Capstone Project-1

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:

Telecommunication Industry

Title:

Customer Churn Prediction for Telecommunication Company: Enabling Targeted Retention Offers using Machine Learning

About:

Telecommunication Company XYZ recognizes the importance of retaining customers to sustain business growth and profitability. To achieve this, accurately predicting customer churn and implementing targeted retention offers is crucial. This capstone project focuses on developing a machine learning-based system for Customer Churn Prediction at Telecommunication Company XYZ, enabling the timely identification of customers at risk of churning, and facilitating personalized retention strategies.

Objectives:

a. Data preprocessing: Preprocess the data by handling missing values, encoding categorical variables, and normalizing numerical features.

b. Feature engineering: Extract relevant features from the dataset to capture various aspects influencing churn, such as call duration, data usage, plan details, customer complaints, network performance, and customer tenure.

c. Model development: Apply advanced machine learning algorithms, such as logistic regression, decision trees, random forests, or gradient boosting, to build accurate predictive models for customer churn.

d. Evaluation: Assess the performance and effectiveness of the predictive models using appropriate evaluation metrics, including accuracy, precision, recall, F1-score, and AUC-ROC.

e. Business impact assessment: Analyze the potential business impact of the predictive churn model, including estimated churn reduction, revenue retention, and return on investment (ROI).

Data Dictionary:

Category	Attribute	Description
Customer Profile	MSISDN	Subscriber MSISDN (Primary key)
	Status	Status is committed or non-committed (Pack activated or not)
	Segment	Customer divided into some segments (Gold, silver etc)
	Age on network	Difference between activation date and current date
	Region Type	2G/3G/4G/5G
	Total no. of complaints	MSISDN wise monthly count of complaint (current month, previous month)
	Is VAS subscriber	0/1, whether customer is VAS subscriber or not
Usage Profile	Total no. of outgoing calls	Any outgoing call count monthly (current month, previous month)
	Total no. of incoming calls	Any incoming call count monthly (current month, previous month)
	Total no. of outgoing SMS	Any outgoing SMS count monthly (current month, previous month)
	Total no. of incoming SMS	Any incoming SMS count monthly (current month, previous month)
	Total free data usage	Total free data usage (current month, previous month)
	Total data usage	Total data usage (current month, previous month)
	Total 4G Data usage	Total 4G data usage (current month, previous month)
	4G Upward Tag	0/1, whether customer 4G data usage increasing month on month or not
	Total 3G Data usage	Total 3G data usage (current month, previous month)
	Total 2G Data usage	Total 2G data usage (current month, previous month)
	Data Social Media Usage	Total social media data usage (current month, previous month)
	Data App Usage	Total app data usage (current month, previous month)
	Total incoming call duration	Total incoming call duration in minutes (current month, previous month)
	Total outgoing call duration	Total out going call duration in minutes
	On net outgoing call count	Total on net call count (current month, previous month)
	Off net outgoing call count	Total off net call count (current month, previous month)
	On net outgoing call duration	Total on net call duration in minutes (current month, previous month)
Revenue Profile	SMS revenue	Monthly SMS Revenue (current month, previous month)
	Call Revenue	Monthly call Revenue (current month, previous month)
	Data revenue	Monthly data Revenue (current month, previous month)
Recharge Profile	Total recharge amount	Monthly total recharge amount (current month, previous month)
	Total recharge count	Monthly total recharge count (current month, previous month)
	Current balance	Present balance of the subscriber
	Current product id	ID of latest used product
	Current top up value	Current top up value
	Validity days	Current validity days of the customer
	Days since last recharge	No. of days since last recharge
	Last bundle sms Purchased	Last bundle sms Purchased
	Last bundle Purchased	Last bundle Purchased
	Last Recharge Channel	Last Recharge Channel

	Data Top up	0/1 whether the customer has done any data top up
	Bundle pack	0/1 whether the customer is on bundle pack in present month
Handset Profile	Handset category	Handset Network Compatibility (2G/3G/4G/5G)
	Sim support	Multi sim supported Handset
	Smart Phone Tag	0/1, whether the phone is smartphone or not
	Handset change	Count of handset change monthly (current month, previous month)
App Profile	App user	0/1, whether the customer is XYZ company's app user or not
	Days on app	difference between registration date and current date
	Days since last app use	difference between last used date and current date
Activity Profile	Days since last Data Session	Difference between last data session date and current date
	Days since last VAS Session	Difference between VAS session date and current date
	Days since last Voice Session	Difference between last voice session date and current date
Dependent Parameter	Churn Flag	0/1, Whether the customer will churn or not in next 30 days

Check Point 1

Task 1.1 (Data Manipulation and Visualization using Python)

Perform data manipulation tasks and visualize the telecommunications customer churn 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 telecommunications customer churn dataset into a Python environment (e.g., using pandas library) and create a dataframe.

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 customer churn.

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 predicting customer churn.

Deliverables:
a. Python code: Provide well-documented Python code showcasing the data manipulation and visualization steps performed on the telecommunications customer churn 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 churn dataset is imbalanced, apply techniques like oversampling or undersampling 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 of churn.

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

Come up with appropriate results and visuals for the following:

1. Which variables are significant in predicting the customer churn?
2. How well do those variables describe the customer churn?
3. Perform relevant hypothesis testing (t, chi-Square, Anova tests)

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

4. Descriptive statistics for both numerical and categorical and draw a few insights from them. (Univariate Analysis)
5. Bi- Variate Analysis and Multi-Variate Analysis
6. Missing values identification and treatment
7. Outlier analysis and treatment
8. Data scaling using min-max and/or Z-score normalization
9. Data transformation
10. Feature Engineering
11. 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 telecommunications customer churn project. This task aims to leverage Power BI's capabilities to visualize and explore the churn dataset, uncover insights, and present the findings in a user-friendly and interactive manner.

Steps:
a. Data import: Import the preprocessed and cleaned telecommunications customer churn 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 customer churn. 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 churn 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 churn 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 churn 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 telecommunications customer churn 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 churn data.

b. Forecasting: Utilize Power BI's forecasting capabilities to create predictive visualizations that project churn trends or predict future churn probabilities 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 churn-related 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.

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 customer churn in the telecommunications industry based on the preprocessed dataset. This task aims to apply various ML algorithms, train and evaluate them to identify the most effective approach for predicting customer churn.

Steps:
a. Data preparation: Split the preprocessed dataset into training and testing sets. Define the features (independent variables) and the churn label (dependent variable) appropriately.

b. Select ML algorithms: Choose a set of ML algorithms suitable for customer churn prediction. Common algorithms include logistic 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 customer churn 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 customer churn. Analyze feature importance, coefficients, or decision rules to understand the variables' impact on churn 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 related to customer churn.

d. Final model documentation: Document the selected ML algorithm, along with the optimal hyperparameter configuration, as the final model for customer churn 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 churn 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 telecommunication 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.

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