**Bank Attrition Analysis**

Sara Nedessa

School of Business, Government, and Economics

Seattle Pacific University

ISM 6353: Python and Machine Learning  
Dr. Andy Chen

November 22, 2022

**Abstract**

This paper seeks to identify the best models for predicting which bank customers are most likely to churn. In an increasingly competitive financial market, customer attrition is a big problem that banks face because it results in reduced revenue. Therefore, it is important for banks to predict customer attrition and take action early enough to retain these customers. This paper is based on a dataset from a single bank with more than 10,000 customers and 21 features describing customer demographics and the financial relations customers have with the bank. Exploratory data analysis was used to narrow down and select features to use in the predictive models. Among several predictive models, random forest and AdaBoost had the best predictive ability with an F-1 score of 98% and 97% respectively and an AUC of 92% and 91% respectively. Recommendations are given on how to increase customer retention.

**Keywords**: Customer attrition/churn, predictive data mining, ensemble models

**1.0 Business Insight**

Customer attrition or customer churn refers to the loss of customers. In this study, it refers to customers who closed all their accounts and left the bank. Customer retention is one of the most important factors for the growth and survival of banks in today’s competitive world (Chitra and Subashini, 2011). Increased competition in the financial arena makes it more difficult for banks to stand out and succeed.

Banks should seek to increase customer retention for several reasons. First, increased customer retention lowers the need to acquire new customers which can be 5 times more expensive than retaining customers (Van den Poel & Lariviere, 2004). Second, long term customers are less affected by competitive markets. Long term customers are easier to serve while generating increased profits and referrals for banks. Third, improving retention rates even marginally can create a substantial increase in profit.

To help with the problem of customer attrition, prediction models can be used to identify customers that are most likely to leave the bank. By gaining insight on previous churning customers’ behaviors, banks can create customized approaches to retain their customers.

The data mining goal of this paper is to find the best classification models for reliably predicting churning customers. The analysis will be based on variables like customer demographics as well as variables that describe the interactions and transactions customers have with the bank. The business goal is to increase customer retention through customized approaches.

**Stakeholders**

There are several stakeholders who are invested in how well a bank does.

Customers are impacted by the products and services provided by the bank as well as any loyalty or incentive programs that the bank creates.

Investors and shareholders care deeply about the success of a bank. The return on the money they have invested depends on the continued profitability of the bank.

Employees and managers are also invested in the success of the bank. Their jobs, bonuses and raises are dependent on the survival of the bank.

Additionally, partner companies and the community at large are also stakeholders because of how banks affect the financial market. All these stakeholders benefit from increased customer retention and the increased profitability of the bank.

**2.0 Data Understanding and Preparation**

The dataset used for this paper was obtained from Kaggle. It has 10,127 data points with 21 columns. The dataset initially had 23 columns, but based on a suggestion that came with the dataset, I removed the last 2 columns before I started working with it. Figure 2.1 shows all the features included in my analysis.

Graphical user interface, table

Description automatically generatedGraphical user interface, text, table

Description automatically generated

Graphical user interface, text, table

Description automatically generated

*Figure 2.1. Dataset display*

The features include client identification, customer attrition, demographics, and customer relationship with the bank. Customer demographics used in the dataset are age, gender, number of dependents, education, marital status, and income category. Customer financial relations with the bank are the following variables: card category, months on book, total relationship count, months inactive in past 12 months, contact counts in past 12 months, credit limit, total revolving balance, average open to buy, total amount change Q4 to Q1, total transaction amount, total transaction count, total count change Q4 to Q1, and average utilization ratio. Features associated with financial relations are explained in table 2.1.

Table 2.1. Feature descriptions

|  |  |
| --- | --- |
| Feature | Description |
| Card\_Category | The type of card a customer has |
| Months\_on\_book | Length of relationship with customer |
| Total\_Relationship\_Count | Number of relationships a customer has with the bank (includes bank accounts and credit cards) |
| Months\_Inactive\_12\_mon | Number of months a customer has been inactive in the past 12 months |
| Contacts\_Count\_12\_mon | Number of times the bank has contacted a customer over the last 12 months |
| Credit\_Limit | Credit limit of a customer |
| Total\_Revolving\_Bal | Portion of credit card spending that is unpaid at the end of a billing cycle |
| Avg\_Open\_To\_Buy | Difference between a customer’s credit limit and the current outstanding balance |
| Total\_Amt\_Chng\_Q4\_Q1 | Change in the total transaction amount of a customer in the fourth quarter compared to the first quarter |
| Total\_Trans\_Amt | Total cash value of all transactions a customer made in the past 12 months |
| Total\_Trans\_Ct | Number of total transactions a customer made in the past 12 months |
| Total\_Ct\_Chng\_Q4\_Q1 | Change in the number of transactions of a customer in the fourth quarter compared to the first quarter |
| Avg\_Utilization\_Ratio | The average percentage of total available credit that a customer is using. This is essentially a debt to credit ratio |

**Exploratory Data Analysis**

The dataset was clean and had no missing values. I created a dummy variable for the target feature, which is customer attrition. Existing customers were assigned 1, and attrited customers were assigned 0. The dataset had skewed proportions with only 16% of total bank customers leaving the bank. I started exploring the data with a heat map to find the features that correlated the most with attrition. This highlighted a few features like total relationship count, total revolving balance and total transaction count that showed a positive correlation with customers that stay with a bank.

I did some exploratory data analysis to understand the customer demographics. The majority of customers are between the ages of 40 and 55 for both existing and attrited customers (figure 2.1). There were slightly more female customers than male customers, with female customers representing 53% of the total.

Chart

Description automatically generated

*Figure 2.1 Age distribution of existing customers and attrited customers.*

In terms of education, 31% of customers are college graduates and 20% are high school graduates. In terms of income, 35% of customers are making less $40,000 annually, 28% are making between $40,000 and $60,000 and the percentage continues to get smaller for higher income. Almost all customers use blue card category, with less than 7% of customers having higher tier cards like silver, platinum, or gold.

For categorical features, I used pie charts and bar graphs to compare the rates of attrition by category. Figure 2.2 is an example of these charts for income category. All the categorical variables (age, marital status, education level, income category and card category) had no noticeable effect on customer attrition. Accordingly, I chose not to create dummy variables for them or use them for the predictive models.

Graphical user interface, chart

Description automatically generated Graphical user interface

Description automatically generatedGraphical user interface, chart, bar chart

Description automatically generated

*Figure 2.2 Income category and customer attrition*

For numerical features, I used boxplots for the comparisons. Figure 2.3 shows these boxplots for the 9 features that showed some difference between the attrition group and existing customers group. I selected the features where the median of attrited customers is outside either the first or third quartile line of existing customers.

Chart, box and whisker chart

Description automatically generated

Graphical user interface, chart, box and whisker chart

Description automatically generated

Chart, box and whisker chart

Description automatically generated

*Figure 2.3 Boxplot comparisons for numerical variables.*

**3.0 Predictive Models**

I selected the 9 features from the boxplots to use for the predictive models. In my analysis, I initially run a logistic regression model through statsmodels to confirm that the p-values of all selected features were below 0.05. Figure 3.1 has all the features that were used for the predictive models. Next, I used sklearn for all the models with a test group of 30%.

Graphical user interface, text, application, email

Description automatically generated

Figure 3.1. Summary table of a logistic regression model using nine features.

**Logistic Regression**: This model was set up with all default parameters.

**Decision Tree**: This model was run with various maximum depths of 5, 10, 15 and 20. The best predictive model was the one with a maximum depth of 10.

**Support Vector Machine**: This model was run with c values of 0.1, 1, 5, 10 and 50, and gamma values of 0.001, 0.05, and 0.5. The models showed a trend where higher values of c and lower values of gamma preformed the best. However, a gamma that is too small results in a model that can’t capture the shape of the data well and a c value that is too high creates very small margins (Pedregosa et al., 2011). Taking these factors into account, I selected a c value of 5 and a gamma value of 0.001.

**Naïve bayes**: This model was set up with default parameters.

I also tried ensemble models to see if these models could be strengthened. Ensemble models combine multiple models to obtain better results compared to the base models (Bolón-Canedo & Alonso-Betanzos, 2019).

**Random Forest**: This model was run with various numbers of estimators (10, 50, 100, 250 and 500). The best model had 100 trees and proved to be an improvement on the decision tree model.

**Bagging Classifier**: This model used SVM as a base model using gamma default values and tried different numbers of estimators (5, 10 and 20). Surprisingly, this model did worse than the regular SVM model. The AUCs couldn’t get higher than 0.51 for any number of estimators that were tested.

**AdaBoost Classifier**: This model was run with various numbers of estimators (5, 10, 20, 50 and 100). The best results came from 100 estimators.

**Stacking Classifier**: This model was run using random forest, AdaBoost, naïve bayes and logistic regression as base models and running them through another logistic regression model as the final estimator. I just wanted to see if the previous models could be improved upon. However, this resulted only in a marginal improvement compared to random forest and AdaBoost. It was not worth it to add another lay of complexity to the predictive models without improving their predictive ability.

The evaluation metrics for all of these preditive models have been summarize in table 4.1.

**4.0 Results and Evaluation**

The dataset is imbalanced because only 16% of the total number of customers are attritted customers. Imbalanced datasets create challenges when training the data. Predictive models dealing with imbalanced datasets are likely to have bias for the majority group and result in misclassification within the minority group (Johnson & Khoshgoftaar, 2019).

Additionally, accuracy score is dominated by the majority group. The accuracy score for an imbalanced dataset can be misleadingly high due to the large number of true positives in the data. Since the primary concern of this paper is churning customers (i.e. the minority group), an evaluation metrics that is very sensitive to the minority group is better suited.

Graphical user interface

Description automatically generated with medium confidence

F-1 score offers an improvement over accuracy by blending precision and recall (Johnson & Khoshgoftaar, 2019). Similarly, maximizing AUC provides better predictions when working with imbalanced datasets (Yuan et al., 2021). Table 4.1 compares F-1 scores, AUCs and accuracies of different models.

Table 4.1 Predictive models along with evaluation metrics.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Model** | **Logistic Regression** | **Decision Tree** | **Naïve Bayes** | **SVM** | **Random Forest** | **AdaBoost** | **Stacking Ensemble** |
| **F-1 Score** | 0.932 | 0.964 | 0.932 | 0.936 | 0.977 | 0.974 | 0.976 |
| **AUC** | 0.713 | 0.891 | 0.784 | 0.751 | 0.920 | 0.915 | 0.916 |
| **Accuracy** | 0.883 | 0.939 | 0.886 | 0.890 | 0.961 | 0.957 | 0.959 |

The results from the random forest model and AdaBoost model have the best F-1 score (97.7% and 97.4% respectively) and AUC (92.0% and 91.5%). I recommend using these two ensemble models as the best option for bank attrition analysis.

**5.0 Recommendations**

Predicting which customers are most likely to churn is only the first step in addressing the problem of customer attrition that banks face. Banks should use these models regularly so they can offer incentive programs to customers that are likely going to leave. I would recommend three options to increase customer retention.

The first recommendation is broad and targets all customers. Figure 3.1 identifies which features have a positive or negative relationship with attrition. Features like number of transactions, average utilization ratio, total revolving balance, and total relationship count have a positive correlation with existing customers. This means churning customers use fewer bank products and they make fewer transactions. Additionally, figure 5.1 obtained from the random forest model identifies total transaction amount as the most important feature.

Graphical user interface

Description automatically generated with medium confidence

*Figure 5.1 Features ranked by importance.*

The bank can be proactive about offering bank products and services to increase customers’ transactions. This includes credit card upgrades to customers who qualify, preapproved new credit card options with higher earnings or discounts based on customer credit as well as options for private loans, auto loans or mortgages based on customer relationship with the bank and demographics like age. This will help customers be aware of the range of services available to them.

This can be expanded to create relationship banking programs that rewards customers for their overall relationship with the bank. Instead of focusing on discounts or rewards for a specific card, the bank can also look at the total products and services a customer uses and give rewards based on that (Ivanauskiene & Auruskeviciene, 2009). For instance, the bank can have a higher tier service for loyal customers who use several bank products. This is particularly useful to build and strengthen long-term relationship with the bank.

The second recommendation is to partner with other companies to offer discounts and rewards to customers based on where they use their cards. The bank can partner with different merchants and offer these benefits on their websites and mobile apps. The bank can even go further if needed. For instance, Capital One offers its customers an online shopping tool that can be added as an extension to a browser and will automatically look for discounts and best prices (Weiss, 2022).

This recommendation works on two levels. First, it will encourage customers to use their cards more often. Second, it gives the bank more insight on what customers are interested in. The information gained from this dataset is limited in terms of understanding customer spending habits. Partnering with retail merchants would help the bank understand customer habits and provide more customized suggestions so customers can have the best experience.

The third recommendation is to ensure customer satisfaction. Number of contacts with the customer is the last 12 months shows a negative correlation with existing customers. This means customers who had more contact with the bank are more likely to churn. This is an opportunity for the bank to make sure that customer service is up to par and the needs of customers are being met whenever a customer calls the bank or vice versa. Management should follow up on conversations to make sure any issues are fully solved. This is supported by previous studies that show that employee attitudes impact customer churn and that customer dissatisfaction is often present before a customer churns (Lappeman et al., 2022).

**6.0 Limitations and Further Research**

The results from this paper are promising; however, it is important to note some limitations. First, these results are for a specific bank and further study is required to see if other banks or even different branches from the same bank follow similar trends and if random forest and AdaBoost are the best predicters for other banks.

It is also important to note that this analysis was done on an imbalanced dataset. Imbalanced datasets are more prone to bias and overfitting (Johnson & Khoshgoftaar, 2019; Onel et al.,2018). Future studies can try balancing the dataset before preforming predictive models.

Additionally, the results from the best predictive models show really high values for the evaluation metrics. This could be due to overfitting, so it is recommended do similar studies with more data from the bank.

**References**

1. (n.d.). BankChurners. Retrieved on Oct 11, 2022 from<https://www.kaggle.com/code/josh1337/bankchurners/data>.
2. Bolón-Canedo, V., & Alonso-Betanzos, A. (2019). Ensembles for feature selection: A review and future trends. *Information Fusion*, *52*, 1-12.
3. Chitra, K., & Subashini, B. (2011). Customer retention in banking sector using predictive data mining technique. In *ICIT 2011 The 5th International Conference on Information Technology*.
4. Diez-Olivan, A., Del Ser, J., Galar, D., & Sierra, B. (2019). Data fusion and machine learning for industrial prognosis: Trends and perspectives towards Industry 4.0. *Information Fusion*, *50*, 92-111.
5. Ivanauskiene, N., & Auruskeviciene, V. (2009). Loyalty programs challenges in retail banking industry. *Economics and Management*, (14), 407-412.
6. Johnson, J. M., & Khoshgoftaar, T. M. (2019). Survey on deep learning with class imbalance. *Journal of Big Data*, *6*(1), 1-54.
7. Lappeman, J., Franco, M., Warner, V., & Sierra-Rubia, L. (2022). What social media sentiment tells us about why customers churn. *Journal of Consumer Marketing*, (ahead-of-print).
8. Onel, M., Kieslich, C. A., Guzman, Y. A., & Pistikopoulos, E. N. (2018). Simultaneous fault detection and identification in continuous processes via nonlinear support vector machine based feature selection. In *Computer Aided Chemical Engineering* (Vol. 44, pp. 2077-2082). Elsevier.
9. Pedregosa et al. (2011). [Scikit-learn: Machine Learning in Python](http://jmlr.csail.mit.edu/papers/v12/pedregosa11a.html). JMLR 12, pp. 2825-2830.
10. Van den Poel, D., & Lariviere, B. (2004). Customer attrition analysis for financial services using proportional hazard models. *European journal of operational research*, *157*(1), 196-217.
11. Verbeke, W., Dejaeger, K., Martens, D., Hur, J., & Baesens, B. (2012). New insights into churn prediction in the telecommunication sector: A profit driven data mining approach. *European journal of operational research*, *218*(1), 211-229.
12. Weiss, R. J. (2022). Capital One Shopping Review: Is It Safe & Does It Work? The Ways to Wealth. <https://www.thewaystowealth.com/reviews/capital-one-shopping-review/>
13. Yuan, Z., Guo, Z., Xu, Y., Ying, Y., & Yang, T. (2021, July). Federated deep AUC maximization for hetergeneous data with a constant communication complexity. In *International Conference on Machine Learning* (pp. 12219-12229). PMLR.