

Predictive Analytics Project Report

Grant Gonnerman, Ryan Greiner, Omar Vargas



Spring 2023

Grand View University

Abstract:

Customer attrition has always been a problem for all types of businesses. Banks and credit card companies specifically have an increased desire to keep their customers loyal. With today’s data collection and modeling strategies it is possible to help any business, including banks and credit card companies, to accurately identify customers who are likely to leave. This allows the business to offer incentives to get these customers to stay loyal.

Table of Contents

[Introduction 2](#_Toc133160393)

[Data Description 2](#_Toc133160394)

[Initial EDA 3](#_Toc133160395)

[Modeling 5](#_Toc133160396)

[Feature Engineering 6](#_Toc133160397)

[Feature Selection 6](#_Toc133160398)

[Hyper-Parameter Tuning 6](#_Toc133160399)

[Ensemble Model 8](#_Toc133160400)

[Conclusion 8](#_Toc133160401)

[Implications and Further Study 8](#_Toc133160402)

# Introduction

Customer churn is defined as a customer who has stopped using a company’s product or service. In the scope of this project a churned customer would be one that has cancelled their credit card. The goal of this project is to build a model that can correctly identify and flag customers about to churn so the company can make efforts to retain their loyalty.

# Data Description

This dataset was found on Kaggle (<https://www.kaggle.com/datasets/thedevastator/predicting-credit-card-customer-attrition-with-m>) but was originally collected and uploaded to Zenodo in December 2020. The dataset contains 10,126 rows and 23 columns with no missing values. Each row represents a customer, and each column represents customer information collected within their credit card portfolio.

|  |  |  |
| --- | --- | --- |
| Column name | Type | Description |
| CLIENTNUM | Numerical | Unique identifier for each customer. |
| Attrition\_Flag | Categorical | Flag indicating whether or not the customer has churned out. |
| Customer\_Age | Numerical | Age of customer. |
| Gender | Categorical | Gender of customer. |
| Dependent\_count | Numerical | Number of dependents that customer has. |
| Education\_Level | Categorical | Education level of customer. |
| Marital\_Status | Categorical | Marital status of customer. |
| Income\_Category | Categorical | Income category of customer. |
| Card\_Category | Categorical | Type of card held by customer. |
| Months\_on\_book | Numerical | How long customer has been on the books. |
| Total\_Relationship\_Count | Numerical | Total number of relationships customer has with the credit card provider. |
| Months\_Inactive\_12\_mon | Numerical | Number of months customer has been inactive in the last twelve months. |
| Contacts\_Count\_12\_mon | Numerical | Number of contacts customer has had in the last twelve months. |
| Credit\_Limit | Numerical | Credit limit of customer. |
| Total\_Revolving\_Bal | Numerical | Total revolving balance of customer. |
| Avg\_Open\_To\_Buy | Numerical | Average open to buy ratio of customer. |
| Total\_Amt\_Chng\_Q4\_Q1 | Numerical | Total amount changed from quarter 4 to quarter 1. |
| Total\_Trans\_Amt | Numerical | Total transaction amount. |
| Total\_Trans\_Ct | Numerical | Total transaction count. |
| Total\_Ct\_Chng\_Q4\_Q1 | Numerical | Total count changed from quarter 4 to quarter 1. |
| Avg\_Utilization\_Ratio | Numerical | Average utilization ratio of customer. |

# Initial EDA

To get a better understanding of our dataset we started by making a few count plots, histograms, and a heat map so see if we could find any noteworthy correlations or trends in the data that could benefit us in the model building process.

Chart, bar chart

Description automatically generatedFigure 1

This data set observes about a 19% attrition rate. That means 19% of customers represented in the data set left the credit card company. This could be considered an imbalanced data set. We will need to keep in mind the recall score of our models.

Chart, bar chart

Description automatically generated

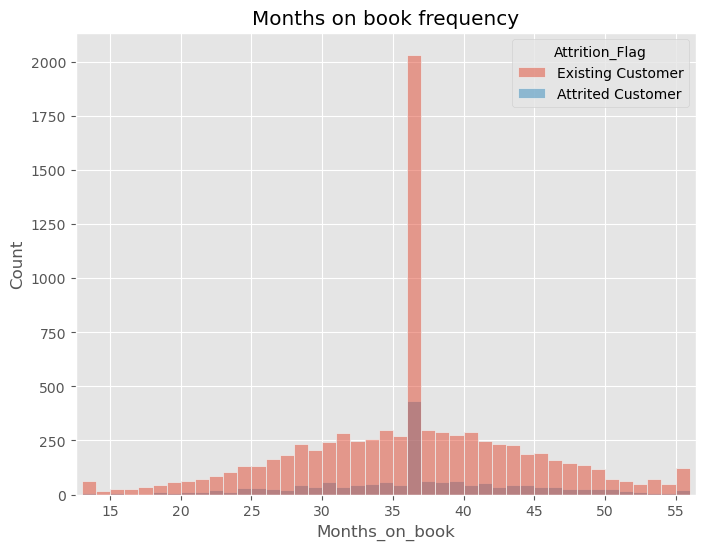
Figure 2

From the count plot we can see a majority of the customers’ highest level of education is graduate followed by High School.

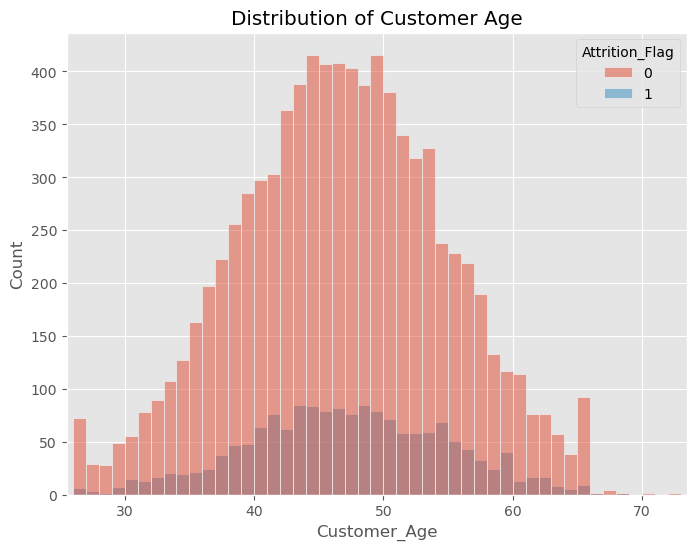
Chart, bar chart

Description automatically generatedFigure 3

The count plot to the left shows that a large majority of customers earn less than $40k per year. We can also see a decent chunk of the customers aren’t reporting an income.

Figure 4

From the above plot we can see that most of the customers have been on books for at least 37 months. The 37 months on book is an extreme outlier being it almost half the customers. Aside from this outlier the distribution is normal.

Figure 5

In the histogram to the left we can see the distribution of customer ages. This chart shows that the company’s customer base is normally distributed.

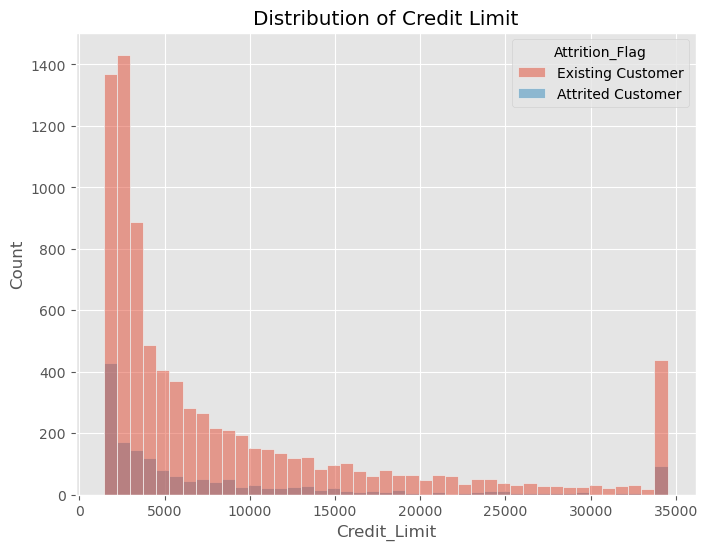
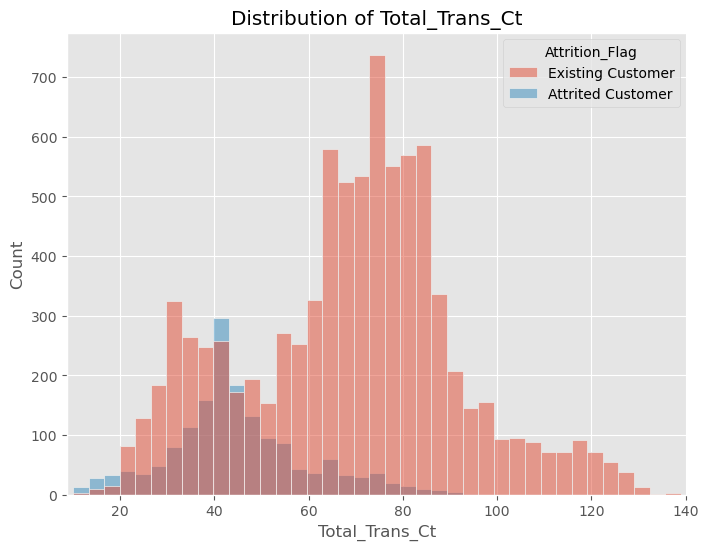


Figure 6

Here we see the distribution of the customer bases credit limit. The distribution is right skewed with a majority of customers having limits below 5,000. There is an interesting group sticking out from the trend with a limit around 34,000

Figure 7

In the count plot to the left we can see an interesting difference in distributions between existing and attrited customers. Existing customers follow a more normal distribution while attrited customers are distributed around 40.

Chart

Description automatically generated

Figure 8

In this heatmap, we highlight the variables that have a negative/positive correlation. We see that there are a number of correlated variables. We could engineer a few features based on these interactions.

# Modeling

The goal of this project is to build a model that can predict customer churn as accurately as possible while minimizing the false negative case. Using customer credit card portfolio information, we built a variety of models to ensure we have the optimal learners to compile into an ensemble. This section will outline our approach to reaching our final model with the best performance.

## Feature Engineering

After our initial EDA we brainstormed some possible features to help boost the performance of our models. We made interaction terms based on variables which had high correlations along with flags based off our visualizations. Those features are described in the table below.

|  |  |  |
| --- | --- | --- |
| Column name | Type | Description |
| Interaction\_1 | Numerical | Customer\_Age \* Months\_on\_book |
| Interaction\_2 | Numerical | Total\_Revolving\_Bal \* Avg\_Utilization\_Ratio |
| Interaction\_3 | Numerical | Total\_Trans\_Amt \* Total\_Trans\_Ct |
| Interaction\_4 | Numerical | Avg\_Utilization\_Ratio \* Avg\_Open\_To\_Buy |
| Months\_on\_book\_36\_flag | Categorical | Flag for Months\_on\_book = 36 |
| Credit\_Limit\_max\_flag | Categorical | Flag for Credit\_Limit = 34516 |
| Total\_Revolving\_Bal\_0\_flag | Categorical | Flag for Total\_Revolving\_Bal = 0 |
| Total\_Revolving\_Bal\_max\_flag | Categorical | Flag for Total\_Revolving\_Bal = 2517 |
| Total\_Trans\_Ct\_under\_60 | Categorical | Flag for Total\_Trans\_Ct >= 60 |

## Feature Selection

In order to decide which features were most important in detecting churn we used RFECV with 5 folds and f1 score as the evaluation metric. We ran through this sequence 5 times and kept track of the percentage of time each feature was selected in each model and selected those used 90% of the time or higher. The following table outlines the number of optimal features for each model and their accuracy score.

|  |  |  |
| --- | --- | --- |
| Model | Number of features | Accuracy score |
| Random Forest | 12 | .9628 |
| Ada Boost | 20 | .9701 |
| Gradient Boosting | 15 | .9707 |
| Hist Gradient Boost | 14 | .9721 |
| XG Boost | 13 | .9718 |
| Light GBM | 22 | .9706 |
| Cat Boost | 25 | .9713 |

## Hyper-Parameter Tuning

To find the optimal hyper-parameters for our data we used the Optuna framework to maximize the f1 score over 50 trials for each of the following models. The optimal hyper-parameter combinations for each model are listed below.

|  |  |
| --- | --- |
| Model | Hyper-Parameters |
| Random Forest | n\_estimators = 974  max\_depth = 10  min\_samples\_split = 5  min\_samples\_leaf = 2 |
| Ada Boost | learning\_rate = 0.08  n\_estimators = 438  max\_depth = 3 |
| Gradient Boosting | n\_estimators = 958  max\_depth = 3  min\_samples\_split = 8  min\_samples\_leaf = 6  learning\_rate = 0.0547 |
| Hist Gradient Boost | max\_iter = 520  learning\_rate = 0.1  l2\_regularization = 0.1  min\_samples\_leaf = 7  max\_leaf\_nodes = 103  max\_depth = 3  max\_bins = 205 |
| XG Boost | learning\_rate = 0.02  n\_estimators = 622  max\_depth = 6  min\_child\_weight = 1  gamma = 0.003011  alpha = 3.305498  colsample\_bytree = 0.51111  subsample = 0.433741 |
| Light GBM | n\_estimators = 966  learning\_rate = 0.02  num\_leaves = 313  min\_data\_in\_leaf = 37  min\_child\_weight = 0.07766  max\_depth = 9  bagging\_fraction = 0.65689  feature\_fraction = 0.52445  lambda\_l1 = 0.9316  lambda\_l2 = 1.0916 |
| Cat Boost | iterations = 699  learning\_rate = 0.1  min\_data\_in\_leaf = 154  depth = 7  random\_strength = 0.76723  bagging\_temperature = 0.34078  border\_count = 79  l2\_leaf\_reg = 64 |

## Ensemble Model

After identifying the optimal features and hyper-parameters for our dataset we picked out our top performing models by accuracy score to ensemble, making sure there was an odd number of models so there are no ties in predictions. The following table describes the top five models and their associated accuracy scores.

|  |  |
| --- | --- |
| Model | Accuracy score |
| Ada Boost | .9744 |
| Gradient Boosting | .9736 |
| Hist Gradient Boosting | .9734 |
| Light GBM | .9749 |
| Cat Boost | .9763 |
| Ensemble model | .9827 |

# Conclusion

In this project we were able to identify churned credit card customers. Through the use of exploratory data analysis, we familiarized ourselves with the data and were able to engineer new features that helped the performance of our models. We also used many different model optimization strategies such as hyper-parameter tuning, feature selection, and ensemble learning to reach achieve our final model with an accuracy of .98 and a recall score of .95 showing high level ability to predict the positive case with very few false negatives.

# Implications and Further Study

With the high level of accuracy our model was able to produce, we believe this model, or ones very similar to it, would be very beneficial to credit card companies or companies where customer attrition is a main priority. Using the results from our model, companies would be able to flag customers that may not be fully satisfied with their services. Companies could then use that feedback to improve the customer’s experience and result in a better product or service for their customers as a whole.

An area of further study we would be interested in looking at is what causes customers to churn from the company. Our dataset focuses on customers’ personal data but does not have any information on credit card features customers could be dissatisfied with. If we had this type of data, we could possibly provide more information to the company on what features and services lead to higher customer satisfaction.