Customer Churn Prediction Report

Objective:

- Develop a machine learning model to predict customer churn based on historical customer data. You will follow a typical machine learning project pipeline, from data preprocessing to model deployment.

Dataset:

(provided)

Machine learning frameworks used:

- scikit-learn

Data analysis and Preprocessing:

- The distribution of the data was checked for class imbalance, and the dataset has a balanced distribution of data for both classes.

- Plotted graphs like Distplot, Heatmap of correlation of features, Sunburst graph of each feature with respect to the target label, and plotted box plots to find any outlier in the data. No outliers were found.
- Used the DictVectorizer() from scikit-learn to encode the categorical variables and to normalize the numerical data.

Models Used:

- Random Forest Classifier with Grid search cross validation
    - Since the problem was a classification task, a Random Forest Classifier was chosen.
    - In order to increase the accuracy of the classifier, grid search cross validation was implemented.
    - However the random forest classifier was able to produce a training set score of **0.54** and test data score of **0.504**

- XGBoost Classifier
    - Due to the lower accuracy of the random forest model, the XGBoost model was chosen next.

    - However, the XGBoost classifier was only able to produce a train data score of **0.5028** and test data score of 0.**499.**

- Multi-Layer Perceptron
    - The same data was MLP Classifier with 500 neurons.
    - Results of the MLP Classifier were similar to both the above classifiers.
    - It gave a training set score of **0.5** and testing data score of **0.49.**
    -

The performances of these three classifiers are present in the attached Jupyter notebook, each evaluated on the testing data to find the F1-score, precision, recall and accuracy. Also the ROC curve and the confusion matrices are plotted for each model for the testing data.

Model Deployment:

- The classifiers are deployed using Streamlit.
- A provision to choose any of the three classifiers is provided
- Docker image is created to help to make the model be deployed irrespective of the environment.

Conclusion:

- With the provided dataset, the classifiers were only able to produce an accuracy of around 50%
- The provided data is not sufficient to produce accurate predictions.