Customer Churn Prediction Report

Introduction

The goal of this project is to develop a machine learning model to predict customer churn based on historical customer data. Customer churn, or customer attrition, refers to the phenomenon where customers discontinue their services or subscriptions with a company. Understanding and predicting customer churn is crucial for businesses to retain customers and improve customer satisfaction.

In this report, we will provide a summary of the approach taken to tackle this problem, including data preprocessing, feature engineering, and model selection decisions.

Data Preprocessing

Data Loading

The first step was to load the dataset from the 'customer\_churn\_large\_dataset.xlsx' file using the Pandas library. The dataset contains various attributes related to customer information and interactions.

Initial Data Exploration

To gain an initial understanding of the data, we called the 'describe' method on the DataFrame. This provided a summary of basic statistics for each numerical column, including mean, standard deviation, minimum, maximum, and quartile values.

Data Cleaning

The 'CustomerID' and 'Name' columns were dropped as they were not expected to provide any predictive value for the churn prediction task. The 'Age' and 'Location' columns were also removed as their influence on customer churn was unclear.

Label Encoding

The 'Gender' column was encoded using Label Encoding, converting the categorical values (Male and Female) into numerical representations (0 and 1).

Handling Missing Data

A check for missing data was performed, and it was found that there were no missing values in the dataset.

Feature Scaling

Feature scaling is important for certain machine learning algorithms, including Gaussian Naive Bayes, which was the chosen model in this case. The 'Total\_Usage\_GB' and 'Monthly\_Bill' columns were standardized using the StandardScaler to have a mean of 0 and a standard deviation of 1. This ensured that all features had the same scale, making them compatible for the chosen algorithm.

Data Splitting

The dataset was split into training and testing sets using the 'train\_test\_split' method from scikit-learn. An 80/20 split was used, with 80% of the data used for training and 20% for testing.

Model Selection

The model selected for this customer churn prediction task was Gaussian Naive Bayes. Naive Bayes is a simple yet effective algorithm for binary classification tasks, making it a good choice for this problem.

Model Training and Evaluation

The selected model was trained on the training data, and predictions were made on the testing data. Model performance was evaluated using various metrics, including:

Accuracy: Measures the overall correctness of the model's predictions.

Precision: Measures the proportion of true positive predictions among all positive predictions.

Recall: Measures the proportion of true positive predictions among all actual positives.

F1 Score: A harmonic mean of precision and recall, providing a balance between the two.

A confusion matrix was also used to visualize the model's performance in terms of true positives, true negatives, false positives, and false negatives.

Conclusion

In this project, we successfully developed a customer churn prediction model using Gaussian Naive Bayes. The model achieved [Accuracy Value] accuracy on the testing data, indicating its ability to predict customer churn.

Further improvements and experimentation could include trying different machine learning algorithms, exploring more advanced feature engineering techniques, and optimizing hyperparameters to potentially enhance the model's performance.