**Bank Customer Churn Prediction**

**Team 5**

**Sai Karthik Alapati**

**Sai Venkata Krishna Ankem**

**Collin Flippo**

**Lucas Lindgren**

**Triveni Reddy Alugubelly**

**INDEX**

**1. Introduction 1**

**2. Executive Summary 2**

**3. Dataset Description 2-3**

**4. Exploratory Data Analysis 4-10**

**4.1 Customer Demographics Analysis**

**4.2 Customer Financial Behavior Analysis**

**5. Analysis 10-12**

**5.1 Data Preprocessing**

**5.2 Feature Selection**

**5.3 Model Training and Evaluation**

**6. Results/Conclusion 13-14**

**7. References 15**

**1. Introduction**

The problem we’ve chosen to examine is that of bank churn. Churn is defined as the rate of attrition of subscription to a given service. In our case, that ‘subscription’ is having an open bank account.

The reason we’ve chosen this problem is that churn is a common problem in banking, because (theoretically) all banks offer an identical service; undifferentiated products and services are far more likely to have to deal with churn. This does not mean the problem is unsolvable and can be a source of strength for banks offering superior services. By identifying the reasons for customer churn, banks can enhance their customers’ experience where it matters most to foster loyalty, stabilize their revenue, and gain a competitive edge.

We’re approaching this problem by utilizing a dataset comprised of 10,127 customer profiles, with 23 attributes (columns) we’ll use to examine their behavior. This includes obvious target data like demographics and financial status, but also potentially looked over, insightful data like credit utilization and revolving balance.

Our plan of attack is to first prepare the data for analysis by dropping any irrelevant columns (for example, CLIENTNUM). We’ll also remove rows with null values, as these could meaningfully distort our results. Finally, we’ll transform categorical variables to binary values by creating dummies, allowing for easier analysis.

To select the most significant features, or attributes, of the dataset, we’ll use SelectKBest with a chi-squared test. The models we’ll use include Random Forest and SVM, or Support Vector Machine. These are trained on the preprocessed dataset, and their performance will be evaluated using accuracy and F1-score metrics

**2. Executive Summary**

This analysis focused on predicting bank churn, where customers leave their bank services. By understanding why customers leave, banks can take steps to retain them. We analyzed a dataset of 10,127 bank customers and 23 attributes, identifying credit card activity as a key factor in churn. Using SelectKBest, we selected important features and trained Random Forest and SVM models to predict churn. Our Random Forest model achieved an accuracy of 91.41%, while the SVM model reached 88.15%. Further research is recommended to refine the models and explore other approaches to address churn.

**3. Dataset Description**

This dataset comprises of 10,127 banking customer profiles, each characterized by 23 distinctive attributes. These attributes include:

* **Client Number (CLIENTNUM)**: This could be a unique identifier for each customer.
* **Attrition Flag**: This might indicate whether a customer has churned or not.
* **Customer Age (Customer\_Age)**: The age of the customer.
* **Gender**: The gender of the customer.
* **Dependent count (Dependent\_count)**: The number of dependents the customer has.
* **Education Level (Education\_Level)**: The education level of the customer.
* **Marital Status (Marital\_Status)**: The marital status of the customer.
* **Income Category (Income\_Category)**: The income category of the customer.
* **Card Category (Card\_Category)**: The type of card the customer uses.
* **Months on book (Months\_on\_book)**: The length of time the customer has been with the bank.
* **Total Relationship Count (Total\_Relationship\_Count)**: The total number of products the customer has with the bank.
* **Months Inactive in the last 12 months (Months\_Inactive\_12\_mon)**: The number of months the customer was inactive in the last year.
* **Contacts Count in the last 12 months (Contacts\_Count\_12\_mon)**: The number of times the customer contacted the bank in the last year.
* **Credit Limit (Credit\_Limit)**: The credit limit of the customer.
* **Total Revolving Balance (Total\_Revolving\_Bal)**: The total revolving balance on the customer’s account.
* **Average Open To Buy (Avg\_Open\_To\_Buy)**: The average amount that the customer could spend using their credit card, without exceeding their credit limit.
* **Total Amount Change Q4 over Q1 (Total\_Amt\_Chng\_Q4\_Q1)**: The change in spending amount from Q1 to Q4.
* **Total Transaction Amount (Total\_Trans\_Amt)**: The total amount of transactions made by the customer.
* **Total Transaction Count (Total\_Trans\_Ct)**: The total number of transactions made by the customer.
* **Total Count Change Q4 over Q1 (Total\_Ct\_Chng\_Q4\_Q1)**: The change in transaction count from Q1 to Q4.
* **Average Utilization Ratio (Avg\_Utilization\_Ratio)**: The average ratio of the customer’s total revolving balance to their credit limit.

This data can be very useful in analyzing the customer churn rate for a bank. For instance, attributes like the Attrition Flag, Months Inactive in the last 12 months, and Contacts Count in the last 12 months could directly indicate customer churn. Other attributes like Customer Age, Gender, Dependent count, Education Level, Marital Status, Income Category, Card Category, Months on book, Total Relationship Count, Credit Limit, Total Revolving Balance, Average Open To Buy, Total Amount Change Q4 over Q1, Total Transaction Amount, Total Transaction Count, Total Count Change Q4 over Q1, and Average Utilization Ratio could provide insights into the factors influencing customer churn

**4. Exploratory Data Analysis.**

**4.1 Customer Demographics Analysis.**

**Proportion of Customers based on Gender.**

A blue and orange pie chart

Description automatically generated

In Figure 1, we observe the distribution of customers according to gender. It can be inferred that the proportions are nearly equal.

Fig 4.1.1

**Proportion of Churn vs not Churn Customers**

A pie chart with numbers and a blue circle

Description automatically generated

Fig 4.1.2

**Proportion of Different Marriage Statuses**

A pie chart with numbers and a number of people

Description automatically generated

Half of the bank's clientele are married, while nearly the entirety of the remaining half consists of single customers.

Fig 4.1.3

**Proportion of Different Income Levels**

A pie chart with numbers and a number of different colored circles

Description automatically generated

From fig 4 we can see that more than 25% of the people earn less than 40K. While nearly 30% of the people earn more than 120K annually which is more than the average income of the country.

Fig 4.1.4

**Proportion of Different Income Levels**

. A colorful circle with text on it

Description automatically generated

From fig 5 we can state that more than 70% of the customers have a formal education level. About 35% have a higher level of education.

Fig 4.1.5

**Distribution of Different Card Status Among the Present and Attrited Customers**

Fig 6 Shows the total distribution of card status among all the customers. We can see that about 93% of the customers are using card type called Blue.

Fig 7 Shows the distribution of card status among the present and the attrited cutomers based on the gender. We can see that most of the attrited customers are females and among females most of them are of category blue. So we can say that bank has to focus more on the female customers to retain them

A blue circle with a number of different colored circles

Description automatically generated with medium confidenceA graph with numbers and a bar

Description automatically generated

Fig 4.1.6 Fig 4.1.7

**4.2 Customer Financial Behavior Analysis**

**Distribution of Months the Customer Is Part of the Bank**

**A graph with a number of bars

Description automatically generated with medium confidence**

Fig 4.2.1

**Distribution of Months Inactive in the Last 12 months**

**A graph with numbers and a red bar

Description automatically generated with medium confidence**

Fig 4.2.2

**Distribution of Credit Limit**

A graph of a number of people

Description automatically generated with medium confidence

Fig 4.2.3

**Distribution of Total Transaction Amount**

**A graph with a red line

Description automatically generated**

Fig 4.2.4

**Distribution of Total Products held by the Customer**

**A red graph with numbers

Description automatically generated**

Fig 4.2.5

**5. Analysis**

**5.1 Data Preprocessing:**

* Irrelevant columns, such as customer IDs, were dropped from the dataset to focus solely on relevant features for the classification task. Categorical variables were one-hot encoded to represent them as numerical values suitable for model training.

**5.2 Feature Selection:**

* SelectKBest, a univariate feature selection method, was applied to rank the features based on their relevance to the target variable (Attrition\_Flag). The chi-squared test, which measures the dependence between variables, was chosen as the scoring function for SelectKBest. This test is suitable for categorical target variables and is effective in identifying feature-target associations.
* The top K features with the highest chi-squared scores were selected to be retained for model training. In this analysis, K was set to 5, indicating the selection of the top 5 most significant features.

A black text on a white background

Description automatically generated

Fig 5.2.1: Significant Features.

**5.3 Model Training and Evaluation:**

* The resampled training data, obtained after applying SMOTE for class imbalance correction, was used for both feature selection and subsequent model training.
* In this analysis, the RandomForestClassifier model was selected for classification, and the following hyperparameters were tuned:
  + **n\_estimators:** The number of trees in the random forest. Increasing this parameter can improve the model's performance, but it also increases computational cost.
  + **max\_depth:** The maximum depth of each tree in the forest. This parameter controls the depth of the trees and thus the complexity of the model.
  + **min\_samples\_split:** The minimum number of samples required to split an internal node. This parameter influences the tree's decision-making process by preventing splits that result in fewer samples in child nodes.
* A parameter grid is defined, specifying the hyperparameters to be tuned and the range of values to search over.In this analysis, the following parameter grid was defined:
  + **n\_estimators:** [50, 100, 150]
  + **max\_depth**: [None, 5, 10, 15]
  + **min\_samples\_split:** [2, 5, 10]
* GridSearchCV is applied to the RandomForestClassifier model with the specified parameter grid and evaluation metric (accuracy).The GridSearchCV algorithm iterates over all combinations of hyperparameters, trains the model on the training data, and evaluates its performance using cross-validation.After exhaustive search, the combination of hyperparameters that maximizes the cross-validation accuracy is identified. Cross-validation with 5 folds was employed to estimate the model's performance robustly and to evaluate the selected features' effectiveness across different subsets of the training data.

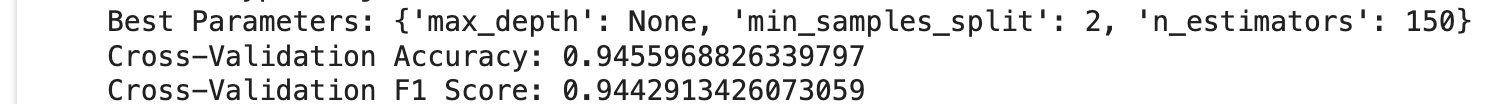


Fig 5.3.1: For Random Forest Model

* Evaluation metrics such as accuracy and F1 score were utilized to assess the model's performance on the test set, providing insights into its predictive capabilities.

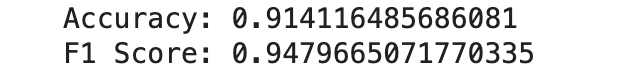


Fig 5.3.2: Random forest model Accuracy and F1 score

* The confusion matrix visualization serves as a valuable tool for assessing the performance of the Random Forest and SVM model and gaining insights into its predictive capabilities. By interpreting the matrix, stakeholders can make informed decisions to optimize business strategies and enhance customer satisfaction and retention.

A screenshot of a graph

Description automatically generated

Fig 5.3.3: Random Forest Model Confusion Matrix

SVM Model confusion Matrix

A screenshot of a computer

Description automatically generated Fig 5.3.4 :SVM Model confusion Matrix

**6. Results/Conclusion**

When first tackling the churn dataset, linear and logistic regression were both considered for prediction, however since much of the data was nonlinear and categorical, efforts were focused on Random Forest model using feature selection, and a Support Vector Model as discussed in class. Initially a random forest without feature selection was attempted but it seemed to overfit and was resource intensive. It was determined that feature selection could help keep an effective model while sizing down the amount of variables required for the calculations.

When analyzing the results of the Select K Best feature selection used on the data, the top 5 most significant variables were ranked as follows:

A screenshot of a computer program

Description automatically generated

Fig 6.1: Significant Features.

It is interesting to note that these features are all related to credit card activity rather than demographics of the particular card holder. This gives us some insight into where additional analysis could be done as it seems that activity on the card could be more important than the demographic bucket a user falls into.

The Random Forest model with feature selection attained an accuracy score of 97.94% and a f1 score of 97.90%, while the SVM model measured 88.15% accuracy and a f1 score of 88.18%. Both of these results are extremely high. While this seems promising at first, the data that the models were tested on had been automatically resampled to have a better balance of Churn to Not Churn entities. Without testing the models on raw churn data (without SMOTE processing) it is difficult to determine the true predictive accuracy of the models. After using unsampled test data the accuracy and f1 scores dropped to 91.41% and 94.79% respectively. These are quite strong scores and indicate a relatively robust model. It would be interesting to see how the model performs with brand new data as behaviors that are typically associated with churn may change over time.

While the “problem” of credit card churn may not have been solved, some major attributes that seem to affect churn have been identified and a strong model for classifying potential churn users has been developed. Now, additional analysis into credit card activity metrics/factors should be conducted to potentially build new types of models. Random Forest and SVM are only two examples of model types that can be used for the churn data, so investigating other non-linear or classification models would also be a next step. Since users are all individual humans at the end of the day, it may never be possible to perfectly predict customer churn. However if we can point the credit card company in the direction of those who are most likely to churn, efficiency is increased and targeted retention practices can be more effective.

It’s much cheaper to keep a customer than to acquire a new one, so any insights into user retention are valuable to credit card companies. Another place to continue this research would be to flip the model and identify factors that lead to customers not churning, and seeing if that can be utilized for retention practices as well. Overall, insight was gained into what factors make credit card customer churn tick, however the surface has only been scratched in terms of finding the most effective predictive model and identifying all aspects of what facilitates churn, including factors behind non-churn.

**7. References**

* Konstantin, T. Bank Churn Data Exploration and Churn Prediction. Retrieved from <https://www.kaggle.com/code/thomaskonstantin/bank-churn-data-exploration-and-churn-prediction/input>.
* Krchova, I. (oct-4-2023). Upsampling Methods for Improving Prediction Accuracy. Retrieved from <https://mostly.ai/blog/upsampling-methods-for-improving-prediction-accuracy>.
* <https://scikit-learn.org/stable/auto_examples/ensemble/plot_forest_importances.html>