Predicting Customer Churn for a Telecommunications Company

Project Overview

Project Title: Predicting Customer Churn for a Telecommunications Company
  Objective: Use a synthetic dataset to predict customer churn, analyze factors influencing churn, and explore customer segmentation.

Project Activities

1. Understanding the Dataset

Activity: Dataset Exploration

- **Description**: Students will load the dataset, understand its structure, and perform initial exploratory data analysis (EDA).

- **Deliverables**: A report detailing the dataset's features, their distributions, and any initial observations.

Guidelines:

- Describe each feature and its possible values.

- Visualize distributions of numerical features (e.g., histograms, box plots).

- Analyze relationships between features using scatter plots and correlation matrices.

2. Data Preprocessing

Activity: Data Cleaning and Preparation

- **Description**: Handle missing values, detect and treat outliers, and encode categorical variables.

- **Deliverables**: A cleaned dataset and a report detailing the preprocessing steps.

Guidelines:

- Choose appropriate methods to handle missing values (e.g., imputation, removal).

- Identify and treat outliers using techniques such as clipping or transformation.

- Encode categorical variables using one-hot encoding or label encoding.

- Normalize or scale numerical features if necessary.

3. Exploratory Data Analysis (EDA)

Activity: In-depth EDA

- **Description**: Perform detailed EDA to understand key factors influencing customer churn.

- **Deliverables**: A comprehensive EDA report with visualizations and insights.

Guidelines:

- Investigate the relationship between features and the target variable (*Churn*).

- Use visualization techniques like bar charts, box plots, and heatmaps.

- Perform hypothesis testing where applicable (e.g., t-tests, chi-square tests).

4. Feature Engineering

Activity: Create New Features

- **Description**: Engineer new features that could improve model performance.

- **Deliverables**: A dataset with new features and a report explaining their creation and potential impact.

Guidelines:

- Combine existing features to create new ones (e.g., interaction terms, ratios).

- Use domain knowledge to add relevant features (e.g., total service usage).

- Justify the inclusion of each new feature with potential benefits.

5. Modeling

Activity: Build and Evaluate Models

- **Description**: Train different machine learning models to predict customer churn.

- **Deliverables**: Trained models, evaluation metrics, and a report comparing model performance.

Guidelines:

- Split the data into training and testing sets.

- Train several models (e.g., logistic regression, decision trees, random forest, gradient boosting).

- Evaluate models using metrics like accuracy, precision, recall, F1-score, and ROC-AUC.

- Perform cross-validation to ensure robust performance estimates.

6. Model Tuning

Activity: Hyperparameter Tuning

- **Description**: Optimize model hyperparameters to improve performance.

- **Deliverables**: Best-tuned models and a report on the tuning process and results.

Guidelines:

- Use techniques like grid search or randomized search for hyperparameter tuning.

- Compare the performance of tuned models against default models.

- Document the tuning process and the chosen hyperparameters.

7. Model Interpretation

Activity: Interpret the Model

- **Description**: Analyze the models to understand which features are most important for predicting churn.

- **Deliverables**: A report on feature importance and model interpretation.

Guidelines:

- Use feature importance scores from models like decision trees or random forests.

- Apply SHAP (SHapley Additive exPlanations) or LIME (Local Interpretable Model-agnostic Explanations) for detailed model interpretation.

- Discuss how the insights align with domain knowledge and business logic.

8. Presentation

Activity: Present Findings

- **Description**: Prepare and deliver a presentation summarizing the project findings and recommendations.

- **Deliverables**: Presentation slides (only in PDF format) and a recorded or live presentation.

Guidelines:

- Summarize the key findings from EDA, modeling, and clustering.

- Highlight important insights and their implications for the business.

- Provide recommendations based on the analysis (e.g., strategies to reduce churn).

Evaluation Criteria

- **Understanding and Exploration**: Clarity and depth of dataset exploration and initial analysis.

- **Data Preprocessing**: Effectiveness of data cleaning, handling of missing values, and outlier treatment.

- **EDA and Feature Engineering**: Thoroughness of EDA and creativity in feature engineering.

- **Modeling and Evaluation**: Accuracy and robustness of models, choice of evaluation metrics.

- **Interpretation and Insights**: Depth of model interpretation and quality of business insights.

- **Clustering Analysis**: Effectiveness of clustering approach and quality of segment analysis.

- **Presentation**: Clarity, organization, and professionalism of the presentation.