Predict the Churning Customer

Abstract

The goal of this project was to use binary classification models to predict churning customers in the bank to know who is going to be churned so they can proactively provide them with better services and influence customers' decisions to use the bank services and improve the performance. I worked with data provided in from Kaggle, did exploratory data analysis using Python and Tableau for better visualization and feature engineering with categorical features, fixed imbalance in the dataset and future selection, exploring models such as logistic regression, KNN and random forest and perform GridSearchCV to choose hyperparameters, to sum up, random forest model was the best model to achieve promising results for this binary classification problem.

Design

This project originates from bank data presents a binary class of attrition flag that is **existing customer** or **attired customer**. Classifying attrition accurately via machine learning models would enable the bank to improve their performance.

Data

The dataset contains 10128 customers with 21 features, 6 are nominal and ordinal categorical feature, such as: Attrition Flag our target variable, Education, Marital Status, Card Category exc.. Also, 15 numerical features highlight Age, Credit Limit, Total Revolving Balance, Average Card Utilization Ratio exc.., To answer the question I selected some of relevant categorical and numerical features that may affect the prediction.

Algorithms

Feature Engineering

1. Encoding categorical features to ordinal and dummy variables.
2. Oversample negative class to balance the dataset.
3. Identify and drop highly correlated features using correlation function.
4. Used SelectKBest and chi2 to calculate relevance scores of each feature to the target variable.

Models

Logistic regression, k-nearest neighbors, and random forest classifiers and perform GridSearchCV to choose hyperparameters were used before settling on Random forest model as it’s the strongest model performance.

Model Evaluation and Selection

Working in split data that is encoded correctly, balanced, using the most relevant futures selection to predict and the official metrics was accuracy and F1 score to focus on the negative class/attired customers for correct prediction.

Final {random forest} scores:

	precision	recall	f1-score	support
0	0.88	0.84	0.86	488
1	0.97	0.98	0.97	2551
accuracy			0.96	3039
macro avg	0.93	0.91	0.92	3039
weighted avg	0.96	0.96	0.96	3039

	Positive	Negative
Positive	409	79
Negative	55	2496

Tools

- Numpy and Pandas for data manipulation
- Scikit-learn for modeling, encoding and future selection and metrics.
- Matplotlib and Seaborn for plotting
- imblearn for oversampling.
- Tableau for interactive visualizations

Samaher turki almalki