Customer Churn Prediction Report

Introduction

This report provides an overview of the data preprocessing, exploratory data analysis (EDA), feature engineering, model selection, and optimization process for predicting customer churn. The goal of this analysis is to understand customer behavior and build a predictive model to identify customers at risk of churning.

Data Preprocessing and Cleaning

The data preprocessing and cleaning steps involve loading the dataset, checking for missing values, and handling duplicate entries.

- The dataset was loaded from the 'customer_churn_large_dataset.xlsx' file.
- A check for missing data was performed using `df.info()`, which showed that there were no null values in the dataset.
- Duplicate values were also checked using `df.duplicated().sum()`, and no duplicate values were found.

Exploratory Data Analysis (EDA)

EDA is a crucial step to understand the data and identify patterns. Key findings from the EDA include:

- Summary statistics for numerical features were calculated, including average age, subscription length, monthly bill, total usage (in GB), and churn rate.
- Histograms and visualizations were created to explore the distribution of age, subscription length, monthly bill, total usage, gender, location, and churn.

Feature Engineering

Feature engineering involves creating new features or transformations of existing features to improve the model's predictive power. In this analysis, we explored two potential features: usage in specific locations and subscription length in specific locations. However, these features did not show strong correlations with churn, and it was decided not to include them in the final model.

Model Selection and Optimization

Several machine learning models were evaluated for predicting customer churn. These models include:

1. Logistic Regression
2. Linear Support Vector Classifier (LinearSVC)
3. Support Vector Classifier (SVC)
4. Random Forest Classifier

5. Artificial Neural Network (ANN)

The models were trained and evaluated using accuracy as the metric. The results showed that the accuracy of all models was around 50%, which suggests that predicting churn based on the given features is challenging due to the balanced nature of the dataset.

Model Deployment

The ANN model was selected as the final model, and it was saved using `joblib.dump()`. While the model's accuracy was not very high, it can still be considered for deployment in a production environment.

Conclusion

In conclusion, this analysis explored the customer churn prediction problem, but the results indicate that predicting churn based solely on the provided features is challenging. The dataset is balanced, and none of the features showed strong correlations with churn. Further feature engineering or the inclusion of additional data may be necessary to improve predictive performance.

The final model, an Artificial Neural Network, achieved the best results among the models evaluated, but it still had limited accuracy. Further efforts to collect more relevant data and refine the features may lead to better predictive models in the future.