Bank Churn Prediction

Introduction

Customer retention is crucial for banks as acquiring new customer is significantly more expensive than retaining existing ones. Customer churn, which refers to customer leaving the bank for competitors, can have substantial financial implications.

Objective

The primary goal of the project is to develop a machine learning model to predict customer churn. By identifying potential churners, banks can implement strategies to retain these customers and reduce churn rate.

Problem Statement

Customer churn in the banking industry refers to the loss of clients or account holders. Predicting churn helps in identifying customer who are likely to leave & try to retain them.

This project focuses on analyzing customer data to predict churn using various machine learning technique. This including data collection, preprocessing, model training, evaluation, and result.

Data Collection and Preprocessing

The data used in this project includes customer demographics, account information , credit score etc. The dataset consists of 10,000 customer record with 12 features each. The target variable is binary indicator of weather the customer has churned(1) or not(0).Key features include customer age, tenure, Balance, number of products and credit score.

The Following Libraries were imported

```
import pandas as pd     (for data manipulation and analysis)
import numpy as np      (For performing Mathematical and Numerical Operation)
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.model_selection import train_test_split, cross_val_score
from sklearn.linear_model import LinearRegression
from sklearn.preprocessing import StandardScaler, LabelEncoder
from sklearn.ensemble import RandomForestClassifier, AdaBoostClassifier
from sklearn.svm import SVC
from sklearn.metrics import classification_report,
confusion_matrix,accuracy_score,precision_score,recall_score,accuracy_score,ro
c_curve
from xgboost import XGBClassifier
```

Data Cleaning

Data cleaning involved Handling missing values. Identifying and handling outliers I detected outliers using a boxplot and visualized them with ‘plt.subplot’ for a clear comparison across different features.

Feature Engineering

Some of columns of the dataset where in categorical form, as we know that our model won’t be able to understand categorical variable, so I needed to convert it in numerical form.

Next there is imbalance data, as the data was imbalance, because the number of customer who are likely to churn were lesser compared to the customer who are still using the services of bank, so I used smote technique to balance the data.

Exploratory Data Analysis (EDA)

EDA help me to understand data better. In EDA I used different visualization technique like KDE plot, HIST plot, with the help of that I were able to discover the patterns, trend of the data.

Feature Selection

As I know that to train an optimal model, all I needed to make sure that I use only those features which were essential. As I have have some features which were not important to train the model could capture unimportant patterns to might learn from noise. I wanted only important features so I carried out feature selection technique.

There are several methods of feature selection. In this case I used filter method(Before model training) & Wrapper method (after model training).

Model Selection & Training

For model building I tried out various algorithms like Logistic Regression, random forest, Ada boost, XG Boost. In model building and evolution I initially explored logistic regression for its simplicity, however due to datasets complexity and nonlinear relationships, then I moved to ada boost, random forest and XG Boost models. Out of them Ada boost and XG Boost gives best accuracy (0.86). Hyper parameter were tuned using grid search. For training process the dataset was split into training and testing sets.

```
              precision    recall  f1-score   support

           0       0.88      0.96      0.92      1991
           1       0.75      0.47      0.58       509

    accuracy                           0.86      2500
   macro avg       0.81      0.72      0.75      2500
weighted avg       0.85      0.86      0.85      2500
```

Model Evaluation

As I aware to classification task then I focus on Precision and recall. Because its bank churn, I focus on recall. The recall is focus on false negative, suppose if certain person suppose to about leave then low recall would be detected as negative person means person could not leave the bank. So because of these is lower number of false positive, so we can tap right kind of person and provide him offers so that he can’t leave the bank.

Conclusion

Developed a machine learning model for bank churn prediction. Ada boost identified as the optimal model, offering customer retention strategies.

This project has great impact for supporting bank in proactively reducing churn rate and optimizing resources allocation for customer retaintion.

.