




Group 9

Detecting Credit Card Churners

Members: Huimin Zhang 505322328,
Julius Castro,
Qinyi Chen 505331238,
Sibo Guo 205119757



Introduction

Credit card churning is a method of gaming bonus incentives offered by banks and credit card companies through the practice of repeatedly applying and closing credit card accounts and only meeting the minimum spending requirements to earn the bonus incentives. These incentives include cash back, airline miles, airline companion passes, hotel loyalty points, and hotel free night certificates. Credit card churners cost banks and companies millions of dollars in profit. As a result, many companies have created systems to detect credit card churners and blacklist them.

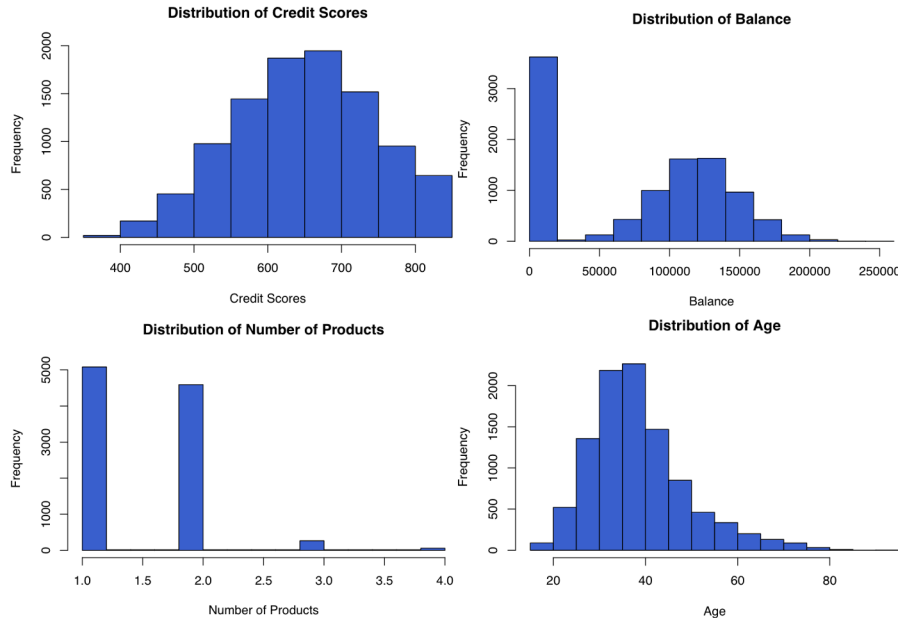
Note: For the purposes of this data set, credit card churners are classified as customers who “Exit” or close their accounts.

Abstract

The goal of this study is to identify important factors that can accurately classify potential credit card churners. Utilizing Chi Squared Tests, ANOVA, checks for multicollinearity, and The Best Subset Method, we identified the best predictor variables, which were credit score, credit card balance, number of products with the company, age, gender, geography, and classification of active and inactive users. Using the KNN model, we achieved the prediction accuracy of 0.8398571 and kappa of 0.3965521. In conclusion, we found that women are more likely to be credit card churners than men and middle aged customers are more likely to be credit card churners than younger and older aged customers.

Exploratory Data Analysis: Numerical Variables

Numerical Variables



→ Numerical Variables

- ◆ Credit Scores
- ◆ Balance
- ◆ Number of Products
- ◆ Age
- ◆ Salary *
- ◆ Tenure *

*NOT SHOWN

Exploratory Data Analysis: Categorical Variables

Gender Frequency Table

	Frequency	Percent	Cum. percent
Male	5457	54.6	54.6
Female	4543	45.4	100.0
Total	10000	100.0	100.0

Geography Frequency Table

	Frequency	Percent	Cum. percent
France	5014	50.1	50.1
Germany	2509	25.1	75.2
Spain	2477	24.8	100.0
Total	10000	100.0	100.0

Active Member Frequency Table

	Frequency	Percent	Cum. percent
1	5151	51.5	51.5
0	4849	48.5	100.0
Total	10000	100.0	100.0

HasCrCard Member Frequency Table

	Frequency	Percent	Cum. percent
1	7055	70.6	70.6
0	2945	29.4	100.0
Total	10000	100.0	100.0

→ Categorical Variables

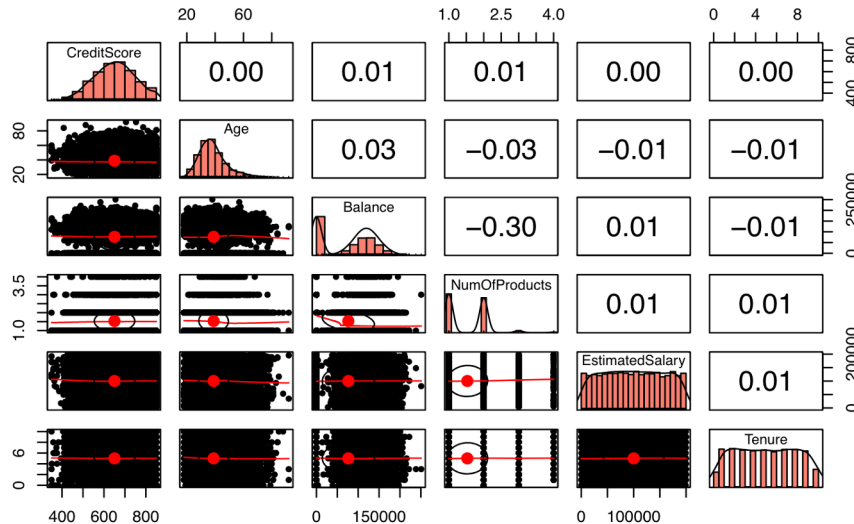
- ◆ Gender
- ◆ Geography
- ◆ Active Member
- ◆ HasCrCard

Research Question

How do factors such as credit score, credit card balance, number of products with the company, age, salary, tenure, gender, geography, and classification of active and inactive users contribute to credit card churning and which groups of customers are more likely to be credit card churners?

Exploratory Data Analysis: Outcome Variable And Correlation Matrix

Correlation Scattor Plots



Exited Frequency Table

	Frequency	Percent	Cum. percent
0	7963	79.6	79.6
1	2037	20.4	100.0
Total	10000	100.0	100.0

- None of the variables are highly correlated with one another.
- The highest correlation is between Balance and Number of Products
 - ◆ Customers who have more credit cards with the company tend to have higher balances.

Note: This plot shows how each numerical variable distributed and their correlation matrix based on Pearson correlation coefficient method.

Exploratory Data Analysis: Insights

- Credit Scores among customers are approximately normally distributed with a mean of 650.53 and a standard deviation of 96.65.
- 36.2% of customers carry a balance of \$0 on their credit cards.
- Age of Customers is skewed right with the average age of 39 years old, with the youngest customer being 18 and the oldest customer being 92.
- About 50% of customers in the data set are in France, about 25% from Spain, and 25% from Germany.
- 51.5% of customers are active users and Gender is split approximately evenly.
- Estimated Salary is uniformly distributed from 0 to 200,000. (Note: Estimated Salary is self reported and is not officially confirmed using formal documents)

Preprocessing: Feature Selection

In order to do the classification, we need to find significant predictors.

During this process, we will not only look at the results, but also will consider the real situation and make it meet our expectations as much as possible.

→ Variables transform

- ◆ Credit score (Category: low, med and good)
- ◆ Balance (Category: only consider whether the accounts are balance or not: Yes/No)
- ◆ Has Credit Card/Is Active Member (Convert numeric to Category)

→ Chisq.test for categorical variables

→ Anova-F-test for numeric variables

→ Check multicollinearity (no high correlation)

→ Best subset method (double check)

- ◆ Exhaustive
- ◆ Forward
- ◆ Backward

Chisq.test

Anova-F-test

Anova-F-test is a good way to test if the numeric predictors is significant or not. Predictors should be put in the anova function one by one, otherwise the order and quantity of predictors we put in anova will affect the result a lot.

The results show that the variable “HasCrCard”, “Tenure” and “EstimatedSalary” are not significant.

Here, we can keep 7 predictors.

Category.pred	P.value	Reject
CreditScore	0.0059043	TRUE
Geography	0.0000000	TRUE
Gender	0.0000000	TRUE
Balance	0.0000000	TRUE
HasCrCard	0.3071143	FALSE
IsActiveMember	0.0000000	TRUE

Numeric.pred	P.value	Reject
Age	0.0000000	TRUE
Tenure	0.2481281	FALSE
NumOfProducts	0.0000251	TRUE
EstimatedSalary	0.9714706	FALSE

Best Subset

method	Adj.R2	CP	BIC
exhaustive	9	7	5
forward	9	7	5
backward	9	7	5

Since this method separates the category variables into different levels and will test each of them individually, this leads to some are significant and others are not. As long as one level makes sense, we will keep the whole variable.

We want to know if we can add some more interesting variables. All three methods show the same result. Therefore, the only thing to think about is how many variables we should choose.

Based on the “CP”, it shows the same result as we did with `chisq.test` and `anova-F-test`.

Based on the “Adj.R2”, it indicates we can keep the variable “Tenure”. Notice, we not only look for how to detect the credit card churners, but also want to know if this relate to their gender or residence. Since “Tenure” is beyond our particular interest to study, we decide not to keep it.

Same with “BIC”, it indicates we should not keep the “NumofProducts”, we will still keep it.

Classification

- **Logistic regression (LR)**
Binary classifier based on maximum likelihood estimation where the response falls into one of two categories, such as yes or no.
- **Linear Discriminant Analysis (LDA)**
Classifier, where we wish to classify an observation into one of K classes (K greater or equal to 2). Based on least squares estimation.
- **Quadratic Discriminant Analysis (QDA)**
QDA is a variant of LDA in which an individual covariance matrix is estimated for every class of observations.
- **K-nearest Neighbors(KNN)**
K Nearest Neighbour is a simple algorithm that stores all the available cases and classifies the new data or case based on a similarity measure.

5 Fold Cross Validation

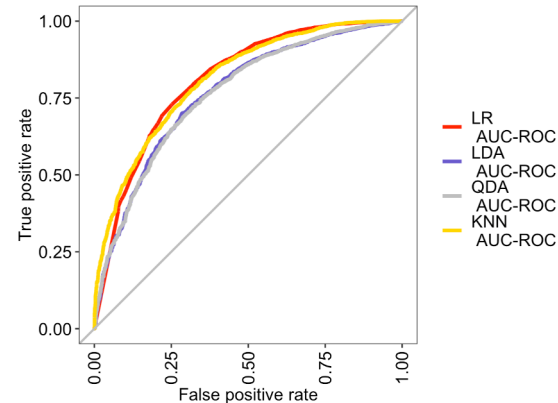
- Used to estimate our machine learning models on unseen data.
- Tend to be less biased.

General steps:

1. Shuffle the data randomly
2. Split into 5 groups
3. For each group, we take that group as test data and the rest of groups as training data
4. Fit models on training data and evaluate with test data
5. Obtain the evaluation

Model Evaluation

- Based on the result, we can see that KNN model performs the best with Accuracy of 0.8398571 and Kappa of 0.3965521. (Fair range for strength of agreement)
- Base on the ROC Curve, we can also see that KNN model performs the best.
- LR and LDA did not perform as well as expected since they are linear boundary.



method	Accuracy	kappa
LR	0.8102859	0.2372941
LDA	0.8089994	0.2493197
QDA	0.8325749	0.3948674
KNN	0.8398571	0.3965521

Confusion Matrix(KNN)

Reference		
Prediction	exit	not
exit	239	101
not	360	2300

- The classifier made a total of 3000 predictions. Out of those cases, the classifier predicted 340 people will exit, and 2660 will not. In reality, 599 people exit, and 2401 not.
- The Misclassification Rate is $(101+360)/3000 = 15\%$ in this prediction.
- However, we want to beware of the people will exit but classify as not exit. We focus on the false negatives rate, which is $360/(239+360) = 60\%$. It is high in this case. We need to improve our prediction later.

False Negative Rate

We will look at the false negative rate for 3 other models:

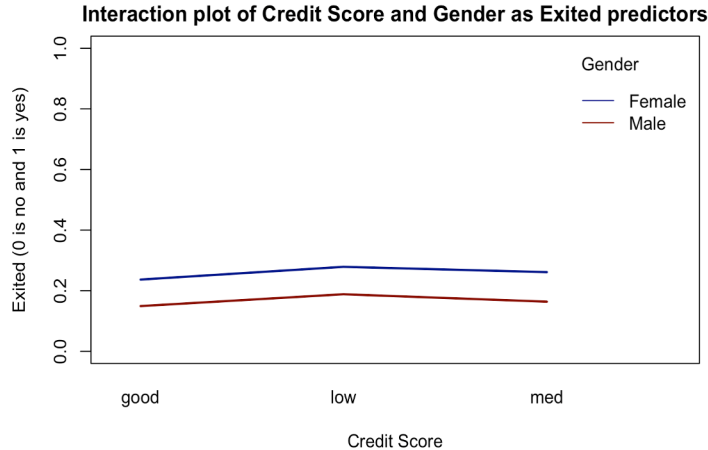
LR: $472/(127+472) = 78\%$

LDA: $461/(461+138) = 76\%$

QDA: $372/(227+372) = 62\%$

We can see that KNN model still has the lowest error rate for people will exit but classify as not exit.

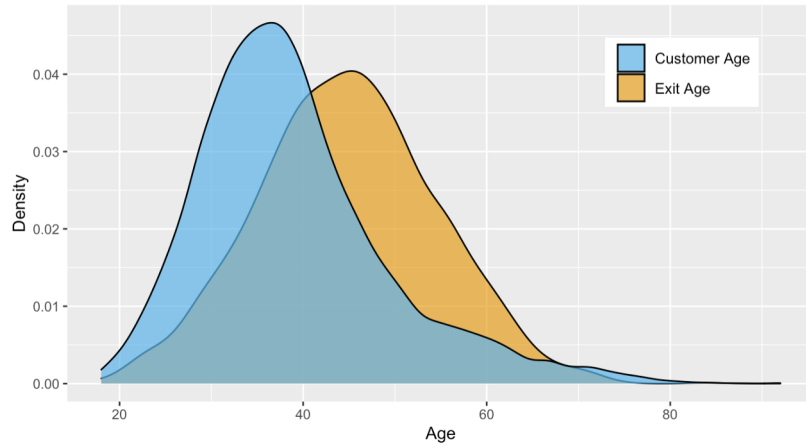
Interaction plot and interpretation



- The interaction plot shows that female customers have a higher chance to exit than male customers because the blue line is always higher than the red line.
- The interaction terms Credit Score and Gender are not significant since both lines are parallel.
- When we add interaction terms Credit Score and Gender to the KNN model, prediction accuracy drops from 0.839 to 0.838 and Kappa drops from 0.39 to 0.38.
- Therefore, the interaction terms Credit Score and Gender are not significant in this model.

Age and Gender

Density plot
Exited by Age and Customer by Age



- The average age of all customers is 39 years old.
- The average age of credit card churners is 45 years old.

Exited By Gender Frequency Table

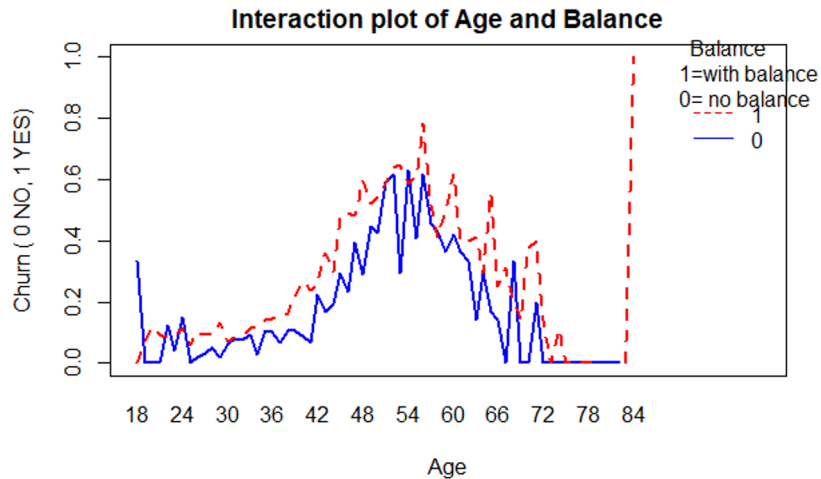
	Frequency	Percent	Cum. percent
Female	1139	55.9	55.9
Male	898	44.1	100.0
Total	2037	100.0	100.0

Gender Frequency Table

	Frequency	Percent	Cum. percent
Male	5457	54.6	54.6
Female	4543	45.4	100.0
Total	10000	100.0	100.0

- Percentage of female customers who are credit card churners is 55.9% while female customers make up only 45.4% of all customers.

Age and Balance



- People with balance on the account seems to be more likely to churn base on the trend.
- Younger and older groups are less likely to churn.
- Churn reached peak around people with age around 50 to 60 for both with balance or not.

Final Insights and Conclusion

→ **Female customers are more likely to be credit card churners than male customers.**

- ◆ Credit card churning can be denoted as a risk as it has the potential to damage your credit score and thus any chances of securing loans and capital. Extant literature concerning risk based on gender, find that men tend to take risks more than women (Harvard Business Review), so it seems odd that our analysis found women to be potentially riskier than men by being more likely to be credit card churners.
- ◆ Although gender is split approximately evenly, the difference of about 10% (about 55% males vs 45% females) is a large portion of the population with a dataset of 10,000, so this could have potential impact on our analysis.

Final Insights and Conclusion

→ **Middle aged customers are more likely to be credit card churners than younger and older customers.**

- ◆ Although, people with balance on the account seems to be more likely to churn based on our analysis, both followed a similar trend based on age as people with no balance and people with balance both peaked in the same area, the only difference was that middle aged customers with balance were likelier to be credit card churners than those with no balance.
- ◆ Possible factors for this include credit history, which is taken into account by lenders when reviewing a person's credit worthiness for a new credit card or multiple cards, as sufficient credit history is required for credit cards that offer larger rewards.

Drawbacks and Recommendations

→ Drawbacks

- ◆ Data set does not specify how Active Users are classified
- ◆ Data set does not specify if Balance is the average revolving balance in the customer's account over a period of time or a balance in a single point in time.
- ◆ Credit Scores may have been recorded at the time of submission of credit card application and not during ownership of the credit card
- ◆ Data set is limited to 3 countries (1 continent)

→ Recommendations

- ◆ Use data from one of the 3 credit bureaus
- ◆ Include what kind of credit cards customers have on file and classify credit cards from most churnable to least churnable.
 - Customers with multiple credit cards that are easily churnable present a greater risk to the company