to validate model stability. The model after convergence struck a balance between training accuracy and predictive power.

MCQ:

How do you avoid overfitting with regression models?

a) Evaluate only using training data

b) Apply cross-validation and testing data

c) Increase predictors without testing

d) Exclusion of error terms of the models

Answer: b) Use cross-validation and testing data

Issue 3: Reading Coefficients for Decision Making

Managers found it difficult to interpret regression coefficients in terms of actionable business strategies.

Solution:

The analytics team explained that:

• The coefficient of advertising spend is positive, meaning when the ad budget was increased, sales increased as well.

• A negative coefficient for competitor price suggested sales were decreasing if competitors reduced prices.

Managers used this version of the interpretation to scalpel budgets and promotions.

MCQ:

What does a positive coefficient mean?

a) Predictor decreases Y

b) Predictor increases Y

c) No relationship exists

d) Predictor has no statistical significance

Answer: b) Predictor increases Y

Conclusion

The retailer used the regression model to predict sales and plan human resources more effectively. Management derived actionable insights by with judicious predictor selection, through train-test validation of the model and coefficient interpretation. The case drives the point for regression models and walking a fine line between accuracy, interpretability and making business decisions.

BUPBM Unit 8 V3.docx

Dynamics of Family Managed Business

Dynamics of Family Managed Business

ATLAS SkillTech University

Document Details

Submission ID

trn:oid:::3618:128395422

Submission Date

Feb 16, 2026, 4:13 PM GMT+5:30

Download Date

Feb 16, 2026, 4:21 PM GMT+5:30

File Name

BUPBM Unit 8 V3.docx

File Size

38.1 KB

19 Pages

4,289 Words

26,685 Characters

*% detected as AI

AI detection includes the possibility of false positives. Although some text in this submission is likely AI generated, scores below the 20% threshold are not surfaced because they have a higher likelihood of false positives.

Caution: Review required.

It is essential to understand the limitations of AI detection before making decisions about a student's work. We encourage you to learn more about Turnitin's AI detection capabilities before using the tool.

Frequently Asked Questions

How should I interpret Turnitin's AI writing percentage and false positives?
The percentage shown in the AI writing report is the amount of qualifying text within the submission that Turnitin's AI writing detection model determines was either likely AI-generated text from a large-language model or likely AI-generated text that was likely revised using an AI paraphrase tool or word spinner.

False positives (incorrectly flagging human-written text as AI-generated) are a possibility in AI models.

AI detection scores under 20%, which we do not surface in new reports, have a higher likelihood of false positives. To reduce the likelihood of misinterpretation, no score or highlights are attributed and are indicated with an asterisk in the report (*%).

The AI writing percentage should not be the sole basis to determine whether misconduct has occurred. The reviewer/instructor should use the percentage as a means to start a formative conversation with their student and/or use it to examine the submitted assignment in accordance with their school's policies.

What does 'qualifying text' mean?
Our model only processes qualifying text in the form of long-form writing. Long-form writing means individual sentences contained in paragraphs that make up a longer piece of written work, such as an essay, a dissertation, or an article, etc. Qualifying text that has been determined to be likely AI-generated will be highlighted in cyan in the submission, and likely AI-generated and then likely AI-paraphrased will be highlighted purple.

Non-qualifying text, such as bullet points, annotated bibliographies, etc., will not be processed and can create disparity between the submission highlights and the percentage shown.

Unit 8: Classification Models (Logistic regression)

Learning Objectives

1. Explain the fundamental concepts of classification and how logistic regression differs from linear regression.

2. Describe the mathematical formulation of logistic regression using the sigmoid function.

3. Interpret the meaning of coefficients and odds ratios in logistic regression models.

4. Differentiate between binary, multinomial, and ordinal logistic regression applications.

5. Apply logistic regression to classify outcomes and predict probabilities of categorical events.

6. Evaluate logistic regression models using metrics such as accuracy, precision, recall, F1-score, and ROC-AUC.

7. Identify and address assumptions, limitations, and potential pitfalls in logistic regression modeling.

8. Use statistical and software tools (e.g., R, Python, SPSS, Excel) to build and validate logistic regression models.

9. Demonstrate business and research applications of logistic regression in areas such as marketing, healthcare, finance, and social sciences.

Content

8.0 Introductory Caselet

"Predicting the Churn Rate of an App using Logistic Regression"

A telecommunications organisation, ConnectPlus has lost many of its subscribers. A large number of customers stop utilizing services within a few months and management wants to identify the factors that lead to customer churn. The company has gathered information about customers' monthly charges, type of contract, internet usage, complaint history and tenure.

Logit The analytics team recommends us to make use of logistic regression because the outcome variable is categorical (churn: Yes/No). Using logistic regression, the team is able to predict the likelihood of churn for every customer and determine which factors are most important. First findings indicate that high monthly charges and many complaints have a strong positive influence on the probability of churning, whereas, lengthy contracts can decrease it.

The model puts out probabilities (i.e. a customer has 75% likelihood to churn). All the customers are assigned to one of two categories, "likely to churn" or "likely to stay", using a given cutoff (e.g. 0.5). This allows the company to develop tailored retention campaigns and provide, for example, reduced pricing or enhanced services to customers who are most at risk of departing.

Critical Thinking Question

If a logistic regression model produces 85% accuracy in terms of predicting churn, but at the same time, it misclassifies many high-value customers as "not at risk," should the company use overall accuracy as an evaluation metric to ensure this is its best approach? What other indicators or methods might offer a more equally weighted perspective for business decision-making?

8.1 Classification Models

8.1.1 Introduction to Logistic Regression

Concept and Purpose

Logistic regression is used when the outcome to be predicted is binary (in two-categories). It doesn't predict the continuous value but it predicts probability of outcome.

• Example: Predict if a customer will churn (1 = Yes, 0 = No).

Unlike linear regression, which could lead to infeasible predictions (like a probability of -0.3 or 1.5), logistic regression makes sure that the predicted probabilities are always between 0 and 1.

The Logistic (Sigmoid) Function

The sigmoid curve is used in the logistic regression model:

• The S-curve can be used to give any value input a probability between 0 and 1.

• Prob (probability) > threshold(if this is 0.5, then use a UserSystemType as yes if it is considered positive otherwise it's no(assuming Positive less than 0)). Otherwise, “No” (0).

Odds and Log-Odds

• Logistic regression is performed on the odds of an event:

This transformation makes it possible to apply linear methods to categorical outcomes.

Types of Logistic Regression

Binary Logistic Regression – It has only two possible outcomes.absent or present, dead or alive etc.

Multinomial Logistic Regression – When there are more than two unordered categories (Transport mode: Car, Bus, Train).

Ordinal Logistic Regression – Categories with order (Customer satisfaction Low, medium and High).

Example

A hospital wishes to predict if a patient has diabetes (Yes/No) based on their glucose level, BMI and age.

• Logistic regression calculates the probabilities of membership in a class; for example, 0.82 means an 82% probability of having diabetes).

• The doctor can again avoid the redundant slave them “Patient had Diabetes” If the confidence crossed cut-off}

8.1.2 Logistic Regression Model Development

The logistic regression is a supervised learning which is applied to the cases with binary and categorical dependent variables. We would like to predict the chance a dependent variable happens based on one or more independent variables. This sectoin walks through a practical step-by-step routine of tasks associated in constructing logistic regression model, theoretical aspects are explained first and the sequence is followed by its Python implementation for demonstration purpose.

Step 1: Problem Definition

When you are ready to train a logistic regression model, this will be your first step: Be Clear on what the problem is and what the categorical outcome variable should look like. Binary

logistic regression is used when the dependent variable has two outcomes, which may be labeled "yes" and "no".

Example: Make a prediction about whether an applicant will default on a loan.

• Dependent Variable: Default (Yes/No)

• Dependent Variable: Employment status• Independent Variables: The credit score, the income level and amount of loan.

This step defines the basis of data selection, features engineering and model design.

Step 2: Data Collection

The second stage is to collect the candidate predictors. These predictors need to be reasonably related to the response. For a default classification problem, predictive variables might be income, credit score, loan term and employment status.

Python Example – load sample data:

import pandas as pd

Sample data

data = pd.DataFrame({

'CreditScore': [600,720,690,580,710],

'Income': [45000, 54000, 50000, 39000, 52000],

'LoanAmount': [10000,15000,12000,9000,13000],

'EmploymentStatus': ['Employed', 'Employed', 'Self-employed', 'Unemployed', 'Employed'], 'Default': [0, 0, 0, 1, 0]

})

Step 3: Data Preparation

The correct statistical treatment of the data is important so that it performs well. This includes managing missing values, handling outliers, encoding categorical data, and scaling of the features.

• Missing values - Replaced with Mean or Median.

• Outliers: Discard (or clamp) extreme values.

• Encoding: The Categorical variables have to be converted by using one-hot-encoding.

•Scaling: Normalize features to ranges of similar scale.

Python Example – Data Preprocessing:

from sklearn. preprocessing import StandardScaler

Encoding categorical variable

data = pd. get_dummies(data, columns=['EmploymentStatus'], drop_first=True)

Feature scaling

scaler = StandardScaler()

scaled_features = scaler. fit_transform(data[['CreditScore', 'Income', 'LoanAmount']])

scaled_df = pd. DataFrame(scaled_features, columns=['CreditScore', 'Income', 'LoanAmount'])

Merge one-hot-encoded features and scaled features, after the completion of their respective operations

processed_data = pd. concat([scaled_df, data[['EmploymentStatus_Self-employed', 'EmploymentStatus_Unemployed', 'Default']]], axis=1)

Step 4: Model Specification

Model specification is the listing of the dependent and independent variables. Predictor choices are based on relevancy to theory, previous findings or exploratory data analysis. Correlation between predictors should be minimized.

Python Example – Defining Variables:

X = processed_data. drop('Default', axis=1) y = processed_data['Default']

Step 5: Parameter Estimation

Model parameters are estimated from logistic regression via Maximum Likelihood Estimation (MLE). MLE determines the parameter values such that the given observations are most probable under the logistic function.

Python Example – Fit Logistic Regression:

from sklearn. linear_model import LogisticRegression

Model initialization & training model = LogisticRegression() model. fit(X, y)

Step 6: Model Evaluation

We measure model performance in terms of classification metrics, not R-squared. Common evaluation tools include:

- Table Confusion : Affiche TP, FP, TN, FN
- Accuracy: (TP + TN) / Total

• Precision: TP / (TP + FP)

• Recall: TP / (TP + FN)

• F1 Score: The harmonic mean of precision and recall

• ROC Curve and AUC: To assess discrimination

Python Example – Evaluation Metrics:

from sklearn. metrics import confusion_matrix, accuracy_score, precision_score, recall_score, f1_scoreroc_auc_scoreroc_curvrom sklearn.

import matplotlib. pyplot as plt

Predictions

y_pred = model.predict(X)

y_prob = model. predict_proba(X)[:, 1]

Confusion matrix and metrics

print("Confusion Matrix: ",confusion_matrix(y,y_pred)) print ("Accuracy:", accuracy_score(y, y_pred)) print("Precision:",precision_score(y, y_pred)) print("Recall:",recall_score(y, y_pred))

print("F1 Score:",f1_score(y, y_pred)) print("AUC:", roc_auc_score(y, y_prob))

ROC Curve

fpr, tpr, thresholds = roc_curve(y,y_prob)

plt. plot(fpr, tpr, label="ROC Curve (area = %0.2f)" % roc_auc_score(y, y_prob)) plt. xlabel("False Positive Rate")

plt. ylabel("True Positive Rate") plt. title("ROC Curve") plt. legend()

plt.show()

Step 7: Model Validation

In order to make sure the model does not simply learn noise, we validate the result by using train-test split or k- fold cross validation.

• Train/Test Split: Usually 70–80% as train and 20–30 % for test

• Cross-Validation : Estimates the strength on various data subsets

Python Example – Train-Test Split & Validation:

from sklearn. model_selection import train_test_split

Split data

X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3) # random_state=42 (test size or radom state must be there)

Fit model on training set model = LogisticRegression() model. fit(X_train, y_train)

Evaluate on test set

y_test_pred = model. predict(X_test)

print("Test Accuracy:", accuracy_score(y_test, y_test_pred))

Step 8: Model Interpretation

Coefficients from the logistic regression are interpreted as odds ratios. These tell you how the odds of having the outcome change for a 1 unit increase in that predictor, given that everything else is kept constant.

For instance, if the coefficient of a predictor is positive (negative), the higher that predictor is, the more likely (less likely) the event is to occur.

Python Example – Coefficient Interpretation:

import numpy as np

Coefficients and odds ratios coefficients = model. coef_[0] odds_ratios = np. exp(coefficients)

for feature, coef, odds in zip(X.columns, coefficients, odds_ratios): print(f"{feature}: Coefficient = {coef:. 3f}, Odds Ratio = {odds:. 3f}")

Step 9: Model Deployment

When validated, interpreted the logistic model able to use as a model for categorizing and decision making. For instance, in a telecom churn model, where the company can see high-risk customers and use retention strategies. For example, in banking, logistic regression can be used to flag high-risk borrowers for a loan.

Example in Python – Use Model to Make Predictions On New Data:

Example new customer data (should be preprocessed in similar way) new_data = pd. DataFrame({

'CreditScore': [685],

'Income': [48000],

'LoanAmount': [11000],

'EmploymentStatus_Self-employed': [0],

'EmploymentStatus_Unemployed': [1]

})

Apply scaler used earlier

new_data_scaled = scaler. transform(new_data[['CreditScore', 'Income', 'LoanAmount']]) new_data_scaled_df = pd. DataFrame(new_data_scaled, columns=['CreditScore', 'Income', 'LoanAmount'])

Final input for prediction

new_input = pd. concat([new_data_scaled_df, new_data[['EmploymentStatus_Self-employed', 'EmploymentStatus_Unemployed']]], axis=1)

Predict probability and class

probability = model. predict_proba(new_input)[:, 1] classification = model. predict(new_input)

print("Predicted Probability of Default :", probability[0]) print("Predicted Class (0 = No default, 1=Default) :", classification[0])

Business Application Example – Banking

A bank uses logistic regression to predict if people will default on loans. The primary predictors are based on an applicant's credit score, income and loan amount.

• Every point higher in credit score is associated with a 5 percent lower probability of default.

• Each 10-percentage-point increase in the loan-to-income ratio increases default probability by 15 percent.

Under this model, the bank modifies lending policies to lower its risk with high-risk-type customers.

Did You Know?

"Did you know logistic regression does not use Ordinary Least Squares like linear regression, but instead relies on Maximum Likelihood Estimation (MLE)? This method finds the parameters that maximize the probability of observing the given data, ensuring more accurate predictions for categorical outcomes such as churn or loan default."

8.2 Assessing Model Performance (Logistic Models)

8.2.1 Confusion Matrix and Metrics, Accuracy, Precision, Recall F1-Score

A confusion matrix is the mutual comparison of the predicted and true features:

Predicted Positive Predicted Negative Actual Positive True Positive (TP) False Negative (FN) Actual Negative False Positive (FP) True Negative (TN) Based on this table, FVA has derived several performance metrics:

Accuracy

• Quantifies the rate of true predictions.

• Formula:

Accuracy = (TP + TN) / (TP + TN + FP + FN)

• Example: For 1000 cases, in which 850 are correct, accuracy would equal 85%.

• Con: Misleading in unbalanced datasets (like fraud detection where 99% are “No Fraud”).

Precision (Positive Predictive Value)

•What proportion of positives predicted are actually positive?

• Formula:

Precision = TP ÷ (TP + FP)

Example: If based on a model’s prediction, 100 customers are expected to churn and if in reality 80 of them churn, precision = 80%.

Recall (Sensitivity or True Positive Rate)

• How many of the true positives did it correctly predict?

• Formula:

Recall = TP ÷ (TP + FN)

• Example: Recall was 75%: if the true value of customers who churned were 100 and our model also predicted that for 75 of them.

F1-Score

• Achieves a compromise between precision and recall (harmonic mean).

• Formula:

F1 = 2 × (Precision × Recall) / (Precision + Recall)

• Example: If precision = 0.80 and recall = 0.75 → F1 = 0.77.

Key Takeaways:

• Accuracy is maximized when classes are balanced.

• We need high precision when false positives are expensive (e.g., marking real users as fraudulent).

• Recall is important when false negatives are expensive (e.g., failure to diagnose cancer).

• F1-Score is applicable when you have an uneven class distribution especially between positive and negative.

8.2.2 ROC Curve and AUC Score

ROC Curve (Receiver Operating Characteristic)

• Graphs on the x-axis True Positive Rate (TP ÷ (TP + FN)) and on the Y-axis False Positive Rate (FP ÷ (FP + TN)) at multiple probability-threshold settings.

• Reveals the trade-off of sensitivity and false alarms.

• Better models will have the curve up and to the left (high TPR, low FPR).

AUC (Area Under the Curve)

• A summary single-level ROC performance metric.

• Range:

o 1.0 → Perfect classifier

o 0.9–1.0 → Excellent

o 0.8–0.9 → Good

o 0.7–0.8 → Fair

o 0.6–0.7 → Poor

o 0.5 →No improvement over random guessing

• Example: An AUC of 0.92 indicates an excellent discrimination by the model.

Business Relevance:

• In credit scoring, a high AUC indicates that the model is good at separating between low-risk borrowers and high risk ones.

• For patients, it promises the accurate separation of healthy and ill in medical practice.

Knowledge Check 1

Choose the correct option:

1. Which metric becomes misleading in highly imbalanced datasets?

a) Precision

b) Recall

c) Accuracy

d) F1-Score

2. What does Precision measure?

a) Correct negatives

b) Correct positives out of predicted positives

c) Missed positives

d) Overall correctness

3. AUC = 0.5 indicates:

a) Perfect model

b) Good model

c) Random guessing

d) Strong classifier

4. Sensitivity is also called:

a) True Negative Rate

b) False Positive Rate

c) Specificity

d) True Positive Rate

8.3 Business Implications of Model Evaluation

8.3.1 Aligning Metrics with Business Goals

Why alignment matters:

Incorrect choice of evaluation metric sometimes result in decision which is statistically right but doesn’t match the business goal.

Examples:

• Health Service: If you are predicting that a patient has cancer, then missing one true case (false negative) could have severe consequences. Sensitivity (recall) is preferred here, even with an increased false positive ratio.

• E-commerce: In product recommendation, precision is more crucial as recommending irrelevant items leads to decrease in user satisfaction.

• Banking: In fraud detection, precision is important (don't flag too many genuine transactions as fraudulent), but so is recall (catch as much fraud that happens as possible).

Business insight: always specify the cost of being wrong (false positives vs. false negatives) before selecting your evaluation metric.

8.3.2 Selecting the Appropriate Metric for the Task

No single measure serves all problems well, and the right one will vary according to context and consequences.

Accuracy

o Suitable when classes are balanced.

o Example: Predicting, whether the patient will survive after 5 years or not given age, time of operation and the number of positive axillary nodes.

Precision

o Significant for "the cost of false positives is high".

o Example: Marking real customers as "probably fraud" would have negative trust implications. Rigorous accuracy means that those flagged cases are truly risky.

Recall (Sensitivity)

o When false negative is expensive.

o Example: In public health, it is better to false alarm than miss a patient.

F1-Score

o Optimal when precision and recall are equally important.

o Example: In a spam detection application, we would like to achieve a high precision (i.e., you do not want to put genuine emails into the category of spam) and at the same time we want high recall (you catch most spams).

AUC (Area Under ROC Curve)

o Will be used for ranking problems or when comparing several models.

o Example: In credit scoring, AUC indicates better model separates high-risk vs. low-risk borrowers.

Key Business Lesson: The "best metric" depends on business trade-offs— not all mistakes cost us the same.

8.3.3 Balancing Accuracy and Interpretability

There are differing in models of the extent to which they can report why they make predictions:

High Precision, Low Interpretability (Black Box Models):

o Models such as ensemble methods (e.g. random forests, gradient boosting) can achieve similar or higher accuracy.

Problem: Difficult to explain responses (such as why a loan application has been rejected).

Moderate Accuracy, High Interpretability (Logistic Regression):

o It will be more difficult to interpret logistic regression. Each coefficient represents the effect of a predictor (e.g.,

"A 1-unit increase in monthly billings raises the probability of churn by 15%").

o Relationship managers love that this is disclosed.

Examples:

• Health & Finance: Laws/regulations require explainable models (e.g., credit denial letters). Interpretability takes priority.

• Retail & Marketing: Interpretability may matter less. Optimizing for accuracy before customer churn can pay off in terms of increased profit.

Key Insight: Enterprises must balance accuracy vs. interpretability given trust, regulation and decision impact.

8.3.4 Communicating Results to Non-Technical Stakeholders

Even the most well-designed model is useless if decision-makers can't understand or trust its output. Good communication can close the divide between data scientist and businessperson.

Best Practices in Communication:

Translate Metrics into Business Language

o Instead of: "Precision = 0.85"

o Say: "Of all the customers predicted to churn, 85 out of a hundred are actually at risk."

Use Visuals

o Heatmaps of confusion matrices, ROC curves, and histograms of probabilities make patterns understandable for non-technical users.

Focus on Business Impact

o Explain rather than only rely on memory:

“If we could increase the recall rate from 70 percent to 85 percent, an additional 1,500 customers a year would be saved,” he said.

Provide Actionable Insights

o Don’t give me a “The churn model got AUC = 0.90.”

o Cases – would say: “ Customers with only short-term contracts and a large number of complaints are at a 70% chance for churn.

— giving discounts might help retention.”

Examples of Stakeholder Communication:

• Execs: Need cost/revenue impact (“this model will save the bank $10 M annually by reducing loan defaults by 15%”).

• Marketing Teams: Need list of action (“Here are 5,000 customers we can target for retention campaigns”).

• Operations Staff: Want efficient rules (“If churn probability > 0.7, flag for intervention”).

Takeaway: Business results must be tied to financial or operational results.

“Activity: Linking Metrics to Business Goals”

“Students will be divided into groups, each representing a different industry (healthcare, banking, e-commerce, marketing). Using sample classification model outputs, they must decide which evaluation metric (accuracy, precision, recall, F1, AUC) best fits their business case and justify choices in a short presentation.”

8.4 Summary

- Logistic regression is employed in classification problems with a categorical dependent variable, where the target can be categorical (1/0, True/False Churn/Stay) and more generally repeated events. It predicts the likelihood of an event and transforms it to be in range [0, 1] by using a sigmoid function.

- Unlike linear regression however, instead of modelling continuous outcomes (e.g. the price), we model log-odds in logistic regression such that predicted probabilities lie within the valid range.
- There are various forms of logistic regression:
  - Binary: The set contains 2 categories (e.g., churn or no churn).
  - Multinomial: More than two, but unordered categories.
  - Ordinal: Categories that has an inherent order.
- Model fitting is outlined to consist of problem definition, data preparation (handling missing values, encoding categories), parameter estimation using MLE and summarizing model performance.
- Performance measurement is based on classification metrics such as:
  - Accuracy
  - Positive predictive value (PPV) = TP / (TP + FP)• Dependence of results on thresholds:ROC (receiver operating characteristic) curve:(FPR, TPR)i.e. True positive rate/Sensitivity for true class vs.Fall-out/1-specificity for another class.
  - Sensitivity (how many true positive is detected)
  - Harmonic mean of Precision and Recall (F1-Score)
- ROC curve and AUC (for classification power)
- Accuracy may be mislead in cases of imbalanced data; precision and recall give more insight when one c lass is much dominant (e.g., fraud recreation or rare disease detection).
- Specificity and sensitivity are typically trade-offs: increasing one tends to interfere with the other. Threshold should be set based on the business cost of mistakes.
    - A confusion matrix is a good way to see the errors of the models (FP, FN) and it serves as an effective tool to derive all the metrics.
- The right evaluation metric varies by business objective. For example:
  - In the medidal domain high recall is very important.
  - Precision may matter more in marketing.
  - In credit scoring, AUC can be used to rank risk.
- Logistic regression models are more interpretable than complex ones such as random forest models, an important factor in regulated industries (e.g., finance, healthcare).
- Needs to communicate results of model in business language: translating the technical metrics and interpreting them into something that stakeholders can understand is important for action.
- Logistic regression is used heavily in the real world for business case scenarios as well – such as in predicting customer churn, detecting fraud or outcomes (as in with medical diagnosis), credit scoring.

8.5 Key Terms

1. Logistic Regression Classifier - A form of predictive analysis that estimates the probability for categorical (binary) outcome based on one or more independent variables.
2. Odds Ratio – A statistic measuring that how a one-unit change of predictor impacts on the odds for occurrence of event.
3. Confusion Matrix – The table which is used to summarise the true positive and true negative across classes, in order to simplify it.
4. Accuracy – The number of correctly predicted AND negatives to the total observations.
5. Recall (Sensitivity) - What proportion of actual positives was identified correctly by the model.
6. F1-Score – The harmonic average of the precision and recall.
7. ROC Curve – Graph showing trade-off between true positive rate and false positive rate for different thresholds.
8. AUC (Area Under Curve) – A number between 0 and 1 that represents how good the model is at distinguishing between classes.

8.6 Descriptive Questions

1. Illustrate logistic regression and the difference of this model with linear regression.
2. Explain the sigmoid function in logistic regression.
3. What is an odds ratio and how do I interpret it in the context of logistic regression?
4. Explain how to develop the logistic regression model in step by step manner?
5. Discuss the significance of confusion matrix in evaluation of logistic regression models.
6. Distinguish between accuracy, precision, recall and F1- score giving an example.
7. What is ROC curve and AUC score? How are they employed in the evaluation of model performance?
8. Describe the trade-off between sensitivity and specificity and give an example.
9. Describe the business downside of choosing the incorrect performance metric as used in logistic regression evaluation?

8.7 References

1. Hosmer, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). Applied Logistic Regression (3rd ed.). Wiley.

2. Menard, S. (2002). Applied Logistic Regression Analysis (2nd ed.). SAGE Publications.

3. Kleinbaum, D. G., & Klein, M. (2010). Logistic Regression: A Self-Learning Text (3rd ed.). Springer.

4. Agresti, A. (2018). An Introduction to Categorical Data Analysis (3rd ed.). Wiley.

5. Pampel, F. C. (2000). Logistic Regression: A Primer. SAGE Publications.

6. James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An Introduction to Statistical Learning: With Applications in R. Springer.

7. Field, A. (2017). Discovering Statistics Using IBM SPSS Statistics (5th ed.). Sage Publications.

8. Freedman, D. A. (2009). Statistical Models: Theory and Practice (Rev. ed.). Cambridge University Press.

9. Wooldridge, J. M. (2016). Introductory Econometrics: A Modern Approach (6th ed.). Cengage Learning.

Answers to Knowledge Check

Knowledge check 1

1. c) Accuracy

2. b) Correct positives out of predicted positives

3. c) Random guessing

4. d) True Positive Rate

8.8 Case Study

Logistic Regression to Predicting Loan Default Risk

Introduction

Lenders incur high risk when approving loans because it financially costs them delinquencies. Properly identifying who will default is critical for reducing risk. A popular classification model that banks used to determine the probability of default is logistic regression which is based on borrower level attributes such as loan amount, income, credit history and employment status.

This case study investigates the implementation of logistic regression in a financial environment, and discusses common pitfalls in model creation, testing, and business interpretation.

Background

A consumer bank was looking to further enhance their lending decisioning through predictive analytics. The bank gathered data on 10,000 applicants who had applied for loans in the past including variables such as income level; age; employment status; credit score and loan to income ratio. The dependent variable was binary: Default = Yes/No.

The analytics team chose to develop a statistical model using logistic regression to predict probability of default and categorize applicants into “Safe to Approve” vs. “High Risk.” The objective wasn’t just accuracy, but rather to make sure the model served business goals — reducing loan defaults for the good of the company while still approving worthwhile customers.

Issue 1: Imbalanced Data Treatment

There were 15% of applicants who defaulted in the dataset, and 85% did not. Such an imbalance might encourage the model to always predict “no default” which can be a problem as it is misleadingly high accuracy and poor on risk depletion.

Solution: apply resampling (oversample defaults, undersample non- defaults) And use precision, recall and F1-score rather than accuracy to test the model.

What is a better performance metric than accuracy for imbalanced datasets?

a) Precision and Recall

b) R^2

c) Mean Squared Error

d) Adjusted R^2

Answer: a) Precision and Recall

Problem 2: How to Select Thresholds

The model generated default probabilities. But making the cutoff (e.g., 0.5) was important. A higher threshold decreased false positives but increased false negatives, and a lower one captured more defaults but turned away safe borrowers.

Solution: The bank tried out thresholds with the ROC curve and AUC score and chose one based on business trade-offs — opting for the recall in order to minimize high-risk defaulters.

MCQ:

Which method do you use to choose a cutpoint for logistic regression?

a) Scatterplot

b) ROC Curve

c) Histogram

d) Boxplot

Answer: b) ROC Curve

Problem 3: Informing Management of the Results

The analytics team first showed metrics that looked like nonsense to the clinicians; coefficients, log-odds, AUC scores. This was less clear to senior managers.

Solution: The findings were re-expressed in terms of business:

• “Applicants with loan-to-income ratio above 40% are twice as likely to default.”

• “The bank can lower default losses by 18% for each year by placing threshold at 0.45.”

This translation was necessary for management to envision the immediate financial ramification.

How would you explain logistic regression to a non-technical person?

a) Present coefficients only

b) Use log-odds values

c) Translate results into business impact

d) Provide only accuracy score

Answer: c) Translate to Business Impact

Conclusion

It helped the bank to use logistic regression model and differentiate risk between high- risk and low-risk applicants while applying for loan. Handling the data imbalance, achieving an appropriate threshold value, and reporting the results in a business understandable language allowed the bank to reduce bad loans through profitable loanes offering. This is a question that actually involves both statistical modelling and business decision making in finance, and I hope demonstrates the role of logistic regression.

BUPBM Unit 9 V3.docx

Building useful Predictive Business models_BBA_3

Building useful Predictive Business models_BBA_3

ATLAS SkillTech University

Document Details

Submission ID

trn:oid:::3618:128400456

Submission Date

Feb 16, 2026, 4:52 PM GMT+5:30

Download Date

Feb 16, 2026, 4:55 PM GMT+5:30

File Name

BUPBM Unit 9 V3.docx

File Size

126.2 KB

20 Pages

4,159 Words

26,466 Characters

0% detected as AI

The percentage indicates the combined amount of likely AI-generated text as well as likely AI-generated text that was also likely AI-paraphrased.

Caution: Review required.

It is essential to understand the limitations of AI detection before making decisions about a student's work. We encourage you to learn more about Turnitin's AI detection capabilities before using the tool.

Detection Groups

1 AI-generated only 0%
Likely AI-generated text from a large-language model.

0 AI-generated text that was AI-paraphrased 0%
Likely AI-generated text that was likely revised using an AI-paraphrase tool or word spinner.

Frequently Asked Questions

How should I interpret Turnitin's AI writing percentage and false positives?
The percentage shown in the AI writing report is the amount of qualifying text within the submission that Turnitin's AI writing detection model determines was either likely AI-generated text from a large-language model or likely AI-generated text that was likely revised using an AI paraphrase tool or word spinner.

False positives (incorrectly flagging human-written text as AI-generated) are a possibility in AI models.

AI detection scores under 20%, which we do not surface in new reports, have a higher likelihood of false positives. To reduce the likelihood of misinterpretation, no score or highlights are attributed and are indicated with an asterisk in the report (*%).

The AI writing percentage should not be the sole basis to determine whether misconduct has occurred. The reviewer/instructor should use the percentage as a means to start a formative conversation with their student and/or use it to examine the submitted assignment in accordance with their school's policies.

What does 'qualifying text' mean?
Our model only processes qualifying text in the form of long-form writing. Long-form writing means individual sentences contained in paragraphs that make up a longer piece of written work, such as an essay, a dissertation, or an article, etc. Qualifying text that has been determined to be likely AI-generated will be highlighted in cyan in the submission, and likely AI-generated and then likely AI-paraphrased will be highlighted purple.

Non-qualifying text, such as bullet points, annotated bibliographies, etc., will not be processed and can create disparity between the submission highlights and the percentage shown.

Unit 9: Time Series Forecasting

Learning Objectives

1. Understand the concept and importance of time series analysis in business forecasting.
2. Identify the key components of a time series: trend, seasonality, and irregularity.
3. Apply smoothing techniques such as moving averages and exponential smoothing.
4. Differentiate between additive and multiplicative time series models.
5. Use decomposition methods to analyze time series data.
6. Apply ARIMA models for advanced forecasting.
7. Evaluate forecasting accuracy using error measures.
8. Interpret forecasting results to support decision-making.

Content

9.0 Introductory Caselet

"Time Series Forecasting"

A popular retail chain TrendMart has many stores in several cities and they, with an extended range of broaband of customer offering. The company has collected monthly sales data for the past five years. Management has observed that sales usually go up during festive seasons, dip a bit in months of the monsoon and show a consistent upward movement altogether because of expansion and consumer demand picking up.

The company has more recently rolled out an inventory management push. Previously, they had challenges with overstock during non-peak months and stockouts during peak times. To address this, the operations team chose to apply time series forecasting algorithms to predict new sales in the future.

They began with moving averages to smooth fluctuations and then added exponential smoothing to place more weight on recent sales patterns. They then tried ARIMA to predict both trend and seasonality. Leveraging these predictions, TrendMart improved purchasing plans, kept leaner stock levels and saved money.

The finance team as well, relied on these forecasts to prepare quarterly revenue projections and the marketing team scoped seasonal campaigns using anticipated demand peaks. The CEO said good forecasting was more than a purely statistical exercise, but rather a decision making tool in the competitive landscape of current business.

Critical Thinking Question

What factors would you consider when choosing which model (moving average, exponential smoothing, or ARIMA)to use for forecasting TrendMart's sales data?: If you were a member of TrendMart's forecasting team, how would you decide whether to use a moving-average, exponential-smoothing, or ARIMA model as the basis for your sales forecasts?

9.1 Introduction to Time Series

9.1.1 Definition and Characteristics of Time Series Data

Definition

A time-series is any time-ordered sequence of observations. It is thus standard that every observation may be affected by time-dependent terms.

Example:

- The monthly sales of a retail store.
- Price of a stock at the day's end.

• Monthly rainfall of a city.

Characteristics of Time Series Data

Time Dependency: Observations are collected over time and each data point depends on past. For example, the stock price today is not independent of yesterday's stock price.

Frequency of Data: The time series can be collected at various frequencies such as daily (temperature), monthly (sales), quarterly (GDP) and yearly (population growth).

Stationarity vs. Non-Stationarity:

o A time series that does not vary over time i.e. mean, variance and covariance is constant at all point of times. Indeed, some types of data might not be able to written as stationary exchange rates in a daily period.

o Non-stationary data show trends or changes in variability over time, e.g. you have an increasing trend for smartphone sales.

Autocorrelation: Time series data is usually correlated with its past values. For example, this month's sales are often very much like last month's.

Patterns and Fluctuations: Time series exhibit trends, cycles, or seasons that make them predictable.

Noise: Apart from the structured patterns, time series usually exhibit some random or irregular variations that are unrelated to systematic factors.

9.1.2 Components of Time Series

A time series analysis entails the study of a data points obtained at sequential time spans. Decompositional analysis would also help if one wants to analzyse the time series and make some interpretation behind it. These factors contribute to the ability to isolate significant patterns, trends and enhanced forecasting accuracy. In general, time series data can be decomposed into the following 4 components :

Trend (Tt)

Seasonality (St)

Cyclic (Ct)

Irregular/Random (It)

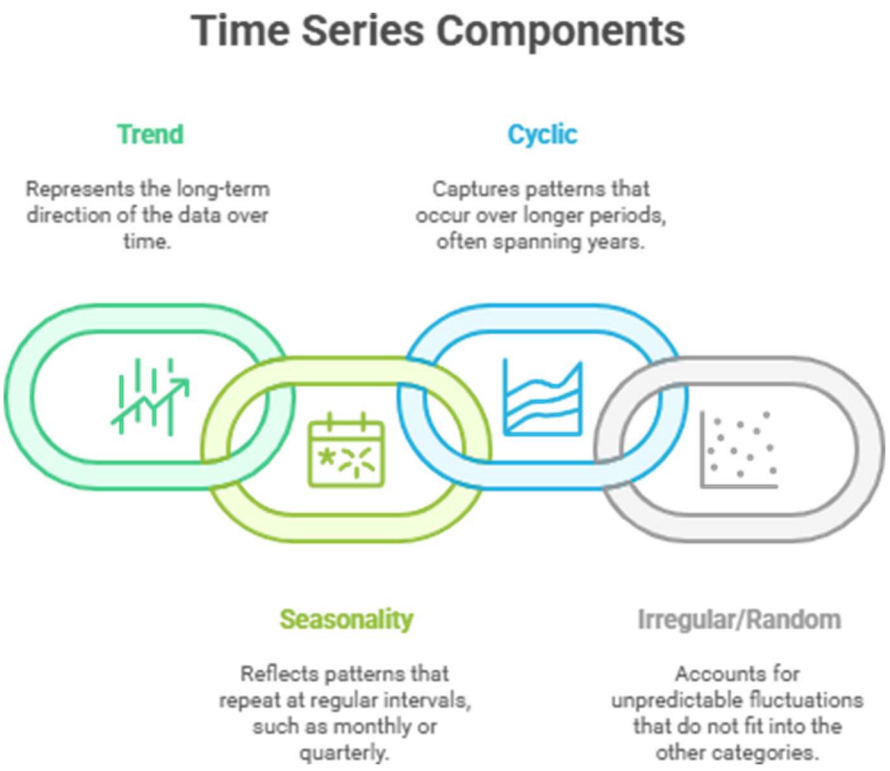

figure: Time Series

The series combine (add or multiply) for some data type.

Trend Component

The trend component is the tendency for a variable to increase, decrease or remain stable over time. It tells you if the data is growing, decreasing or staying constant over a long period. The shaped by trends are often attributed to some structural change, whether in economic circumstances, population growth or technology diffusion.

For example, real estate prices might be going up; more people are using mobile phones than before or fewer households have a landline.

Python Example — Simulating and Plotting the Trend:

import numpy as np import pandas as pd

import matplotlib. pyplot as plt

Simulated trend data np. random. seed(42)

time = pd. date_range(start="2010", periods=100, freq = 'M') trend = np. linspace(50, 150, 100) # Linear increasing trend noise = np. random. normal(0, 5, 100)

series = trend + noise

Plot plt. figure(figsize=(10, 4))

plt. plot(time, series, label='Observed Series') plt. plot(time, trend, label='Trend', linestyle='--') plt. title("Time Series with Trend Component") plt. legend()

plt.grid(True) plt.show()

Seasonality Component

Repetitive behavior that is related to a specific time of year, month or week is referred to as seasonality. These patterns are driven by things like weather, holidays and school schedules.

Examples might be retail sales higher in December or electricity use lower in the spring.

Python Example – Adding Seasonality:

Simulate seasonality

seasonality = 10 np. sin(2 np. pi * np. : meta with seasonality # + np.cos(2 * np.(np.arange(100), 0.5*np.pi)) series_with_season = trend + seasonality + noise

Plot plt. figure(figsize=(10, 4))

plt. plt.plot(time, series_with_season, label='Series with Seasonality') plt. title("Time Series with Seasonal Component")

plt. legend() plt. grid(True) plt. show()

Cyclic Component

The cyclical component is a longer term oscillations that do not repeat in a regular seasonal fashion. These oscillations have a duration of several years and are usually linked to business cycles or economic periods, such as growth, recession or recovery.

Cycles, on the other hand, 13are of irregular length and amplitude and are consequently more difficult to 14Monday (day k) in March of year i is denoted dki. Example: 5-10 year economic boom and busts.

Python Example – How to simulate Periodic Behaviour:

Simulate some cyclic pattern with a low-frequency sine wave cycle = 15 np. sin(2 np. pi * np. arange(100) / 50) series_with_cycle = trend + cycle + noise

Plot plt. figure(figsize=(10, 4))

plt. plot(time, series_with_cycle, label='Series with Cycle') plt. title("Cyclic Time Series") plt. legend()

plt.grid(True) plt.show()

Irregular (Random) Component

The short range component represents unpredictable and unstructured variations of the time series (random or irregular). Those might be the results of unexpected factors, e.g. strikes, pandemics and natural disasters or technical problems. This factor basically behaves like white noise which cannot be predicted.

These might be stock market downturns or geopolitical events that impact a supply chain.

Example in Python – Separating Random Noise:

Let's only add some random noise random_series = noise

Plot plt. figure(figsize=(10, 4))

plt. plot(time, random_series, label='Random Component')

plt. title("Irregular (Random) Component") plt. legend()

plt.grid(True) plt.show()

Models based on additive or multiplicative time series

Time series may be represented with two models:

• Additive Model:

And for yt, It can hold,true as long as the Fuelcell always in fluctuation state.

• Multiplicative Model:

Suitable for when variation in series grows as a multiple of level of the series.

It depends on the organization of the data. The addi-tive model is applicable when seasonal and irregular variations donot change over time; themultiplicative model would be more appropriate if fluctuations increasedwith the level of the trend.

Python Example – Decomposition with statsmodels:

from statsmodels. tsa. seasonal import seasonal_decompose

Convert to DataFrame

series_df = pd. Series(series_with_season, index=time)

Decompose (Additive)

additive_result = seasonal_decompose(series_df, model='additive', period=12) additive_result. plot() plt. suptitle("Additive Decomposition", fontsize=14)

plt.show()

Decompose (Multiplicative)

("multiplicative_result = seasonal_decompose(series_df+100, model='multiplicative', period=12)\r " " multiplicative_result. plot()

plt. suptitle("Multiplicative Decomposition", fontsize=14) plt. show()

9.1.3 Common Forecasting Methods

Prediction is concerned with the future values of a time series based on the information in past observations. Several methods are commonly used:

Moving Average Method

• A technique of smoothing by averaging a fixed number of most recent observations to forecast future values.

• Assists in smoothing out short-term movements and highlighting longer-term trends.

• Example: To develop a 3-month moving average forecast for sales, one uses the sales of the previous three months.

• Drawbacks: Fails to capture seasonality or trends well, and forecasts could trail actual data.

Exponential Smoothing Method

• Weights past observations by exponentially decreasing amounts. Newer data is weighted more highly.

• Simple Exponential Smoothing (SES): optimal for data without trend and/or seasonality.

• Holt's Linear Method: Generalizes SES to capture trends.

• Holt-Winters Method: Generalizes it to introduce the capability of handling trend as well as seasonality.

• Example: A business forecasting peak monthly electricity demand would use the Holt-Winters method because demand is based on long-term growth (trend) as well as seasonal factors such as summer peaks.

The ARIMA (Auto-Regressive Integrated Moving Average) Model

• A strong statistical model applied when autocorrelation in the data is observed.

• Components:

o Auto-Regressive (AR): Utilizes the previous values in order to forecast the current value.

o Differenced (I): Further differencing is applied to detrended the series and make it stationary.

o Moving Average (MA): Makes forecasts based on previous forecast errors.

• ARIMA(p, d, q) where p = AR order, d = differencing degree and q = MA order.

• Example: Utilized in predicting the stock market indices, exchange rates and other macroeconomic indicators.

Strength: Extremely flexible in modeling a broad class of time series patterns.

• Shortcoming: It relies on statistical knowledge and fine-tuning parameters.

Teaching Note:

• Moving Average is most appropriate for short-term smoothing.

• Comparing with weighted moving average, Exponential Smoothing is sensitive to change.

• ARIMA works best for non-seasonal time series that are not too long, with autocorrelation.

“Activity: Spot the Pattern in Time Series”

Students are given monthly sales data of a retail store for three years. They must identify the underlying trend, seasonal peaks, and irregular fluctuations. In small groups, they classify components into trend, seasonality, cyclic, or irregular. Each group presents findings, justifying their reasoning with evidence from the dataset.

9.2 Applications of Time Series in Business Industry It plays a crucial role in several business industries such marketing, sales, business Research and decision making form the Management with respect to time series forecasting applications (Arora et al., 2013).

9.2.1 Sales and Revenue Forecasting

Projections on sales and revenue enable companies to predict customer demand and future cash flows, playing a critical role in business survival.

Key Applications:

Demand Planning: Companies look at what has historically sold to anticipate future demand Thereby preventing stockouts or overages.

Production planning: Manufacturers level production with environmental sales, thereby avoiding underutilized or over-revenue.

Planning of Financials: Goes to make positive income forecasts, for budgeting purposes, as well as how the asset can be most effectively utilized.

Strategic Marketing–Companies target periodic highs (e.g., holidays, fests) to get marketing rewards.

Salesforce Management : Forecasts help in planning for resources such as manpower, distributors and logistics.

figure: Key Applications:

Practical Example:

You have been given the monthly sales data for an FMCG (Fast-moving consumer goods) company and you are tasked with analyzing them. For drinks seasonal spikes towards the summer are highly pronounced in the time series. The company forecasts sales next quarter using Holt-Winters exponential smoothing and ensures that warehouses are filled with an artisan's attention to detail in supplies before the peak customer demand.

Teaching Note:

Students can work through this presentation by graphing some real or simulated sales data and breaking it down into trend, seasonality, and irregular components to see how these methods forecast.

9.2.2 Stock Market and Financial Forecasting

When compared with forecasting other time-dependendent data, such as economic or agricultural quantities, financial forecasting has been popular throughout the recent years.

Key Applications:

Stock Price Prediction: Time series techniques like ARIMA and ARCH/GARCH account for the volatility and the autocorrelation present in stock returns.

Risk Management: Predicting volatility also allows traders to hedge portfolios against risk.

Currency Exchange Forecasting: Firms forecast exchange rate fluctuations in order to manage risks associated with international trade.

Interest Rate Forecasting Banks and investors use forecasts to adjust lending policies and bond investments.

Macroeconomic Forecasting: Governments predict GDP growth, inflation and employment figures to inform policy making.

Practical Example:

A financial investment analyst applies ARIMA(1,1,1) to predict the daily close price of a certain stock. Although specific forecasts are challenging because of the presence of random shocks, the model replicates trend and volatility behaviours. This provides a guide for portfolio managers who may wish to overweight or underweight the stock.

Teaching Note:

As data on financial variables typically exhibits irregular shocks, learners could be exposed to the Favor of simple models and how crucial model validation are when based accomplishment of statistical tests.

Did You Know?

“Stock market forecasting is one of the most challenging applications of time series because prices are influenced by countless economic, political, and psychological factors. Models like ARIMA and GARCH help capture trends and volatility, but even small unexpected events can cause dramatic shifts in financial markets.”

9.2.3 Energy Consumption Forecasting

Accurate energy forecasting is crucial for operational planning and sustainable development, since energy consumption affects the cost of production, security of supply and environmental policy.

Key Applications:

Load Forecasting (Short-term): Utilities predict hourly or daily electricity consumption for balancing supply to prevent blackouts.

Long-Term Capacity Planning: Governments and companies create forecasts to plan for new plants, the integration of renewable energies, or expansion of the grid.

Cost Reduction: The more accurate the weather forecast, the less waste and better buy of fuel sources such as coal, gas and oil.

Renewable Energy Integration: Forecasting mitigates variable resources such as solar and wind with common demand behaviors.

Policy and Sustainability: National energy forecasts inform policy, such as efficiency targets and carbon reduction.

Practical Example:

A state electricity board is analyzing daily consumption over five years, as well as temperature readings. It forecasts summer peak load with the use of Holt-Winters exponential smoothing. That enables the utility to buy additional power from independent producers in advance, which prevents shortages when demand is high.

Teaching Note:

It is also possible to show the students some real public datasets for electricity consumption (see that many are available online) where they can continue exercises fitting models and valuing forecast errors.

Knowledge Check 1

Choose the correct option:

1. Which application of time series helps businesses avoid stockouts?

a) Energy forecasting

b) Sales forecasting

c) Stock price analysis

d) Risk management

2. ARIMA models are widely used in which area of business forecasting?

a) Energy load forecasting

b) Stock market forecasting

c) Sales promotions

d) Inventory tracking

3. Load forecasting in the energy sector is mainly used to:

a) Plan marketing campaigns

b) Reduce staff turnover

c) Balance electricity supply

d) Increase GDP

4. Which component is critical in financial forecasting for managing portfolio risks?

a) Trend analysis

b) Volatility forecasting

c) Seasonal variation

d) Random shocks

9.3 Practical Work with Time Series Data

9.3.1 Importing and Exploring Time Series Datasets

Importance:

Datasets need to be well prepared before the prediction process. Importing makes data usable in a format, and exploration keeps the quality of data.

Key Steps in Practice:

Import data: Input datasets from CSV, Excel or databases.

Datetime Conversion: Transform date columns into datetime objects so the series are properly indexed.

Frequency Check: Do I have data every day, week, month or year?

Missing Data: Discover and manage missing or corrupted data (interpolation, forward fill, deletion).

Summary Statistics: Calculate mean, variance, max, min to get a sense of the behavior overall.

Business Example:

A retail firm imports monthly sales information from their ERP into CSV files. For forecasting, analysts used to validate input data checking if all months are present (clients not reporting or technical issues) adjusting odd items (i.e., no sales reported as there was a mistake on uploading the file), validating that dates are ordering correctly.

Python Illustration:

import pandas as pd

Import dataset

data = pd. read_csv("sales_data. csv", parse_dates=['Date'], index_col='Date')

Inspect first few rows print(data. head())

Basic summary print(data. describe())

print(data. isnull(). sum()) # check missing values

9.3.2 Visualizing Time Series Data

Importance:

Visualization is the essence of time-series analysis. It serves to reveal underlying patterns that nothing but numbers can make clear.

Key Visualization Techniques:

Line Charts: Display general trends and variations.

Seasonal Plots Compare trends per year or month.

Histogram/Boxplots: Reveal distribution and outliers.

Decomposition Plots: Split out the time series into trend, seasonality and residuals for better understanding.

Business Example:

A drink manufacturer forecasts 5 months of monthly sales. The time series shows a continuous increasing trend of the market, seasonal variation during summer and the inconsistence related to unespected strikes.

Python Illustration:

import matplotlib. pyplot as plt

from statsmodels. tsa. seasonal import seasonal_decompose

Line chart plt. plot(data['Sales'])

plt. title("Monthly Sales Over Time") plt. xlabel("Date")

plt. ylabel("Sales") plt. show()

Decomposition

decomposition = seasonal_decompose(data['Sales'], model='additive', period=12) decomposition. plot()

plt.show()

9.3.3 Building Forecasting Models in Python/Google Colab

Importance:

After cleaning the data and discovering trends, forecasting models are employed to forecast future values.

Common Models for Classroom Use:

MA: Effectively filters short term changes; appropriate for stable series.

Exponential Smoothing with a Sliding Window: Prior inferences decay over time; faster to adapt to changes.

Holt-Winters Method: Smoothing factorization that deals with trend and seasonality as well.

ARIMA Models: Good for complex dataset having autocorrelation and non-stationarity.

Business Example:

An energy company forecasts daily demand for electricity using Holt-Winters, whereas a bank predicts interest rates with ARIMA.

Python Illustration (Holt-Winters):

from statsmodels. tsa. holtwinters import ExponentialSmoothing

model = ExponentialSmoothing(data['Sales'], trend="add", seasonal="add",seasonal_periods=12) model_fit = model. fit()

forecast = model_fit. forecast(12) # 12 months ahead

plt. plot(data['Sales'], label="Actual") plt. plot(forecast, label="Forecast") plt. legend()

plt.show()

Python Illustration (ARIMA):

from statsmodels. tsa. arima. model import ARIMA

model = ARIMA(data['Sales'], order=(1,1,1)) model_fit = model. fit()

forecast = model_fit. forecast(steps=12)

print(forecast)

9.3.4 Evaluating Forecast Accuracy (MAPE, RMSE)

Importance:

The model evaluations are necessary to check the certainty of models. The forecast errors are a measure of how well the model predicts the actual values.

Common Metrics:

MAPE (Mean Absolute Percentage Error):

o Percentage of error as measured against actual reading (actual values).

o Intelligible and comparable over series.

o Ideally when actual values are not near zero.

RMSE (Root Mean Squared Error):

o Penalizes larger errors more heavily.

o Great to find out when the predictions go horribly wrong.

Business Example:

• Let's take a look at two models being considered by a retailer to forecast sales during the holiday season.

• Model A has better relative accuracy (lower MAP E) but Model B has fewer large deviations (lower RMSE).

• Choice among decision-makers is based on business objectives: high-quality accuracy vs. big forecast misses averse.

Python Illustration:

from sklearn. metrics import mean_absolute_percentage_error, mean_squared_errorimport numpy as np

Example True vs Predicted Vlaues y_true = [100, 120, 130, 150]

y_pred = [110, 125, 128, 145]

mape = mean_absolute_percentage_error(y_true, y_pred) * 100 rmse = np. sqrt(mean_squared_error(y_true, y_pred))

```
print("MAPE:", mape)

print("RMSE:", rmse)
```

9.4 Summary

- Time series is an ordered set of values observed along the time.
- Characteristics Time dependence, frequency of collection, stationarity and autocorrelation are the characteristics.
- Time series can be decomposed in four components viz. trend, seasonal variation, cyclical variation and residual fluctuation.
- Trend describes long term movement, while seasonality explains short term repeating patterns.
- And the cyclic components are long-term oscillating, and the irregular components are unpredictable.
- Organizations use time series predication for sales, revenue, finance and energy planning.
- Sales forecast assists in demand planning, control of stock level and budgeting.
- Stock market prices, exchange rates and risk management are NBA naïve trend prediction applied in future forecasting of financial data such as loans.
- Energy prediction permits the efficient load allocation, capacity planning and renewable energy integration.
- Applying the theory, this hands-on forecasting requires data importing, cleaning and exploration.
- ii) Visualization methods such as line charting and decomposition identify trends.
- Models that have been explored in the literature are moving averages, exponential smoothing, Holt-Winters and ARIMA.
- Accuracy of prediction is tested based on the error estimators such as MAPE and RMSE to select reliable model.

9.5 Key Terms

1. Time Series: a sequence of data points measured, recorded or collected at equally spaced time intervals.
2. Trend: The gradual increase or decrease of a time series over time.
3. Seasonality: Steady, predictable fluctuations that repeat over a consistent interval of time, such as one year.
4. Cyclic Variation: Long term fluctuations in the dynamic data series that there is no fixed period of time and is frequently related to business cycles.

5. Noise: Random fluctuations in a signal that is introduced by the various stages of processing and transmission.
6. Stationarity: Property of a time series for which the mean, variance, and autocovariance are constant over time.
7. Autocorrelation: The correlation between an observed time series data and a past value of itself.
8. Moving Average: A smoothing process involving averaging some number of most recent observations.
9. Exponential Smoothing: A method of forecasting that gives greater importance to recent observations.
10. Holt-Winters Exponential Smoothing: An extension of the exponential smoothing method that also takes trend and seasonality into account.
11. ARIMA Model : A predictive model with features of Autoregression, Differencing and Moving Average.
12. 3) MAPE (Mean Absolute Percentage Error): the average magnitude of percentage breakdown in unit root condition.
13. RMSE (Root Mean Squared Error): A method of calculating the error by penalizing large errors more.

9.6 Descriptive Questions

1. Define time series. Discuss its salient features with the help of one or two illustrations.
2. Explain the four elements of a time series with business examples.
3. Distinguish between additive and multiplicative models of time series decomposition.
4. Discuss the role of sales and revenue forecasting in business decisions.
5. Explain how time series forecasting is used in the context of stock market and financial analysis.
6. What is energy consumption forecasting? Emphasises its importance in planning and sustainability.
7. Discuss the effect of visualization in time series. Illustrate with suitable charts.
8. Talk to me about how you built forecasting models in Python/Google Colab.
9. Discuss MAPE and RMSE as the measures for forecasting accuracy with some examples.
10. Compare and contrast moving average, exponential smoothing, and ARIMA models based on when it is appropriate to use each one? What are the limitations of each model?

9.7 References

1. Box, G. E. P., Jenkins, G. M., & Reinsel, G. C. (2015). Time Series Analysis: Forecasting and Control. Wiley.

2. Brockwell, P. J., & Davis, R. A. (2016). Introduction to Time Series and Forecasting. Springer.

3. Chatfield, C. (2003). The Analysis of Time Series: An Introduction. Chapman & Hall/CRC.

4. Hyndman, R. J., & Athanasopoulos, G. (2018). Forecasting: Principles and Practice. OTexts.

5. Wei, W. W. S. (2006). Time Series Analysis: Univariate and Multivariate Methods. Pearson.

6. Shumway, R. H., & Stoffer, D. S. (2017). Time Series Analysis and Its Applications: With R Examples. Springer.

7. Makridakis, S., Wheelwright, S. C., & Hyndman, R. J. (1998). Forecasting: Methods and Applications. Wiley.

8. Hamilton, J. D. (1994). Time Series Analysis. Princeton University Press.

9. Montgomery, D. C., Jennings, C. L., & Kulahci, M. (2015). Introduction to Time Series Analysis and Forecasting. Wiley.

Answers to Knowledge Check

Knowledge check 1:

1. b) Sales forecasting
2. b) Stock market forecasting
3. c) Balance electricity supply
4. b) Volatility forecasting

9.8 Case Study

Time Series Prediction Analysis in Retail Industry for Sales Force Optimization

Introduction

Forecasting is important in the field of retail business because it helps decide inventory, minimize cost as well as satisfying customer needs. Time series forecasting methods, which includes moving averages, exponential smoothing and Arima modeling – assist managers in identifying patterns in historical sales data and predicting future trends. When they don't have accurate forecasts, businesses subject themselves to overstocking, stockouts and financial loss.

Background

ShopSmart, a medium-sized store chain, was struggling to stabilize its sales performance. While the increase in demand over time was consistent, stockouts were common during the festive season and excess inventory occurred during the off-peak months resulting in lost revenue. Management found that relying on intuition-based forecasting did not work, and it adopted scientific time series methods.

The company collected five years of monthly sales data and observed:

- A naked, upward graph due to expansion of stores.
- Cyclical peaks at festive and holiday times.
- Short-term, erratic changes influenced by promotions or local events.

To meet these challenges, the ShopSmart analytics team employed various forecasting models and compared predictions.

Question (1): Efficient Inventory Management

ShopSmart would typically be overstocked in slow months, but short on stock when demand spiked for holidays.

Solution: The client leveraged Moving Average and Exponential Smoothing techniques to mitigate demand volatility and maintain correct procurement schedules. That lowered excess holding costs and increased product in stock.

Problem Statement 2 (Capturing Seasonal Demand Patterns)

Festive seasons always experienced high traffic but there was no efficient way for the management to predict these spikes.

Solution:S hopSmart used the Holt-Winters Exponential Smoothing method, which is based on trend and seasonality. This allowed the company to stock up, in advance of festivals, and increase its sales and customer satisfaction.

Problem 3: Long-Term Forecasting Accuracy Not as Good as it Should Be

For long-term planning, the company required accurate revenue projections for investors and financial planning.

Solution: The team utilized ARIMA (Auto-Regressive Integrated Moving Average) models to account for autocorrelation and forecast long-term sales. Forecasts were utilized to create annual budgets and growth plans.

Outcome

Through the combination of time series forecasting with decision-making:

• Warehousing costs reduced as overstocking decreased 18%.

• Stockouts during period of high demand fell 25 percent.

• Prediction accuracy was greatly enhanced and resulted in improved marketing and financial planning.

• Consumer satisfaction grew as the company was able to keep up with seasonal demand.

Critical Thinking Questions

If you were the manager of ShopSmart, which model would you choose for predicting shortrun and long-run sales?

What are some external factors (inflation, competition, pandemics, etc) that could make time series forecasting models unreliable?

How can companies go beyond conventional models to enhance forecast accuracy?