1. Data collection and wrangling summary
2. Exploratory data analysis summary (visualization and inferential statistics)
3. Results and In-depth analysis using machine learning

The Dataset is about bank customers churning and can be found on Kaggle:

https://www.kaggle.com/barelydedicated/bank-customer-churn-modeling

Disclaimer: The dataset above is simulated.

Introduction

Banks need to have satisfied customers for it to be successful and this report looks at bank customers and our ability to correctly predict if a bank customer will exit the bank or not. The bank is concerned about potentially losing current customers. They want to be able to predict which customers are at risk of churning and be able to do outreach (in the form of promotions, etc) to convince them to retain their services with the bank.

Data Wrangling

The dataset for the bank customers was obtained from Kaggle and was in a csv format. Before starting any analysis, the shape of the dataset needs to be known which in this case was 10000 rows and 14 columns. Out of the 14 columns in the dataset, 13 are feature columns and the 14th column, **'Exited',** is the response column. The dataset also needed to be checked for null values and fortunately, our dataset did not contain any null values as it was obtained from Kaggle and was already clean. Though, this is not often the case with real world data.

The redundant columns, like **'RowNumber'**, were dropped from the dataframe as they do not provide any useful information and the new shape of the dataset became 10000 rows and 13 columns. Another important step that needs to be carried out is converting the categorical columns into numerical values. For this dataset, there were two such columns **'Geography'** and **'Gender'** that were converted. The 'Geography' column contained values France, Germany and Spain which were converted to 0, 1 and 2 respectively. Similarly, the Gender column contained values Male and Female that were converted to 0 and 1 respectively.

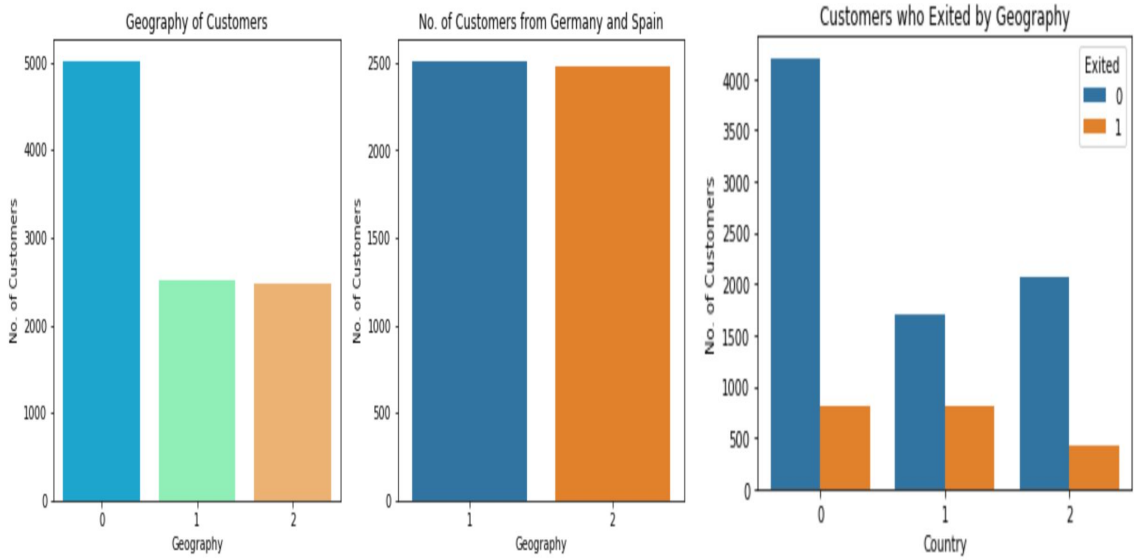
The column ‘Exited’ was moved to the leftmost side of the table as this allows viewing the data quickly and also makes splitting the dataset into train/test sets easier at a later stage. The data was checked for outliers by looking for any extreme values in the min and max fields of the columns of the dataframe. Fortunately, the data does not contain any outliers.

After performing data wrangling on the dataset, the next step was to investigate, by visualizing the data, to uncover interesting information about the data and learn more about the customers in the bank who are exiting or not. The ‘Geography’ and ‘Gender’ categorical columns had been previously converted into numerical values using the following relationship:

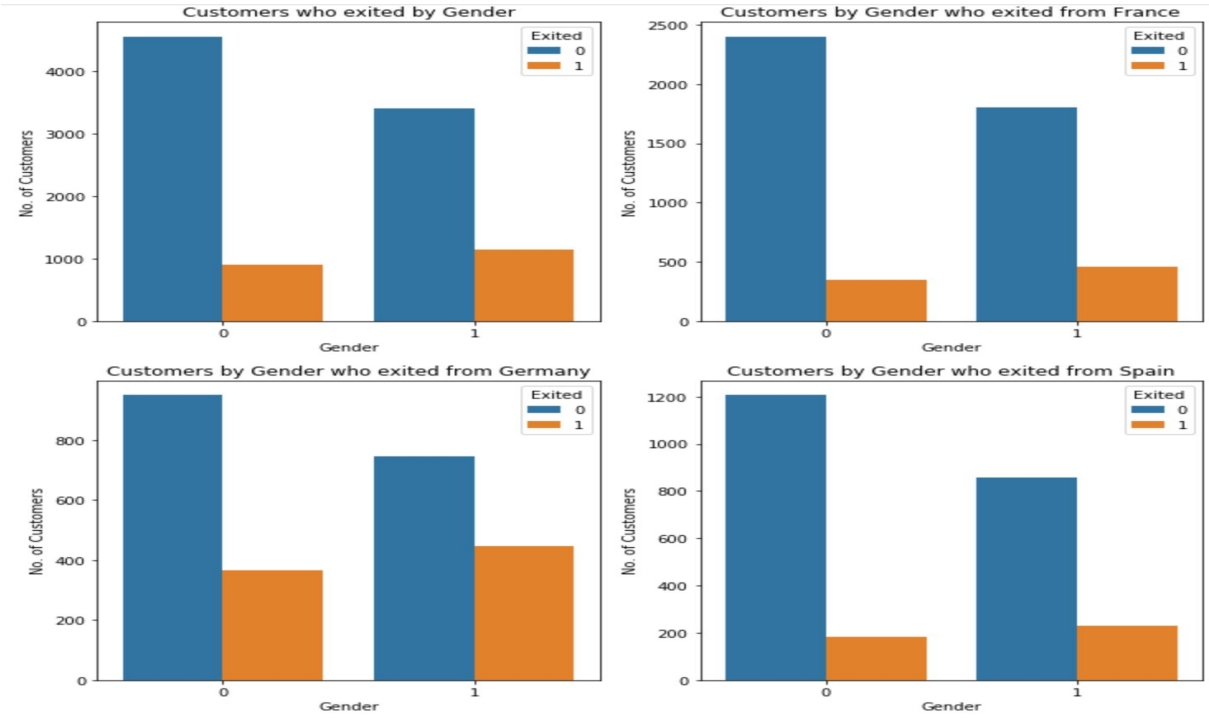
- For the ‘Geography’ column, the values of **France**, **Germany** and **Spain** are represented by **0**, **1** and **2** respectively.
- For the ‘Gender’ column, the values of **Male** and **Female** are represented by **0** and **1** respectively.

Data Visualization

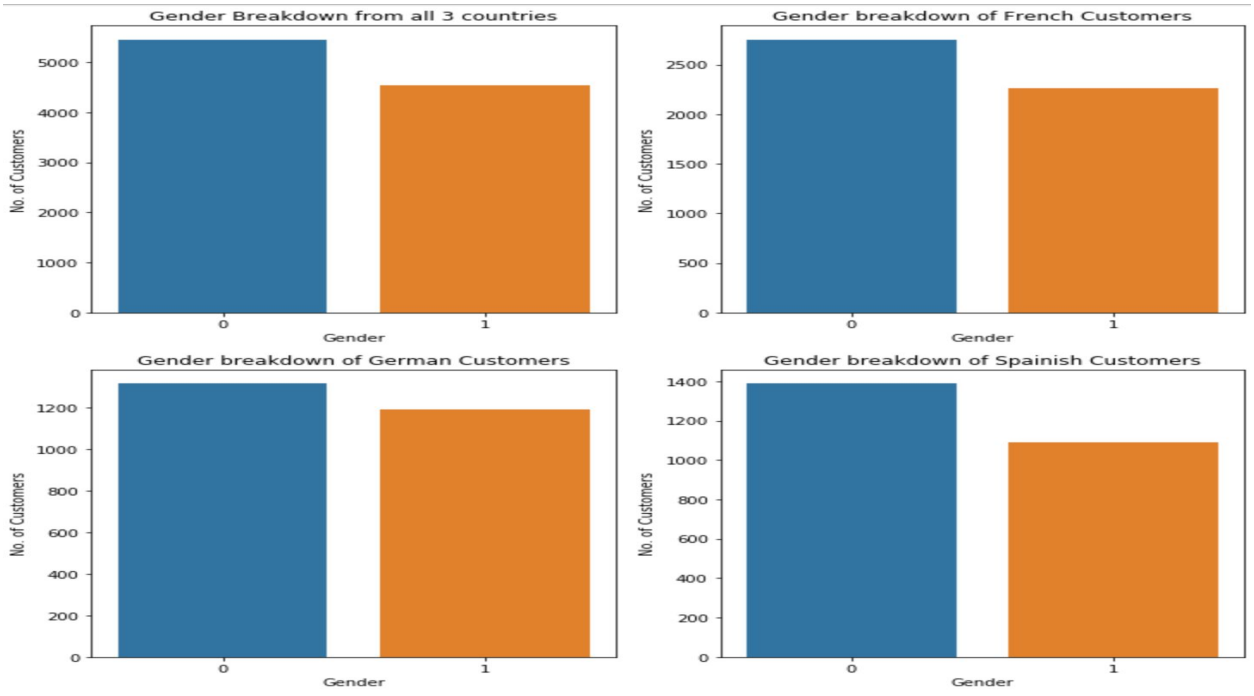
After visualizing where the customers are from, it was discovered that there were twice as many French Customers than from Germany and Spain. Germany also had the highest number of customers who exited the bank followed by France and Spain. These inferences can be seen in the 3 plots below.



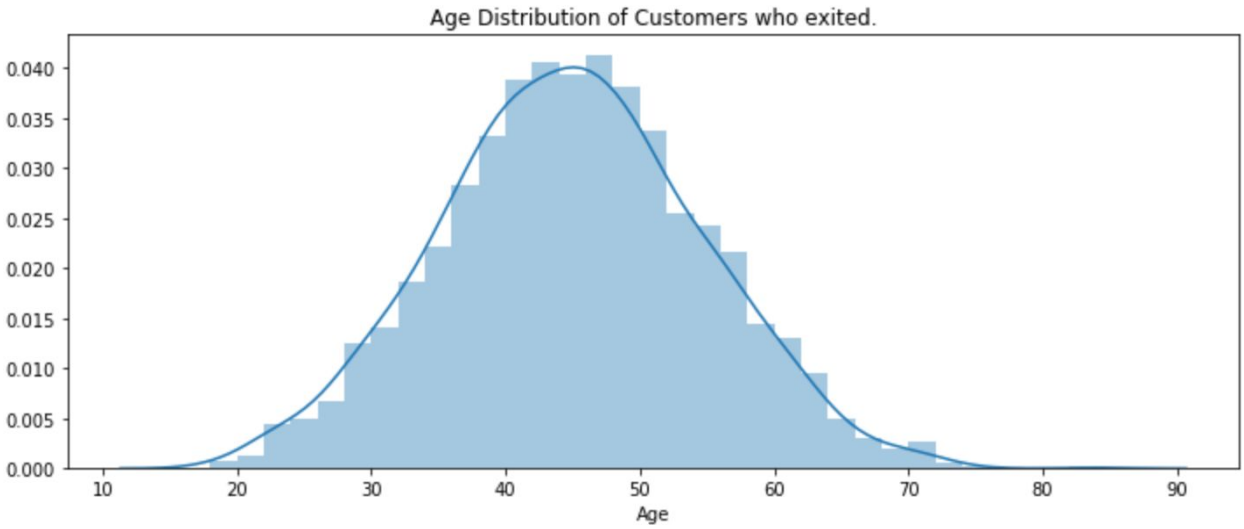
Further visualization of the data showed that female customers left the bank more than male customers, France had the highest number of female customers who exited the bank and Germany had the highest number of customers who left the bank as well as the highest number of males who exited the bank.

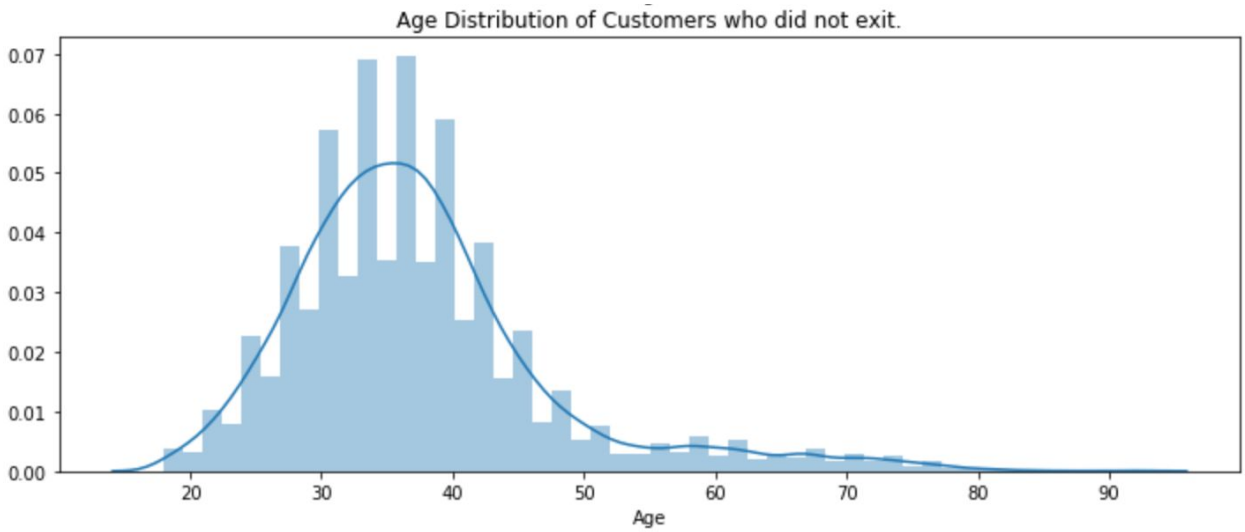


When the gender breakdown of the bank customers was visualized, it showed that France had the highest number of female and male bank customers.

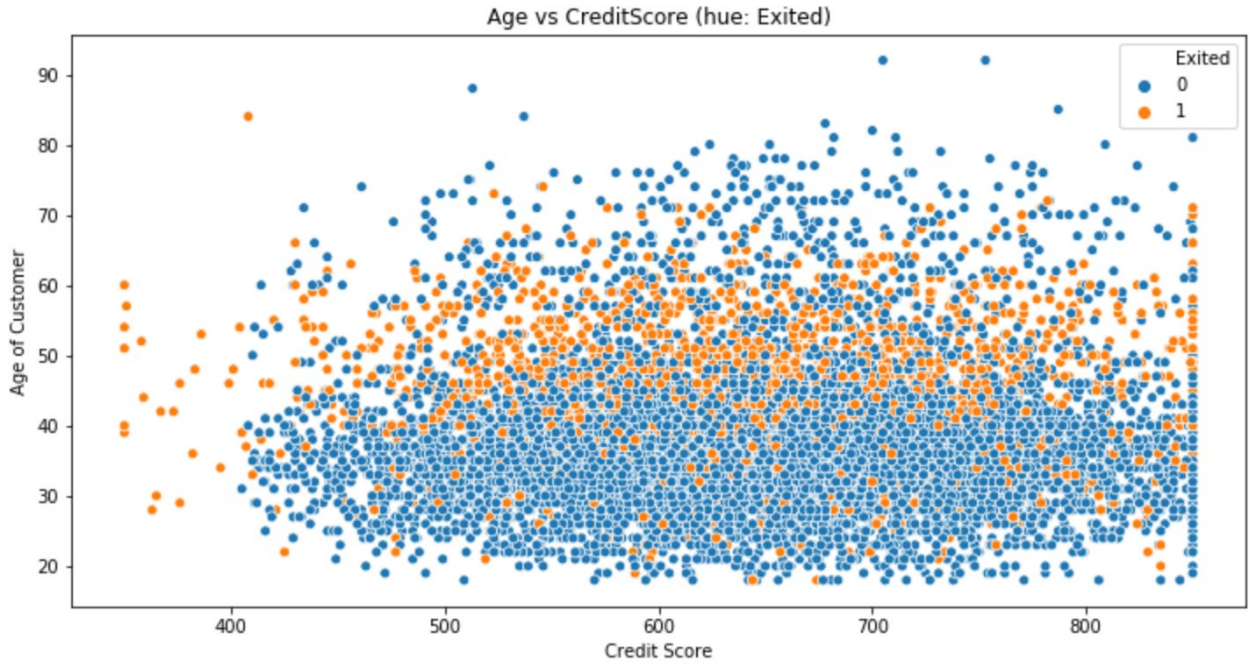


The age distribution of the customers who exited the bank looks similar to a normally distributed distribution.

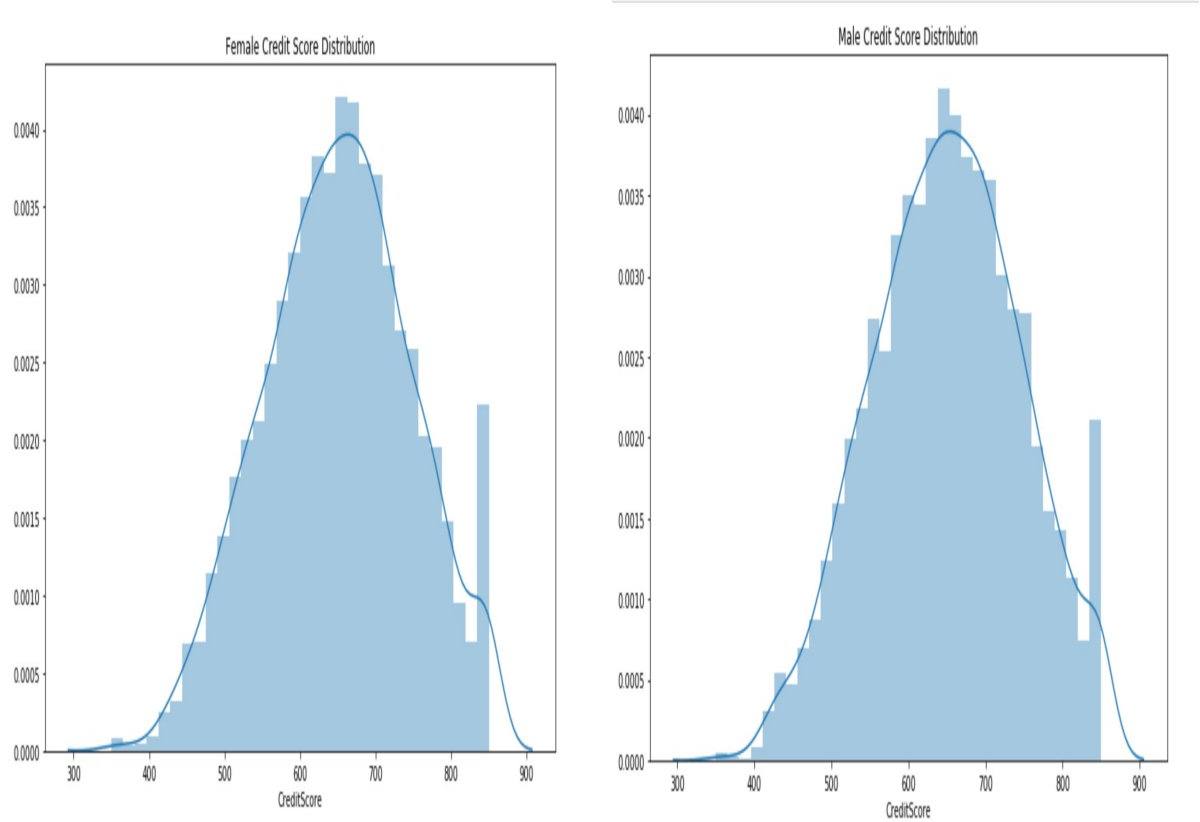




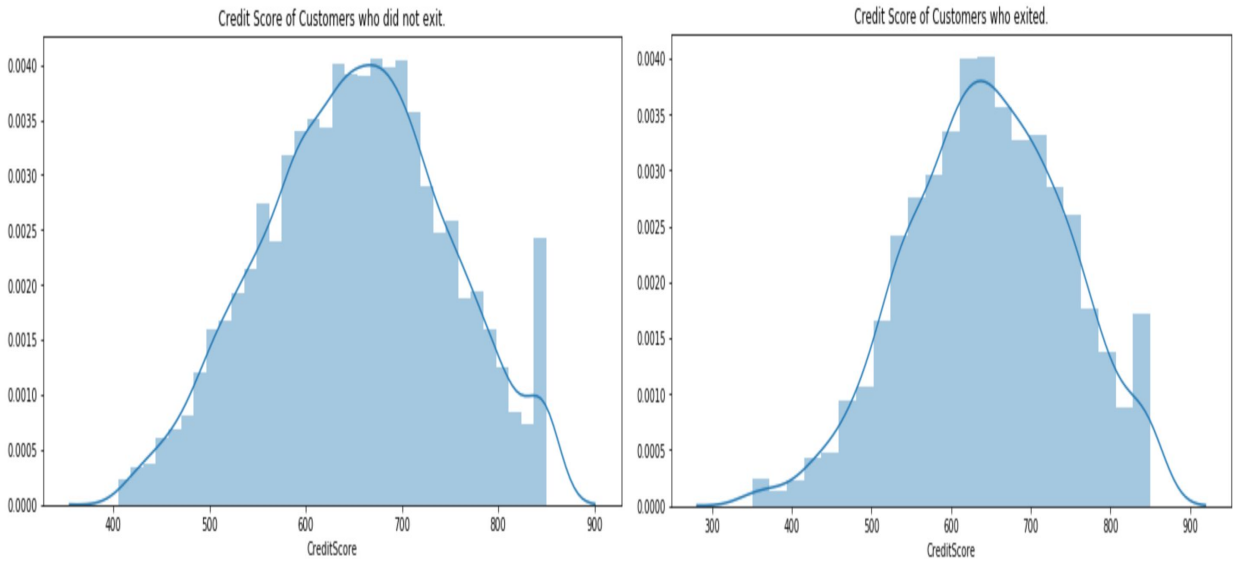
A scatterplot between credit score and age revealed that the majority of the customers who exited were between 40 years and 60 years of age. It would be interesting to see the reason why this particular group of people aged between 40 and 60 years old are exiting the bank? Are they getting better offers from other banks?

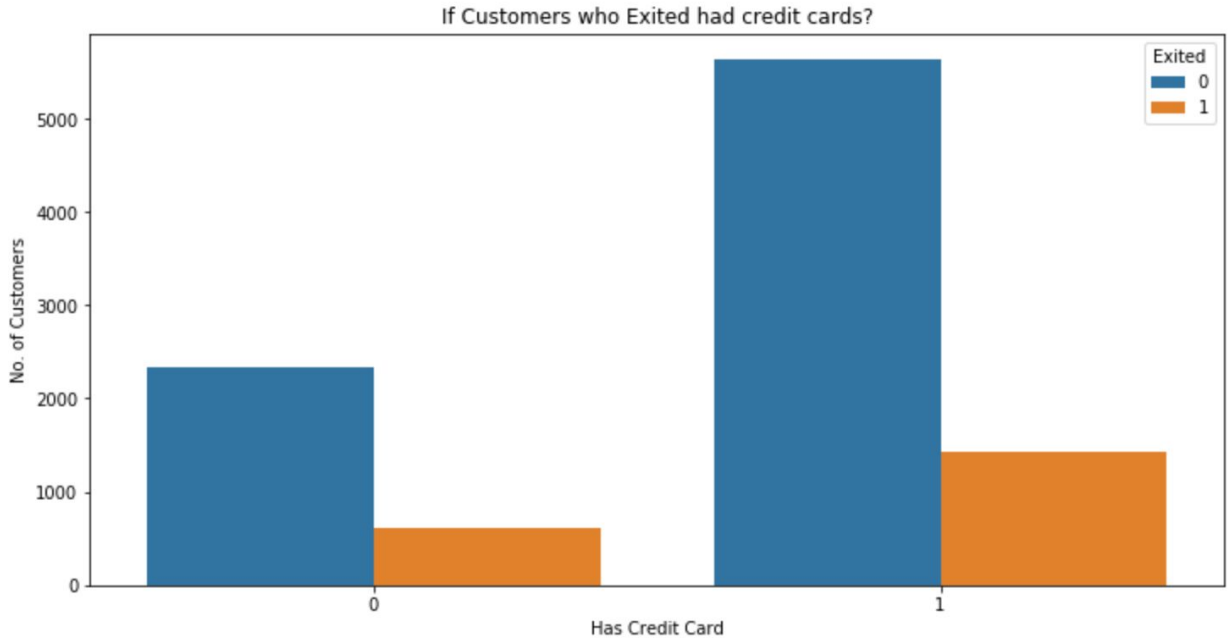


Visualizing the male and female bank customers credit score distributions ,showed no major significant difference in the distribution of both distributions.



It was interesting to find that customers who exited had a roughly lower credit score and customers who had a credit card exited more than customers who didn’t own a credit card.





These are interesting insights about the bank customers and allows us to further ask questions: if the customers could get better offers based on their low credit score from other banks that is causing them to exit? Are other banks offering lower APR on their credit cards to lure customers?

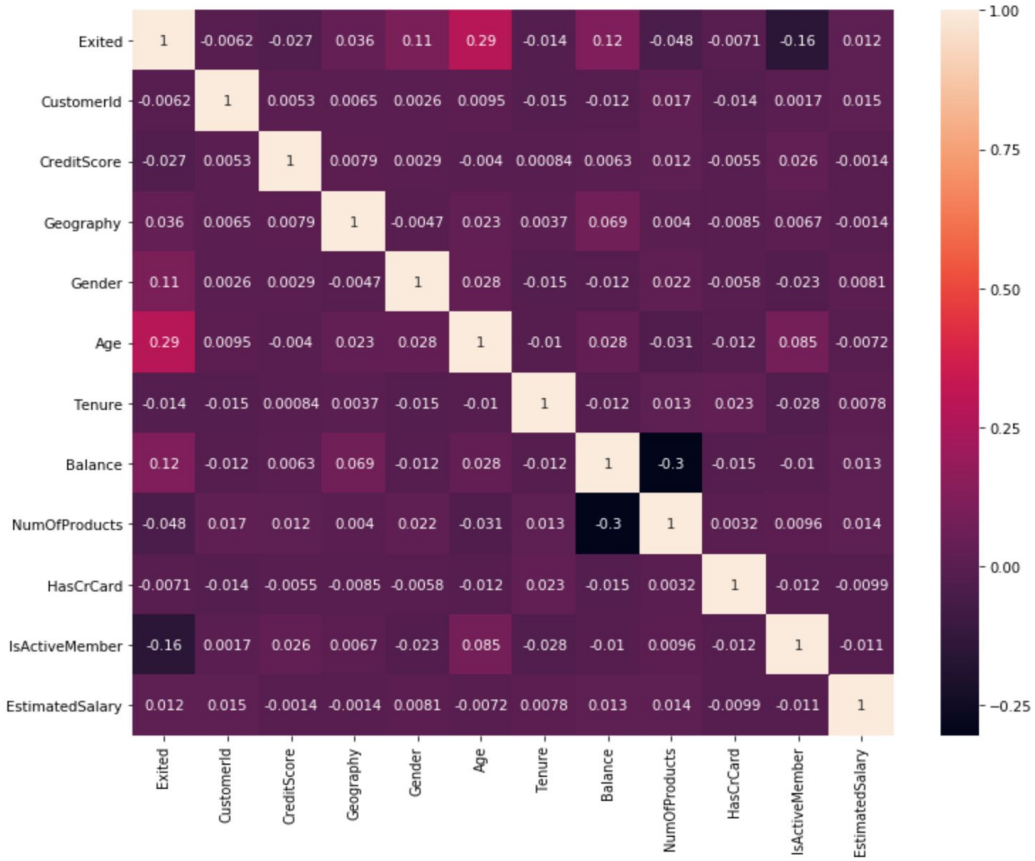
Statistical Analysis

After gaining further information about the dataset by performing exploratory data analysis, the next step is to investigate the dataset through a statistical lens. The main aim is to look for any significant correlations between variables and performing a test to see if there is any statistical

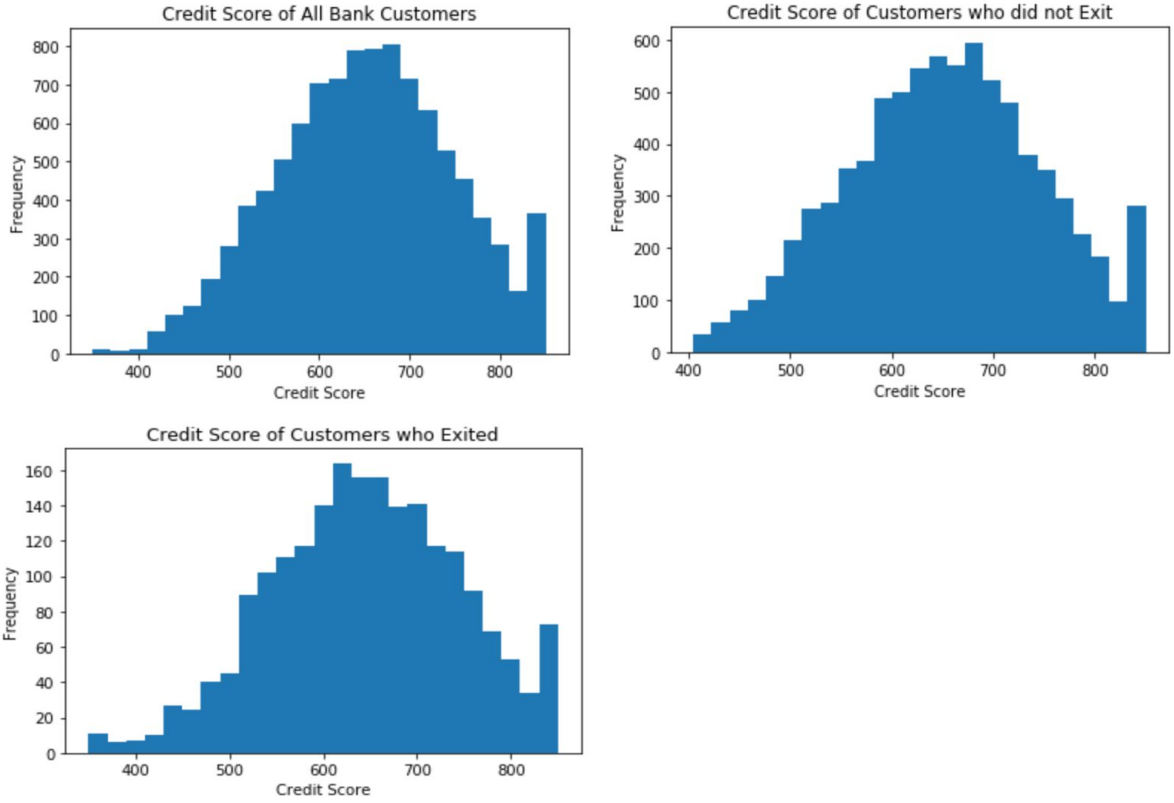
difference in the means of credit score of customers who exited compared to the mean credit score of the entire bank customer population.

The summary statistics of the dataset provide us with some initial information about customers who had stayed with the bank and customers who had exited the bank. They showed that **79.6%** of customers stayed at the bank while **20.4%** of customers exited the bank. The mean credit score of customers who stayed was 651.85 compared to 645.35 for customers who exited the bank.

It is important to check if the variables in the dataset have any meaningful correlations between them. A correlation value close to -1 indicates strong negative correlation, values close to 0 indicate no/very weak correlation and values close 1 indicate a strong positive correlation between variables. Looking at the figure of the heatmap below, the **'Exited'** and **'Age**' variables had the highest correlation value of 0.29, which does not suggest strong correlation between the pairs as the value is close to 0. All the other variables had very weak correlation values as can be seen below in the heatmap.



The figures below show how the credit score distributions looked like for customers who had exited, who had stayed and the entire bank customer population. The credit score distribution of customers who exited appears different than the other two distributions.



Next, a one-sample t-test was performed to test whether a population mean is significantly different from some hypothesized value. In this case, the test was performed to check whether the average credit score of customers that exited differs from the average credit score of the entire employee population. For this test, the Null and Alternate hypothesis were:

Null Hypothesis: The null hypothesis would be there there is no difference in credit score between employees who exited and all bank customers.

Alternate Hypothesis: The alternative hypothesis would be that there is a difference in credit scores between customers who exited and all bank customers.

The one sample t-test was performed at a 95% confidence interval to check if we could correctly reject the null hypothesis based on our result of the test.

The t-distribution left quartile range was -1.96112925575354 and right quartile range was 1.9611292557535396. The test result showed the test statistic 't' is equal to -2.329. T is simply the calculated difference represented in units of standard error and tells us how much the sample mean deviates from the null hypothesis. The null hypothesis can be rejected if the

t-statistic lies outside the quantiles of the t-distribution corresponding to the chosen confidence level and degrees of freedom.

A p-value of 0.019946347165310532 meant that they would expect to see data as extreme as their sample due to chance way less than 5% of the time if the null hypothesis was true. In this case, the p-value is lower than our significance level α (equal to 1-conf.level or 0.05) so the null hypothesis should be rejected.

Based on the statistical analysis of a one sample t-test, there seems to be some significant difference between the mean credit score of bank customers who exited and the entire bank customer population. The low P-value of 0.019946347165310532 at a 5% confidence interval is a good indicator to reject the null hypothesis.

The results of our test showed that there is some significant difference between the mean credit score of bank customers who exited and the entire bank customer population. Further experiments can be conducted or more data can be collected about the bank customers to develop our understanding of the data and potentially get more refined insights.

Results and In-depth analysis using machine learning

After performing statistical analysis in the last step, the next thing to do was choosing and training different machine learning classifiers on the data. First and foremost, the class balance of the response variable in the dataset was checked and it turned out that the response classes were imbalanced in the dataset. What this means is that 'Exited' (the response variable) has two classes 0 and 1 that are imbalanced in the dataset. In this particular case, there were 20.4% '1' and 79.6% '0' instances of the two classes in the 'Exited' field. This information is important as now we know that we should use a class balancing technique later on before training our models.

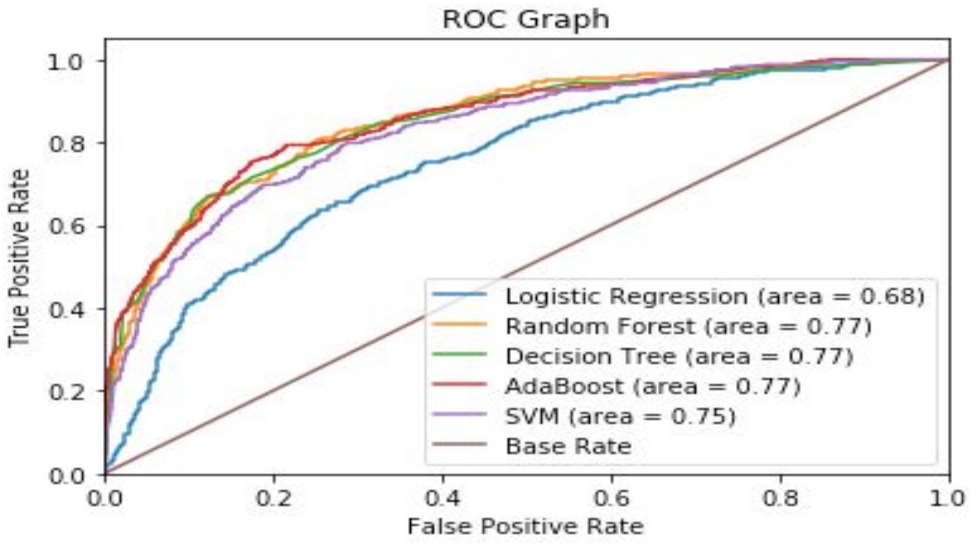
Another important step was to get rid of feature columns that do not give us any valuable information. In this case, we dropped the columns 'Surname' and 'CustomerId' with the rest of the data saved as 'X'. I wanted the independent variables to occupy the same scale so I used the robust scaler method on my data. The robust scaler removes the median and scales the data according to the quantile range while also being robust to outliers, hence the name robust scaler. After scaling the data, it was split into training and test sets with an 80-20 train test split.

There was a need for a base classifier, which in this case I picked the Logistic Regression Classifier, so that the result of this classifier could serve as a baseline to check other models against. One thing that struck out about the logistic regression classifier was that its AUC - ROC score improved by 10% after changing the class_weight parameter to 'balanced'. This could be explained due to our data being imbalanced. The Logistic Regression Classifier got an AUC - ROC score of **0.68**, which now serves as our baseline AUC - ROC score to check against the other classifiers. The AUC - ROC curve is a performance measurement for classification

problems at various threshold settings. ROC is a probability curve and AUC represents degree or measure of separability. Basically, it tells how much a model is capable of distinguishing between classes. The higher the AUC, the better the model is at predicting 0s as 0s and 1s as 1s. The ROC curve is plotted with True Positive Rate (TPR) against the False Positive Rate (FPR) with the TPR on the y-axis and FPR on the x-axis. Another thing that is important to note here is that we are using the AUC - ROC score as our metric of choice to assess the quality of the models and not the accuracy because we had an imbalanced dataset.

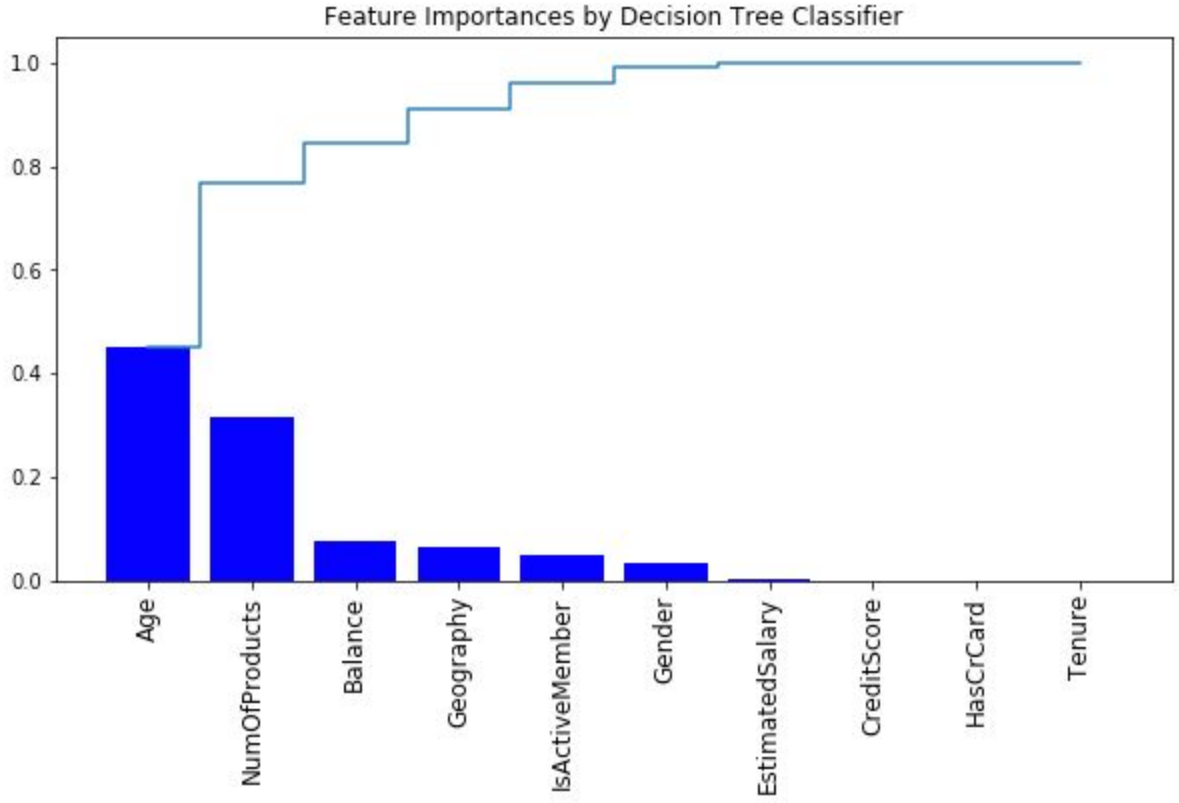
Next, it was time to handle the class imbalance in the dataset using SMOTENN which combines over- and under-sampling using SMOTE and Edited Nearest Neighbours. After transforming the data using SMOTEENN, there were 5544 **'1'** samples and 4282 **'0'** samples. It is important to mention that this transformation was only performed on the training set as we would like our test set to be as pure as possible. After getting the resampled data, hyper-parameter tuning was carried out for Decision Tree Classifier, AdaBoost Classifier, Random Forest Classifier and the Support Vector Machine Classifier to find the best hyper-parameters of each model. Though, for my own curiosity I did play around with the suggested values and in one instance was able to get a higher AUC - ROC score for the Decision Tree classifier. The suggested values gave me an AUC - ROC score of 0.75 for the Decision Tree classifier and after tweaking around, I was able to bump up the AUC - ROC score to 0.77 as it can be seen below in the ROC graph.

All the 4 machine learning classifiers performed much better than the Logistic Regression base classifier. Subsequently, 3 of the 4 machine learning classifiers were tree based methods and they all performed equally, with an AUC score of **0.77**. The Support Vector Machine Classifier performed the poorest out of the 4 classifiers and got an AUC score of **0.75.**



One of the things that Decision Trees allows you to do is assess which features in your dataset were the most important and we can visualize each features contribution as it can be seen below. It appears that 'Age' and 'NumOfProducts' contributed more than 70% in terms of feature

importance. ‘Tenure’, ‘HasCrCard’, and ‘CreditScore’ did not contribute at all which is surprising as one would think they would be good predictors of bank customers exiting or not.



Below is a table of the classifiers used and their respective AUC-ROC scores.

Classifier	**AUC-ROC Score**
Decision Tree	0.77
AdaBoost	0.77
Random Forest	0.77
SVM	0.76
Logistic Regression	0.68

The best model performance out of the 5 classifiers (Logistic Regression, AdaBoost, Decision Tree, Random Forest, Support Vector Machine) was of the three tree based methods (AdaBoost, Decision

Tree, Random Forest). I would recommend using the **Decision Tree model** as it is the least computationally expensive. Furthermore, one of the things that Decision Trees allows you to do is assess which features in your dataset were the most important and we can visualize each features contribution as it can be seen below in the notebook. For our dataset, the features 'Age' and 'NumOfProducts' contributed to almost **80%** of all features' importance.

Base Rate Model

- The Logistic Regression Classifier was chosen as the base rate model. A base rate model is a model that is used for reference to compare how better another model is against it.

Model Evaluation

This dataset is an example of a class imbalance problem because of the uneven distribution of customers who did and did not exit the bank. To handle this, use the SMOTEENN method which combines over- and under-sampling using SMOTE and Edited Nearest Neighbours

In this case, evaluating our model's algorithm based on accuracy is the wrong thing to measure. We will have to take into consideration the False Positive and False Negative Errors and use that as a metric to evaluate our model's performance.

False Positives (Type I Error): You predict that the customers will leave, but do not.

False Negatives (Type II Error): You predict that the customer will not leave, but does leave.

Now that we have our model and we can predict a customer exiting or not, how do we use that information to positively impact the bank's business? It is important that rather than just predicting if the bank customer will exit the bank or not, we would rather have an estimate of the probability that the customer will exit the bank or not. We could rank the customers by their probability of leaving, and allocate an incentive budget to the highest probabilities in hopes of retaining them.

We can run into two sorts of problems with this approach. Firstly, consider that the customer is given incentives by the bank because they think the customer will leave at the end of the month, but in reality the customer doesn't leave. This is called a false positive and this mistake could be expensive, inconvenient and time consuming for all parties. It is not all negative though as it can be seen as a good investment for relational growth between the bank and the customer.

The opposite scenario would be where the bank does not provide any incentive offer to the customer and they do leave. This is called a false negative and this error is more costly because the bank lost a customer, which translates to lost revenue for the bank. It also means that the more false negative cases the bank endures, the more it will have to spend on its marketing to attract new customers. Depending on these errors, the different costs are weighed based on the type of customer being

treated. If it is a high-net worth customer then the loss is greater than losing a low-net worth customer. Hence, the cost for each error is different and this should be taken into account.

Solution 1:

- We can rank the customers by their probability of leaving the bank and set aside a budget to incentivize those customers not to leave the bank.
- Flipping this approach around, we can rank the customers by highest expected loss to the bank if they lose their business and reach out to those customers.

Solution 2: Provide training to the bank staff to improve their customer services. Track and measure their performances using analytics. Some ways to improve customer services include:

- Empowering employees
- Allow consumers to self-service
- Stay consistent across all touch points
- Educate customers on financial literacy
- Embrace financial technology
- Become an advisor, not just a lender, for small businesses
- Segment your client base and create personalized customer experiences
- Keep iterating on processes

Where do we go from here?

This problem is about equipping the bank with actionable knowledge regarding their customers. When modeling the data, we should not use the predictive metric as our final solution. Instead, we should use the information we get from modeling and arm the bank staff so they can carry out informed decision making.

Another thing the bank could do is start collecting more data on more features for e.g. how long the customer has been with the bank, satisfaction score with the bank etc. These things might help us improve our model, especially collecting more data as this bank customer dataset was relatively small. Once we have more data, we can go back and improve our predictions as well as gain further insights to see if anything has changed now that we have more customer information.

After our attempts to understand why customers are leaving the bank, we can flip the problem around and ask ourselves:

- What features contribute most to customers retaining their services with the bank?

- What features cause employees not to quit?
- What is the most valued thing about the bank by the customers?