DEPI – Graduation Project

Sales Forecasting and Demand Prediction

Milestone-1: Data Collection, Exploration, and Preprocessing

Tasks Distribution

Milestone Description

1.Data Collection:

- Acquire a churn dataset from sources like Kaggle, UCI Repository, or generate synthetic data.

- Ensure the dataset includes key features such as customer demographics, usage patterns, subscription details, etc.

2. Data Exploration:

- Conduct exploratory data analysis (EDA) to understand the dataset's structure and identify potential relationships between features.

- Check for missing values, duplicates, and outliers. Summarize data distributions and basic statistics.

3. Preprocessing and Feature Engineering:

- Address missing data through imputation or removal.

- Handle outliers and ensure data consistency.

- Transform features using techniques like scaling, encoding categorical data, and creating interaction features relevant to churn prediction.

4. Exploratory Data Analysis (EDA):

- Create visualizations (heatmaps, pair plots, histograms) to detect patterns, correlations, and outliers.

- Document key patterns and relationships in the data.

Deliverables:

• EDA Report: A document summarizing key insights from data exploration and preprocessing decisions.

• Interactive Visualizations: An EDA notebook showcasing visualizations that reveal key patterns and relationships.

• Cleaned Dataset: A dataset that is cleaned and prepared for machine learning.

Task Distribution for 6 Developers

Developer 1: Data Collection

Tasks:

1. **Acquire the Dataset:**
    - **Identify and download a churn dataset from sources like Kaggle, UCI Repository, or generate synthetic data.**
    - **Ensure the dataset includes key features such as customer demographics, usage patterns, subscription details, etc.**
    - **Verify the dataset's integrity (e.g., file format, completeness, and relevance to the problem).**
2. **Dataset Documentation:**
    - **Document the dataset's source, features, and any initial observations about its structure.**
    - **Share the dataset with the team in a shared repository (e.g., Google Drive, GitHub).**

Deliverables:

- **A clean, accessible dataset ready for exploration.**
- **A brief document summarizing the dataset's source, features, and initial observations.**

Developer 2: Data Exploration – Basic Analysis

Tasks:

1. **Initial Data Exploration:**
    - **Load the dataset into a Python environment (e.g., Jupyter Notebook).**
    - **Perform basic exploratory data analysis (EDA) to understand the dataset's structure:**
        - **Check the number of rows and columns.**
        - **Identify data types (numeric, categorical, etc.).**
        - **Summarize basic statistics (mean, median, standard deviation, etc.).**
2. **Missing Values and Duplicates:**
    - **Identify missing values and duplicates in the dataset.**
    - **Summarize the percentage of missing values per feature and decide on a strategy (e.g., imputation or removal).**

Deliverables:

- **A notebook with basic EDA, including summary statistics and missing value analysis.**
- **A report summarizing the dataset's structure and missing value insights.**

Developer 3: Data Exploration – Outlier Detection and Handling

Tasks:

1. **Outlier Detection:**
    - **Use statistical methods (e.g., IQR, Z-score) or visualization techniques (e.g., boxplots) to detect outliers in numeric features.**
    - **Summarize the findings and decide on a strategy for handling outliers (e.g., capping, removal).**
2. **Data Consistency Checks:**
    - **Ensure data consistency by checking for logical errors (e.g., negative values in age, unrealistic subscription lengths).**
    - **Document any inconsistencies and propose solutions.**

Deliverables:

- **A notebook with outlier detection and handling techniques.**
- **A report summarizing outlier findings and proposed solutions.**

Deve

loper 4: Preprocessing - –issing Data Handling and Feature Transformation

Tasks:

1. **Missing Data Handling:**
    - **Implement strategies for handling missing data (e.g., mean/median imputation, removal of rows/columns).**
    - **Document the chosen strategy and its justification.**
2. **Feature Transformation:**
    - **Scale numeric features (e.g., using Min-Max scaling or Standardization).**
    - **Encode categorical variables (e.g., one-hot encoding, label encoding).**
    - **Create interaction features (e.g., combining usage patterns and subscription details).**

Deliverables:

- **A notebook with code for missing data handling and feature transformation.**
- **A report summarizing the preprocessing steps and their impact on the dataset.**

Developer 5: Preprocessing - –eature Engineering

Tasks:

1. **Feature Engineering:**
    - **Create new features that may be relevant for churn prediction (e.g., average usage per month, customer tenure).**
    - **Perform feature selection to identify the most important features for the model.**
2. **Data Consistency Checks:**
    - **Ensure that the engineered features are consistent and free from errors.**
    - **Document the new features and their relevance to the problem.**

Deliverables:

- **A notebook with code for feature engineering and selection.**
- **A report summarizing the new features and their relevance.**

Developer 6: Exploratory Data Analysis (EDA) and Visualization

Tasks:

1. **Data Visualization:**
    - **Create visualizations to explore patterns, correlations, and outliers in the data:**
        - **Heatmaps for correlation analysis.**
        - **Pair plots for feature relationships.**
        - **Histograms for distribution analysis.**
    - **Use tools like Matplotlib, Seaborn, or Plotly for interactive visualizations.**
2. **EDA Report:**
    - **Summarize key insights from the visualizations.**
    - **Document any patterns or relationships that could inform the churn prediction model.**

Deliverables:

- **An EDA notebook with interactive visualizations.**
- **A report summarizing key insights from the visualizations.**

Final Deliverables for the Milestone

1. **EDA Report: A comprehensive document summarizing insights from data exploration and preprocessing decisions.**
2. **Interactive Visualizations: An EDA notebook showcasing visualizations that reveal key patterns and relationships.**
3. **Cleaned Dataset: A dataset that is cleaned, preprocessed, and ready for machine learning.**

Addressing the Requirements

The task distribution aligns with the requirements of the Data Collection, Exploration, and Preprocessing milestone:

- **Data Collection: Developer 1 ensures the dataset is acquired and documented.**
- **Data Exploration: Developers 2 and 3 handle basic analysis, missing values, duplicates, and outliers.**
- **Preprocessing and Feature Engineering: Developers 4 and 5 focus on handling missing data, transforming features, and creating new features.**
- **EDA and Visualization: Developer 6 creates visualizations and summarizes key insights.**

This distribution ensures that each developer has a clear, manageable set of tasks while contributing to the overall milestone deliverables.