

CMR INSTITUTE OF TECHNOLOGY

Affiliated to VTU, Approved by AICTE, and Accredited by NBA, by NAAC with A++

VTU- SoE

VTU 3rd Semester MBA Degree Examination Dec'24/ Jan '25

Exploratory Data Analysis for Business

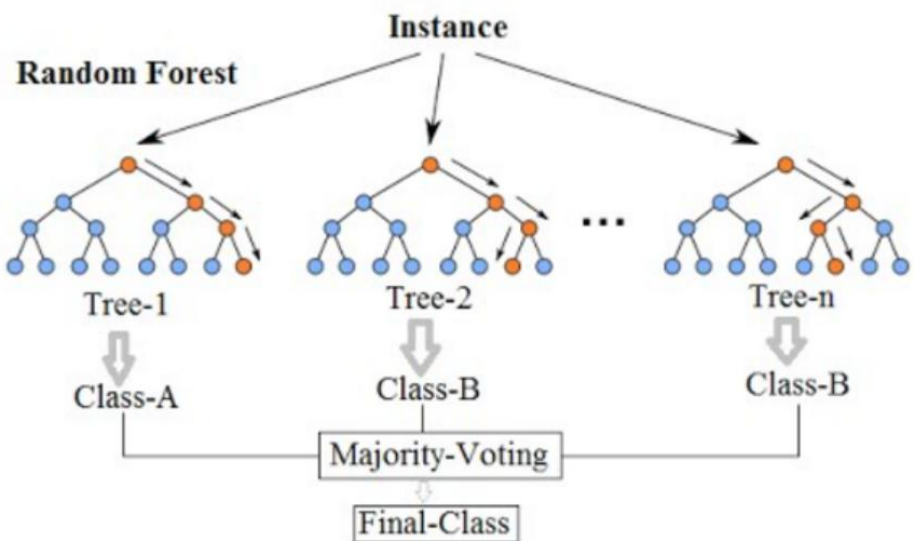
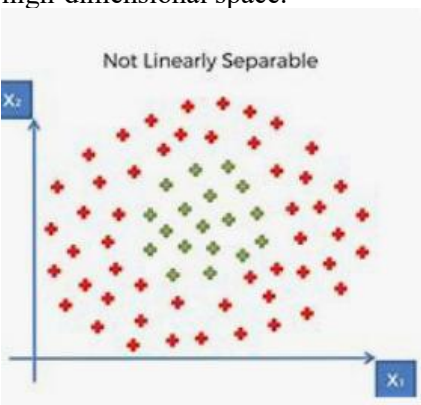
22MBABA304

Questions	Solution
1. a.	<p>What is Exploratory Analysis? EDA is the process of analysing data sets to summarize their main characteristics, often using visual methods. It Detects patterns, trends, outliers, and relationships. Helps in selecting appropriate models and preprocessing methods. Ensures better understanding of the data before applying data mining techniques.</p>
b	<p>Illustrate the applications of datamining. Data Mining is the process of discovering patterns, correlations, and useful information from large sets of data using statistical and computational methods. Uses:</p> <ul style="list-style-type: none"> • Fraud detection • Market basket analysis • Customer segmentation • Predictive maintenance
c	<p>Explain the classification problem in real life Spam Detection: Classifying emails as spam or not. Emails are analyzed based on their content, sender, subject, and metadata. A classification model learns from labelled examples (spam vs. not spam). It identifies patterns like certain keywords or suspicious links. The goal is to automatically filter out unwanted emails. This helps users manage their inbox more efficiently. Medical Diagnosis: Identifying diseases based on symptoms and test results. Doctors use symptoms, medical history, and test results as input. A model is trained to classify whether a patient has a specific disease. It helps identify conditions like diabetes, cancer, or infections. Accurate classification can assist in early diagnosis and treatment. This improves patient outcomes and supports clinical decision-making.</p>
2a	<p>Define prediction error Prediction error refers to the difference between the actual (true) value and the value predicted by a model. It measures how well or poorly a model's predictions match real outcomes. In simple terms: Prediction Error = Actual Value – Predicted Value There are two common types:</p> <ul style="list-style-type: none"> • For classification problems: Error is usually counted as incorrect predictions (e.g., misclassifying spam as not spam). • For regression problems: Error is often measured using metrics like Mean Absolute Error (MAE), Mean Squared Error (MSE), or Root Mean Squared Error (RMSE). <p>Smaller prediction errors indicate better model performance.</p>
b	<p>Examine the concept of bias-variance trade-off 1. Bias:</p> <ul style="list-style-type: none"> • Bias refers to errors due to overly simplistic assumptions in the model.

	<ul style="list-style-type: none"> • High bias means the model underfits the data — it misses important patterns. • Example: Using a linear model for data with a nonlinear relationship. <p>2. Variance:</p> <ul style="list-style-type: none"> • Variance refers to errors due to model sensitivity to small fluctuations in the training data. • High variance means the model overfits — it captures noise as if it were signal. • Example: A complex decision tree that fits training data too closely but performs poorly on new data. <p>Trade-Off:</p> <ul style="list-style-type: none"> • Reducing bias often increases variance, and reducing variance can increase bias. • The goal is to find a balance that minimizes total prediction error on unseen data. • This is achieved through methods like cross-validation, regularization, or model selection.
c	<p>Explain different methods of cross validation.</p> <p>Cross-validation is a model evaluation method where the dataset is split into several parts, and the model is trained and tested multiple times. Cross-validation Advantage: Reduces variability. Provides a more accurate estimate of model performance.</p> <p>Cross-Validation: A technique where a dataset is repeatedly split into training and validation sets to assess model performance. It Prevents Overfitting by: Ensuring model is not evaluated on training data, identifying models that perform well across multiple data subsets, helps in hyperparameter tuning with reliable metrics.</p> <ul style="list-style-type: none"> • K-fold cross-validation and its advantages: <p>Example (K=5): Data is split into 5 parts. • Model is trained on 4 parts and tested on 1, repeated 5 times (rotating test fold each time). • The average accuracy across all folds is reported.</p> <p>Advantages: • Better generalization estimate. • Efficient use of data. • Reduces overfitting risk.</p> <p>Leave-One-Out Cross-Validation (LOOCV): • For dataset with n observations: Train on n–1 data points. Test on the 1 left-out point. Repeat for all n points. Average all test errors for final estimate.</p> <ul style="list-style-type: none"> • Limitations of Holdout Sample: May give unreliable results if data is not representative.
3 a	<p>What is Principal Component Analysis?</p> <p>Principal Components: Linear combinations of original variables that capture maximum variance in the data. • Interpretation: Each principal component represents an orthogonal axis in the feature space along which data variability is greatest. The first principal component captures the most variance, followed by the second, and so on.</p>
b	<p>Explain linear regression model.</p> <p>A linear regression model is a fundamental statistical and machine learning technique used to predict a continuous outcome based on one or more input variables.</p> <p>Key Concepts:</p> <ol style="list-style-type: none"> 1. Goal: To find the best-fitting straight line (in 2D) or hyperplane (in higher dimensions) that models the relationship between the input(s) and the output. 2. Equation (Simple Linear Regression):

	$y = \beta_0 + \beta_1 x + \varepsilon$ <ul style="list-style-type: none"> • y: predicted value (dependent variable) • x: input (independent variable) • β_0: intercept (value of y when $x = 0$) • β_1: slope (change in y for each unit change in x) • ε: error term (difference between predicted and actual values) <p>The model learns the best values for β_0 and β_1 by minimizing the sum of squared errors between actual and predicted values.</p> <p>3. Assumptions:</p> <ul style="list-style-type: none"> ○ Linearity between variables ○ Independence of errors ○ Constant variance of errors (homoscedasticity) ○ Normally distributed errors <p>4. Use Cases:</p> <ul style="list-style-type: none"> ○ Predicting house prices from size and location ○ Forecasting sales based on marketing spend ○ Estimating student scores based on study hours
c.	<p>Explain the methods for variable selection in linear regression.</p> <p>Variable Selection Methods:</p> <ul style="list-style-type: none"> • Forward Selection: Start with no variables; add predictors one by one based on improvement in model fit (e.g., lowest AIC). • Backward Elimination: Start with all predictors; remove the least significant one iteratively. • Stepwise Selection: Combination of forward and backward methods. • Lasso Regression: Regularization method that shrinks coefficients and sets some to zero. • Ridge Regression: Shrinks coefficients but doesn't eliminate variables; helps in multicollinearity. • Elastic Net: Combines Lasso and Ridge for flexibility.
4 a	<p>How to handle missing values?</p> <p>Remove Missing Data</p> <p>Impute (Fill In) Missing Values</p> <ul style="list-style-type: none"> • Mean/Median/Mode Imputation: • Forward/Backward Fill: • K-Nearest Neighbors (KNN) Imputation: • Regression Imputation: <p>3. Use Algorithms that Handle Missing Data</p> <p>4. Add Missing Value Indicators</p>
b	<p>Demonstrate geometric interpretation and properties of ridge regression.</p> <p>Traditional Linear Regression (OLS): Minimizes the sum of squared residuals:</p> $\text{Loss} = \sum (y_i - \hat{y}_i)^2$ <p>No penalty on coefficients, so coefficients can become large, especially with multicollinearity.</p> <ul style="list-style-type: none"> • Ridge Regression: Minimizes squared residuals plus a penalty term on the size of coefficients: $\text{Loss} = \sum (y_i - \hat{y}_i)^2 + \lambda \sum \beta_j^2$

	The penalty term shrinks coefficients to reduce overfitting and variance, improving prediction on new data.
c	<p>Explain in brief formation of right sized tree via pruning</p> <p>When the tree overfits training data, capturing noise rather than patterns.</p> <ul style="list-style-type: none"> • When the tree is too large or deep, making interpretation difficult. • To improve generalization by reducing complexity. • In datasets with small sample size, pruning helps avoid fitting noise. • When computational resources or model simplicity is important. <p>Pruning considerations:</p> <ul style="list-style-type: none"> • Complexity vs. accuracy trade-off. • Use cross-validation or a validation set to decide pruning extent. • Pruning improves model generalization and interpretability.
5 a	<p>What is a kernel function?</p> <p>Kernel functions implicitly map data into a higher-dimensional space where it may become linearly separable.</p> <ul style="list-style-type: none"> • Common kernels include: Linear kernel: For linearly separable data. Polynomial kernel: Captures polynomial relations. Radial Basis Function (RBF) kernel: Handles complex nonlinear patterns.
b	<p>Explain Bagging with algorithm</p> <p>Bagging (Bootstrap Aggregating): Creates multiple versions of a model by training on different bootstrap samples (random samples with replacement) of the data. 25 • The predictions from these models are aggregated (averaged for regression, majority vote for classification). • Role: Reduces variance and overfitting common in single decision trees, leading to more robust and accurate predictions.</p> <div style="text-align: center;"> <h2>BAGGING Algorithm</h2> <h3><i>Bootstrap Aggrigating</i></h3> <hr/> <pre> graph LR Data[Data] --> B1[B1] Data --> B2[B2] Data --> B3[B3] B1 --> M1((M1)) B2 --> M2((M2)) B3 --> M3((M3)) M1 --> Agg[Agg. / Vote] M2 --> Agg M3 --> Agg Agg --> Output[Output] </pre> <p>Training Data Bootstrap samples Model Aggregation / Voting Outcome</p> </div>
c	<p>Explain the steps of process of decision tree</p> <p>Start with all data in the root node.</p> <ul style="list-style-type: none"> • At each step, consider splitting the node based on predictor variables. • For each candidate split, calculate the impurity (e.g., Gini index, entropy for classification, or mean squared error for regression) in the resulting child nodes. • Choose the split that maximally reduces impurity (best separation/homogeneity). • Repeat recursively for child nodes until a stopping criterion (like minimum node size or max depth) is met.
6 a	What is random forest?

	<p>Random Forest is a powerful and popular ensemble machine learning algorithm used for both classification and regression tasks. It builds multiple decision trees and combines their results to improve accuracy and control overfitting.</p> <p style="text-align: center;">Random Forest Simplified</p>  <p>The diagram illustrates the Random Forest process. An 'Instance' is fed into a 'Random Forest' which consists of multiple decision trees (Tree-1, Tree-2, ..., Tree-n). Each tree outputs a class prediction (Class-A or Class-B). These predictions are combined via 'Majority-Voting' to determine the 'Final-Class'.</p>
b	<p>Examine the concept of linear discriminant analysis</p> <p>LDA assumes classes have identical covariance matrices but different means.</p> <ul style="list-style-type: none"> • Finds a linear combination of features maximizing class separability. • Classifies by assigning points to the class with closest mean in transformed space. <p>Steps:</p> <ol style="list-style-type: none"> 1. Estimate class means and covariance matrices. 2. Compute linear discriminant functions. 3. Assign new observations to classes based on these functions.
c	<p>Identify the measures of similarity and dissimilarity in detail.</p> <ul style="list-style-type: none"> • Similarity: Quantifies how alike two data points are. Higher value = more similar. Examples: Cosine similarity (for text), Jaccard index (for binary attributes) • Dissimilarity (Distance): Measures how different data points are. Lower value = more similar. Examples: Euclidean distance, Manhattan distance, Hamming distance <p>Application: Used in clustering, nearest neighbor classification, and anomaly detection.</p>
7 a	<p>How to deal with data when it is not linearly separable?</p> <ul style="list-style-type: none"> • Kernels allow SVM to build nonlinear decision boundaries without explicit computation in high-dimensional space.  <p>The scatter plot shows two classes of data points (red '+' and green 'x') in a 2D space defined by axes x_1 and x_2. The data is labeled 'Not Linearly Separable' because a straight line cannot separate the two classes.</p>

b

Compare logistic regression with linear regression on indicators.

Here's a comparison between Logistic Regression and Linear Regression based on key indicators:

Indicator	Linear Regression	Logistic Regression
Purpose	Predicts continuous outcomes	Predicts categorical outcomes (usually binary: 0 or 1)
Output	Real-numbered predictions (e.g., salary, price)	Probability (between 0 and 1) which is then classified (e.g., spam or not)
Equation	$y = \beta_0 + \beta_1 x + \varepsilon$	$P(y = 1) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x)}}$
Error Metric	Mean Squared Error (MSE), RMSE, MAE	Log Loss, Accuracy, Precision, Recall, F1-Score
Linearity	Assumes a linear relationship between input and output	Assumes a linear relationship between input and log-odds
Interpretation	Output directly predicts value	Output gives probability of a class
Range of Output	$-\infty$ to $+\infty$	0 to 1
Use Case Examples	Predicting house prices, sales, or temperature	Classifying emails, detecting fraud, diagnosing disease

c

Outline the tools for displaying relationship between single, two, and more than two variables.

Single variable:

• Histogram: Shows distribution and frequency.

• Boxplot: Highlights median, quartiles, and outliers.

Two variables:

• Scatter Plot: Shows correlation and distribution.

• Box Plot: Compares distribution and outliers.

• Line Graph: Visualizes trends over time.

• Correlation Matrix: Displays strength of variable relationships.

A scatter plot between income and age may reveal a positive correlation, while a boxplot of income alone may show income disparities.

Feature	Single Variable Tools	Multi-Variable Tools
Purpose	Understand distribution, spread, and central tendency	Examine relationships, correlations, and interactions
Examples	Histogram, Bar chart, Boxplot, Density plot	Scatter plot, Heatmap, Pair plot, 3D plot
Insight Focus	Frequency, outliers, skewness	Patterns, clusters, correlation, causation
Simplicity	Easier to interpret	Can be complex but more informative

8

Case Study

ABC corporation a leading manufacturing of consumer electronics faces a critical business problem regarding a declining market share and profitability in the Smartphone segment. Several key elements contribute to this challenges necessitating exploratory business analysis.

Firstly ABC Corp, observe a steady decline in Smartphone sales over the past fiscal quarters, despite increased marketing expenditure and product innovation efforts.

Secondly, market research indicates shifting common preference towards competitor brands offering advanced features and lower price points.

Thirdly, internal data reveals inconsistencies in product performance metrics and customer satisfaction rating, highlighting potential quality and design issues impacting brand perception. Moreover, supply chain descriptions and production delays further increase the problem leady to inventory stock outs and missed revenue opportunities.

To address these pricing issues, ABC corporation embarks on an exploratory business analysis initiative, aiming to uncover underlying patterns, identify root causes, and formulate data-driven strategies for business revitalization and market repositioning.

a. Analyze the Key Factors for the Decline in Market Share and Profitability in the Mobile Segment

The following factors are contributing to ABC Corp's declining performance in the smartphone market:

1. Declining Sales Despite Marketing and Innovation: Despite investing in marketing and product innovation, smartphone sales have steadily dropped, indicating a disconnect between offerings and market needs.
2. Shifting Consumer Preferences: Customers are moving toward competitor brands that offer better features at lower prices, suggesting ABC Corp's value proposition may be outdated or overpriced.
3. Product and Quality Issues: Internal data highlights inconsistencies in product performance and customer satisfaction, signaling potential flaws in product design and quality.
4. Supply Chain Problems: Delays and poor inventory management have led to stock-outs and missed revenue opportunities.
5. Brand Perception Challenges: Quality and performance issues are damaging brand reputation, reducing customer trust and loyalty.

b. Explain the Benefits of Exploratory Business Analysis (EDA) for ABC Corp

Exploratory Data Analysis (EDA) can offer ABC Corp several strategic advantages:

1. Better Understanding of Business Data: EDA helps uncover hidden patterns, relationships, and anomalies in data before building predictive models.
2. Identifying Trends and Patterns: It reveals insights into customer behavior, product performance, and revenue fluctuations over time.
3. Early Detection of Data Issues: EDA highlights incomplete, inconsistent, or misleading data, allowing corrective actions early in the analysis process.

	<ol style="list-style-type: none"> 4. Supports Informed Decision-Making: Through visualizations and summaries, EDA makes it easier to communicate insights to stakeholders. 5. Example: A retail company may use EDA to identify underperforming product categories and refocus promotional strategies accordingly — ABC Corp can do the same for its smartphone portfolio. <p>c. Analyze the Role of Consumer Behaviour and Market Dynamics in Positioning Products in the Mobile Market</p> <p>Consumer behavior and market dynamics significantly influence product positioning in the smartphone industry:</p> <ol style="list-style-type: none"> 1. Feature and Price Sensitivity: Consumers today prefer smartphones that offer advanced features at competitive prices — value-for-money is a key driver. 2. Brand Perception and Trust: Customer reviews and satisfaction levels impact how the brand is perceived in the market, influencing new purchases. 3. Technology Trends and Peer Influence: Social proof, influencer marketing, and the rapid adoption of tech trends shape consumer choices. 4. Competitive Benchmarking: Competitor offerings often redefine customer expectations, making it essential to continuously align positioning with evolving preferences. 5. Buying Journeys and Experience: The ease of purchase, after-sales support, and overall user experience are becoming critical in winning and retaining customers. <p>d. How ABC Corp Can Leverage Insights from Exploratory Analysis to Regain Market Leadership</p> <p>ABC Corp can take the following steps using insights from EDA:</p> <ol style="list-style-type: none"> 1. Refine Product Strategy: Identify which product features resonate most with customers and optimize the smartphone lineup accordingly. 2. Targeted Marketing: Use customer segmentation data to tailor marketing campaigns to specific demographics or user behavior. 3. Improve Quality and Satisfaction: Analyze performance metrics and feedback to fix product flaws, enhance design, and improve customer satisfaction. 4. Optimize Pricing and Offers: Use historical sales and competitor analysis to adjust pricing strategies and develop competitive bundles or promotions. 5. Streamline Operations: Address supply chain inefficiencies by identifying bottlenecks in production and distribution, minimizing stock-outs. 6. Enhance Customer Experience: Focus on user experience, from product design to post-sale service, based on data-driven insights. <p>By aligning data insights with strategic decisions, ABC Corp can rebuild customer trust, differentiate in a competitive market, and regain market leadership.</p>
--	---