Statistical Classification Applied to Credit Card Consumers

Marcelo Amaral

July 5, 2023

Case study for the Statistical Modeling course submitted as the A2 coursework in the 5th semester of Data Science and Artificial Intelligence.
Professor: Luiz Max

Abstract

This study examines statistical classification methods applied to credit card consumers to understand factors that influence customer attrition. With increasing competition in the financial services market, retaining customers has become a priority for any organization that provides credit based services. The aim is to provide credit for as many active customers as possible, while avoiding the ones more likely to churn. This paper aims to identify the most significant predictors of customer attrition, developing a simple model and focusing on interpretability. The thought process is developed and demonstrated within this case study based on data from over 10000 consumers who hired a credit service.

Keywords: multi-level model, statistical modelling, classification, credit card, consumers, churn

1 Introduction

Credit cards have become an indispensable tool in modern commerce, shaping the financial behavior of consumers around the globe. As of 2022, over 2 billion credit card transactions took place worldwide on a daily basis. The use of credit cards affords consumers with the flexibility of instantaneous transactions without immediate outflow of cash, leading to profound impacts on consumer spending habits. However, the conveniences afforded by credit cards come with their complexities and risks, both for the cardholders and issuing institutions.

At the heart of these complexities are the individual behaviors and decisions of the millions of consumers who utilize credit cards. Each cardholder interacts with their credit card in unique ways, influenced by a multitude of factors, from demographic characteristics to financial capacity, personal preferences, and even macroeconomic trends.

In the rapidly evolving landscape of the financial services sector, understanding consumer behavior has become essential. Companies are now utilizing advanced statistical models to predict consumer behavior, thereby informing business decisions and strategic planning. One area of considerable interest is the prediction of credit card customer attrition. Understanding its predictors is critical to implementing effective retention strategies.

In this context, the analyzed dataset is comprised of information on 10.000 consumers found on *Analyticca*, a EdTech platform for Data Science, Analytics, Machine Learning and AI. There are 20 variables and the goal is to predict a possible future class for a customer, given two existing outcomes:

- Existing Costumer: if the account is open.
- Attrited Costumer: if the account is closed.

Several factors can influence the likelihood of a customer leaving the company, such as health related emergencies and . The demographic information can provide a snapshot of the customer's life stage and financial stability, and information regarding the card category and credit limit can reflect both the customer's financial behavior and the level of trust and commitment between the customer and the company.

Customer engagement is another crucial aspect that can be gauged through variables like the total relationship count, months inactive, and contact count. These variables indicate the breadth of the customer's relationship with the company.

Analyzing the customer's financial behavior can provide additional insights. Variables like the total revolving balance and change in transaction amount provide a comprehensive picture of the customer's financial habits.

Table 1: Dataset [1] Variables

Variable	**Description**	**Data Type**
CLIENTNUM	Client number. Unique identifier for the customer holding the account	Categorical
Attrition_Flag	Whether the customer is existing or attrited	Binary
Customer_Age	Customer's age in years	Integer
Gender	Sex of the account holder	Categorical
Dependent_count	Number of dependents (people relying on customer's financial support)	Integer
Education_Level	Educational Qualification of the account holder	Categorical
Marital_Status	Married, Single, Divorced, Unknown	Categorical
Income_Category	Annual Income Category of the account holder	Ordinal
Card_Category	Type of Card (Blue, Silver, Gold, Platinum)	Ordinal
Months_on_book	Duration in months of relationship with bank	Integer
Total_Relationship_Count	Total number of products held by the customer	Integer
Months_Inactive_12_mon	Number of months inactive in the last 12 months	Integer
Contacts_Count_12_mon	Number of contacts in the last 12 months	Integer
Credit_Limit	Credit limit on the credit card	Float
Total_Revolving_Bal	Total revolving balance on the credit card	Float
Avg_Open_To_Buy	Open to buy credit line (Average of last 12 months)	Float
Total_Amt_Chng_Q4_Q1	Change in Transaction Amount (Q4 over Q1)	Float
Total_Trans_Amt	Total Transaction Amount (Last 12 months)	Float
Total_Trans_Ct	Total Transaction Count (Last 12 months)	Integer
Total_Ct_Chng_Q4_Q1	Change in Transaction Count (Q4 over Q1)	Float
Avg_Utilization_Ratio	Average Card Utilization Ratio	Float

All these variables can reflect the customer's financial health, the dynamism of their financial behavior, their dependency on credit, and overall activity levels.

In summary, the present study aims to investigate these various factors to predict credit card customer attrition. Through the application of statistical classification methods, it aims to provide meaningful insights to explore predictive techniques in the financial services sector for statistical models.

This analysis will allow us to identify patterns in credit card consumption and attrition, potentially informing more effective strategies for customer retention and service improvement. Through this, we hope to contribute to a more robust understanding of consumer behavior in the realm of credit card usage, and to provide somewhat actionable insight regarding the subject.

2 Metodology

2.1 Data Cleansing

The dataset had some unwanted columns, representing data that was supposed to be removed before publicizing the dataset. Some of the categorical columns contained values named "Unknown", which should be interpreted as missing values instead of a separate class, so they were replaced.

Other than these aspects, the data was well structured and didn't require tempering with.

2.2 Categorical/Binary Models

The R programming language was utilized for model development. Binary models are used for mapping the input data to a probability, indicating whether the data point belongs to one class or the other. In this case, 0 denotes failure and 1 represents success, which means that the customer didn't churn.

The main motivation behind the initial models was to gain a better understanding of the data. The analysis relied on statistical measures such as standard deviation and the p-value obtained for each estimated parameter of the model.

The initial prediction accuracy demonstrated a notable performance, reaching approximately 90%. However, it should be noted that this accuracy is inconsistent and heavily influenced by the random split of the data during the model fitting process. To enhance the predictive performance further, the best model was selected based on K-fold cross-validation and a refined feature selection approach. As a result, a marginal yet consistent improvement of two percentage points in the prediction accuracy was observed.

2.3 Binomial Distribution for Multi-Level Model

2.4 Evaluation Metrics

When examining other metrics, it became clear that the increased accuracy came at the cost of a 35% increase in the AIC and BIC metrics. Ultimatelty, Less features imply less complex models, usually followed by high interpretability.

Other than the aforementioned accuracy metric, several metrics were considered for evaluating the models, namely: precision, recall, F1 score, Cross Entropy, Akaike Information Criterion (AIC), Bayesian information criterion (BIC) and the Area Under the ROC Curve (AUC-ROC).

Most of these appeared to be very good at first glance, although, upon further inspection, the influence of the distribution of the data become apparent. The prevalence of persisting customers, the ones without attrition, skewed the values of these metrics. With few failure cases to classify, the model had hidden deficiencies that were captured mostly by the AUC-ROC.

In this context, it is especially important to consider the precision metric of the predictions, seeing as wrong decisions are likely to lead to financial loss.

2.5 Choice of covariables

During the model fitting process, an issue with the dataset was encountered. One of the coefficients could not be estimated due to singularity, suggesting that a column was a linear combination of one or more other columns. Specifically, this column was the average open-to-buy credit line, which was found to be, for over 99% of the data, the total revolving balance subtracted from the total credit limit. Thus, it was removed from the model.

In order to filter out some of the less descriptive features of the dataset, a refinement step was undertaken by examining the initial logistic regression model. Any feature with a p-value less than 0.1 was inspected more closely. Some features such as income category were expected to have a significant influence in the target variable, but all of the inspected features were eventually deemed less descriptive and subsequently discarded.

3 Results

3.1 Exploratory Data Analysis

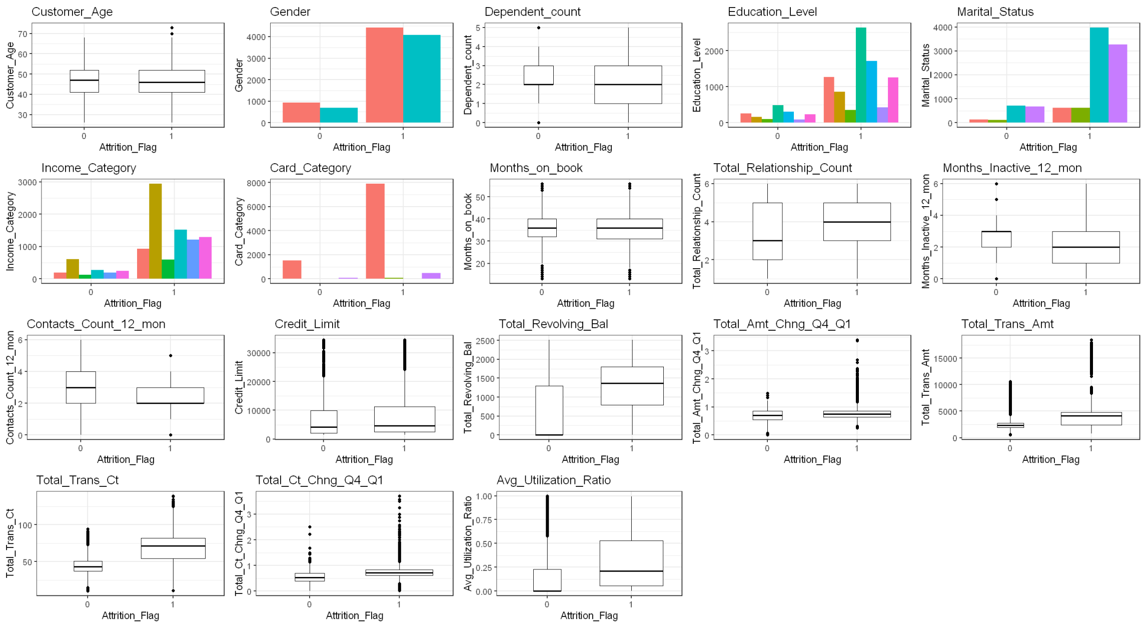

Figure 1: Overview of the dataset with all used columns plotted against the target. 0 means attrition and 1 means existing customer.

Because of the high dimensionality of the dataset, it wouldn't be feasible to visualize every pair of covariables, so the analysis was mostly focused on trying by intuition and plotting against the target column.

Many of these variables are similarly distributed when grouped by the target variable, which means the process wasn't as insightful as the actual modeling.

3.2 Predictions

After performing the standard data split between the training and test datasets, the model demonstrates consistent performance on the test data, often with a slight advantage compared to the training data.

Considering this information alongside an accuracy of over 90%, it can be reasonably inferred that the model does not suffer from underfitting (i.e., high bias) or overfitting (i.e., high variance) issues. Although it may not be the optimal model specifically tailored for this problem domain, it effectively serves its purpose as a classifier.

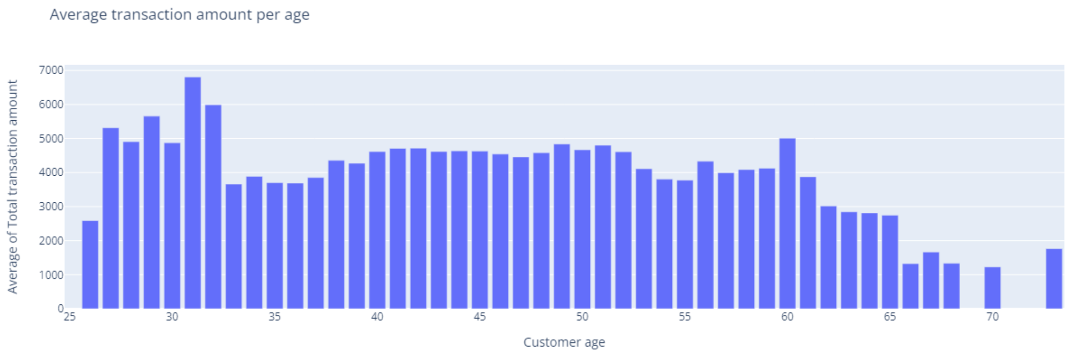

Figure 2: Exploratory bar chart.

4 Discussion

The best performing model had most of the computed metrics above 0.9, but it didn't do so well in the AUC-ROC metric and the information criteria, most likely due to the higher density of one of the classes and its lack of compatibility with the nature of the data. In spite of this, the results were satisfactory.

4.1 Limitations

There are two main approaches when it comes to improving the current results:

- Improving the logistic regression: a deeper analysis of the problem domain could be made by interviewing specialists and reading more case specific information; it is also possible to try a Bayesian logistic regression;
- Trying a different model: the initial proposal was to compare multi-level models to the simpler and more interpretable logistic regression.

These ideas weren't developed due to a lack of a critical resource, time.

4.2 Conclusion

This study proposes a process of credit card churn prediction focusing on the execution of well established methods and technologies. The thought process and development is exposed in an effort to make the process more understandable, without delving into what the best performing model would be for this context.

The developed model, similar to other models in the field, may not capture the intricacies of the data as comprehensively as more advanced techniques. However, the model's performance metrics remain reasonably competitive, albeit not at the cutting edge of research.

Previous studies [2] have indicated that advancements in the field of credit card churn prediction will likely arise from the utilization of more sophisticated models and larger datasets. In light of these insights, this research primarily aimed to enhance the didactic value by demonstrating the thought process and development stages rather than solely focusing on achieving the highest predictive performance.

References

[1] Sakshi Goyal. Credit card customers, Nov 2020. URL `https://www.kaggle.com/datasets/sakshigoyal7/credit-card-customers`.

[2] D. J. Hand and W. E. Henley. Statistical classification methods in consumer credit scoring: A review. *Journal of the Royal Statistical Society Series A: Statistics in Society*, 160(3):523–541, September 1997. `doi:10.1111/j.1467-985x.1997.00078.x`. URL `https://doi.org/10.1111/j.1467-985x.1997.00078.x`.