

Abstract

The goal of this project is to build a classification machine learning model to predict a bank's customer churn. The term 'churn' in this analysis is to see if customers are likely to leave from a bank. The model will give banking institutions an understanding of their customers traits and behaviors allowing them to take measures to improve customer retention or gain new customers.

Design

Banking institutions often spend large amounts of budget to grow customers tenure because the cost of gaining a new customer is often more expensive. Therefore, they implement customer retention strategies to increase market share to increase revenue. The need for new customers is as important as keeping existing customers. The data science solutions implemented in this machine learning classification model will enhance the banks customer base.

Data

This bank customers churn dataset can be found at [Kaggle Datasets](#). It contains customers and their characteristics as well as if they churned. Each observation represents a unique customer and information such as age, gender, name, location, tenure, balance and many more. To access and view a detailed description of the dataset, click [here](#).

Algorithms & Models

Feature Engineering

- Converting categorical features to binary dummy variables
- Creating interaction terms

Measuring Metric:

F2_Beta was used because it puts more attention on minimizing false negatives than minimizing false positives.

Model Testing:

The data was split using stratified train/test with 10 K-Folds.

The Algorithms used for this analysis include:

- XGB
- AdaBoost
- RandomForest
- ExtraTrees
- Bagging
- DecisionTree
- LogisticRegressionCV
- KNeighbors
- SVC
- Bernoulli
- Gaussian

The top 5 model results:

Machine Learning Algorithm	F_Beta 2 Score
XGB	52.09155
Decision Tree	49.85373
Gradient Boosting	49.75858
Random Forest	49.01696
AdaBoost	48.94638

Tools

The following tools were used in this project:

1. Python & Pandas to:
 - Clean & Explore
 - Feature Engineering
2. SKLearn to implement various classification models
3. Matplotlib and Seaborn to visualize the data and model outputs

Communication

The findings and slide deck accompanying this project's presentation are accessible in this GitHub repository.