MIS528: Customer Churn Analysis

Nick Sherman, Kishor Kumar Sridhar, Sridhar, Adjoa Adanledji, and Andrew Smith

12/20/19

Introduction

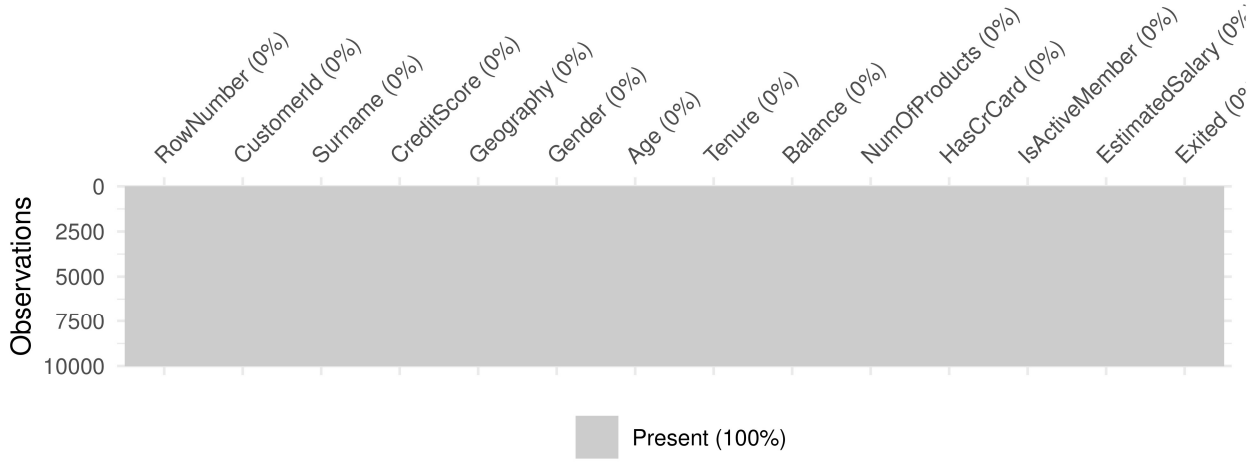
The data set that we chose to explore is a customer churn data set for a bank in Europe, found through a Kaggle data set. The goal of our exploration was to see which variables in the data set played the biggest role in a customer exiting the bank. Being able to identify which customers will leave or those that are at a potential risk of leaving plays an important role in any marketing plan the bank wishes to pursue. We not only examined our data through various plots we also performed some classification models to learn more about the the bank's customers and what influences them leaving. Before we could perform these graphs or models we had to perform some data cleaning.

The Data

The first steps we took in the data cleaning were to change our variables from their current types to the correct type for modeling and data exploration.

```
## # A tibble: 5 x 14
##   RowNumber CustomerId Surname CreditScore Geography Gender   Age Tenure
##       <dbl>      <dbl> <chr>         <dbl> <chr>     <chr>  <dbl>  <dbl>
## 1         1   15634602 Hargra~         619 France    Female    42      2
## 2         2   15647311 Hill            608 Spain     Female    41      1
## 3         3   15619304 Onio            502 France    Female    42      8
## 4         4   15701354 Boni            699 France    Female    39      1
## 5         5   15737888 Mitche~         850 Spain     Female    43      2
## # ... with 6 more variables: Balance <dbl>, NumOfProducts <dbl>,
## #   HasCrCard <dbl>, IsActiveMember <dbl>, EstimatedSalary <dbl>,
## #   Exited <dbl>
```

We also checked to see if there were any null values in our data, and found no missing values.



After these initial check we decided to remove Surname due to potential privacy, and legal issues. We also removed Customer Id and Row Number, as they did not need to be used in our data exploration or modeling.

Finally, we duplicated some of the variables and transformed them (e.g. adding words to factor levels, and bins for continuous variables) to aid in the data exploration.

Create new fields to aid in the data exploration.

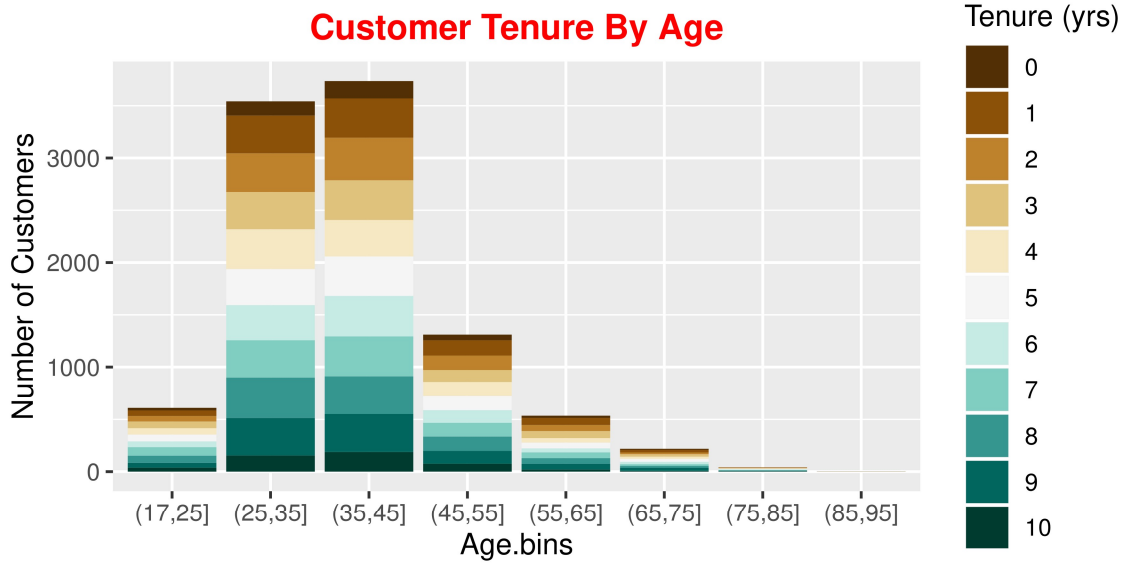
```
## # A tibble: 5 x 18
##   CreditScore Geography Gender   Age Tenure Balance NumOfProducts HasCrCard
##         <int> <fct>     <fct>  <int>  <int>   <dbl> <fct>         <fct>
## 1         619 France    Female    42      2       0  1             1
## 2         608 Spain     Female    41      1   83808. 1             0
## 3         502 France    Female    42      8  159661. 3             1
## 4         699 France    Female    39      1       0  2             0
## 5         850 Spain     Female    43      2  125511. 1             1
## # ... with 10 more variables: IsActiveMember <fct>, EstimatedSalary <dbl>,
## #   Exited <fct>, HasCrCard.words <fct>, IsActiveMember.words <fct>,
## #   Exited.words <fct>, Tenure.bins <ord>, CreditScore.bins <fct>,
## #   Age.bins <fct>, EstimatedSalary.bins <fct>
```

Data Analysis

In order to gain understanding of the data set and to determine significant variables to use for customer churn modelling, we have compared several variables against each other using different types of graph.

Comparing age to tenure.

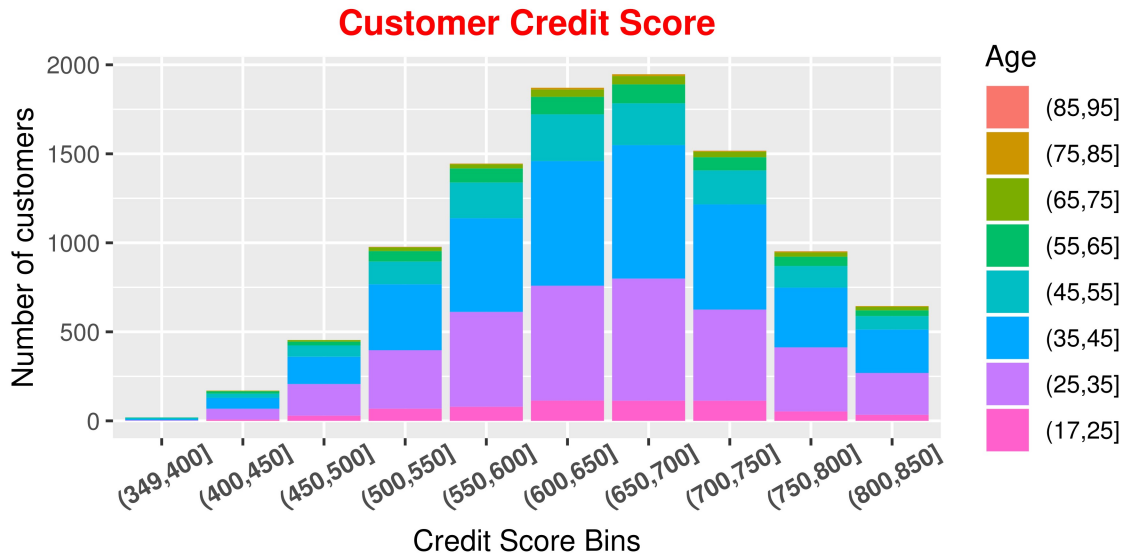
The histogram presented below compares customer age to tenure. As shown on the graph, shades of brown indicates customers that have been with the bank for less than 4 years and shades of green more than 6. From the comparison we can conclude that, majority of customers are between ages 25 to 50 and have been with the bank for at least 5 years.



Comparing age to credit score

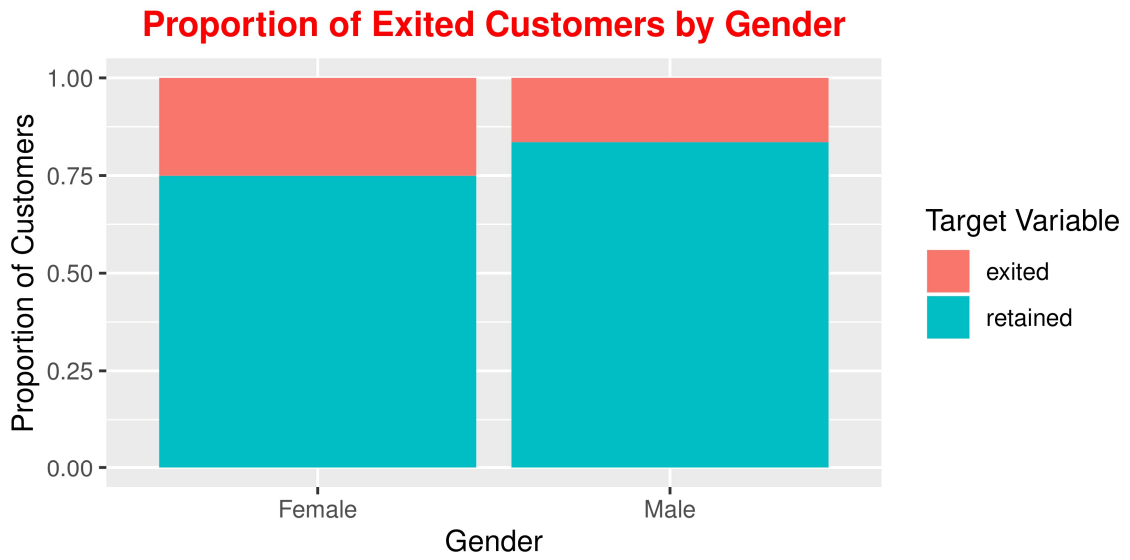
In order to conduct this analysis, groups of credit score and age were formed. A comparison of age to credit score does not indicate anything special about the data. However, the shape of the bar graph indicates normality of the credit score variable.

```
##
##  (17,25]  (25,35]  (35,45]  (45,55]  (55,65]  (65,75]  (75,85]  (85,95]
##      611     3542     3736     1311      536      219       42        3
## (95,105]
##        0
```



Gender Comparison

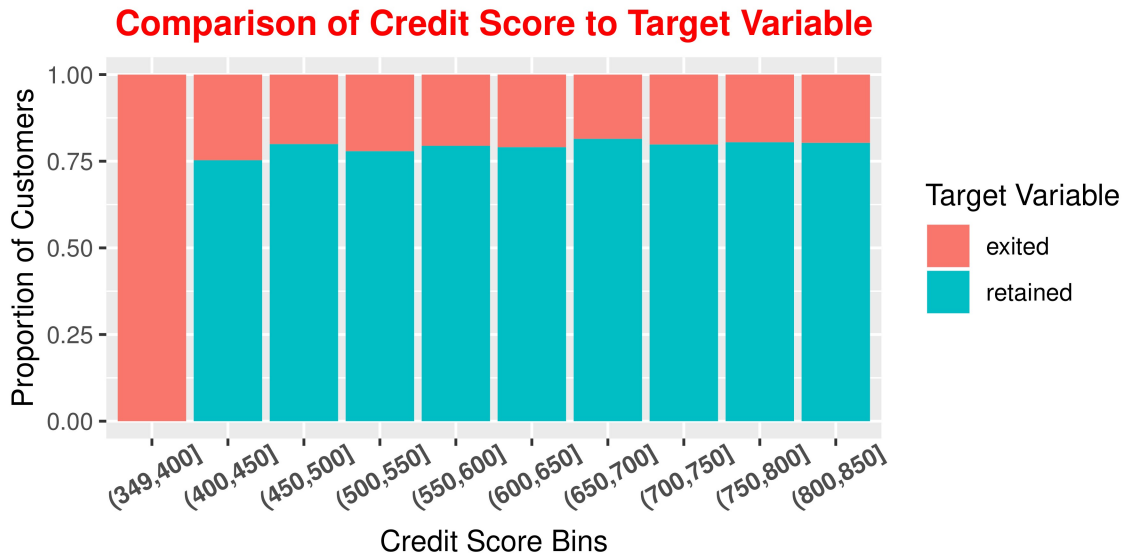
This graph presents the customer churn proportion by gender. 25 percent of female exited versus approximately 14 percent of male. The fact that we observe a higher exit rate for female is not a mere result of distribution of the population of interest. Males actually accounts for 55% of the data. The bank should look closely at this and develop marketing strategies to reduce female exit rate.



Comparing Credit Score to Target variable

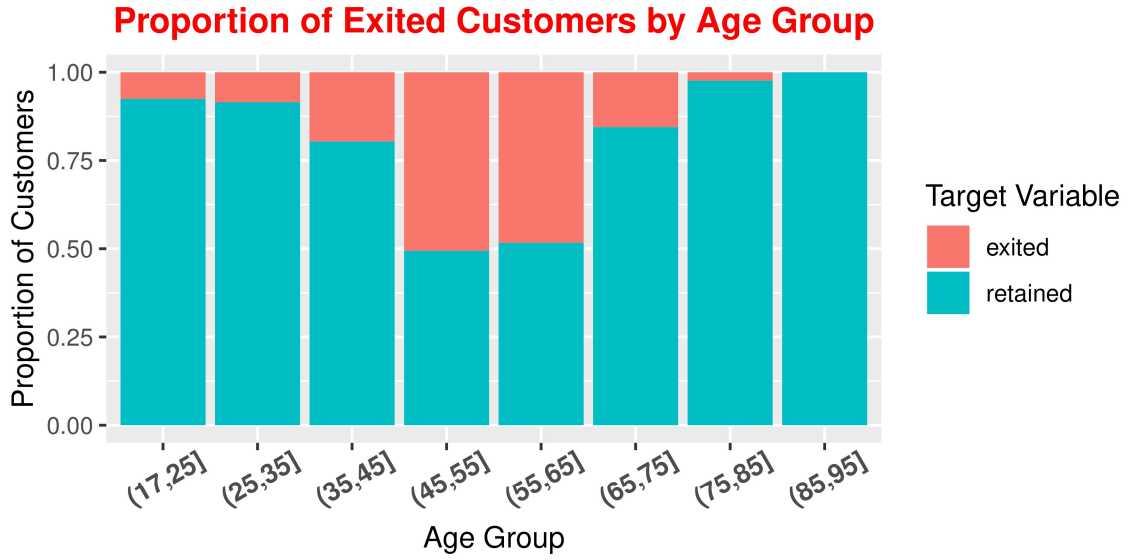
Customers with credit score less than 390 exited the bank. These customers left either because customers with low credit score are charged higher fees or the bank put some restriction forcing customers who fall into

this category to leave the bank. Looking at customers with credit score greater than 390, we notice that in average 25% of customers left the bank. Thus, credit score does not seem to have an impact on customers decision to stay or exit the bank.

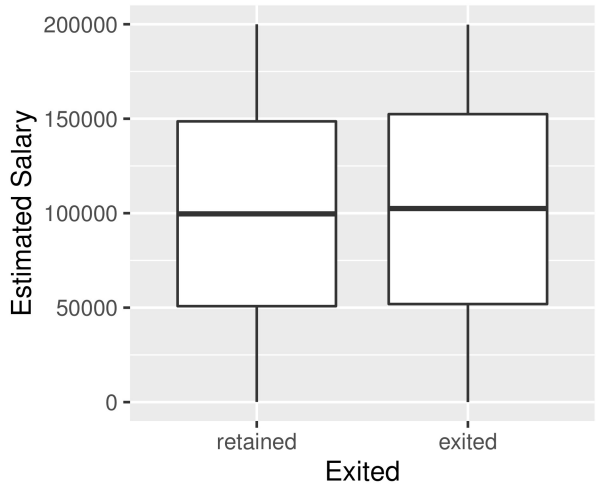


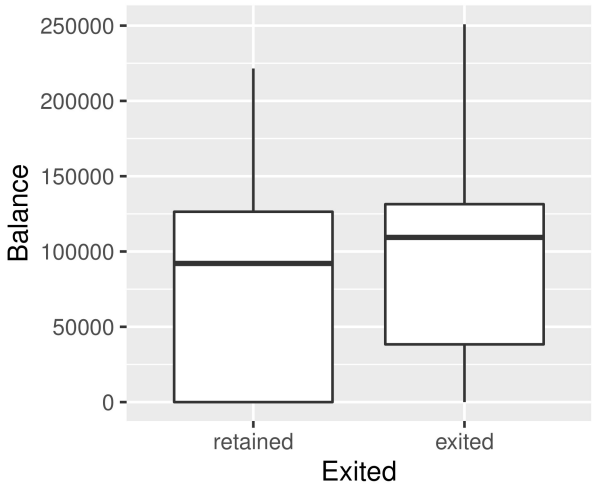
Comparing Age to Target variable

This graph shows some relation between age and the target variable. Exit rate increases as age increases until it reaches the maximum exit rate of approximately 55 percent for age group 50 t0 55 then decreases after. The bank should look into what is causing customers between ages 40 and 65 lo leave the bank. Perhaps, the bank does not have good investments product to keep these age group who are trying to maximize their 401K value and prepare for retirement.

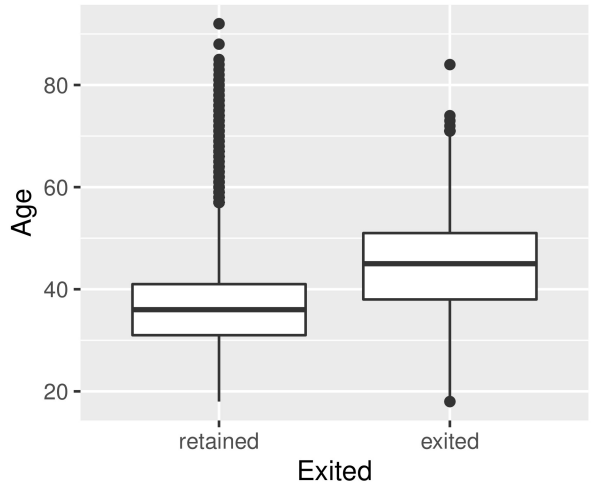


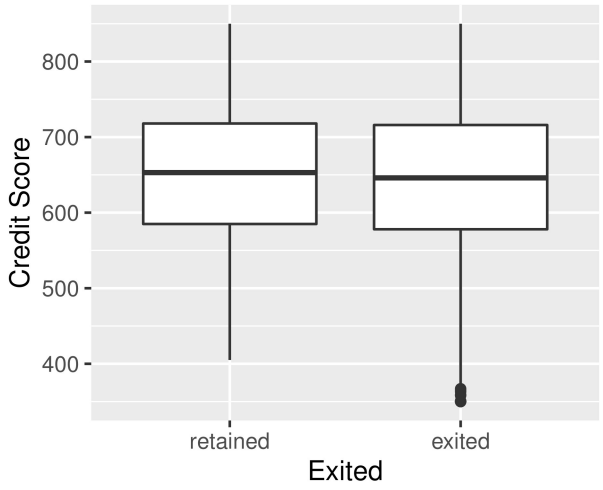
To understand more about the customers in the bank based on various parameters like Estimated Salary, Balance, Age, Credit Score, we created a few box plots of each of the variable with respect to the target variable “Exited”.



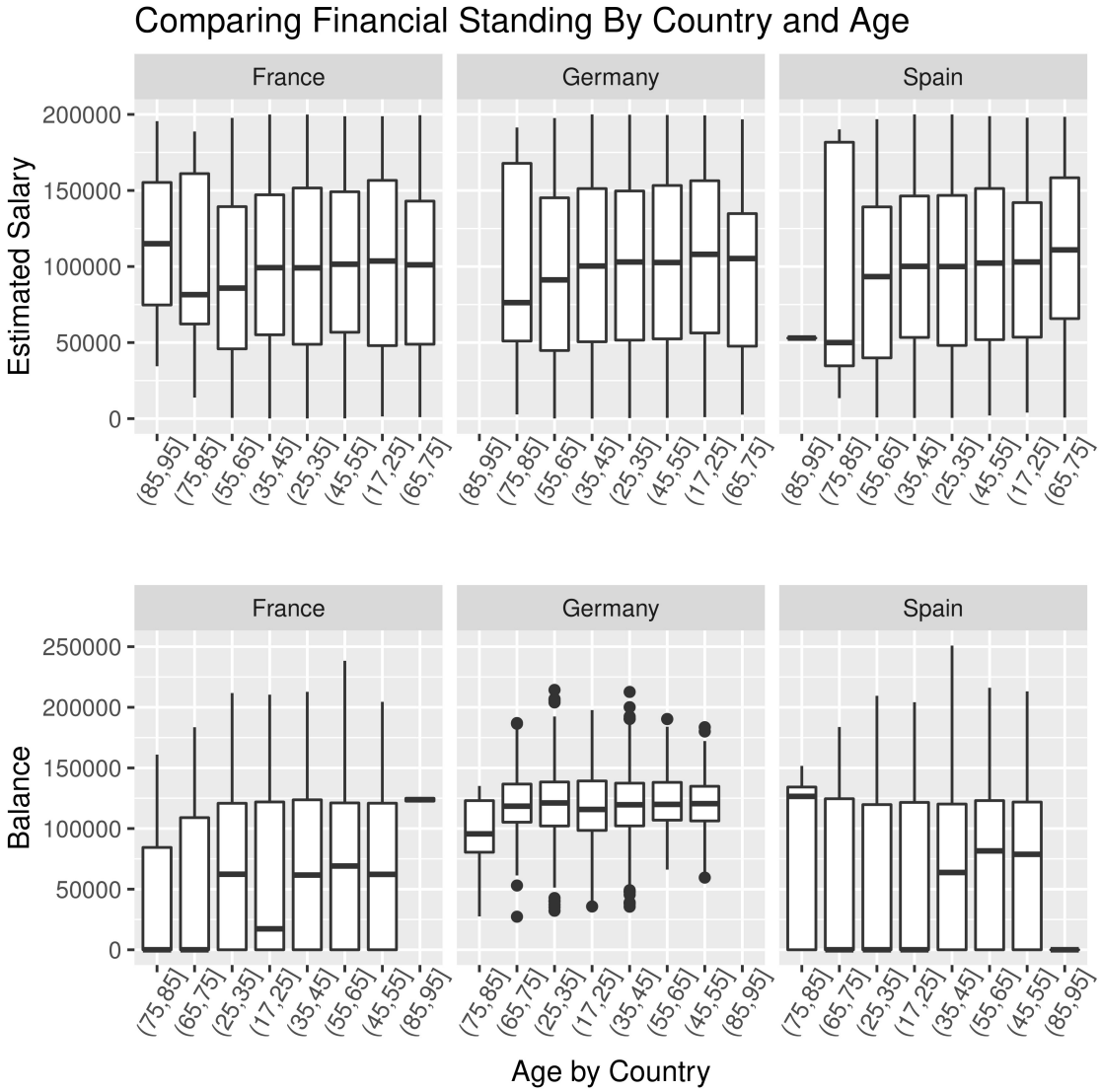


From the above two box plots, it is visible that the median Estimated Salary of the people who have exited and the median Estimated Salary of the people who have been retained by the bank almost the same. And, in the box plots of Balance with respect to the target variable shows that median Balance of the people who have exited and the median Balance of the people who have been retained by the bank are almost the same, the spread of Balance of the people retained by the bank is comparatively larger than the spread of Balance of the people who have exited. This is surprising because, this may imply that people who have had relatively more Balance in their account have exited.

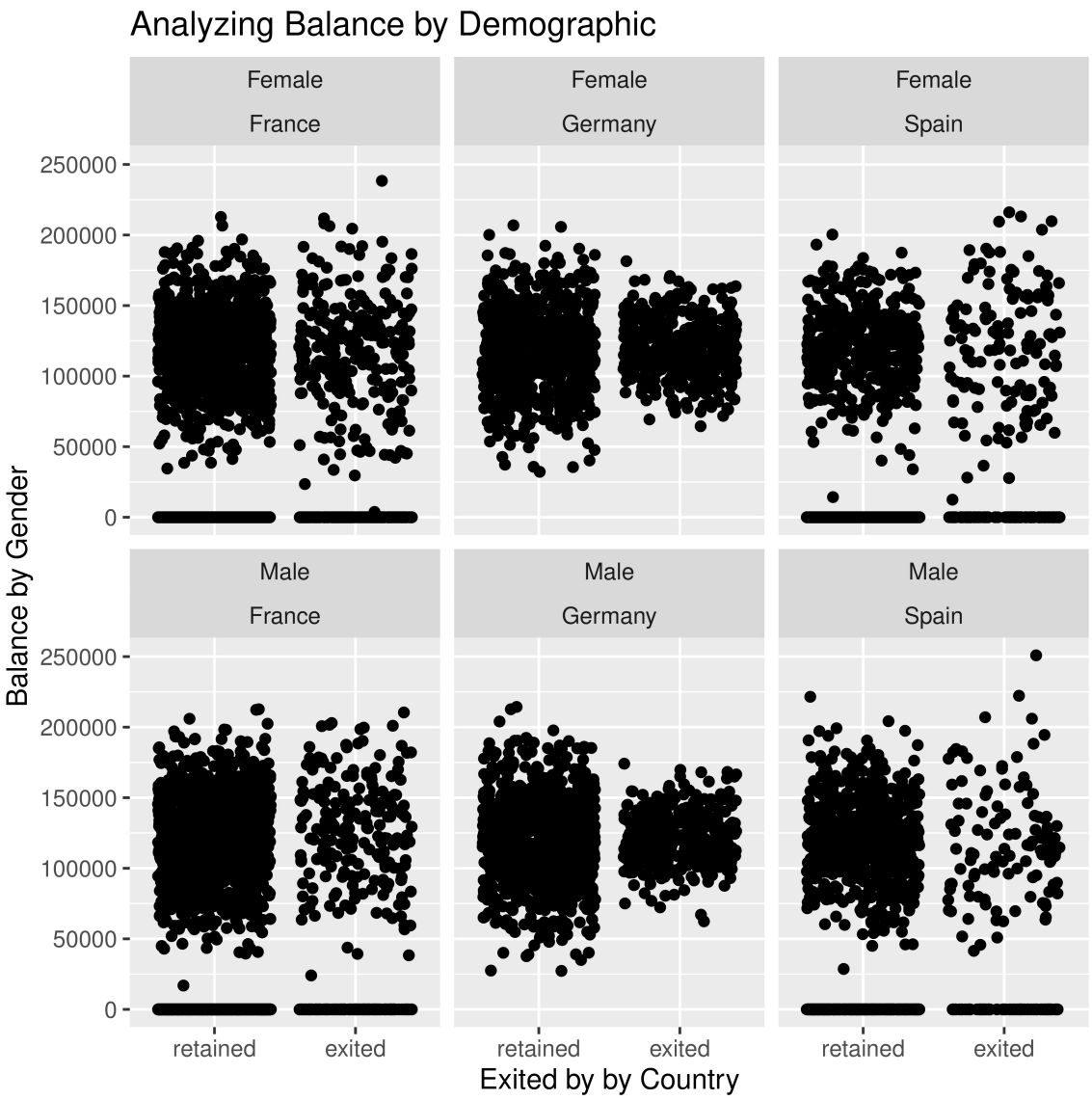




Here, the median Age of the people who have exited is higher the median Age of the people who have been retained by the bank. This could mean that the bank is relatively better at retaining younger customers as opposed to older customers. Nevertheless, there are a lot of outliers for the people who have been retained by the bank which could suggest that these customers may be some privileged customers with some senior citizen perks. Yet, the median Credit Score of the people who have exited and the median Credit Score of the people who have been retained by the bank almost the same indicating that Credit Score doesn't play a pivotal role in determining whether or not the customer exits the bank.



As we can see from the above box plots, while the Estimated Salary is spread almost evenly across all the countries, the Balance in Germany is way too concentrated between 100,000 and 150,000 with a lot of outliers.



From the above scatter plot, we can see that both Males and Females in Germany show a similar distribution of Balance with respect to the customers who have exited the bank. Although, females have a higher spread in balance when compared to males.

Modeling Analysis

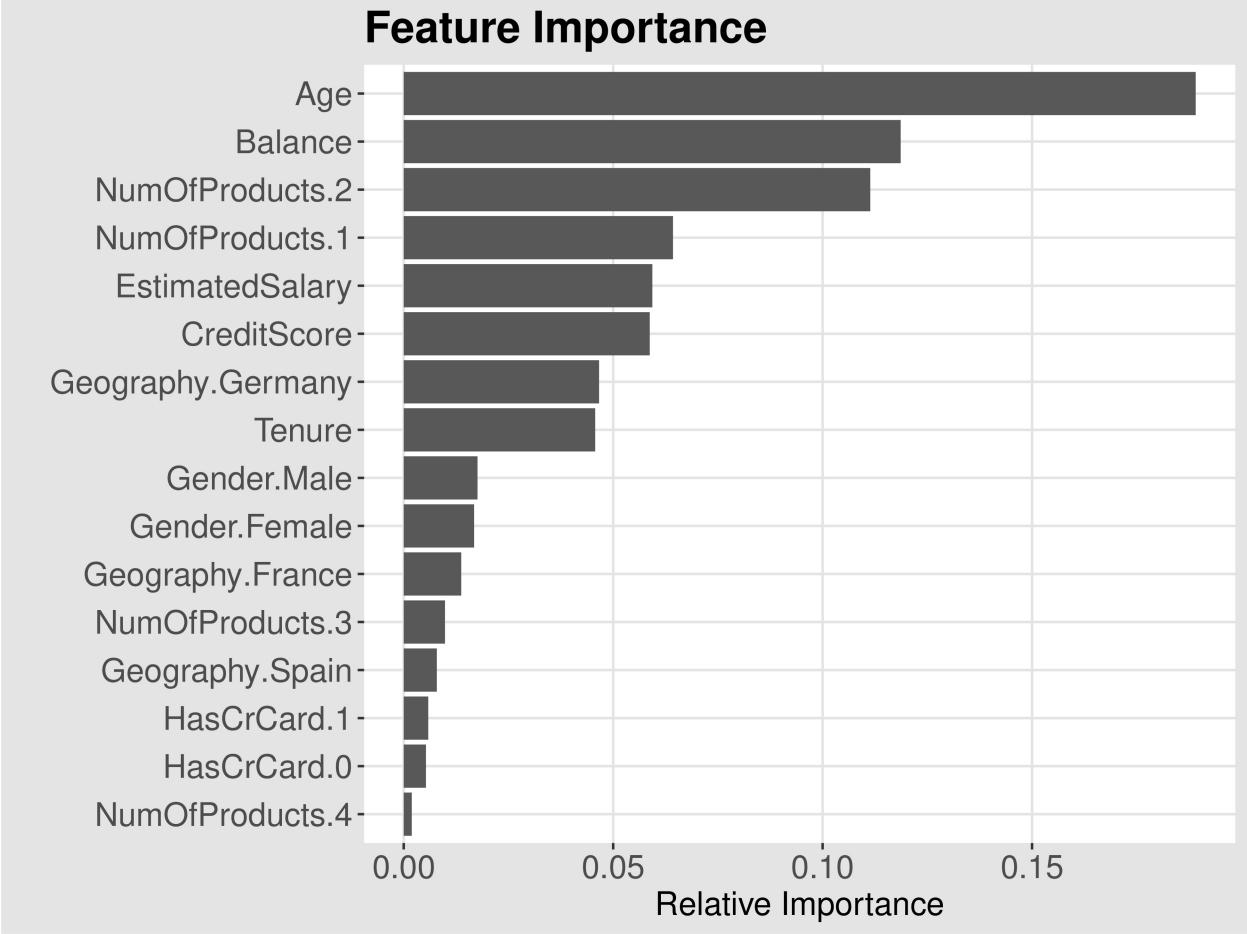
Our ultimate goal for this analysis is to help the bank predict customers who are going to exit the bank, presumably so that the bank could offer those customers rebates, or reduced rates in order to entice them to stay. Therefore, we chose to focus on minimizing the false negative rate, or the number of times a customer exited the bank, and our model did not predict so. Theoretically, this could be achieved by always predicting that a customer was going to exit, but in the business environment this would result in awarding everyone the rebate, improved rate, etc., and wouldn't help the bank improve profits. For this reason, we also looked at total model accuracy as a secondary metric.

Also, our data set contained a class imbalance, with only ~20% of the instances being exited customers. Therefore, after randomly splitting the data into the 70/30 training/test sets, we randomly over sampled the training set until we had an class balance. This helped prevent the model from over fitting to the "retained"

classification.

```
##
## retained   exited
##     7963     2037
```

Before modeling, we quickly looked at feature importance. The plot below reiterates themes we saw in our data exploration, with age and balance being ranked as valuable features for classification.



Baseline Models: Logistic Regression and Naive Bayes

These two models were used as our starting point for model performance. We found that after verifying that our team was working with the same version of the data set, the performance for logistic regression was much better than previously reported, and the Naive Bayes actually fit to the negative, "retained" class, showing poor performance for the sensitivity.

Logistic Regression Confusion Matrix:

```
##     predicted
## true 0         1
##    0 1791      597       tpr: 0.69 fnr: 0.31
##    1 188       424       fpr: 0.25 tnr: 0.75
##      ppv: 0.42 for: 0.09 lrp: 2.77 acc: 0.74
##      fdr: 0.58 npv: 0.91 lrm: 0.41 dor: 6.77
```

Naive Bayes Confusion Matrix:

```
##     predicted
## true 0         1
##    0 2374      14        tpr: 0.11  fnr: 0.89
##    1 546       66        fpr: 0.01  tnr: 0.99
##      ppv: 0.82 for: 0.19 lrp: 18.39 acc: 0.81
##      fdr: 0.18 npv: 0.81 lrm: 0.9   dor: 20.5
```

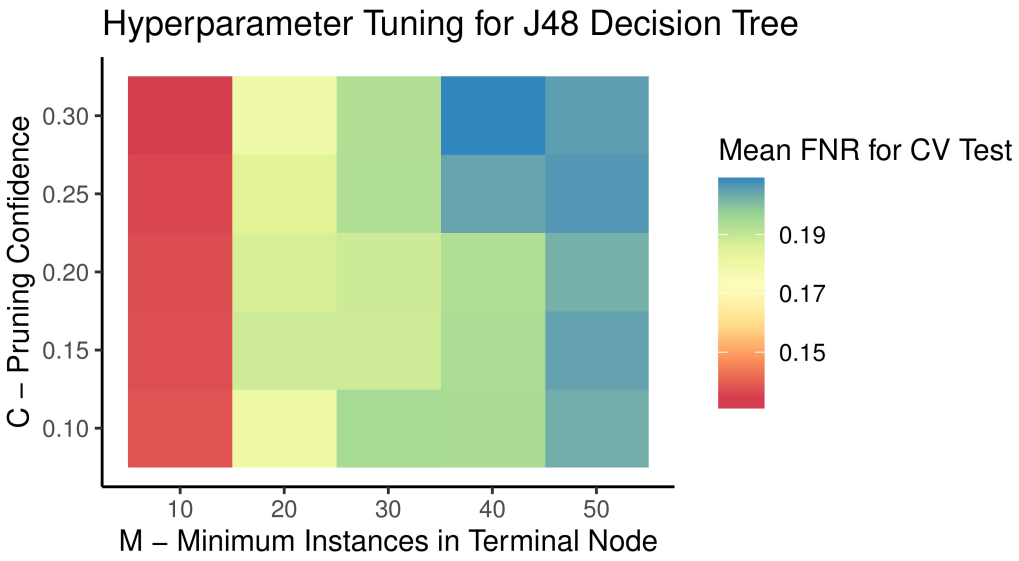
A low value for sensitivity was considered poor, as 1 - sensitivity = the false negative rate, our primary metric.

J48 Decision Tree

Next we tried a decision tree model. It's one of the more basic concepts for classification modeling, but has some parameters that are able to be tuned. After tuning, we found the overall accuracy of this model to be higher than logistic regression, but a worse sensitivity, and the opposite when compared to Naive Bayes.

J48 Model Confusion Matrix after Hyperparameter Tuning:

```
##     predicted
## true 0         1
##    0 1936      452       tpr: 0.58 fnr: 0.42
##    1 255       357       fpr: 0.19 tnr: 0.81
##      ppv: 0.44 for: 0.12 lrp: 3.08 acc: 0.76
##      fdr: 0.56 npv: 0.88 lrm: 0.51 dor: 6
```



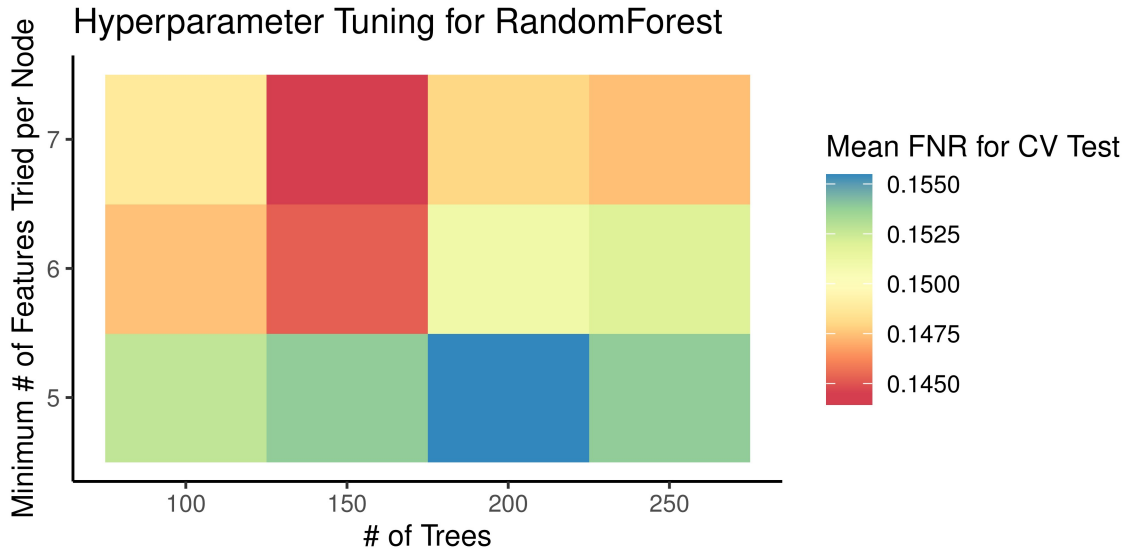
The hyperparameters for this model were tuned via the grid search method. Each combination of hyperparameters within the grid was evaluated for its mean false negative rate over a 5-fold cross validation. From the figure above, you can see that the smaller number of allowed instances in the end node showed improved performance, however we found that the lower value we went, the model began to over fit to our training set, and thus capped the minimum number of instances in a leaf node to 10.

Random Forest

The random forest algorithm uses a large number of simpler decision trees to eliminate noise, and better perform on the test hold out set. We explored tuning the number of trees used by the model, as well as the number of features to try at each node.

Random Forest Model Confusion Matrix after Hyperparameter Tuning:

```
##     predicted
## true 0        1
##    0 1987     401      tpr: 0.66 fnr: 0.34
##    1 211      401      fpr: 0.17 tnr: 0.83
##      ppv: 0.5 for: 0.1 lrp: 3.9  acc: 0.8
##      fdr: 0.5 npv: 0.9 lrm: 0.41 dor: 9.42
```



This model had a better sensitivity and accuracy compared to the J48 model, but the logistic regression still maintains the highest pure sensitivity value

Support Vector Machine

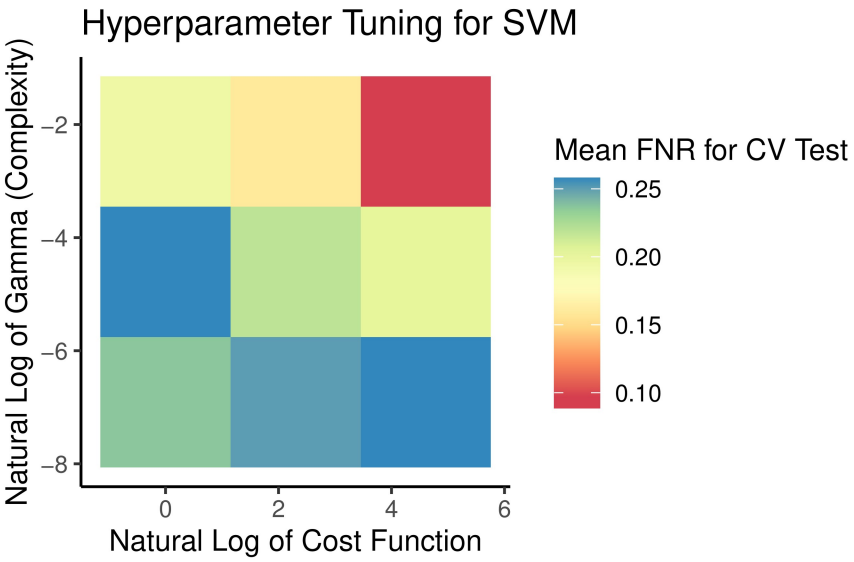
Lastly, we trialed a support vector machine model. This model surprisingly performed worse after the parameter tuning. We suspect that this was due to over fitting. That being said, the default model parameters resulted in our highest sensitivity, albeit a lower accuracy than the tuned random forest model

Before Tuning:

```
##     predicted
## true 0         1
##    0 1843      545       tpr: 0.7  fnr: 0.3
##    1 181       431       fpr: 0.23 tnr: 0.77
##      ppv: 0.44 for: 0.09 lrp: 3.09 acc: 0.76
##      fdr: 0.56 npv: 0.91 lrm: 0.38 dor: 8.05
```

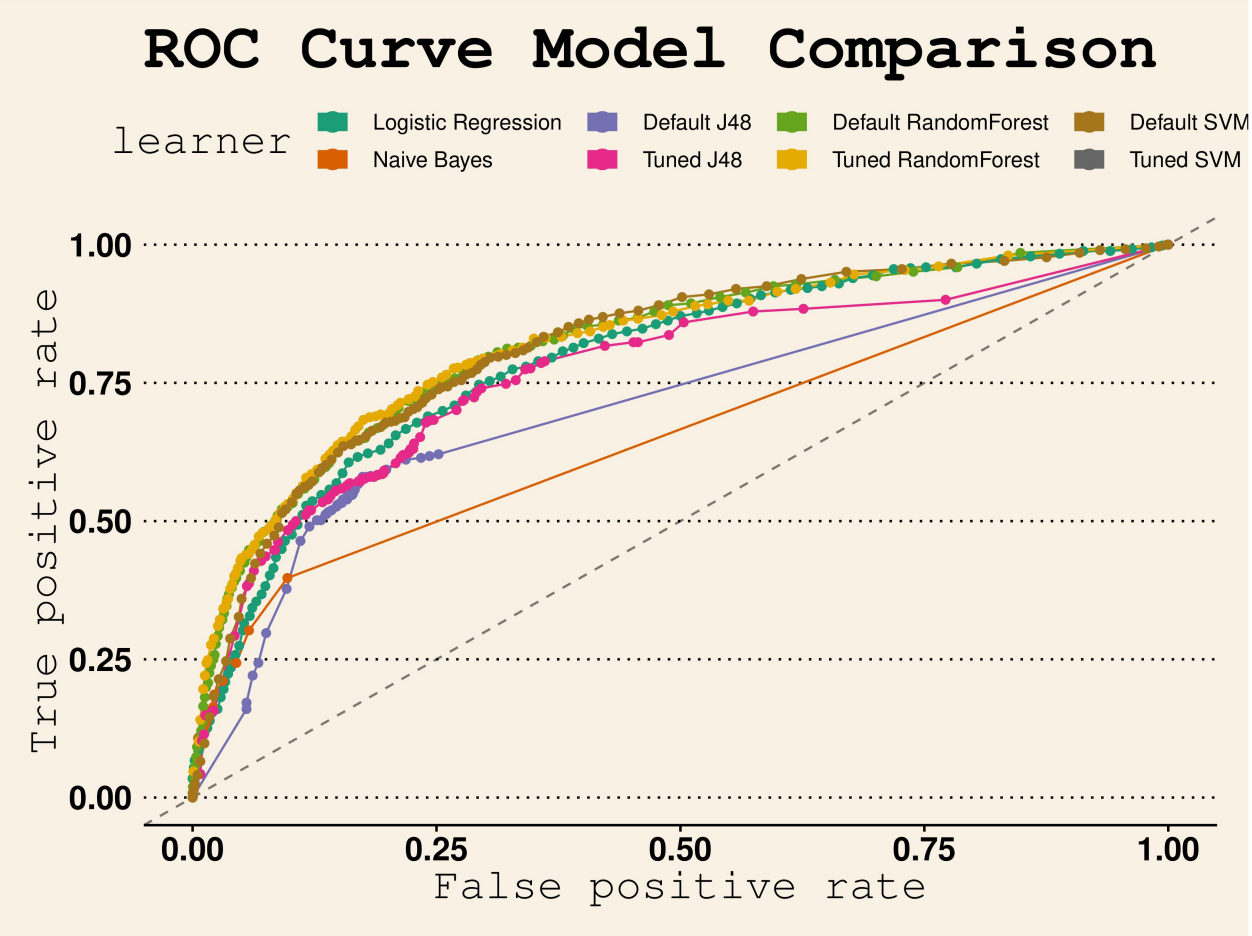
After:

```
##     predicted
## true 0         1
##    0 1919      469       tpr: 0.58 fnr: 0.42
##    1 257       355       fpr: 0.2  tnr: 0.8
##      ppv: 0.43 for: 0.12 lrp: 2.95 acc: 0.76
##      fdr: 0.57 npv: 0.88 lrm: 0.52 dor: 5.65
```



ROC Curve

We Also compared each model via their ROC curves, looking at their prediction accuracy across different thresholds of false positive error rates. We were looking for the largest area under the curve to rate our models, and the two random forest models, as well as the default SVM model, fit this criteria was intuitive considering their confusion matrix results.



Conclusion

In conclusion, we were able to dive deeper into identify some key variables and create some models useful for retaining customers. From our exploration we were able to see that the majority of the banks business are males between 25 and 50. Males also had a smaller exit percentage then females, giving the bank an opportunity to improve the interactions with female customers. While the customer base is mainly between 25 and 50 the rate of customers leaving increases as customers as age increases up until age 55. We also were able to identify that Germany has a higher Exit rate than the other countries in our data set. From our random forest feature importance we were able to identify age, balance in accounts, 2 or more credit products, and credit score. Another key takeaway from our models is if decision makers want to avoid false positives, thus avoiding spending on customers that they would have retained anyway they should look at using an SVM. When trying to predict identify which customers will exit with the higher accuracy the decision maker should use a random forest model.