Project

(Churn for bank customers)

Eunmi Kim
EXK180015

Contents

Executive Summary

Nowadays, there are many competitors coming out on market. It becomes important to hold customers. There must be many reasons to make customers leave the bank company. In this paper, I analyze what kind of customers are more likely to churn or not. For this analysis, I used a large database maintained by the bank company. This database contained relevant information of churn rate for bank customers at a specific time. I used a number of multivariate statistical techniques for our analysis.

I will create awareness about credit score, balance, tenure, possession of credit card, gender, and other factors leading to churn. I can suggest the company to focus on marketing, advertising, providing some options, and others to the customers who will churn in the future, which in turn will prevent the company loss of profit.

Introduction

The bank company would like to notice which customers tend to leave. It is better for the bank company to prevent the customers leaving. Acquiring new customers cost more than retaining existing customers. I would like to suggest that we should focus on marketing on those who are likely to churn.

To figure out how to reduce the influence of churn, I would like to observe each factor which cause customer leaving the bank company. I will use predictive analytics with machine learning models to predict whether the customers churn or not.

Through exploratory analysis, data pre-processing, and creating machine learning models (Logistic Regression model, Decision Tree model, Random Forest model, and Support Vector Machine) utilizing R, the goal of my project is to uncover patterns that might be hidden in data and gain insight into possible predictors of churn.

Dataset Overview

Attribute:

- RowNumber
- CustomerId
- Surname
- CreditScore
- Geography
- Gender
- Age
- Tenure
- Balance
- NumOfProducts
- HasCrCard
- IsActiveMember
- EstimatedSalary
- Exited

Data Description

I found this dataset from Kaggle.com. There is no additional source to explore. This dataset does not include many information. Especially, it does not show when they measure this for the given data. Regarding tenure, there is no information of year base or month base. I assume it is the year base tenure.

This dataset contains details of a bank's customers and the target variable is a binary variable reflecting the fact whether the customer left the bank (closed his account) or he or she continues to be a customer.

It has 14 columns and 10,000 observations from 3 countries (Spain, France, and Germany). There are 9 categorical variables and 5 numeric variables.

Preprocessing Data

I reviewed the dataset to identify what attributes will be necessary to be used to analyze.

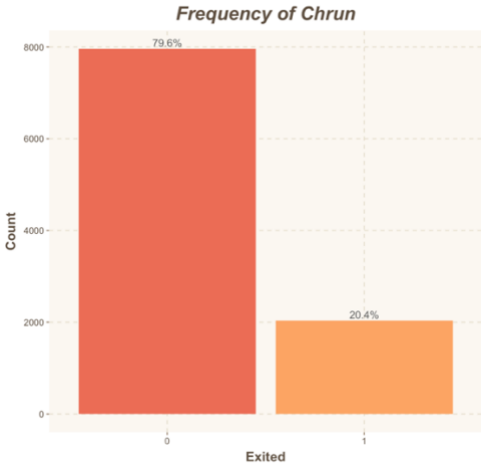
There is no missing value in this dataset. I removed the unnecessary attributes, RowNumber, CustomerId which is unique identifier for a given customer and Surname, which are not the important features to explain about who tends to leave the bank company from the dataset.

Also, I used a categorical variable for tenure instead of a numerical variable due to the explanatory power.

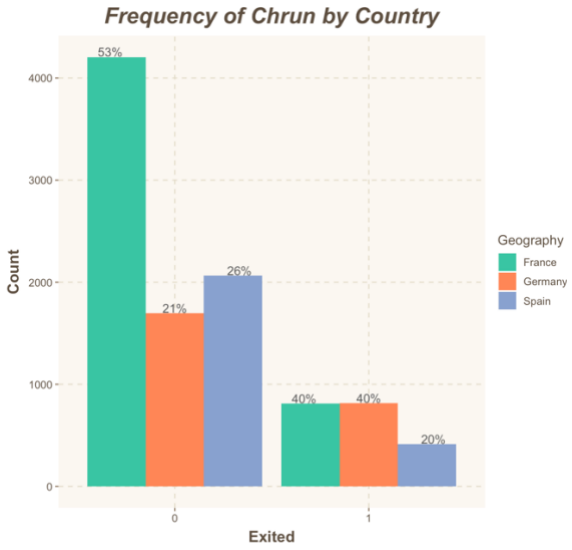
Exploratory Data Analysis

1) Exited (Target variable)

The below plot shows the frequency of churn. As can be seen below, we can see that about 80% of customers tend to stay at the bank company and 20 % of customers tend to leave the bank company. For now, I see this dataset with unbalanced target variable.

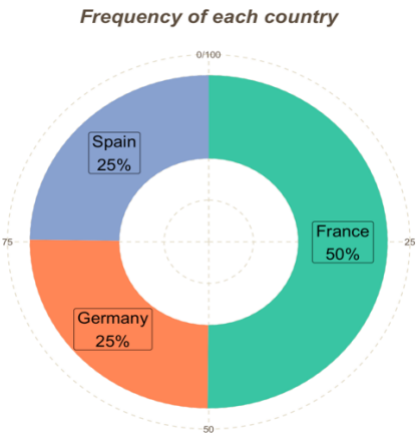


I compare the frequency of churn by country. There are customers living in France who are more likely to stay than churn, followed by customers in Spain with 26% and customers in Germany with 21%. Compared to this, out of churners, customers in France accounts for about 40 % of churners, customers in Germany accounts for 40% of churners, and customers in Spain accounts for 20% of churners.



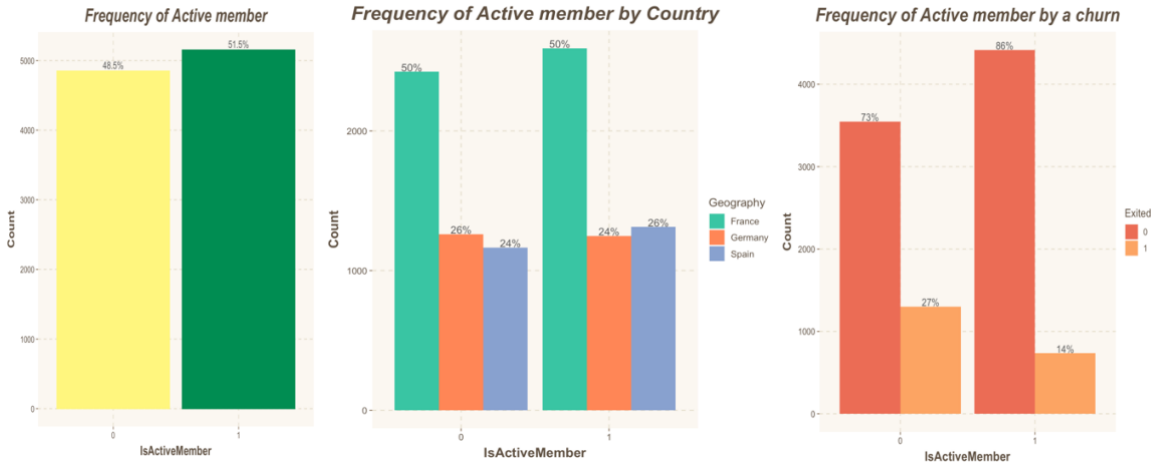
2) Geography

Around 50 % of customers live in France, followed by Spain and Germany.



3) Active member

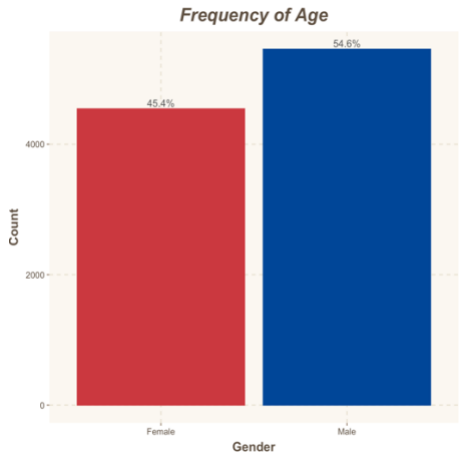
There are 51.5 % Active members and 48.5% non-active members at the bank company. Secondly, I would like to analyze how many active members there are each country. Out of active members, there are 50% of customers in France, 26% of them in Germany, and 24% of them in Spain. Out of non-active members, there are 50% of customers in France, 24% of them in Germany, and 26 % of them in Spain. Since customers in France accounts for around 50 % of customers, it would be common to explain this result.

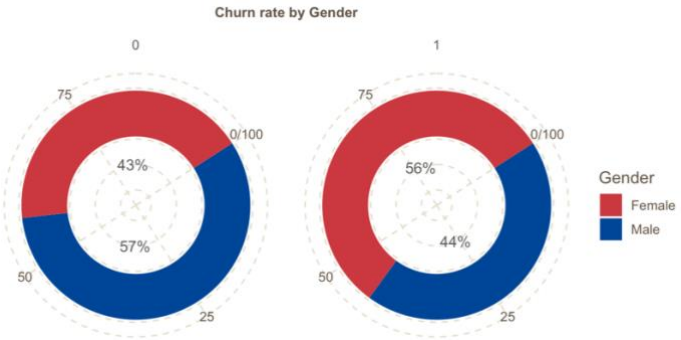


The above third graph talks about how many active members by churn. It shows that non-active members are more likely to leave the bank company compared to active members. The rate of churners when they are an active member is 14%. On the other hands, the rate of churners when they are a non-active member is 27% which is 13% higher.

4) Gender

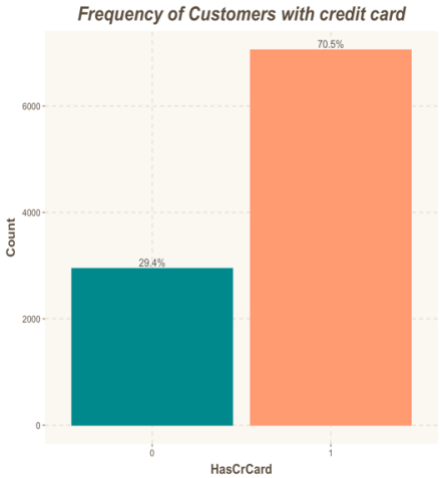
There are more male than female of customers at the bank company. When it comes to churn rate each gender, female customers are more likely to churn than male customers do. Male customers. The proportion of male customers staying at the bank company is also greater than that of female customers.

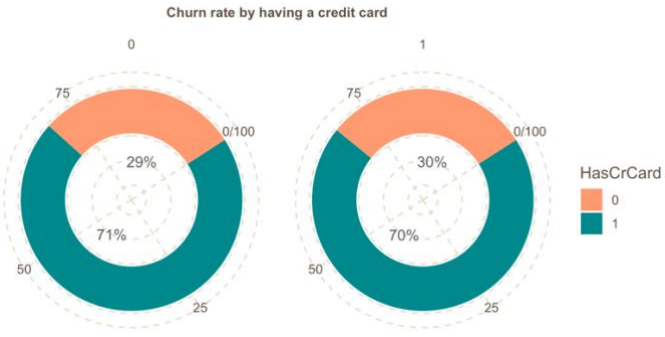




5) Credit card

There are 70.5% customers with a credit card and 29.4% customers without a credit card. When it comes to churn rate by having credit card, no matter who have a credit card or not, the rate of churn is similar. It shows that having a credit card is not the important feature to analyze the churn rate.

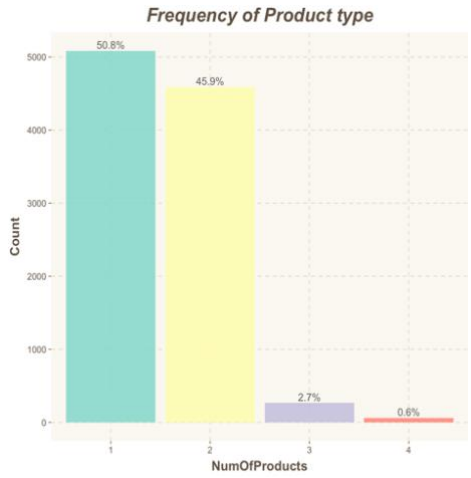


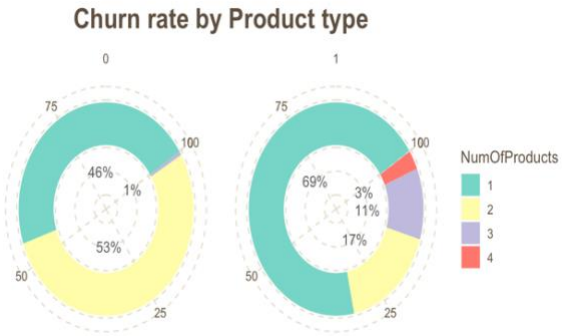


6) Product type

The graph below shows that customers at the bank company possess one or two products. Customers barely have more than two products. The rate of those who use only one product is the highest with 50.8%, followed by with two products with 45.9%, three products with 2.7%, and four products with 0.6%.

The right graph shows that I can see that the customers tend to leave the bank company when they have more than two products. Almost all of customers with two or three products tend to churn.

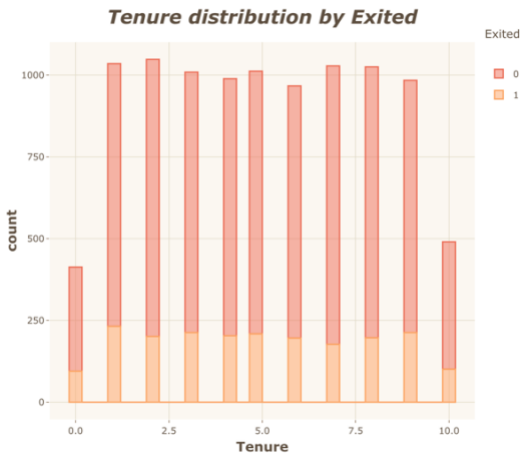


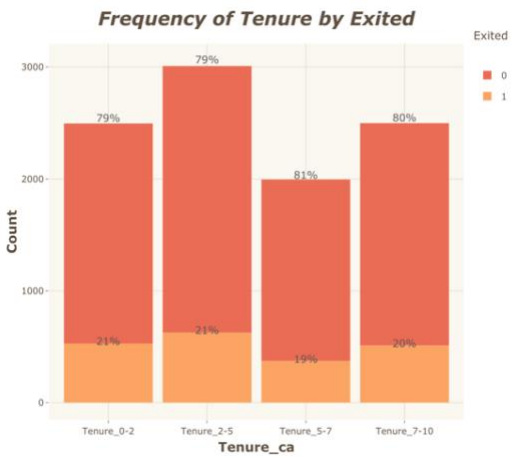


7) Tenure

I assume that the year-base tenure is the term of having been with the bank company.

I compare the numeric variable of and the categorical variable of tenure attribute. The left chart shows that tenure has no a distinct character on the likelihood of churn. After I divide tenure attribute into 4 levels, tenure0-2, tenure2-5, tenure5-7, and tenure 7-10. I can see that customers with below 5 tenure tend to churn more than customers with above 5 tenure do.

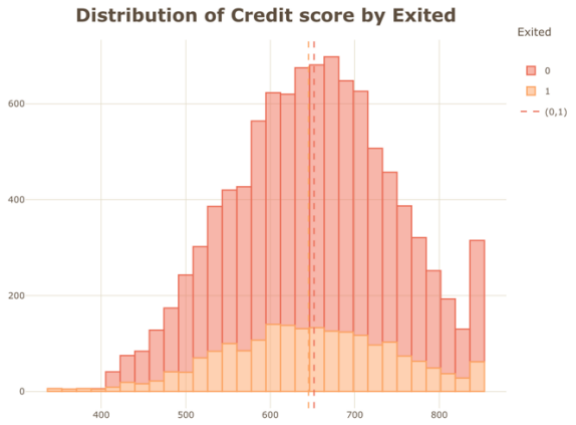




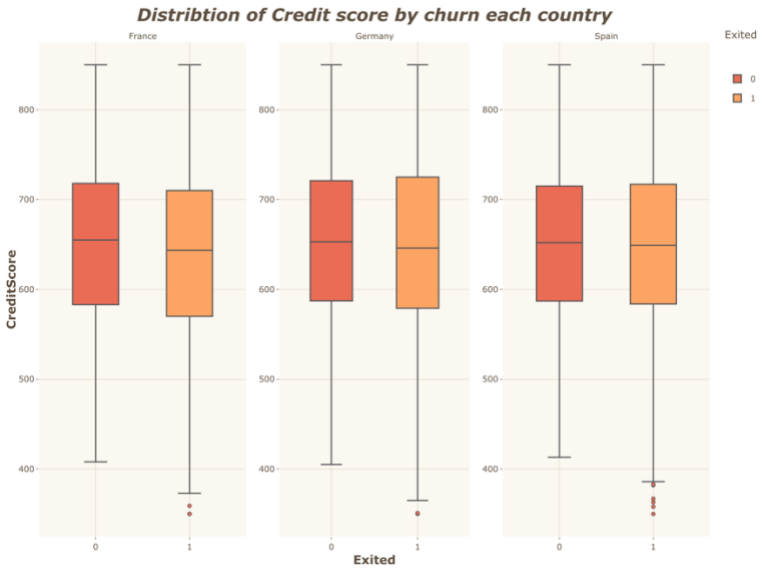
8) Credit score

Credit score for customers the bank company is distributed from 350 to 850. The customers with low credit score tend to churn. Also, customers having around 600 credit score churn the most.

The mean of credit score for non-churn customers is slightly higher than the mean of credit score for churn customers.

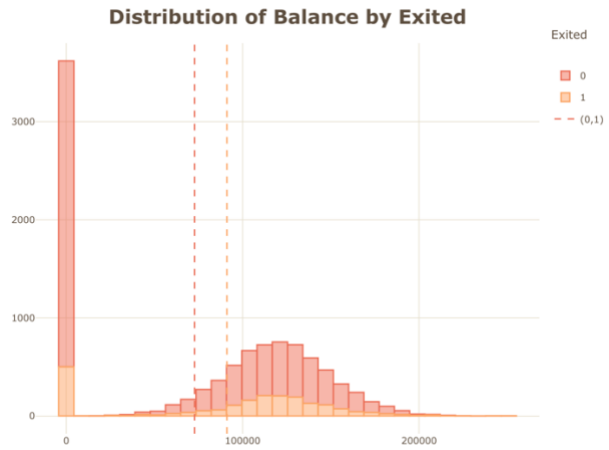


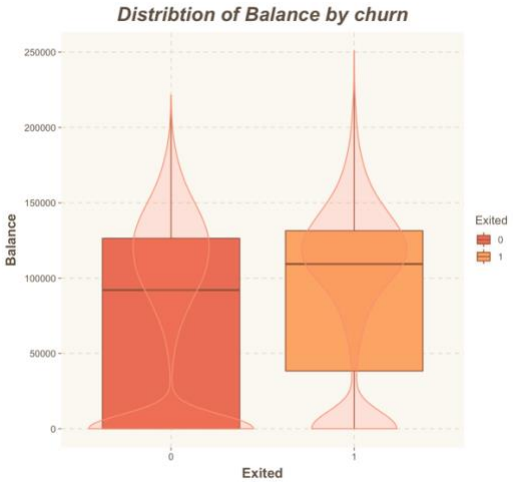
Customers in Germany have higher credit score than customers in Spain and France. All customers who do not churn each country have higher credit score than who do churn do.



9) Balance

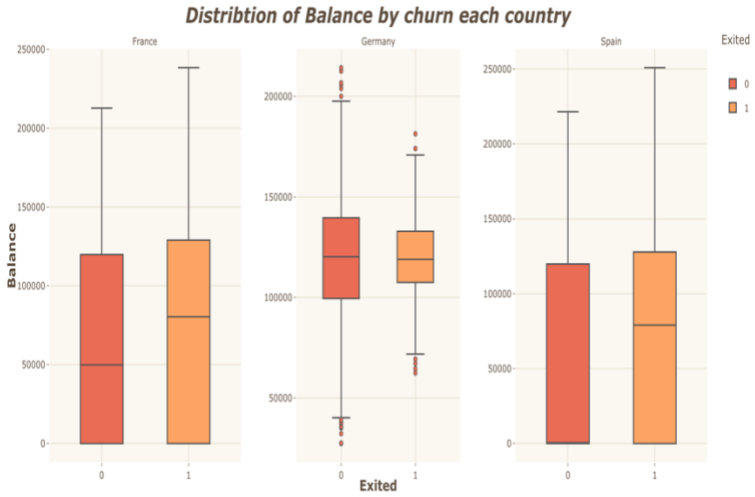
The charts below show that the balance has effect on the chance of a customer churning. There are many customers with 0 balance in their account. The chance of churning with when they have higher balance in its own account is higher than with low balance. The mean of balance with customers who churn is higher than the mean of balance with customers who stay at the bank.





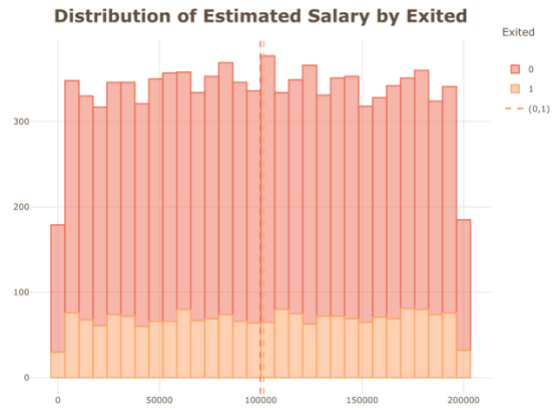
Mostly customers in Germany have higher balance than customers in other countries have. Customers in Germany tend to churn even though they have higher average balance. Unlike the result above, the average balance of non-churn customers is a little bit higher than one of churn customers in Germany.

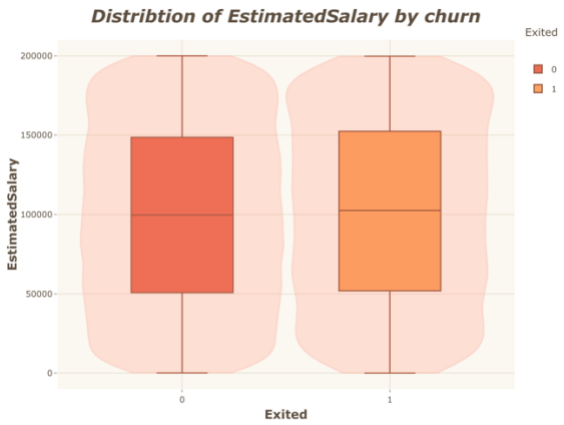
Customer in Spain have the lowest balance among these countries.



10) Estimated salary

Estimated salary has little effect on the chance of a customer churning. There is no difference between salary with churn customers and non-churn customers. Average salary of with churn customers and non-churn customers is almost same. However, customers with higher salary are more likely to churn than others.



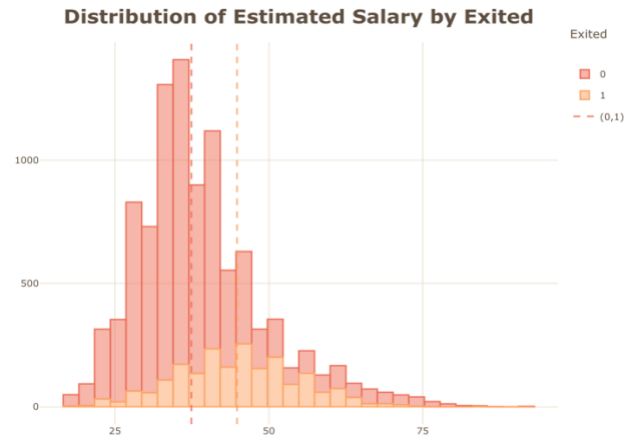


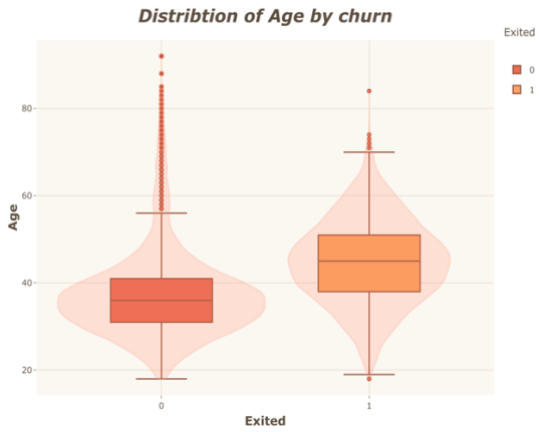
The chart below explains that each country has the different salary range for churn and non-churn customers. Customers in Germany with lower salary are more likely to churn. On the other hands, the average salaries of churn-customers are higher than on of non-churn customers in France and Spain.

11) Age

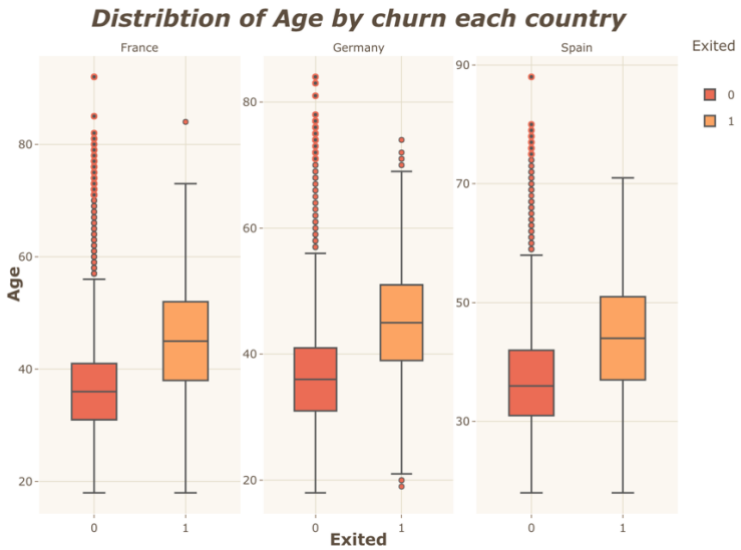
The chart below shows that the most distinguishable attribute is Age among all attributes. The older customers are churning more than the youger customers. The mean of age of churn customers, 36 years old, is much higher than the mean of age of non-churn customers, 45 years old. Around 46 years old customers churn the most and customers over 75 years old mostly churn. Also, customers below 40 years old are more likely to stay at the bank not churn.

Thus, we can say Age has a significant effect on the likelihood to churn.





There is no special feature on the distribution of age by country. The means of age of churn customers and non-churn customers are almost same across all countries.

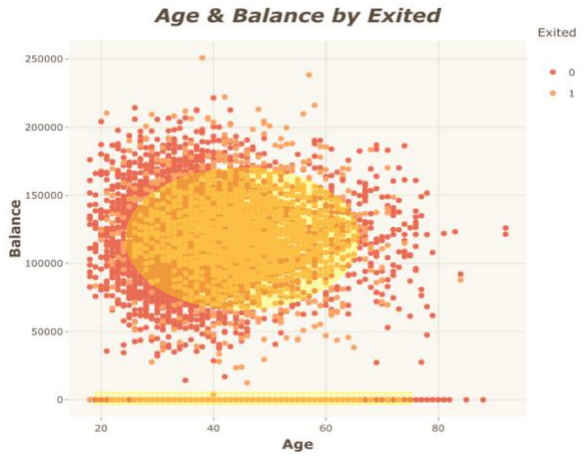


Comparison of two variables (Scatter plot)

Other than estimated salary, I make scatter plots to display values for two variables, which affect the churn rate. Since Salary has no effect to analyze the likelihood of churn.

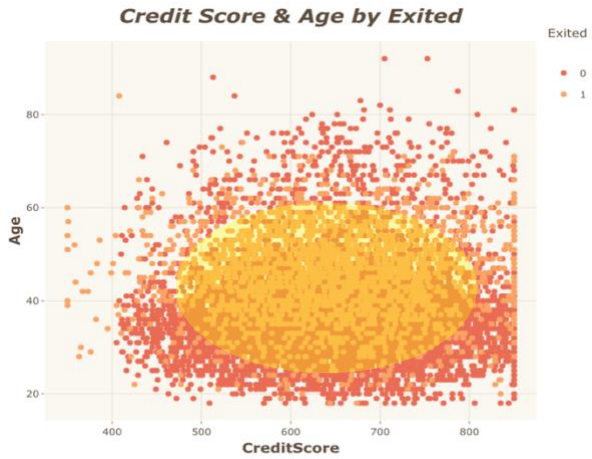
The chart is the scatter plot of credit score and balance. It does not show the linear relationship. However, there are certain patters on the graph.

It shows that no matter what credit score customers have, when they have 0 balance in their account, they are likely to churn.



The chart is the scatter plot of Age and Balance. It does not show the linear relationship. However, there are certain patters on the graph.

Those who are 30 to 60 age and have 100,000 to 150,000 balance are more likely to churn.



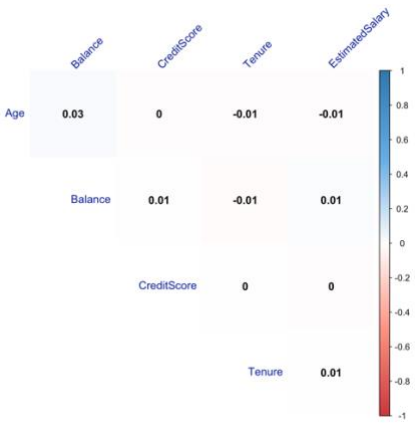
The chart is the scatter plot of Age and Balance. It does not show the linear relationship. However, there are certain patters on the graph.

Those who are 30 to 60 years old and 500 to 700 credit score are more likely to churn.

Customers with less than 400 credit score churn and their age are around 40 to 60 years old.

Correlation of numeric variable

There is a little bit negative relationship between Age and balance. The older people have the lower balance. Others do not have any correlation each other.



The chart below shows the correlogram to analyze the relationship between each pair of numeric variables of the dataset. I can see the overall distribution of these variable and the correlation of variables. Also, I can see which customers with a specific credit score, age, tenure, balance, and salary tend to churn or stay at the bank. These are all similar result as I explained before.

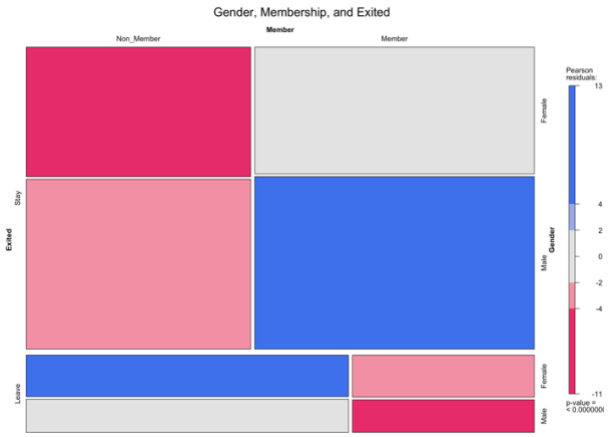
Almost all of these variables do not have a significant effect on churn.

Correlation of Categorical variables (Mosaic plots)

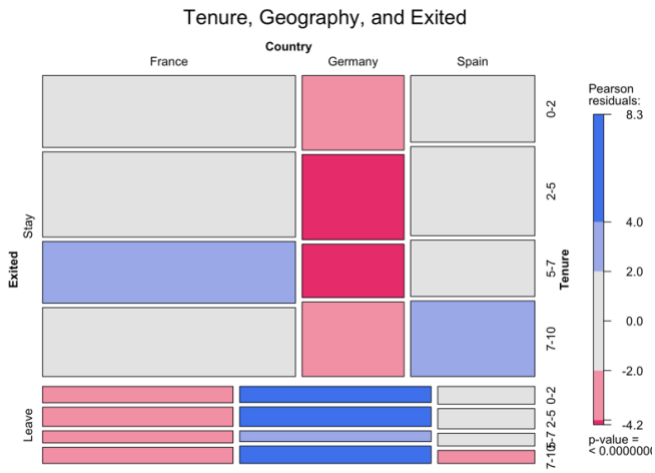
Mosaic charts can display the relationship between categorical variables using rectangles whose areas represent the proportion of cases for any given combination of levels. The color of the tiles can also indicate the degree relationship among the variables. The size of the tile is proportional to the percentage of cases in that combination of levels.

Clearly more customers stay at the bank, than churn. Those who are male and member stay at the company the most. If customers are member of the bank, they are more likely to stay rather than churn. The cell for male member customers has more observations that would be expected under the null model.

Female non-member customers are more likely to churn and the cell for female member has more observations of customers that would be expected under the null model.

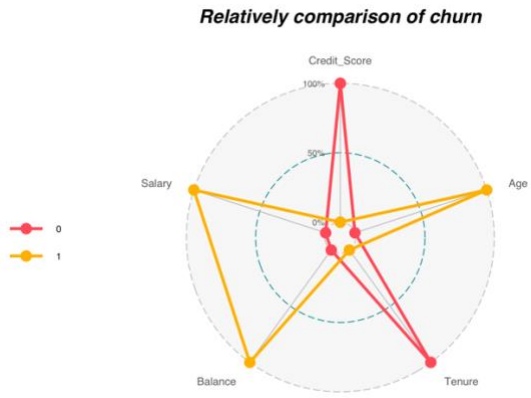


The chart below shows that French with 5-7 tenure have a tendency to not churn and Spanish with 7-10 tenure have a tendency to not churn and more observation in that cells that would be expected under the null model. On the other hands, Germans with all tenure period have a tendency to churn. The cell for customers in France who leave the bank has fewer observations that would be expected under the null model.



 # Radar chart comparison

The chart below explains which mean of the variable is higher between with churner and non-churner. If the mean of Salary for churner is higher than the mean of Salary for non-churner, then the salary is marked for churner on axes. Salary, Balance, Age are marked for churner which means that churners have the higher average of Salary, Balance, and Age than non-churners have. Non-churners have the higher average of Credit score and tenure than churner have.



E. Empirical Analysis

Machine Learning Models

For my predictive modeling I chose to deploy 4 different models: Logistic Regression model, Decision Tree model, Random Forest model, and Support Vector Machine model. The categorical nature of my target variable either if a customer churn or not.

After run these models, I would evaluate the model performance based on the results. In order to find the best model, I think that how accurately we can distinguish customers who are more likely to churn. To do so, the criteria we should consider is recall.

After I decide the best model based on the recall of all ML models, I will find the important variable.

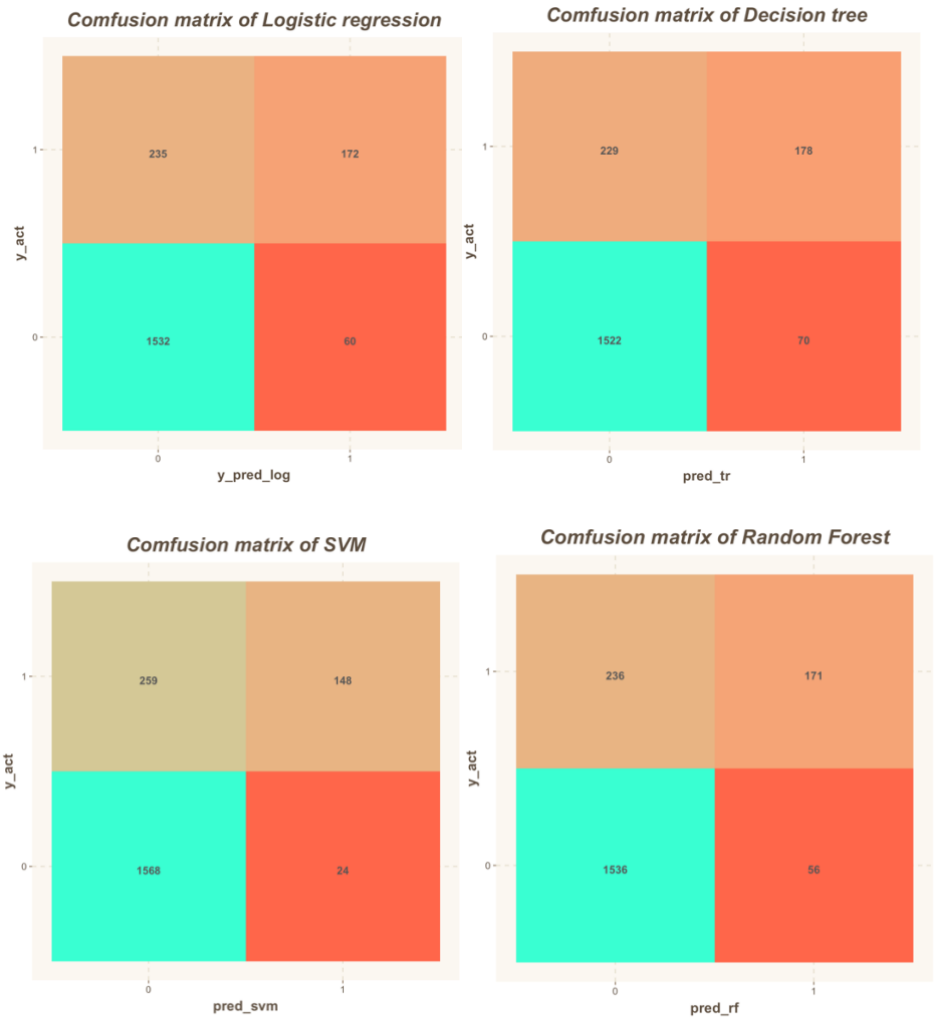
1) Logistic regression

2) Decision tree

3) Random forest

4) Support Vector Machine

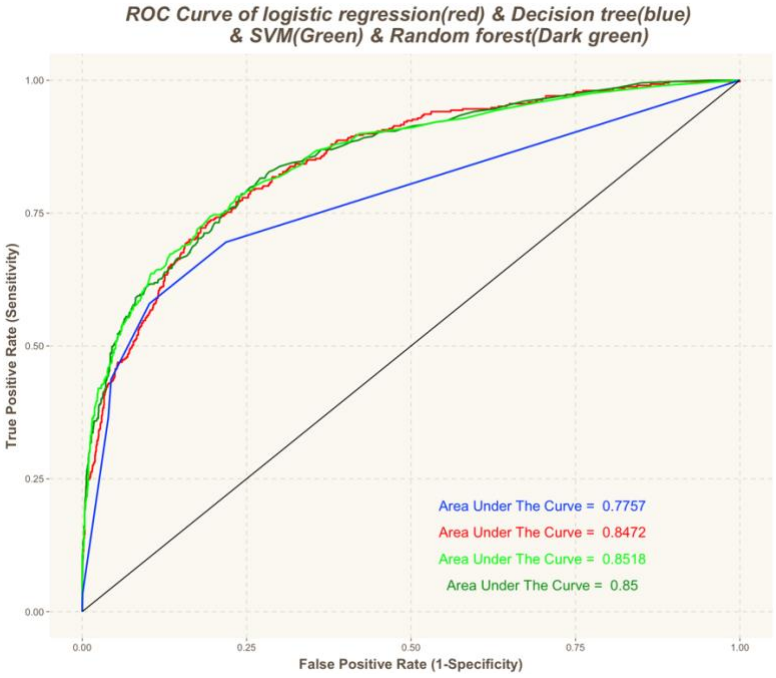
Confusion Matrix



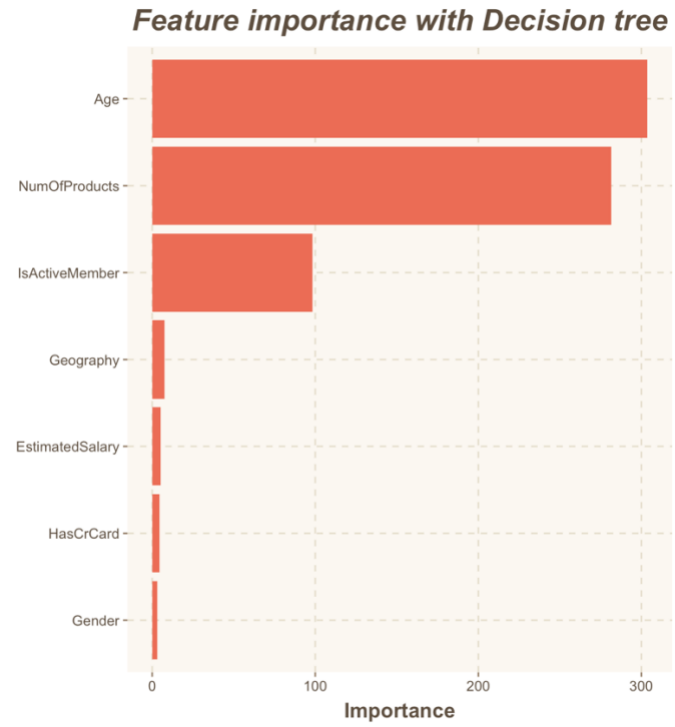
	Logistic regression	Decision tree	SVM	Random forest
Accuracy	0.852	0.85	0.858	0.854
Sensitivity	0.423	0.437	0.363	0.42
Specificity	0.962	0.956	0.985	0.965
Recall	0.423	0.437	0.363	0.42

ROC Curve

The AUC score ROC Curve of all ML models were compared. SVM has the highest AUC scores and better ROC curve for majority of the runs.



Feature Importance



Conclusion

I analyzed the dataset of churn customers at the bank company. It shows whether customers at the bank churn or not. By doing this analysis, I can predict the customers that will possibly churn and also see which features affect the customer's decision for leaving the bank or stay at the bank.

Decision tree being the most accurate ML model with highest recall score was used in determining the significant predictors of churn customers. Below are the most significant predictors derived with the help of Decision tree feature importance: Age, NumsofProducts, IsActiveMember, Geography, and so on.

Through Exploratory Data Analysis and ML models, I can see that Age is the most significant effect on the likelihood to churn. The older customers are churning more than the younger customers.

Sources

www.kaggle.com

https://bookdown.org/lyzhang10/lzhang_r_tips_book/preface.html

https://topepo.github.io/caret/index.html