Predicting Bank Churn

Scott Rushford

2024-12-04

Introduction

The ABC Multinational Bank, operating in France, Germany and Spain, is looking to improve customer retention. They have provided customer data of account holders found on Kaggle to use for this purpose. (Topre, n.d.) The object of this project is to develop a classification algorithm using R programming to predict bank churn.

The data frame consists of 10,000 rows of bank customers containing 12 columns of customer information.

1. customer ID - a number used to identify the customer
2. Credit Score - a numeric value indicating the customer's current credit score
3. Country - a factor of 3 country categories: France, Germany, Spain indicating the customer's country of residence.
4. Gender - a factor of 2 categories: Male, Female.
5. Age - a numeric value indicating the customer's age.
6. Tenure - the number of years the customer has been with the bank.
7. Balance - numeric value indicating the current bank balance.
8. Number of Products - a numeric value indicating the number of bank products the customer holds.
9. Credit Card - a factor of 2 categories represented by 0 = no card, 1 = card.
10. Active Member - a factor of 2 categories represented by 0 = not active, 1 = active with a mean of 0.5151.
11. Estimated Salary - the customer's salary as estimated by the bank.
12. Churn - Churn is a factor of 2 levels represented by 0 = customer, 1 = churn with a mean of 0.2037 and is the target for prediction.

The key steps involved in this project included:

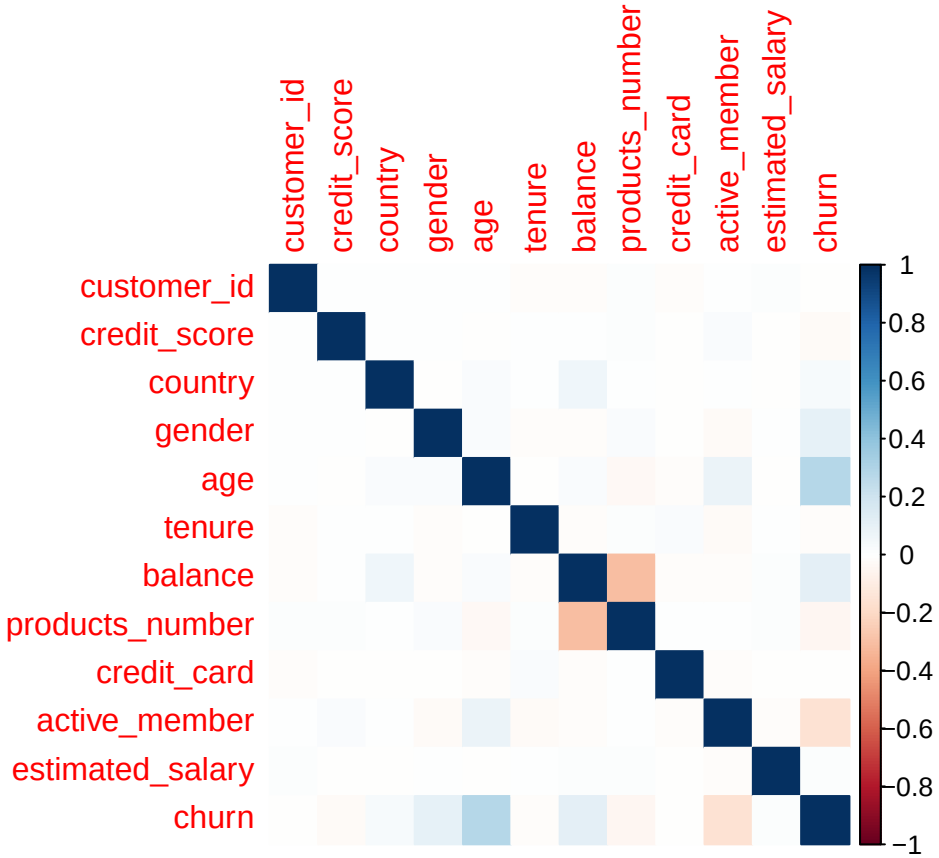
1. Correlation and Principal Component Analysis
2. Data Visualization
3. Establish Baseline Measurements
4. Test, Train and Validate machine learning algorithms.
5. Review Results

Methods & Analysis

Correlation and Principle Component Analysis

To measure correlation amongst the variables and component importance in the data set columns a Correlation Analysis (Soetewey, n.d.) and a Principle Component Analysis was run on the data. The correlation analysis was plotted to visualize the results. (Wei and Simko 2024). The plot does not show any correlation greater than 0.5, but does suggest some correlation between churn, gender, age and bank balance.

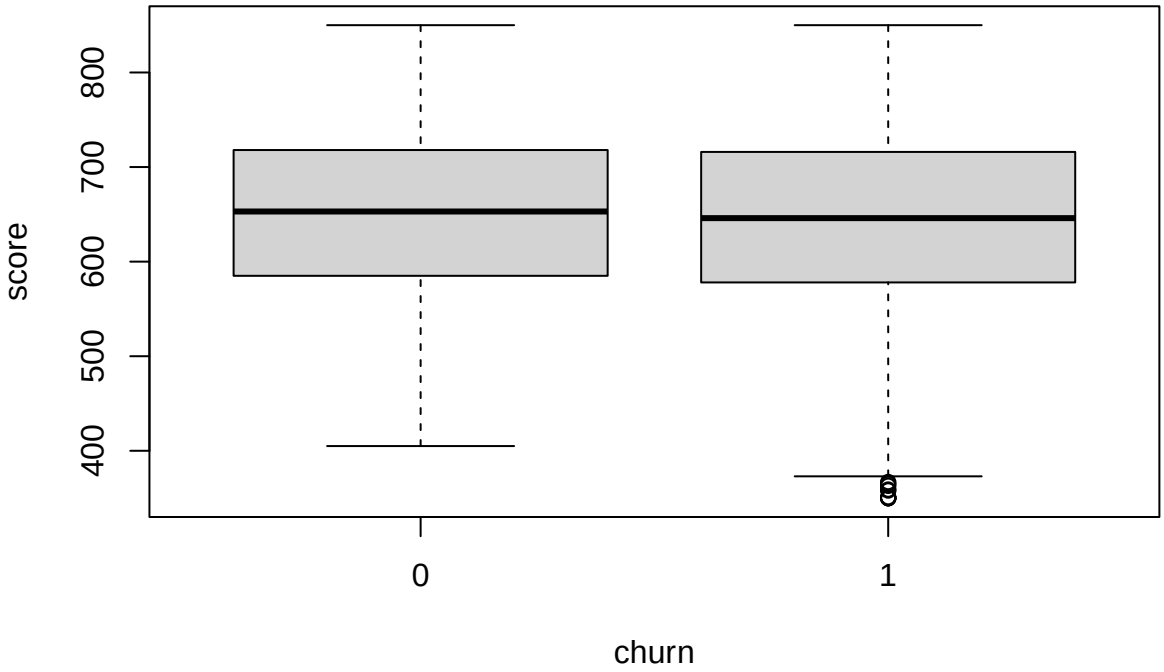
The proportion of variance in the principal component analysis ranges from 5.5% to 13.07%, suggesting that there is no strong component that can explain the majority of the data.



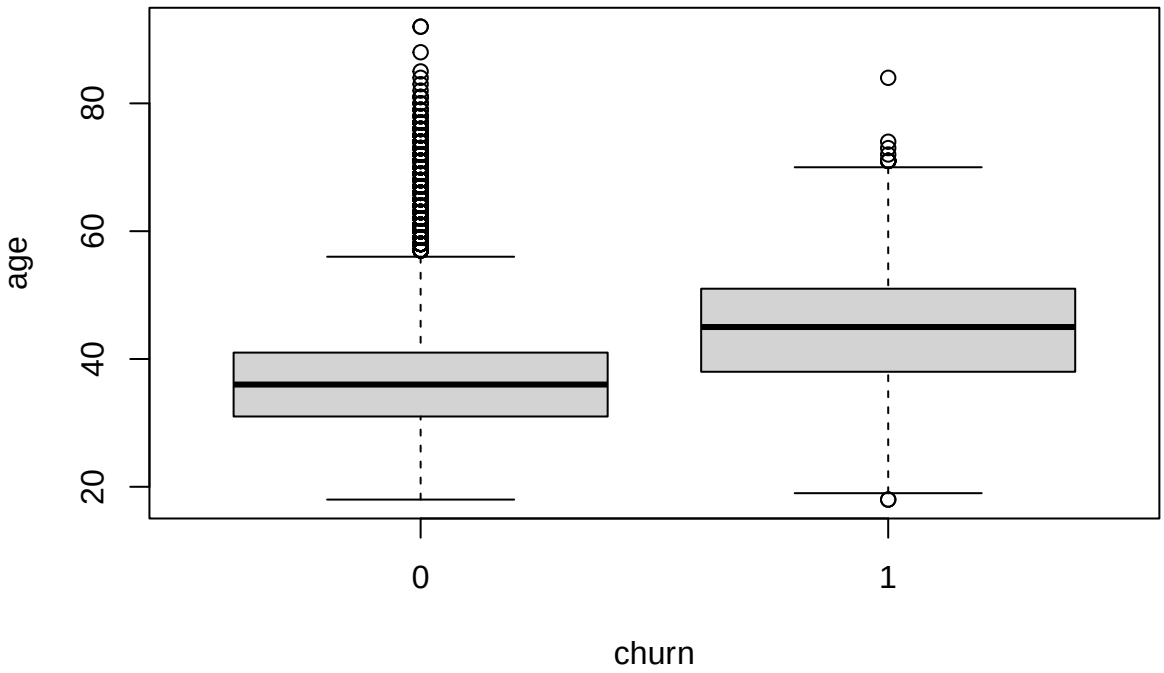
```
## Importance of components:
##                            PC1    PC2     PC3     PC4     PC5     PC6     PC7
## Standard deviation     1.1992 1.1049 1.04718 1.01018 1.00746 0.99756 0.99512
## Proportion of Variance 0.1307 0.1110 0.09969 0.09277 0.09227 0.09047 0.09002
## Cumulative Proportion  0.1307 0.2417 0.34141 0.43418 0.52645 0.61692 0.70694
##                            PC8     PC9    PC10    PC11
## Standard deviation     0.98566 0.97220 0.83490 0.78096
## Proportion of Variance 0.08832 0.08592 0.06337 0.05545
## Cumulative Proportion  0.79526 0.88118 0.94455 1.00000
```

Data Visualization

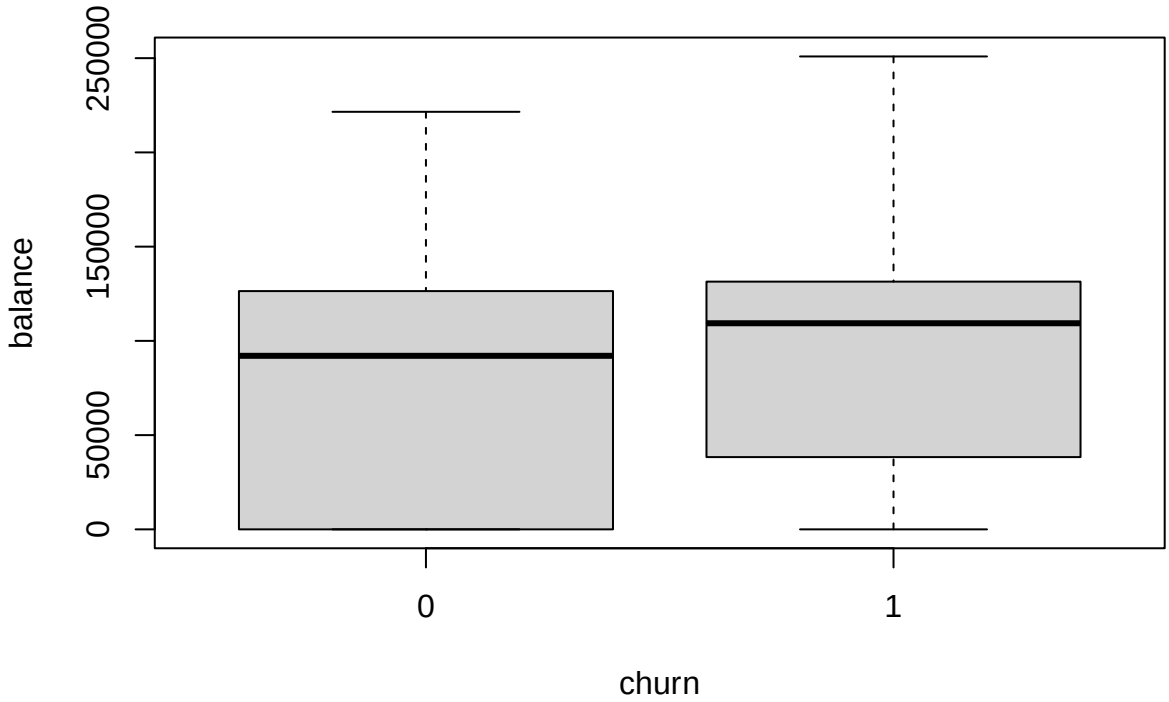
Box plots, histograms, bar graphs and line plots were created to visualize the data to gather additional insights.



The box plots for churn and credit score show similarities in median, and upper and lower quartiles. The lower extreme for churn (1) extends below that of customers (0) with outliers which would indicate that a credit score of less than that 400 would increase the probability of churn.



The median, upper and lower quartiles and upper extreme are higher for churn (1) than customer (0). The box plot suggests that the probability of churn increases as customers become older than 40 and that customers over 60 are considered outliers.

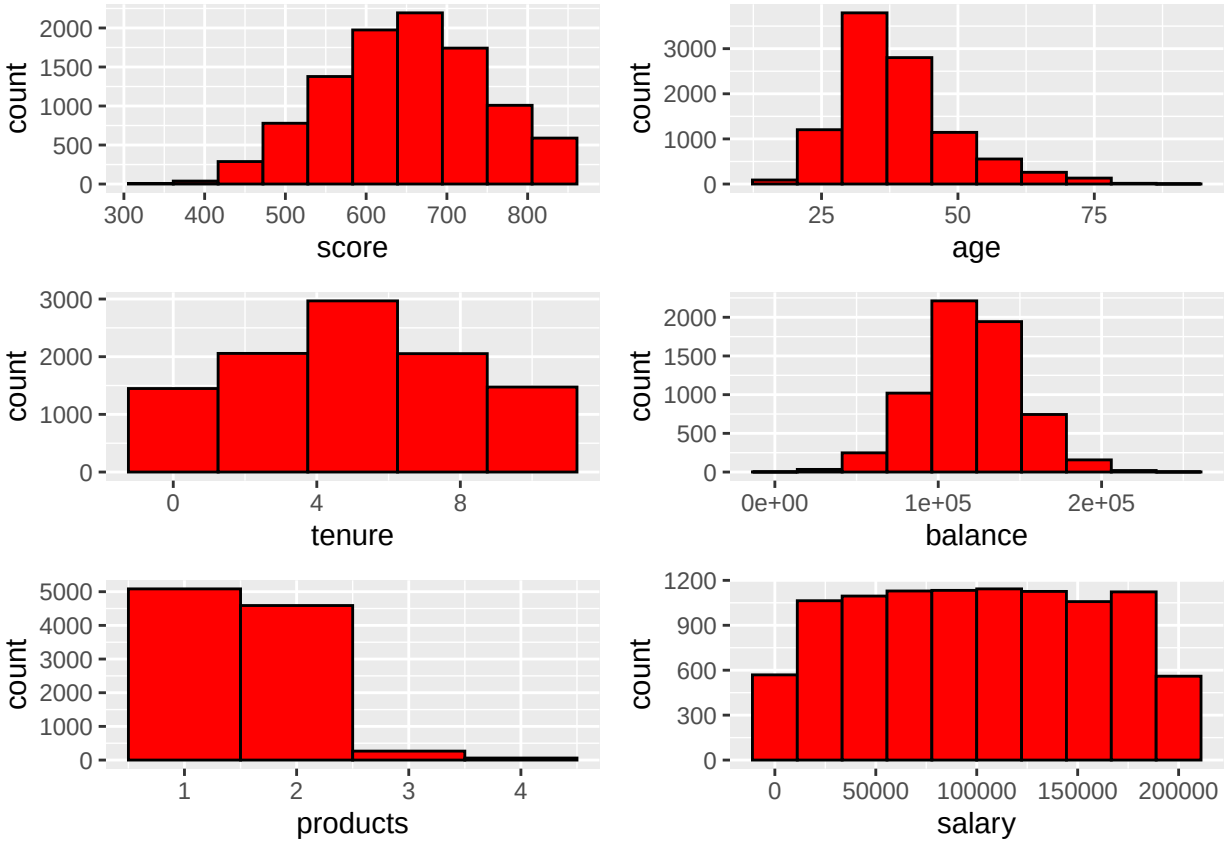


The median, upper and lower quartiles and upper extreme are higher for churn (1) than customer (0). The box plot suggests that the probability of churn increases as customers begin to carry a bank balance that exceeds 150,000.

(2023)

Histograms

Histograms were created for 6 variables (credit score, age, tenure, bank balance, products and salary) to visualize their distribution among the bank's customers.



(Auguie 2017)

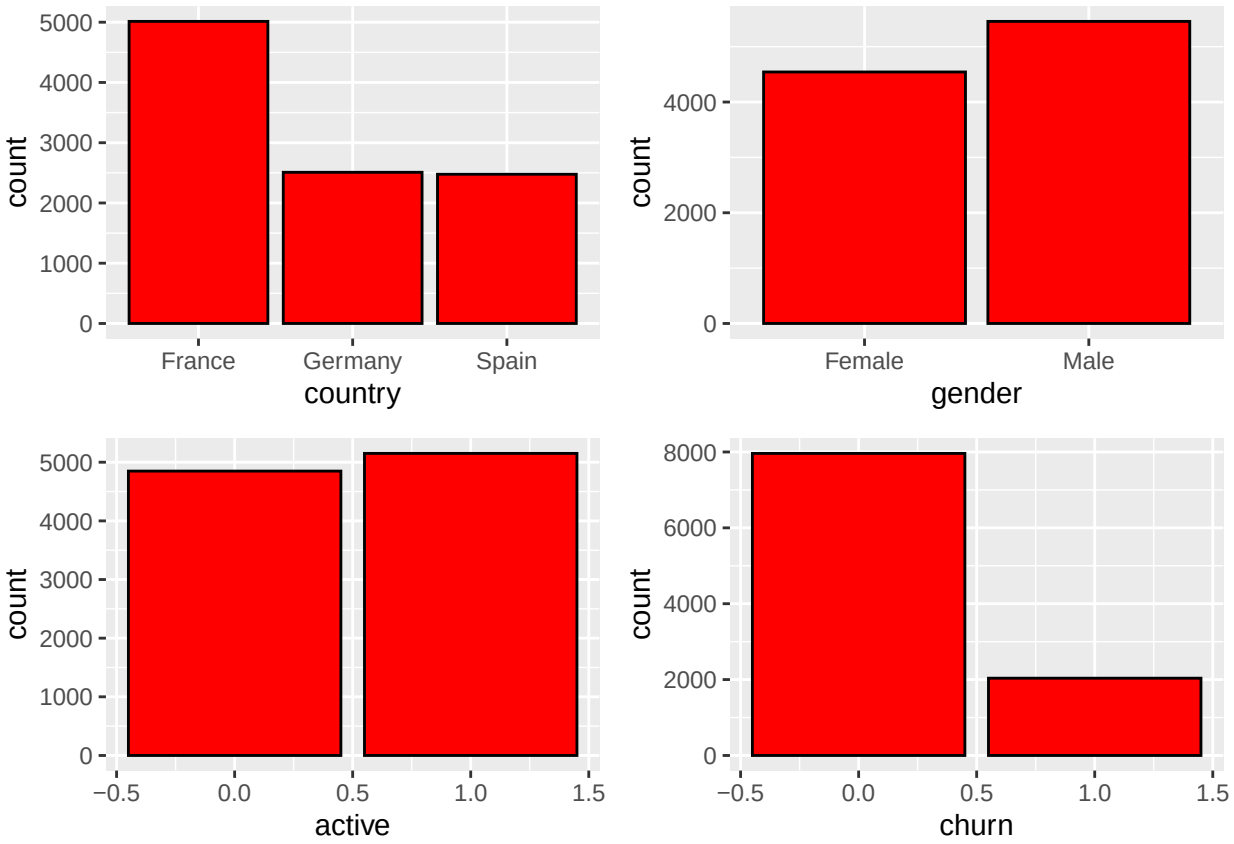
(Wickham 2016a)

From examining the histograms a profile of the typical ABC Bank customer begins to emerge:

1. a credit score greater than 500
2. under the age of 50
3. a tenure with the bank of 4 to 6 year duration
4. maintains a bank balance of 100,000 to 150,000
5. has 1 or 2 of the bank's products.

Bar Graphs

Bar graphs were created to provide additional insight into the customer base.



1. Approximately 5,000 customers reside in France with 2,500 each living in Spain and Germany.

2. Males make up 55% of the customer base and 45% are Female.

3. Active members make up over 50% of the data set.

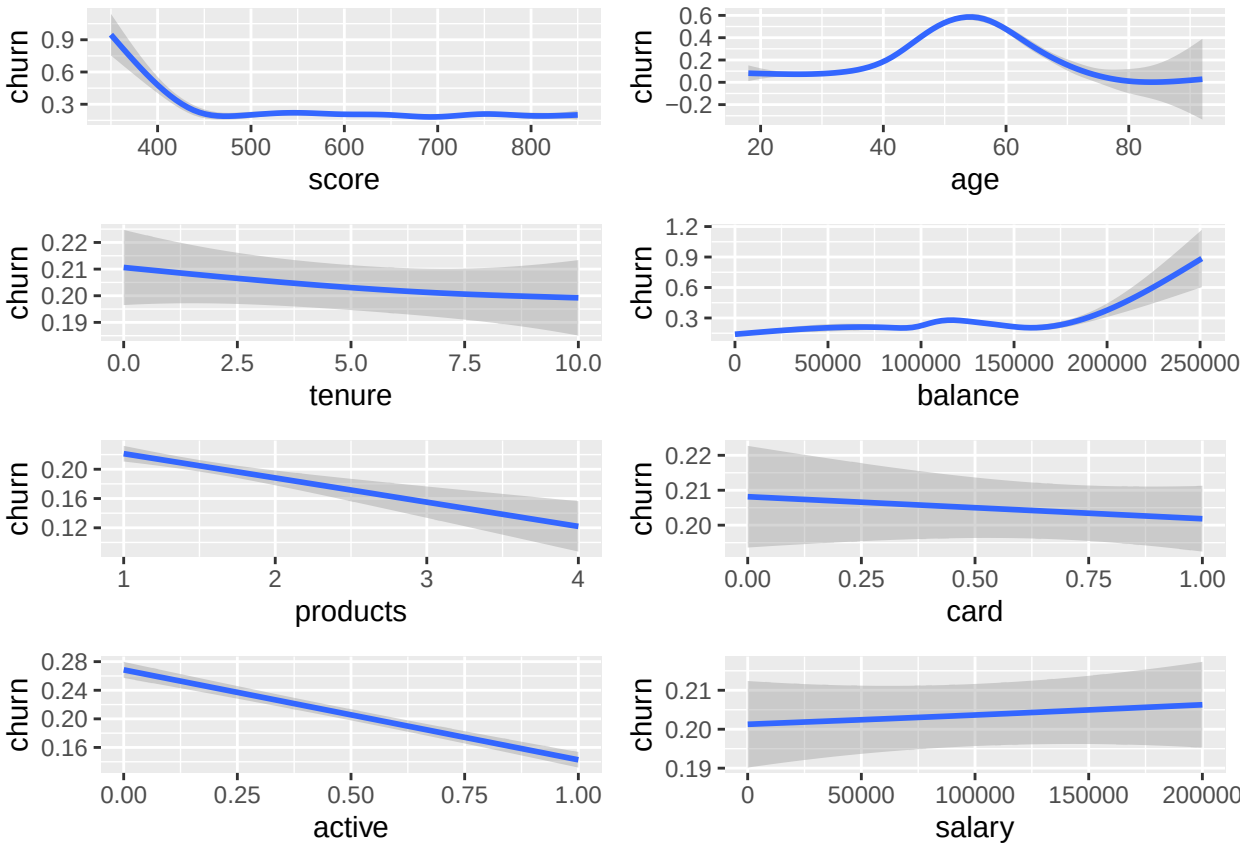
4. The churn rate is over 20%.

Line Plots

Line plots were created to compare churn, on the x axis, against all key variables, on the y axis. Since this is a classification problem, the GAM (General Additive Model) method was used for most of the plots except for the active, products and card variables where the LM (Linear Model) method was used. (Wickham 2016b)

```
## `geom_smooth()` using formula = 'y ~ s(x, bs = "cs")'
## `geom_smooth()` using formula = 'y ~ s(x, bs = "cs")'
## `geom_smooth()` using formula = 'y ~ s(x, bs = "cs")'
## `geom_smooth()` using formula = 'y ~ s(x, bs = "cs")'
## `geom_smooth()` using formula = 'y ~ x'
## `geom_smooth()` using formula = 'y ~ x'
```

```
## `geom_smooth()` using formula = 'y ~ x'
## `geom_smooth()` using formula = 'y ~ s(x, bs = "cs")'
```



The line plots confirm the observations seen in other visuals: churn rate increases for customers with a credit score less than 500. The probability of churn increases for customers over 40 years of age, peaks at age 50 and then decreases at a similar rate. The probability of churn increases as customers begin carrying account balances greater than 150,000.

Baseline

In order to compare the models developed in this project, 2 baseline models were created. The first model consisted of a random draw of 0s and 1s producing a random list of 5000 zeros and 5000 ones. When the random list was compared to the actual data set it predicted the actual at an accuracy rate of 50.02%. This model is referred to as: r_n.

The second model is a random set of 0s and 1s in proportion to what appears in the data set. This produced a random list with 79.63% of the numbers being 0 and 20.37% of the numbers being 1. When this model was compared to the actual data set it was accurate to 67.33%. This model is referred to as: r_p.

Data Preparation

In order to use the machine learning algorithms the data was changed in the following ways:

1. Columns were renamed for simplification.
2. Country and Gender were changed from character strings to factors.
3. The variables active, card and the predicted value churn were converted from numeric to factors. (Bobbitt 2023)
4. Customer ID and Estimated Salary were removed from the data set. The ID number is a unique or independent variable so it has no predictive value. As salary is an estimated value and there is no data indicating how that estimate was made or its accuracy, the variable was removed.
5. The data was normalized for classification.

Optimize, Train, Test and Validate

In order to provide enough data to the training set without creating an issue with over training, 60% of the data was allocated to training with 20% saved for each of the testing and validation sets.

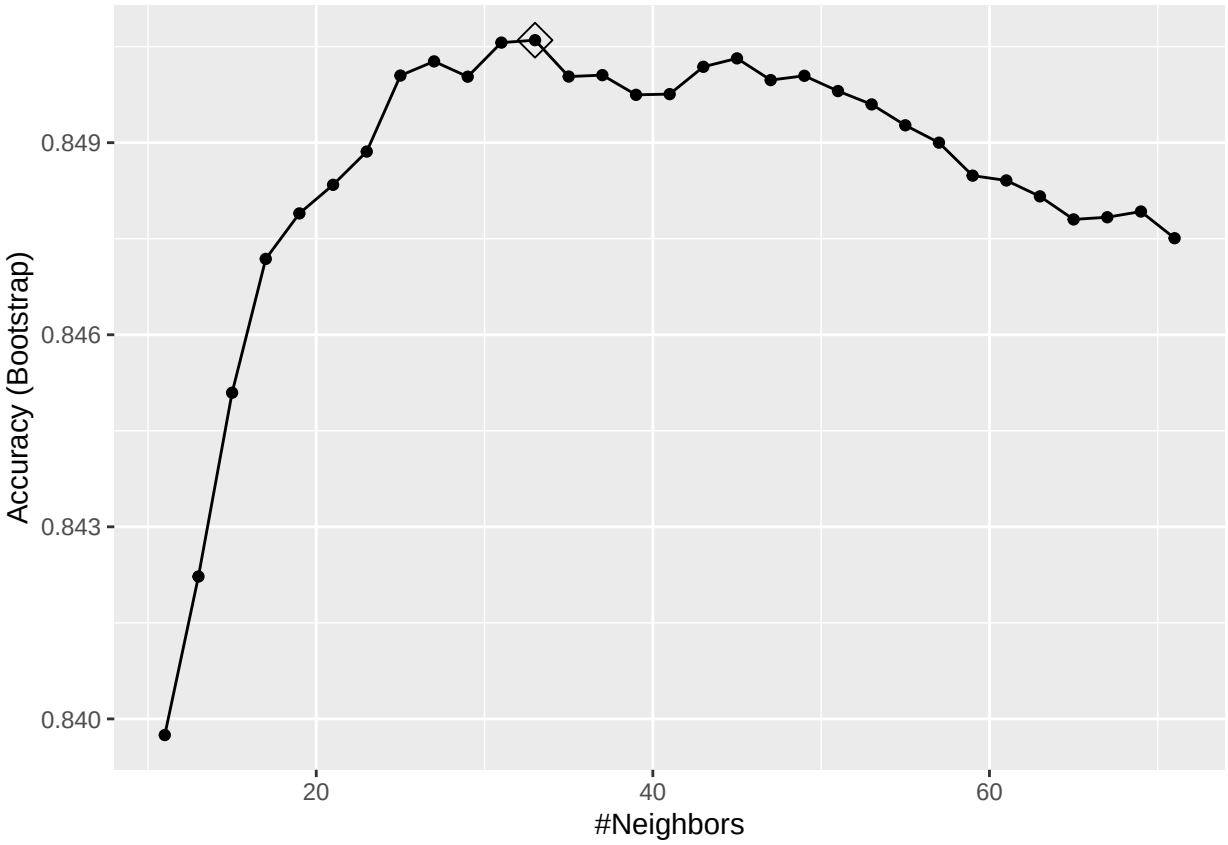
The following process was followed for each model

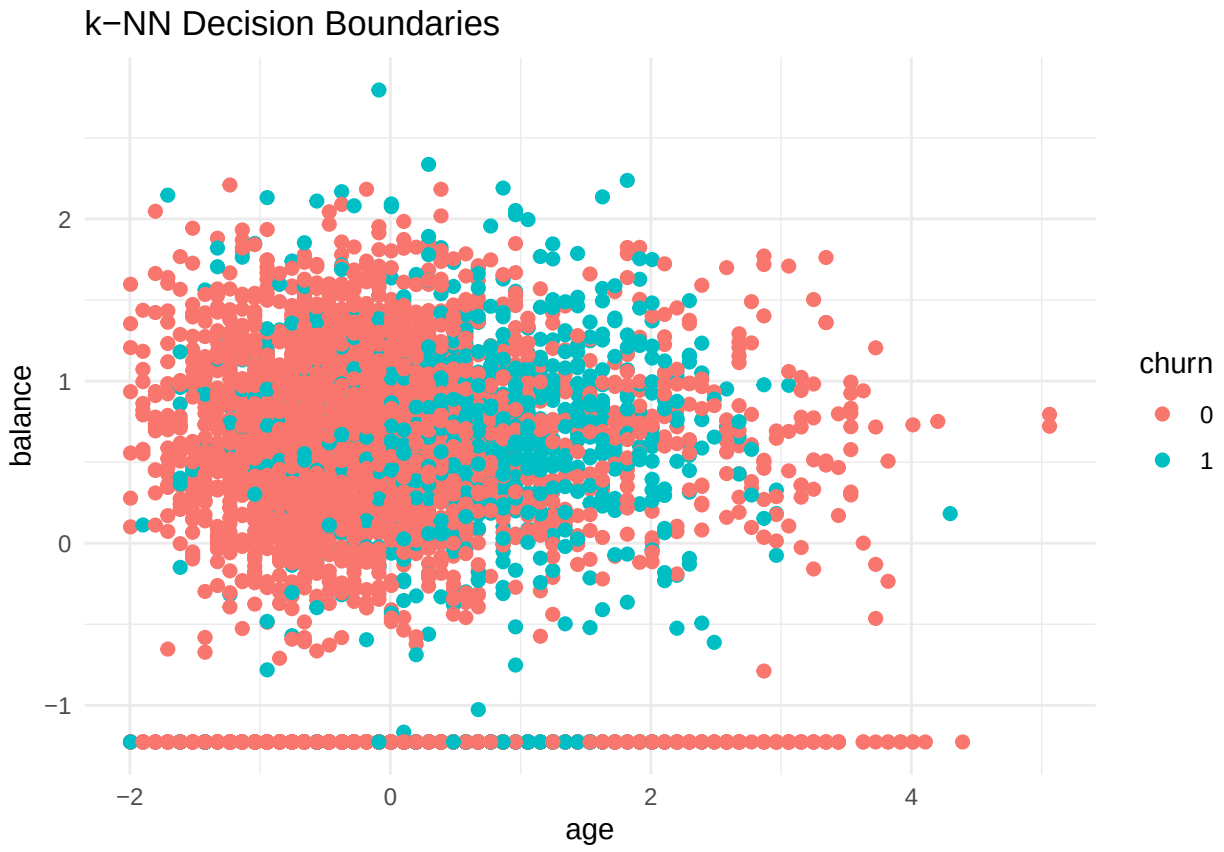
1. Optimization of the model parameters using cross validation and bootstrapping methods
2. Train Model
3. Test Model
4. Validate Model
5. Evaluate each model based on its overall accuracy and its positive and negative predictive values. The positive value in this instance is 0, being a current customer and the negative value is 1, churn or not being a customer.

k Nearest Neighbor

In order to predict the class of customer, the k Nearest Neighbor algorithm looks at the k nearest values in the training set to predict the new value. (Adler 2012). By using the train function in the caret package it was determined that k was optimized at 33. (Kuhn 2007). This means that each point is compared to the next closest 33 points, the neighborhood, and computes the average of 0s and 1s in each neighborhood.

The following graph demonstrates how the various test of k increase the accuracy of the model until its optimization point at 33. Increasing k beyond 33 decreases the model accuracy.






```
##      Model  Accuracy  CI_Lower  CI_Upper       PPV       NPV
## 1     Test 0.8577141 0.8447002 0.8700262 0.8627959 0.8114478
## 2 Validate 0.8556667 0.8425781 0.8680574 0.8622222 0.7966667
```

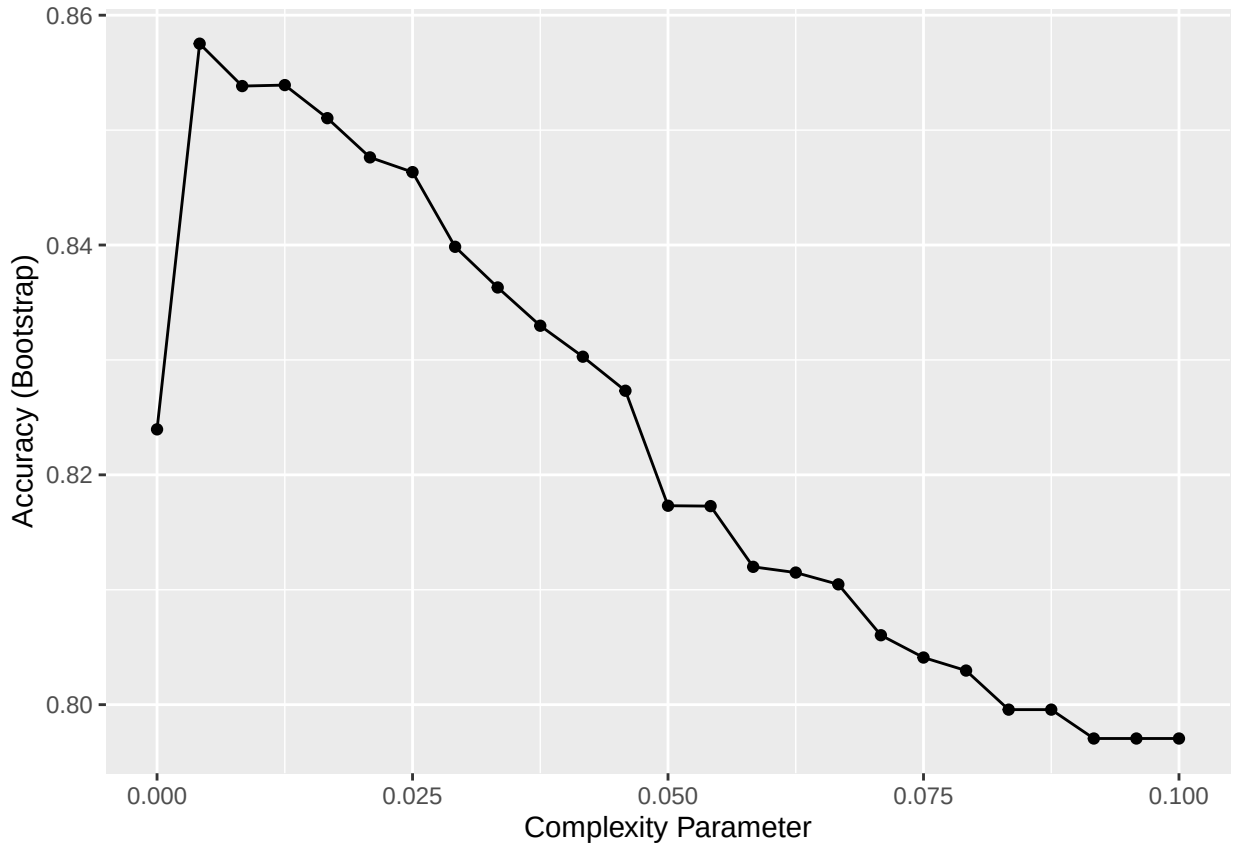
The decision graph confirms that as age and balance increase the algorithm was assigning that as 1 (churn). ("Contour of KNN Model in Ggplot Using r" 2024)

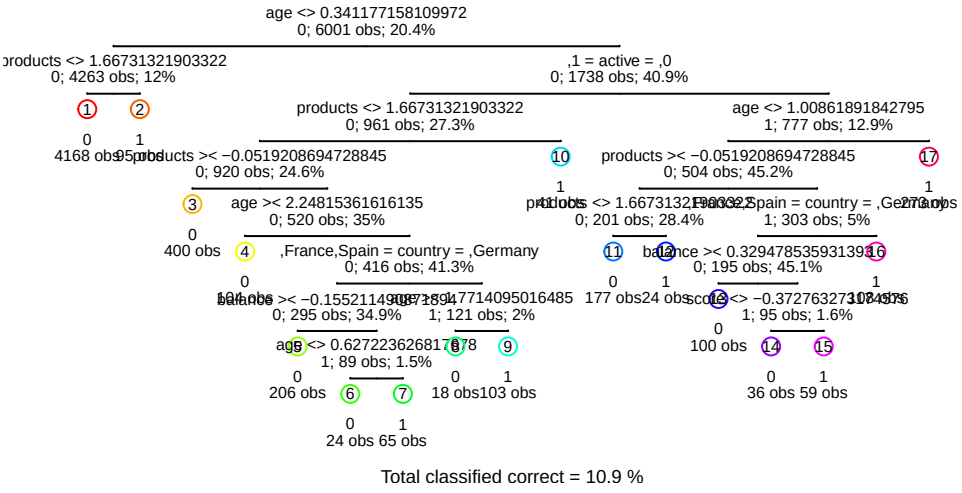
The variance in the accuracy (0.002) of the Test and Validation sets do not show any signs of over-training or over-smoothing.

This algorithm has the ability to predict new values.

Classification Trees

A Classification Tree (Therneau and Atkinson 2023)is created using a complexity parameter which is the minimum improvement required for the tree to create a new node or branch. (Adler 2012). The complexity parameter for this tree is optimized at 0.004166667, as seen in the following graph. The variables used in tree construction are: active, age, balance, country, products and score. The first split of the model occurs at the normalized value for age at 0.34. The model made 1223 correct predictions out of 6001 observations for a root node error of 0.2038.



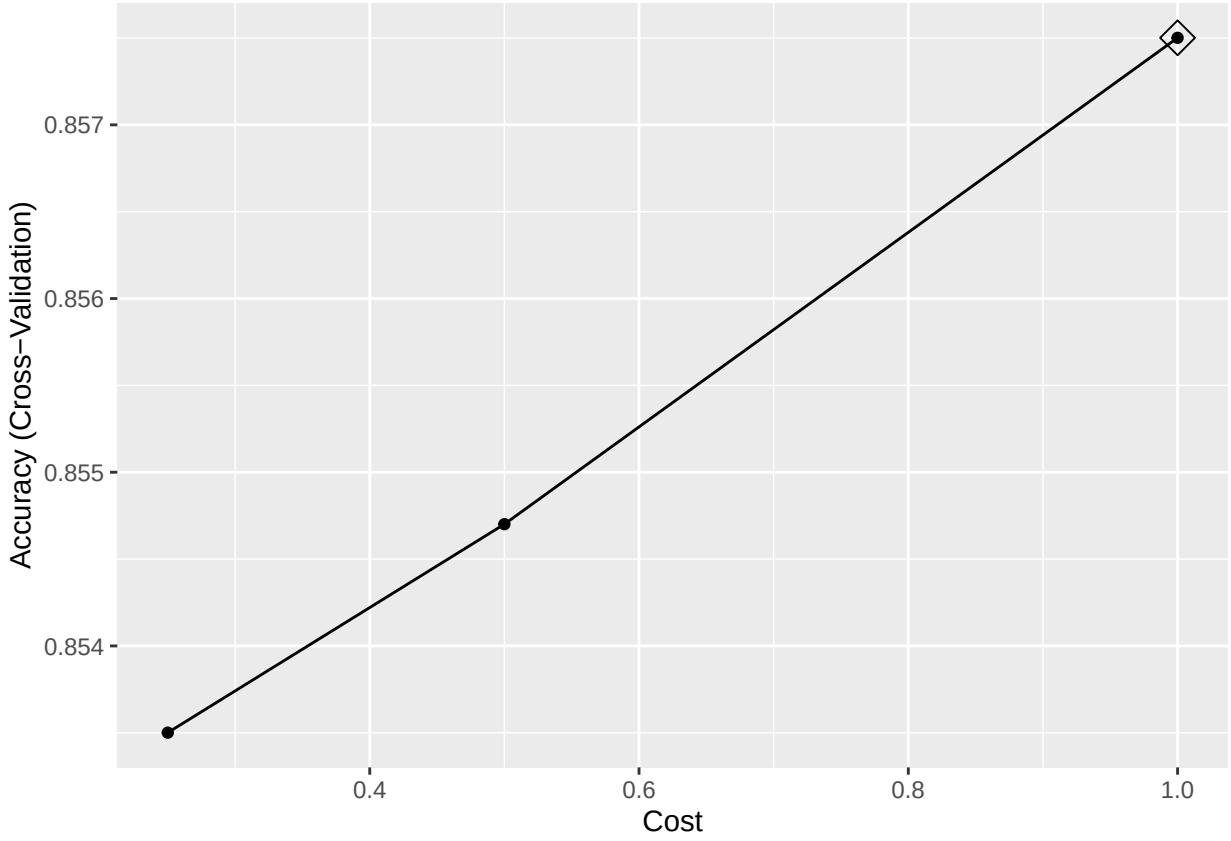



```
##      Model  Accuracy  CI_Lower  CI_Upper       PPV       NPV
## 1     Test 0.8647118 0.8519522 0.8767556 0.8627959 0.8114478
## 2 Validate 0.8636667 0.8508662 0.8757532 0.8622222 0.7966667
```

The tree was created with 17 nodes as illustrated in the above visual. (White and Gramacy 2022)The Test and Validation accuracy variance is 0.001 which would not indicate any issues with the Classification Tree model and it ability to predict new values.

SVM - Support Vector Machine

Support Vector Machines (Meyer et al. 2023) (Karatzoglou, Smola, and Hornik 2024)do not necessarily use all the data to train the model, it may only use some observations which are the support vectors. (Adler 2012). The radial method was used to create the SVM model with the parameters set to sigma = 0.06574296 and Cost = 1 from the results of optimization which created 2154 support vectors.



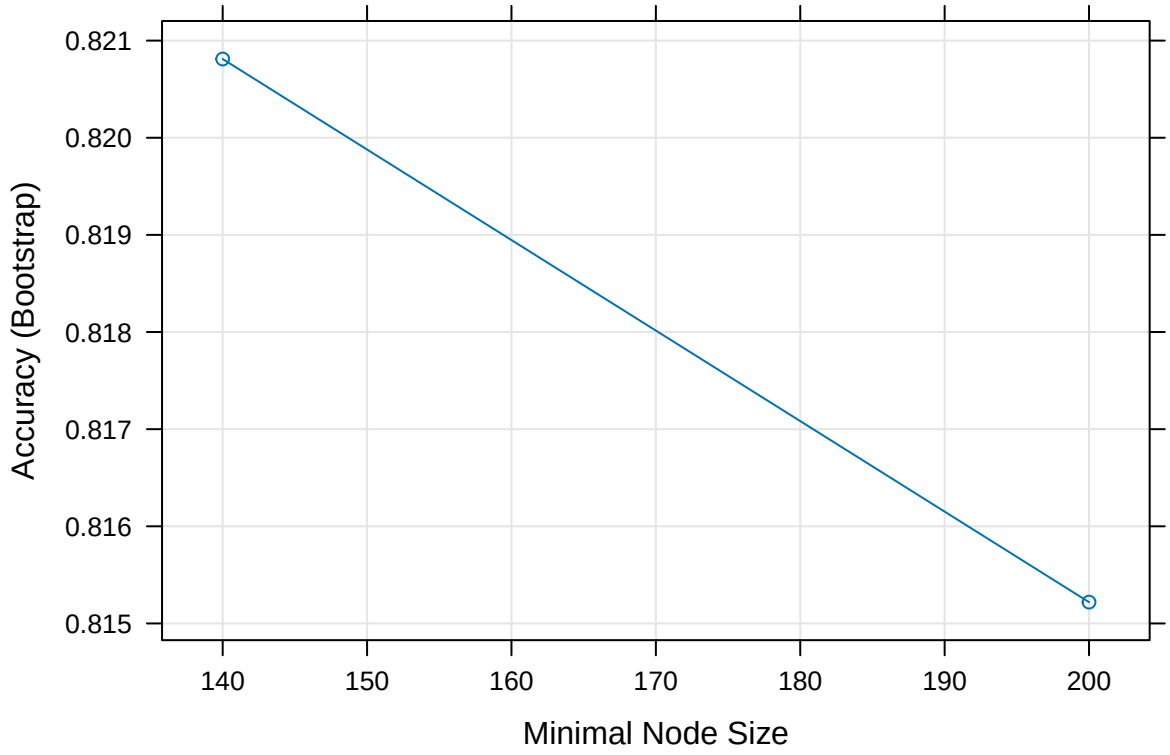

```
##      Model  Accuracy  CI_Lower  CI_Upper       PPV       NPV
## 1     Test 0.8600467 0.8471164 0.8722705 0.8628824 0.8333333
## 2 Validate 0.8543333 0.8411980 0.8667734 0.8590876 0.8085106
```

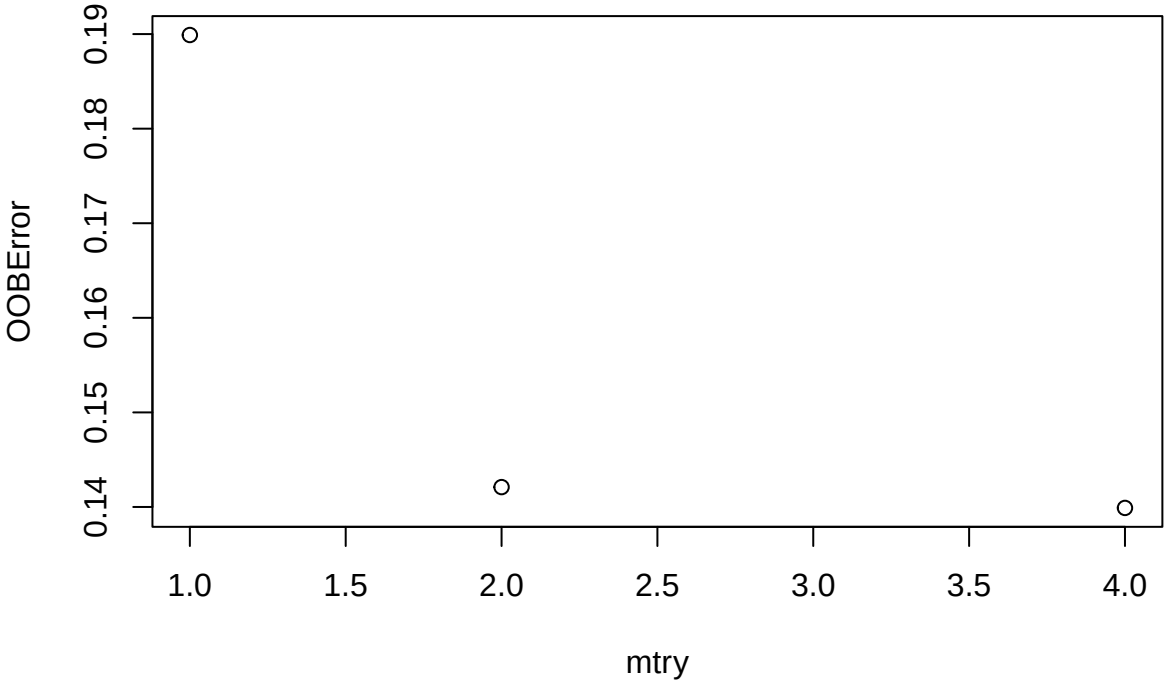
The accuracy variance in the SVM model is 0.007, certainly the most variance seen so far, but not significant enough to think that the model is problematic.

Random Forest

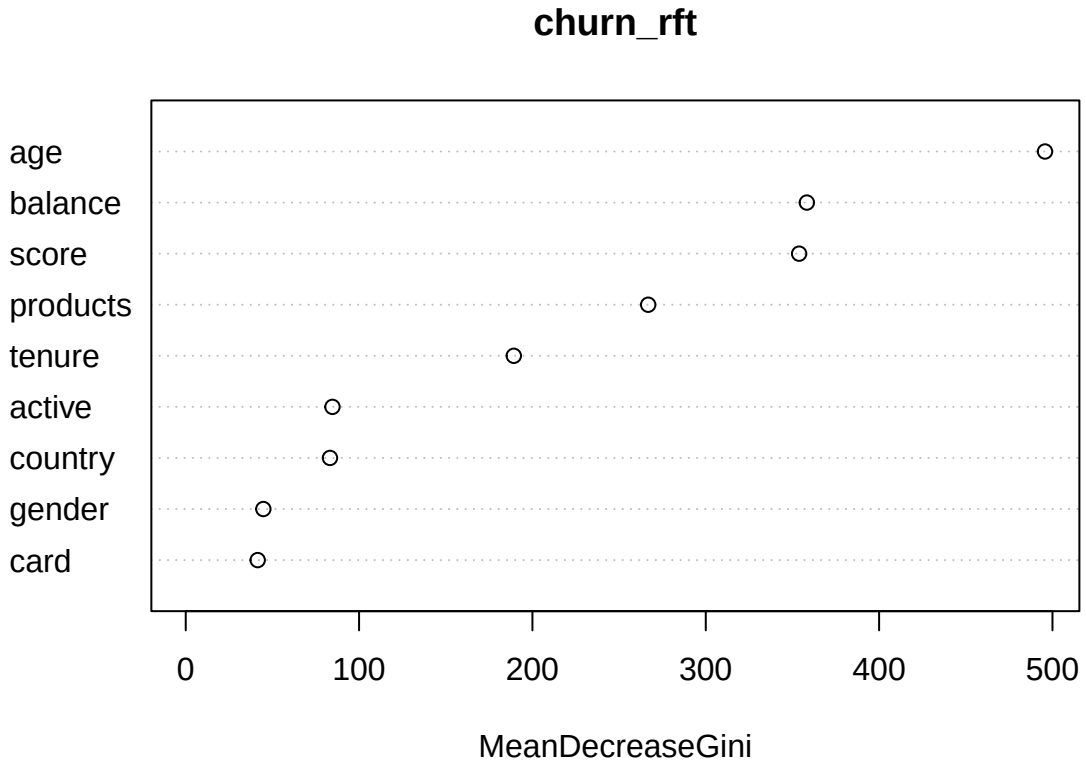
The Random Forest algorithm (Liaw and Wiener 2002) (Seligman 2024)is used to "build a series of trees from a random sample of the observations in the test data, random forests build trees from a random sample of the columns in the test data." (Adler 2012). After optimization, the model was built using an minimum node of 140 and mtry set at 4. The mtry argument controls the number of variables to sample randomly as candidates at each

split. The following graphs illustrate how accuracy decreases as the minimum node becomes greater than 140 and how the Out-of-Bag (OOB) error decreases as mtry increases from 1 to 4.






```
##       Model  Accuracy  CI_Lower  CI_Upper       PPV NPV
## 1      Test 0.9996668 0.9981448 0.9999916 0.9995816   1
## 2 Validate 1.0000000 0.9987711 1.0000000 1.0000000   1
```



As seen in the above graph, age, balance and score are the 3 most important variables in the Random Forest model. The model is nearly 100% accurate with an accuracy variance of just -0.0003. The NPV of both the test and validation sets was 1, which is important in the ability for the model to predict bank churn.

Results

Each model had their test model validated and the results compared. There was no indication in any of the models that there was any overfitting or biases.

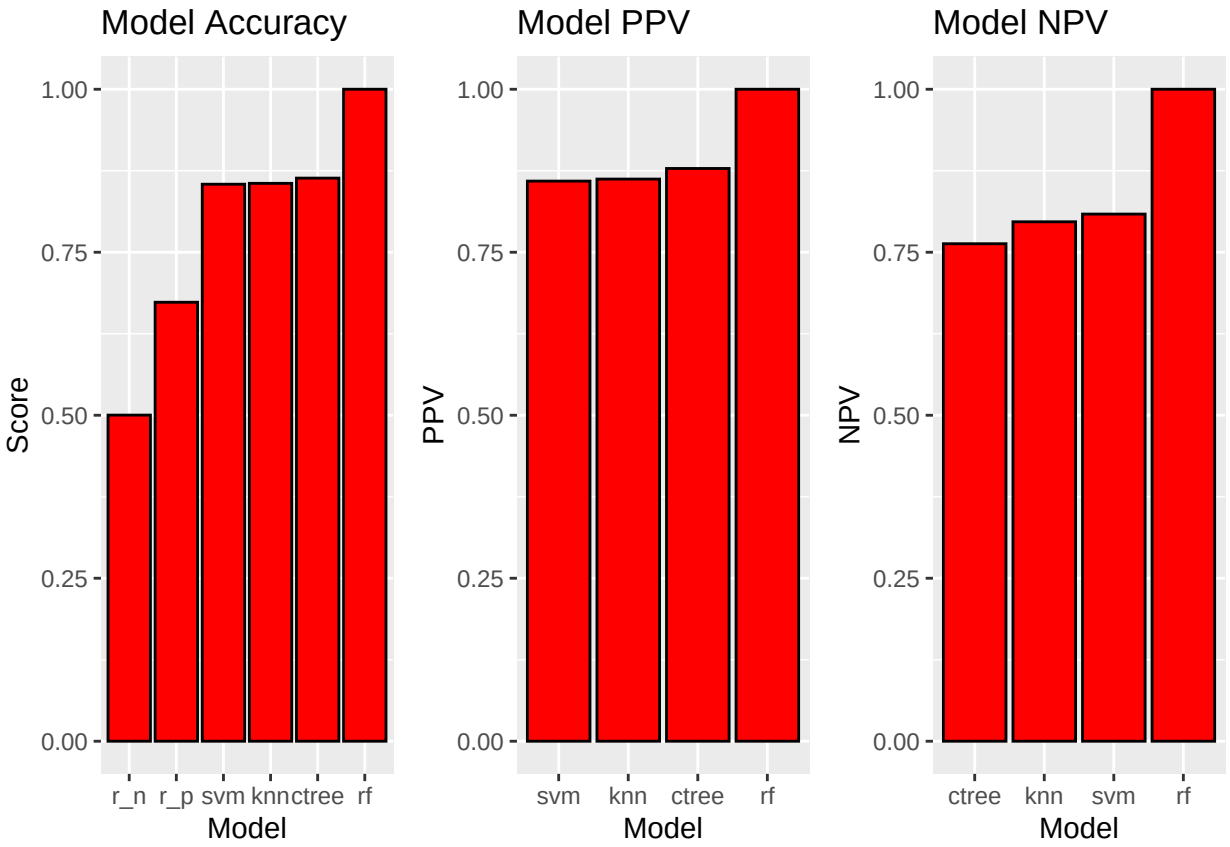
```
##   model     score       PPV       NPV
## 1   r_n 0.5002000 0.0000000 0.0000000
## 2   r_p 0.6733000 0.0000000 0.0000000
## 3   knn 0.8556667 0.8622222 0.7966667
## 4 ctree 0.8636667 0.8784404 0.7630208
## 5   svm 0.8543333 0.8590876 0.8085106
## 6    rf 1.0000000 1.0000000 1.0000000
```

The results indicate that the k Nearest Neighbor (kNN), Support Vector Machine (SVM), and Classification Tree (ctree) performed to withing 0.9% of one another ranging from 85.4% to 86.3%, in regards to overall accuracy.

Classification Tree has the second highest positive predictive value (PPV) of all the models at 86.4%, but the lowest negative predictive value (NPV) at 76.3%.

SVM and kNN have similar PPV and NPV scores.

Random Forest outperformed all models in every aspect and produced results near or at 100%



Conclusion

The purpose of this exercise was to create a machine learning algorithm to predict bank churn using the Bank Churn Dataset. This is best achieved using the Random Forest model developed in this study. Random Forest has an overall accuracy of 100% in predicting bank churn.

The ability to predict churn gives the ABC Multinational Bank a tool to uncover customers who are in danger of leaving the Bank. If the goal is to prevent customer churn, a model that may overestimate churn is preferable to one that would underestimate it.

While this model can predict churn it, does not have the ability to prevent churn. This study does provide insights into the effects that age, bank balance and credit score have on churn. Further investigation on these items may lead the Bank to develop different services to prevent churn. A breakdown of the types of products customers hold would also add value to any future study.

```
## [1] "R version 4.3.2 (2023-10-31 ucrt)"
```

References

Adler, Joseph. 2012. *R in a Nutshell*. 2nd ed. Sebastopol, CA: O'Reilly Media Inc.

Auguie, Baptiste. 2017. "gridExtra: Miscellaneous Functions for "Grid" Graphics." https://CRAN.R-project.org/package=gridExtra.

Bobbitt, Zach. 2023. "How to Convert Multiple Columns to Factor Using Dplyr." https://www.statology.org/dplyr-convert-multiple-columns-to-factor/.

"Contour of KNN Model in Ggplot Using r." 2024. https://www.geeksforgeeks.org/contour-of-knn-model-in-ggplot-using-r/.

Karatzoglou, Alexandros, Alex Smola, and Kurt Hornik. 2024. "Kernlab: Kernel-Based Machine Learning Lab." https://CRAN.R-project.org/package=kernlab.

Kuhn, Max. 2007. "Caret: Classification and Regression Training," October. https://doi.org/10.32614/CRAN.package.caret.

Liaw, Andy, and Matthew Wiener. 2002. "Classification and Regression by randomForest" 2: 18–22. https://CRAN.R-project.org/doc/Rnews/.

Meyer, David, Evgenia Dimitriadou, Kurt Hornik, Andreas Weingessel, and Friedrich Leisch. 2023. "E1071: Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071), TU Wien." https://CRAN.R-project.org/package=e1071.

R Core Team. 2023. "R: A Language and Environment for Statistical Computing." https://www.R-project.org/.

Seligman, Mark. 2024. "Rborist: Extensible, Parallelizable Implementation of the Random Forest Algorithm." https://CRAN.R-project.org/package=Rborist.

Soetewey, Antoine. n.d. "Correlation Coefficient and Correlation Test in r." https://statsandr.com/blog/correlation-coefficient-and-correlation-test-in-r/.

Therneau, Terry, and Beth Atkinson. 2023. "Rpart: Recursive Partitioning and Regression Trees." https://CRAN.R-project.org/package=rpart.

Topre, Gaurav. n.d. "Bank Customer Churn Dataset." https://www.kaggle.com/datasets/gauravtopre/bank-customer-churn-dataset.

Wei, Taiyun, and Viliam Simko. 2024. "R Package 'Corrplot': Visualization of a Correlation Matrix." https://github.com/taiyun/corrplot.

White, Denis, and Robert B. Gramacy. 2022. "Maptree: Mapping, Pruning, and Graphing Tree Models." https://CRAN.R-project.org/package=maptree.

Wickham, Hadley. 2016a. "Ggplot2: Elegant Graphics for Data Analysis." https://ggplot2.tidyverse.org.

———. 2016b. "Ggplot2: Elegant Graphics for Data Analysis." https://ggplot2.tidyverse.org.