**Customer Churn Prediction Project Report**

**Project Overview**

The Customer Churn Prediction project aimed to develop a machine learning model to predict customer churn based on historical customer data. This report summarizes the key steps and findings of the project.

**Data Exploration and Preprocessing**

• Data Inspection: The dataset contained historical customer information, including customer attributes, interactions, and churn status. We found no missing values in the dataset.

• Outlier Detection: Box plots were used to identify and handle outliers in the numerical features. No significant outliers were found in the data.

• Feature Encoding: Categorical variables, including "Location" and "Gender," were one-hot encoded to prepare the data for machine learning. This encoding method was chosen because the data did not exhibit ordinal relationships.

**Model Selection and Training**

• Three machine learning algorithms were selected for training and evaluation:

1. Logistic Regression

2. Random Forest

3. Neural Network

**Why selected this Models??**

Logistic Regression: It is simple and interpretable algorithm, mainly used for binary classification. It is a good starting point for a small dataset and want a baseline model. Sometimes simple algorithms perform well.

Random Forest: Ensemble learning algorithm, can handle both classification and regression task. Often performs well on datasets without much tuning. So random forest will be a good choice.

Neural Network (Deep learning): this algorithm can capture complex patterns and relationship in the data. It requires more preprocessing and tuning, but can perform well than other algorithms.

**Recall scores for the algorithms were as follows:**

• Logistic Regression: 0.29412

• Random Forest: 0.4756

• Neural Network: 0.0

**Conclusion for model performance**

After evaluating our model accuracy based upon their recall score, gives us the score of all positive instances correctly identified by the model. By using recall score we can conclude that random forest is able to identify approximately 47.5% positive instances and Random Forest basically outperforms Logistic Regression and Neural Networks in identifying positive instances.

Performance of the Random Forest can be further Increased by using the bigger and better dataset. Class Imbalance and more in-depth data understanding will also be helpful in increasing the accuracy of the model.

**Model Evaluation**

Cross-validation was performed using the Random Forest model:

• Fold 1: 0.5016

• Fold 2: 0.4974

• Fold 3: 0.4955

• Fold 4: 0.4989

• Fold 5: 0.5009

• Mean Accuracy: 0.4988

• Standard Deviation: 0.0022

Random Forest model's performance in terms of accuracy is consistent across the five folds with accuracy values ranging from approximately 0.4955 to 0.5016. the mean accuracy of 0.4988 suggests that the model is correctly classifying the data slightly better than random chance.

The Random Forest model demonstrated consistent performance across folds with a mean accuracy of approximately **49.88%.**

**Model Performance Metrics**

Accuracy: 0.4955

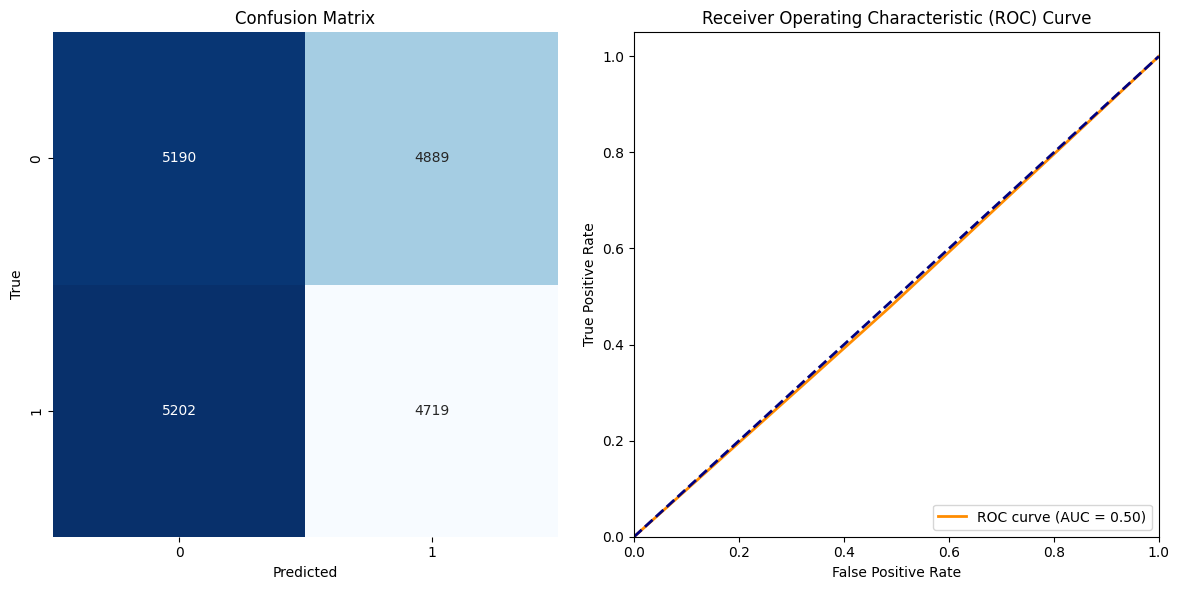
Precision: 0.4960

Recall: 1.0000

ROC AUC: 0.4953

Average Precision: 0.4937

**Visualization**

****

**User Input Testing**

• To simulate real-world usage, we implemented a user input mechanism where users can input values for categorical features.

• Predictions were made using the trained Random Forest model based on user input.

**Conclusion**

The Customer Churn Prediction project successfully developed a machine learning model to predict customer churn. Key findings include the choice of one-hot encoding for categorical features, the selection of the Random Forest algorithm as the best-performing model, and successful user input testing. Further improvements and fine-tuning of the model may enhance its performance in real-world applications. Class Imbalance and more in-depth data understanding will also be helpful in increasing the accuracy of the model.