Churn Prediction Report

Problem Statement:

Decreasing the Customer Churn is a key goal for any business. Predicting Customer Churn (also known as Customer Attrition) represents an additional potential revenue source for any business. Customer Churn impacts the cost to the business. Higher Customer Churn leads to loss in revenue and the additional marketing costs involved with replacing those customers with new ones.

In this challenge, as a data scientist of a bank, you are asked to analyze the past data and predict whether the customer will churn or not in the next 6 months. This would help the bank to have the right engagement with customers at the right time.

Objective

Our objective is to build a machine learning model to predict whether the customer will churn or not in the next six months.

Understanding Business Problem:

Churn is defined in business terms as 'when a client cancels a subscription to a service they have been using.' In this case members who are not satisfied with bank services or due to their financial problems are likely to churn.

There are any factors influence the reasons for a customer to Churn. It may be the fact that there's a new competitor in the market offering better prices or maybe the service they are getting has not been up to the mark, or customers financial problems so on and so forth.

My approach in this case study is to find out the factors affecting churn by doing EDA on each feature comparing it to target and take the relevant features and build a choose best model for prediction.

Approach:

Churn Prediction is Binary Classification Problem.

Data Modelling and Model Selection is done in Four Steps:

- Data-Preprocessing
- Data Visualization
- Feature Engineering
- Building Machine Learning Model

Data-preprocessing-

1)Dusing data analysis we found that there was **no missing value in data**

2) we have 5 object type elements - **'ID' is unique and irrelevant for analysis** so we are not going to use it for analysis.

3)Categorical values like Gender Product_Holdings, Credit_Category, Income effects the churn rate - which we are going to analyze.

- **we can encode this categorical data to numerical values for model building** 4)

the Is_Churn column is int64 making my target binary feature

Data Visualization:

It shows that distribution that almost 23% of our total training data members are churned (1) due to some factors lets see the analysis to find the reason.

From this we can say that data is **imbalanced and need to up-sampling before creating model**

- Derive some relevant insights out of the given data using different approaches (Such as using Seaborn/Matplotlib.)
- Detecting Outliers
- Relevant Insights
- For Numerical columns plotting boxplot to Detect outlier in the dataset.
- For Categorical columns plotting Barplot to detect which Column is more effecting the Target Column (Is_Churn).
- Descriptive Analysis of Dataset

Feature Engineering:

- In the DataSet, the Classes are Imbalance (it means the Class '0' is Majority Class and Class '1' is Minority Class). So to Deal with Imbalance Classes Problem the Sampling has to be done. To balance these Imbalance Classes used Resampling (Over Sampling) to match the Count of Majority Class. After Resampling the Majority and Minority Classes are approximately equal.

- In the dataset there are two types of Columns:

Categorical Columns

- For Categorical Columns, applied LabelEncoder for Two Class Columns and for more than two applied OrdinalEncoder (Here OrdinalEncoder because the Categorical Columns has the data in a Order Form Like Income Column).

Numerical Columns

- After Encoding Categorical Columns, check for correlation using heatmap and removing the columns which are highly Correlated.
- And the Balance Column Contains Outliers. But since we are using random Forest classifier it does not matter.

Model Selection:

- Build Model Using Decision Tree Classifier, Random Forest Classifier, XGB Classifier individually
- Performed hyperparameter tuning on each of this classifier
- Performed Model Selection according to the Model which perform best Prediction Score and with Best F1 (Macro) Evaluation Metrics.
- And for Model Selection, Selecting the Model with High Prediction Score on Test Data which is Random Forest Classifier have Prediction Score of 0.7172

	precision	recall	f1-score	support
0.0	0.80	0.84	0.82	1030
1.0	0.35	0.30	0.32	300
accuracy			0.72	1330
macro avg	0.58	0.57	0.57	1330
weighted avg	0.70	0.72	0.71	1330

and f1-score Evalution matrix.

Conclusion:

For predicting the Churn value after 6 months we need time series data which is not into our dataset but Random Forest Classifier gives us predict_proba() function to predict the probabilities for churn for upcoming future, So that we can take action to retain the customer.