Analysis of Churn Rates at ABC BANK using Machine Learning Algorithms

Martin Ha and Ryan Nguyen

COMP 562

MAY 2023

Abstract

Within this paper we aim to look at customer retention rates. Specifically we will look at churn rates - the percentage of customers that stop using a product or service. We need to understand customer churn risk due to how crucial it is within many industries. The industry we aim to look at is the banking industry. Due to the high cost of customer acquisition, it is important to focus on customer retention. This study will look at different categorical variables such as credit score, country, age, gender, balance, etc. in order to predict the likelihood of a customer churning. Often competitors will offer short term bonuses which can lead to churning. This further emphasizes the need to understand which customers are at the highest risk of churning to retain customers. Additionally, unlike other subscription based industries where each customer may be worth relatively the same (I.E streaming services), the importance of a customer's balance will impact what customers a bank may allocate more of its resources towards.

Introduction

Due to different business metrics and how they affect the banking industry, customer retention is critical. It is often understood that retaining a customer is much cheaper than acquiring a new customer. This study references that customer acquisition on average 5 times the cost of customer retention.[1] Additionally, customer retention is vital to a bank's reputation, which in turn generates more customers. In order to consider the profitability of a bank customer, their churn rates along with variables such as tenure need to be considered.

Within this study we use a logistic regression model due to its efficiency in training and how well it works on datasets that are linearly separable. Additionally, it can handle outliers, is easily interpretable, and its flexibility. Along with a logistical regression model we will use a gradient boosting model. Our decision to use this model was based on how a gradient boosting model can help determine the important features within a dataset along with its scalability to larger datasets.

Related Work

Generally bank retention and other forms of retention across multiple subscription based companies use the term “churn”. There have been previously been studies related to this using an artificial neural network. Different categorical values are used such as customer age and lifetime in a genetic algorithm. A study by researched bank churn used, Logistic Regression, C 5.0, CHAID, ANN, XG-BOOST, and Decision Tree techniques within their algorithm to understand bank churn.[2] However discussed how “there is not yet a coherent conclusion on the application effect of these models”.[2] A different study reached a 90% accuracy rate using an ensemble learning algorithm. [3]

Methods

To provide an accurate model and dataset our methodology is displayed below.

1. Data Acquisition/Preparation
    a. Acquisition: Our model and research will be done on data from ABC BANK. We acquired the dataset from kaggle. [4]
    b. Deleting Missing Data: Due to the scale of the dataset it was inevitable that there would be missing data that we would have to drop. Among the 10000 columns that were presented there was roughly 9600 points of data that we acquired.
    c. Data Validity: On our data, we create boundaries for our data for impossible values such as credit scores above and below 850 and 300 respectively.
    d. Unnecessary Data: Columns that didn’t have value to our research were dropped such as a client’s id. Additionally, we dropped categorical
    e. Variables Used:
        i. Credit Score
        ii. Country
        iii. Gender
        iv. Age
        v. Tenure

vi. Balance

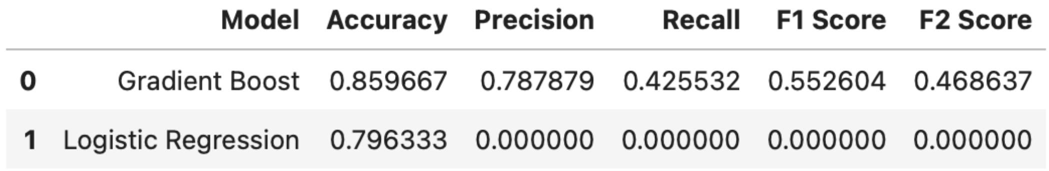
vii. Credit Card with Bank

viii. Estimated Salary

2. Models Used
    a. Logistic Regression:
    b. Gradient Boosting:
3. Software Used:
    a. Scikit-Learn

Results

By using the dataset we sourced from kaggle we were able to use the gradient boosting algorithm and logistical regression to predict the rate of churning. After cleaning our data set by getting rid of null values and outliers, our results are displayed below.



	Model	Accuracy	Precision	Recall	F1 Score	F2 Score
0	Gradient Boost	0.859667	0.787879	0.425532	0.552604	0.468637
1	Logistic Regression	0.796333	0.000000	0.000000	0.000000	0.000000

As displayed above, we found that the gradient boosting function had ~7% more accuracy in predicting if someone was going to churn. However, our F1 and F2 scores depicted that results weren’t extremely strong with a respective .569 and .491 score.

We had difficulty when calculating the precision, recall, f1 score, and f2 score, which can be seen in the figure above. We believe the issue might have resulted because of a function issue, despite coming to this conclusion, we found it very difficult to locate the issue as the function worked for gradient boost. After trying to debug the function as well as looking at surrounding code to see if there was an input error, we decided to rely on a built-in function to calculate these missing values. The built-in function gave us the same accuracy value as presented above, which gave us some insight on the obstacle at hand since we know that the function is producing the right accuracy value, so now we just have to figure out while the other values are missing.

This study used different learning algorithms to develop a model for predicting customer churn. In the future, we aim to focus on different variables and enhance those features to develop a more comprehensive model. Specifically, we want to delve deeper into age as it was one of the highest variables associated with churning.

References

1. Analysis and prediction of Bank User Churn based on ensemble learning ... (n.d.-a). https://ieeexplore.ieee.org/abstract/document/9362520/
2. Naturalspublishing.com. (n.d.-b). https://www.naturalspublishing.com/files/published/yt9r868jnrv116.pdf
3. Pfeifer, P. E. (2005, January 1). *The optimal ratio of acquisition and Retention Costs - Journal of targeting, Measurement and analysis for marketing*. SpringerLink. https://link.springer.com/article/10.1057/palgrave.jt.5740142
4. Topre, G. (2022, August 30). *Bank Customer Churn Dataset*. Kaggle. https://www.kaggle.com/datasets/gauravtopre/bank-customer-churn-dataset