Capstone Project-7

This Case Study has 3 (three) checkpoints defined in it.

Check Point Topics	Remarks	Max Marks
• Data manipulation and Visualization using Python (30 marks) • Statistical Analysis and Exploratory Data Analysis (50 marks)	**Checkpoint 1**	**80**
• Visualization using Power-BI Dashboard (40 marks) • Model Building using ML algorithms (80 marks)	**Checkpoint 2**	**120**
Final Presentation and Viva (50 marks)	**Checkpoint 3**	**50**

Domain:

Retail Analytics

Title: Exploratory Data Analysis and Customer Purchase Behavior Analysis in Retail

About:

Black Friday is an informal name for the Friday following Thanksgiving Day in the United States, which is celebrated on the fourth Thursday of November. The day after Thanksgiving has been regarded as the beginning of the United States Christmas shopping season since 1952, although the term "Black Friday" did not become widely used until more recent decades.

Many stores offer highly promoted sales on Black Friday and open very early, such as at midnight, or may even start their sales at some time on Thanksgiving. Black Friday is not an official holiday, but California and some other states observe "The Day After Thanksgiving" as a holiday for state government employees, sometimes in lieu of another federal holiday, such as Columbus Day. Many non-retail employees and schools have both Thanksgiving and the following Friday off, which, along with the following regular weekend, makes it a four-day weekend, thereby increasing the number of potential shoppers.

Black Friday has routinely been the busiest shopping day of the year in the United States since 2005, although news reports, which at that time were inaccurate, have described it as the busiest shopping day of the year for a much longer period. Similar stories resurface year upon year currently, portraying hysteria and shortage of stock, creating a state of positive feedback.

A retail company "ABC Private Limited" wants to understand the customer purchase behaviour (specifically, purchase amount) against various products of different categories. They have shared purchase summary of various customers for selected high-volume products from last month.
The data set also contains customer demographics (age, gender, marital status, city_type, stay_in_current_city), product details (product_id and product category) and Total purchase_amount from last month.

Objectives:

a. Data preprocessing: Handle missing values, encode categorical variables, and perform feature engineering to prepare the data for analysis.

b. Exploratory data analysis (EDA): Analyze the data to understand customer purchase behavior, identify patterns and trends, and explore relationships between variables.

c. Descriptive analysis: Calculate descriptive statistics, generate visualizations, and conduct statistical tests to summarize and highlight key findings related to customer purchase behavior.

d. Customer segmentation: Segment customers based on their purchase behavior to identify distinct groups and understand their characteristics and preferences.

e. Predictive modeling: Build predictive models to forecast customer purchase amounts based on customer demographics and product details using appropriate machine learning algorithms.

f. Evaluation and recommendations: Assess the performance of the predictive models, provide insights on factors influencing purchase behavior, and offer recommendations for improving sales and marketing strategies.

Data Dictionary:

Variable	Definition
User_ID	User ID
Product_ID	Product ID
Gender	Sex of User
Age	Age in bins
Occupation	Occupation (Masked)
City_Category	Category of the City (A,B,C)
Stay_In_Current_City_Years	Number of years stay in current city
Marital_Status	Marital Status
Product_Category_1	Product Category (Masked)
Product_Category_2	Product may belong to other category also (Masked)
Product_Category_3	Product may belong to other category also (Masked)
Purchase	Purchase Amount (Target Variable)

Check Point 1

Task 1.1(Data Manipulation and Visualization using Python)

Perform data manipulation tasks and visualize the dataset using Python. This task aims to explore and understand the dataset, clean the data, and create visualizations for further analysis.

Steps:
a. Load the dataset: Import the dataset into a Python environment (e.g., using pandas library) and create a data frame.

b. Data exploration: Perform initial exploration of the dataset to gain insights into its structure and content. Use functions such as .head(), .info(), .describe(), and .shape to understand the data's dimensions, variable types, and summary statistics.

c. Data cleaning: Identify and handle missing values, outliers, and inconsistent data. Implement appropriate techniques to clean the data, such as dropping or imputing missing values, removing outliers, and addressing inconsistent entries.

d. Data transformation: Apply necessary transformations to the data to make it suitable for analysis. This may include feature scaling, encoding categorical variables, creating derived variables, or aggregating data as required.

e. Data visualization: Utilize Python's data visualization libraries, such as matplotlib or seaborn, to create informative visualizations. Generate various types of plots, such as histograms, bar charts, scatter plots, or box plots, to understand the distribution, relationships, and patterns within the dataset.

f. Exploratory Data Analysis (EDA): Perform EDA techniques to uncover meaningful insights and relationships within the data. Conduct analyses such as correlation analysis, frequency analysis, or segmentation analysis to understand the factors influencing the price.

g. Data summary: Summarize the key findings from the data manipulation and visualization tasks, including notable data trends, patterns, and potential variables of interest for prediction.

Deliverables:
a. Python code: Provide well-documented Python code showcasing the data manipulation and visualization steps performed on the dataset.

b. Visualizations: Include visualizations generated during the data exploration and EDA processes, such as plots, charts, or graphs, that provide insights into the dataset.

c. Data summary: Prepare a concise summary highlighting the important findings and observations derived from the data manipulation and visualization tasks. Summarize any data cleaning or transformation steps undertaken to ensure data quality.

Optional Enhancements:
Depending on the dataset and specific project requirements, you can consider additional data manipulation and visualization techniques, such as:

a. Handling imbalanced data: If the dataset is imbalanced, apply techniques like oversampling or under sampling to balance the classes for better Modeling.

b. Interactive visualizations: Utilize libraries like Plotly or Bokeh to create interactive visualizations that allow for deeper exploration and interactivity.

c. Dimensionality reduction: Apply techniques like Principal Component Analysis (PCA) or t-SNE to visualize high-dimensional data in reduced dimensions.

d. Geospatial visualization: If the dataset contains location information, create geospatial visualizations using libraries like GeoPandas or Folium to understand the geographical patterns.

e. Temporal analysis: Analyze temporal patterns and trends by creating time series plots or heatmaps to identify seasonality or changes in the behavior over time.

Note: The specific data manipulation and visualization techniques may vary depending on the dataset and project requirements. Adapt the steps and enhancements accordingly.

Here are some indicative types of analysis you can perform. Please note that this is not an exhaustive list, you may add more

- Come up with appropriate results and visuals for the following:
- Maximum, spend in different categories of products
- Based on above data which set of customers can be offered personalised discount vouchers
- And in which category the voucher should be offered
- Or it should be on the total amount.

TASK 1.2 (Exploratory Data Analysis & Statistical Analysis)

Data Preparation/Analysis tasks include (but are not limited to) the following.

1. Descriptive statistics for both numerical and categorical and draw a few insights from them. (Univariate Analysis)
2. Bi- Variate Analysis and Multi-Variate Analysis
3. Missing values identification and treatment
4. Outlier analysis and treatment
5. Data scaling using min-max and/or Z-score normalization.
6. Data transformation

7. Feature Engineering
8. Perform relevant hypothesis testing (t, chi-Square, Anova tests)

Checkpoint 2

TASK 2.1 (Visualization using Power-BI Dashboard)

Objective:
Create an interactive and visually appealing Power BI dashboard for the project. This task aims to leverage Power BI's capabilities to visualize and explore the dataset, uncover insights, and present the findings in a user-friendly and interactive manner.

Steps:
a. Data import: Import the preprocessed and cleaned dataset into Power BI. Connect to the appropriate data source and load the data into the Power BI environment.

b. Data modeling: Perform any necessary data modeling tasks within Power BI to define relationships between tables, create calculated columns, or apply other transformations required for analysis.

c. Dashboard design: Design the layout and structure of the Power BI dashboard. Select appropriate visualizations, arrange them logically, and customize their appearance to ensure a cohesive and visually appealing dashboard.

d. Key performance indicators (KPIs): Identify and define relevant KPIs related to the project. Create visualizations, such as KPI cards or gauges, to track and display these key metrics prominently on the dashboard.

e. Exploratory data visualizations: Utilize various Power BI visualizations, such as bar charts, line charts, scatter plots, or treemaps, to explore different aspects of the dataset. Create interactive visualizations that allow users to drill down, filter, or highlight specific data points for deeper analysis.

f. Cross-filtering and slicing: Implement cross-filtering and slicing functionalities within Power BI to enable users to interactively filter and slice the data based on different criteria. This allows for dynamic exploration and comparison of patterns across different dimensions.

g. Insights and storytelling: Create narrative-driven visualizations and storytelling elements within the Power BI dashboard. Use text boxes, images, or tooltips to provide context, highlight key findings, and guide users through the insights derived from the dataset.

h. Dashboard interactivity: Set up interactions between different visualizations within the Power BI dashboard. Define how one visualization affects or filters another to create a seamless and interactive user experience.

i. Testing and refinement: Test the Power BI dashboard functionality, responsiveness, and user experience. Refine and optimize the visualizations, interactions, and overall performance as needed.

Deliverables:
a. Power BI dashboard: Provide the Power BI dashboard file (.pbix) containing the interactive visualizations, KPIs, and storytelling elements created for the project.

b. Documentation: Document the design decisions, visualizations used, and any notable insights or observations derived from the Power BI dashboard. Include a brief guide explaining how to navigate and interact with the dashboard for other users.

Optional Enhancements:
Depending on the specific project requirements and dataset, consider additional enhancements for the Power BI dashboard, such as:

a. Advanced calculations: Incorporate advanced calculations and measures using Power BI's DAX (Data Analysis Expressions) language to derive custom metrics or perform complex calculations based on data set.

b. Forecasting: Utilize Power BI's forecasting capabilities to create predictive visualizations based on historical data.

c. Natural language querying: Implement natural language querying functionality within the Power BI dashboard, allowing users to ask questions and receive visualizations or insights in response.

d. Data alerts: Configure data alerts within Power BI to notify stakeholders or users when specific metrics or thresholds are met or exceeded.

Note: Adapt the steps and optional enhancements according to the specific requirements of the project and the available features and capabilities of Power BI.

Connect the data with the Power BI desktop and perform Data Manipulation using Power Query Editor. Perform the below tasks in Power BI Desktop.

- Which Age group is purchasing the highest products from the store? What is the purchase amount of the age group between 0-17?
- Which gender is purchasing the highest products from the store?
- Display the total purchase happening in the store
- Visualize top 5 occupations by purchase
- Does marital status have any impact on the purchase amount?
- Display the maxim purchase that happened from the store
- Which product category was sold in maximum from the store.
- Display the top 5 product categories sold in maximum.
- Built a decomposition for purchase amount with appropriate variables affecting the purchase amount.
- Visualize a Key influencer visual for the purchase amount to explain it by product category. Identify the impact of product category on purchase amount

NOTE: Results and graphs must be backed with appropriate inferences and insights.

TASK 2.2 (Model building using ML algorithms)

Objective:
Build machine learning models to predict the price of the Marketing Campaign on the pre-processed dataset. This task aims to apply various ML algorithms, and train and evaluate them to identify the most effective approach for predicting price.

Steps:
a. Data preparation: Split the pre-processed dataset into training and testing sets. Define the independent variables and the dependent variable appropriately.

b. Select ML algorithms: Choose a set of ML algorithms suitable for prediction. Common algorithms include regression, decision trees, random forests, gradient boosting, or support vector machines (SVM).

c. Model training: Train each selected ML algorithm using the training dataset. Fit the models to the training data and adjust the hyperparameters, if necessary, to optimize performance.

d. Model evaluation: Evaluate the trained models using the testing dataset. Calculate evaluation metrics such as accuracy, precision, recall, F1-score, and AUC-ROC to assess the models' predictive performance.

e. Model comparison: Compare the performance of different ML algorithms based on the evaluation metrics. Identify the most effective algorithm(s) for prediction.

f. Hyperparameter tuning: Fine-tune the hyperparameters of the selected ML algorithm(s) to further improve their performance. Utilize techniques such as grid search, random search, or Bayesian optimization to find optimal hyperparameter configurations.

g. Model interpretation: Interpret the trained ML models to gain insights into the factors contributing to the prediction. Analyze feature importance, coefficients, or decision rules to understand the variables' impact on prediction.

h. Final model selection: Select the best-performing ML algorithm based on the evaluation metrics, interpretability, and business requirements.

Deliverables:
a. Model building code: Provide well-documented code showcasing the implementation of ML algorithms, including data preparation, model training, evaluation, hyperparameter tuning, and interpretation.

b. Evaluation results: Present the evaluation metrics, such as accuracy, precision, recall, F1-score, and AUC-ROC, for each trained model. Compare and summarize the results to identify the best-performing algorithm.

c. Model interpretation summary: Summarize the key insights derived from the interpretation of the ML models, including feature importance, coefficients, or decision rules.

d. Final model documentation: Document the selected ML algorithm, along with the optimal hyperparameter configuration, as the final model for price prediction. Explain the rationale behind the model selection and its potential implications for the telecommunication industry.

Optional Enhancements:
Depending on the project requirements and available resources, consider the following enhancements:

a. Ensemble Modeling: Explore ensemble techniques, such as stacking, bagging, or boosting, to combine multiple ML algorithms for improved prediction accuracy.

b. Feature selection: Implement feature selection techniques, such as recursive feature elimination or feature importance ranking, to identify the most relevant features for prediction and refine the model accordingly.

c. Model deployment: Deploy the final selected ML model into a production environment, allowing real-time or batch predictions on new customer data. Ensure scalability, reliability, and compatibility with the company's existing infrastructure.

d. Performance monitoring: Set up a system to monitor the performance of the deployed model, track prediction accuracy, and recalibrate or retrain the model periodically to account for changes in customer behavior or market dynamics.

Note: Adapt the steps and optional enhancements based on the specific requirements of the project and the available ML algorithms and resources.

1. Build an appropriate ML model/s on the data.
2. Compare various ML models with appropriate regularization and/or hyperparameter tuning.
3. Evaluate the performance of the model.
4. Identify the right metric to evaluate the performance of the model.
5. Identify issues and concerns on the given data and suggest the best technique/s to overcome the issues.

Checkpoint 3

Prepare a crisp Final presentation including all the Checkpoint achievements and appear for the Q&A session.

The above three Checkpoints completes the Capstone Project