Predicting Banking Customer Churn

Project Report

Group 31

Ajay Parthiban Senthilvel

Monisha Prasad

857-437-9303

617-516-9675

senthilvel.a@northeastern.edu

prabhuprasad.m@northeastern.edu

Percentage of Effort Contributed by Student 1: 50

Percentage of Effort Contributed by Student 2: 50

Signature of Student 1: Ajay Parthiban Senthilvel

Signature of Student 2: Monisha Prasad

Submission Date: 25th April 2022

Problem Setting

Customer churn is an important metric for most businesses, losing customers almost always equals loss of revenue, which is why it is important to prevent it. The problem lies in the fact that it is uncertain why a customer would choose to leave and when. This project aims at predicting when a customer might churn (leave) and suggest ways that could possibly prevent this.

Problem Definition

In a financial setting, like in banks, customer loyalty is of great concern, customers who leave a bank are unlikely to recommend it to anybody causing a cascading loss of revenue. However, they are unaware of issues a client is facing as they seldom ask for routine customer feedback. To prevent losing a customer, data mining techniques can be applied on existing customer data to find patterns such as existing trends within groups who choose to stay with the bank vs. groups who leave. This knowledge can also be used to find answers to questions such as:

- Are there aspects of customer service that can be improved to prevent this?
- Is there anything that can be done to decrease the overall exit rates of customers?

This project aims at finding the most data mining models to predict whether a customer would leave a bank by classifying them based on other factors and provide insightful solutions.

Data Source

The dataset used for the project is the publicly available Churn Modeling dataset from Kaggle

- https://www.kaggle.com/adammaus/predicting-churn-for-bank-customers/metadata

Data Description

The dataset has 10,000 observations with 14 attributes - 1 index, 7 numerical, 3 categorical, and 3 logical variables. It provides demographic as well as banking information. The "Exited" attribute will serve as the target variable as it tells us whether a customer has left the service or not (0 = No, 1 = Yes). A test set will be set aside to evaluate the machine learning models. Attributes such as the index, "Customer ID" etc. will be dropped as part of feature

selection. Features such as “Geography”, “Credit Score”, “Has Credit Card” etc. will be used to assess the customers and develop the machine learning model.

Data Collection

The dataset picked from Kaggle has the following attributes, describing the attributes helps to get a better understanding of the data before processing:

RowNumber	Index indicating the total number of customers
CustomerId	Unique ID for each customer
Surname	Customer’s surname
CreditScore	Customer’s credit score
Geography	Customer’s location (country)
Gender	Male/Female
Age	Customer’s age
Tenure	Time period the customer has had account (in months)
Balance	Customer’s account balance
NumOfProducts	Number of bank products customer uses
HasCrCard	Binary value indicating if a customer has a credit card or not
IsActiveMember	Binary value indicating if a customer is active or not
EstimatedSalary	Customer’s estimated salary
Exited	Binary value indicating if a customer has churned or not

The descriptions aid feature selection.

Data Processing

Missing Values and Outliers: No missing values were found in the dataset. The attributes ‘Tenure’, ‘CreditScore’, ‘Balance’, and ‘EstimatedSalary’ were checked for outliers and although there were values below the lower limit in ‘CreditScore’ it was not significant and hence it was not removed.

Feature Selection: Attributes like 'RowNumber' and 'CustomerId' do not provide any insight into our data or our goal of predicting churn, hence they are dropped. The 'Surname' attribute is also dropped because it is fairly meaningless as multiple customers can have the same name without any relationship.

Encoding: Features like 'Geography' and 'Gender' could possibly have an impact on churn so they are encoded into a numerical form to be more useful when implemented into a data mining model.

For 'Geography': The three geographic locations represented in the dataset are 'Germany', 'France', and 'Spain'. This attribute is encoded using One Hot Encoding. The get_dummies() function gives 'Geography_Germany', 'Geography_France' and 'Geography_Spain'

For 'Gender': Using the same function for one hot encoding we get 'Gender_Male' attribute and 'Gender_Female'

Scaling: To normalize the range of features in the dataset, feature scaling has been done on the continuous variables using sklearn's StandardScaler().

After implementing the above changes we have a dataset with 14 attributes and 10,000 instances.

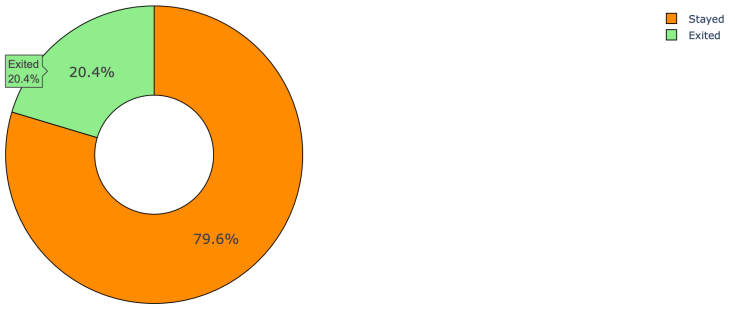
Data Exploration and Visualization

The data is explored to find insights primarily in the following three ways:

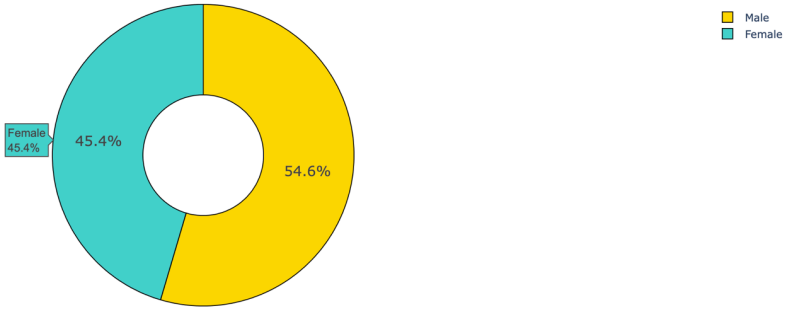
- Overview of Churn
- Overview of Customers
- Overview of Churn with respect to the relationship of the customer with the bank and its various services

Churn Distribution - The percentage of customers who exited was found to be 20.4% while that of customers who stayed was 79.6% (indicating class imbalance)

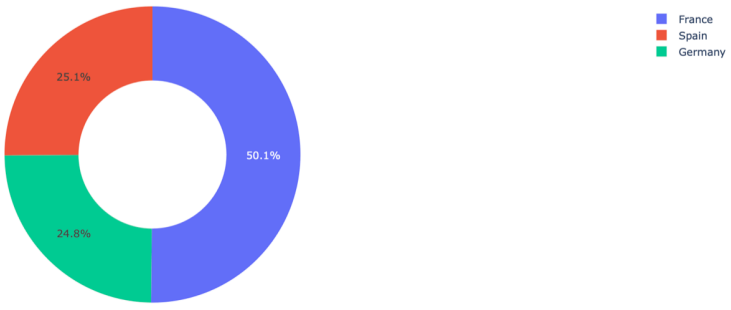
Churn Distribution



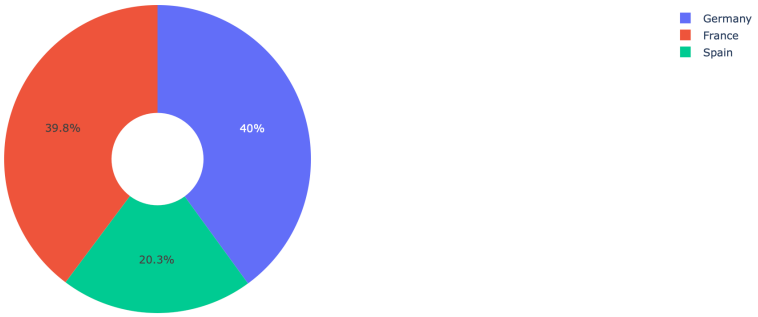
Customer Gender Distribution - The number of male banking customers were greater than the number of female banking customers



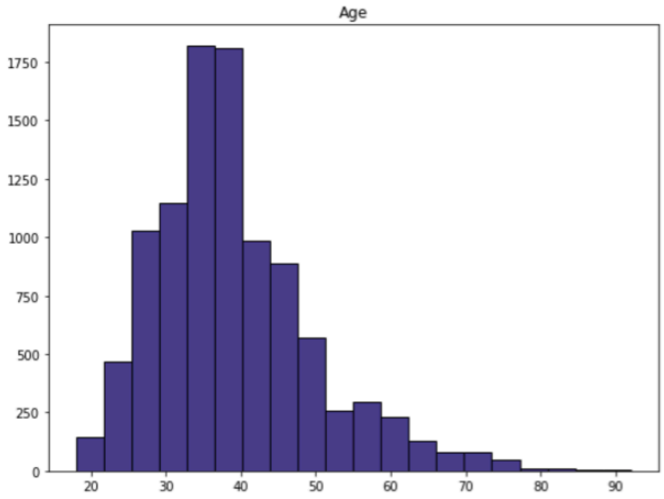
Geographical Distribution - The three geographic locations recorded in this dataset are France, Spain and Germany, with France having the highest percentage of customers and Germany having the least.



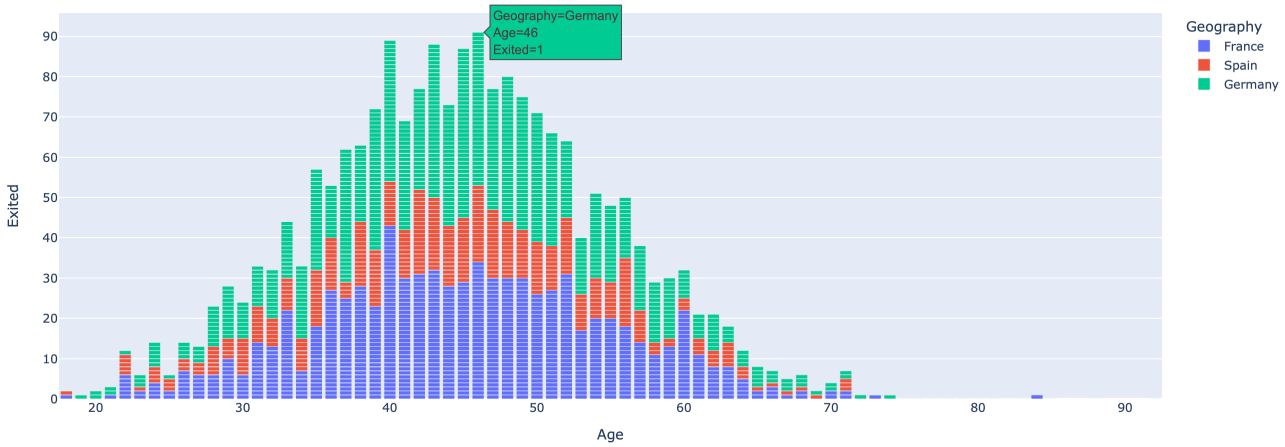
Churn wrt. Geography - Shows the percentage of customers who churned in each country. Germany had the highest percentage of customer churn even though Germany had the least number of customers. This is closely followed by France.



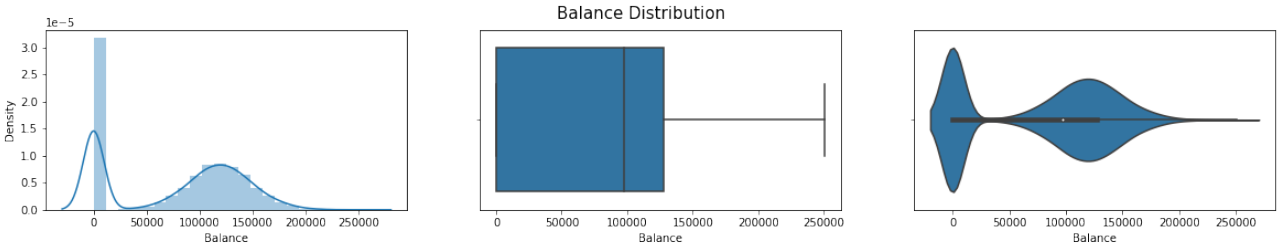
Age Distribution - On plotting a Histogram of age (20 bins) it can be seen that the most customers were between the ages of 30 and 40



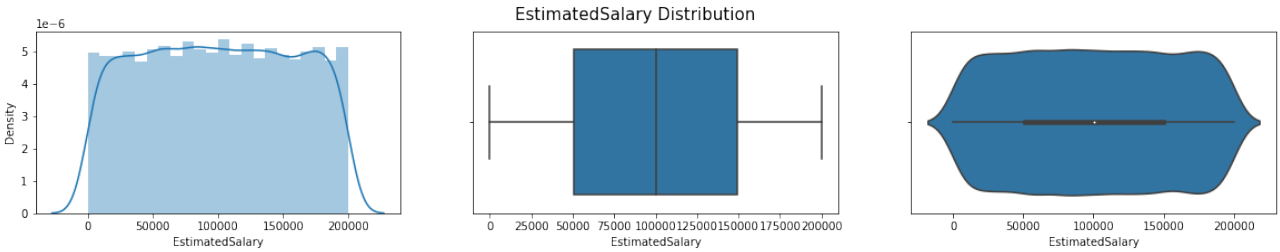
Age and Geography wrt. Churn - The highest number of people exiting from the bank, in total, are 46 years old (although as seen from the previous graph, the most number of customers were between 30 and 4



Finding the distributions of some other continuous variables -

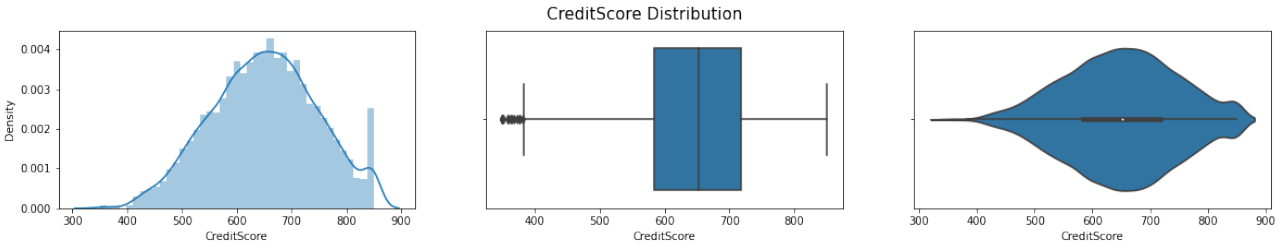


A surprising number of customers had a balance of 0, while the rest of the customer data followed a normal distribution

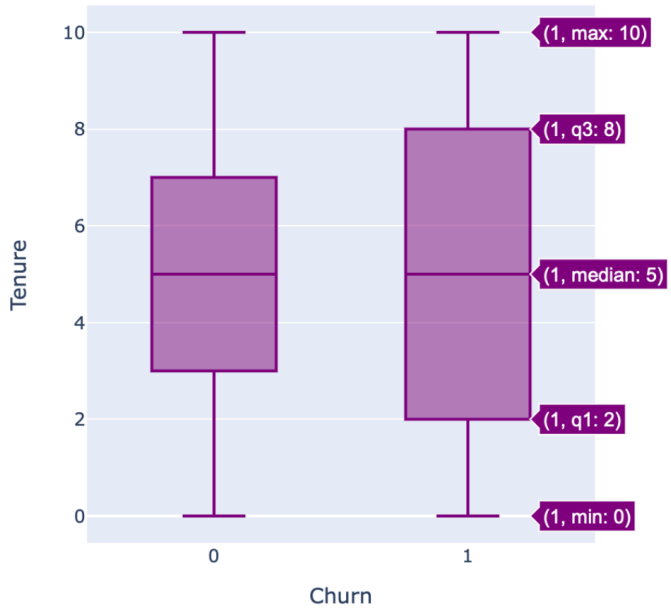


The estimated salary distribution indicates that the dataset has information about a similar number of customers with a salaries of across the board (uniform distribution)

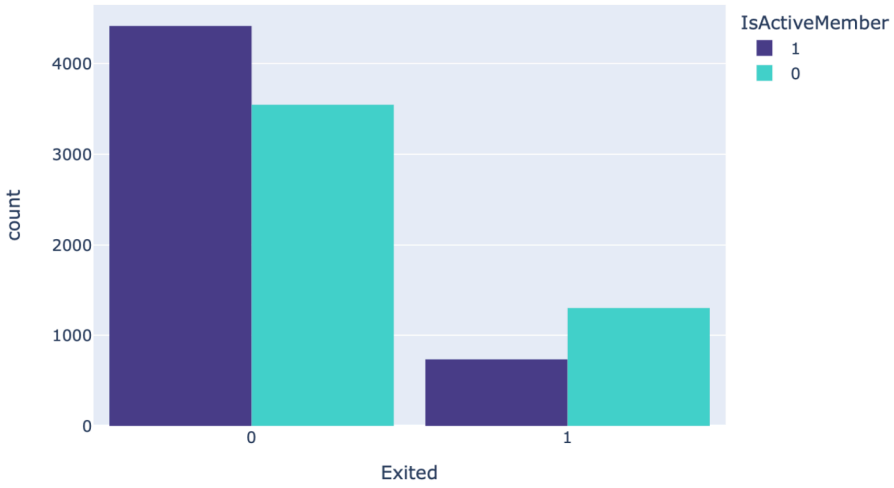
Credit score mostly follows a normal distribution, with a slight peak towards the end



Churn wrt. Tenure - Customers who exited and customers who stayed had a very similar tenure, (both with a median of 5), so it seems Tenure had little effect on churn, however, the larger inter quartile range of Exited customers indicates that those customers have a higher variability than ones who did not exit.

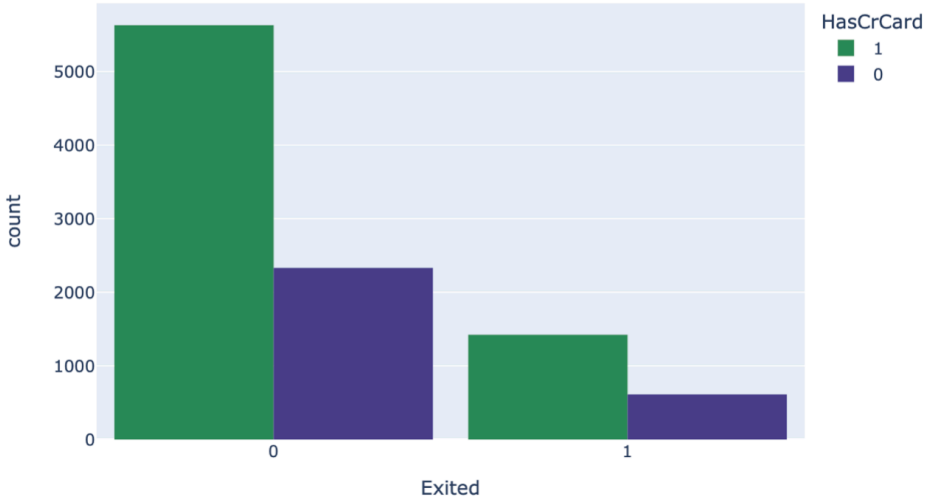


Churn wrt. Active Member - Unsurprisingly, customers who stayed with the banking service had a higher number of active members while customers who exited had a higher higher number of inactive members

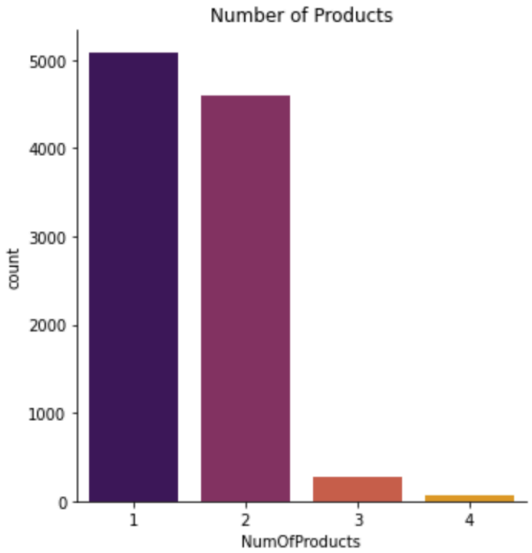


Churn wrt. Credit Card Owner - Customers who stayed as well as ones who left had higher number of people with a credit card.

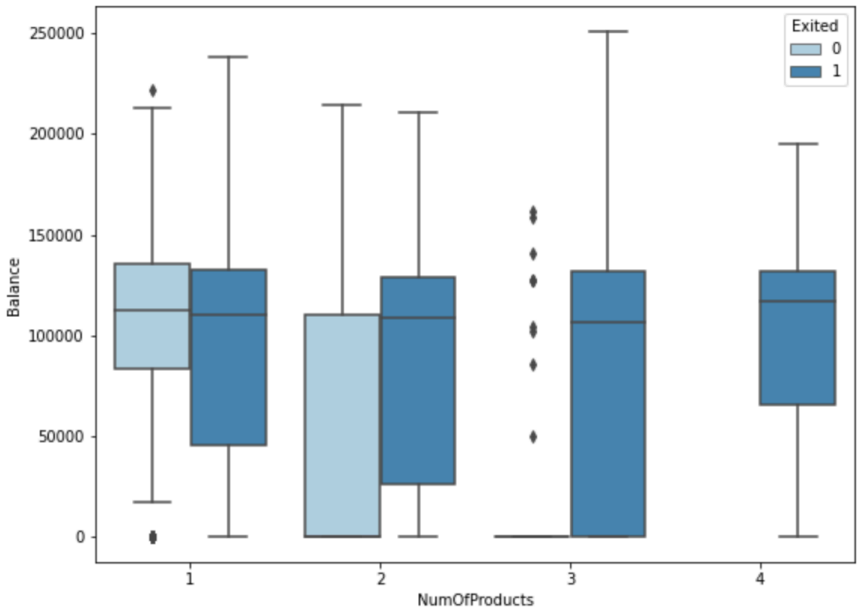
Further inspecting the data it is seen that 69.9% of customers who exited had a credit card while 70.9% of customers who stayed had a credit card. Hence, they show similar distributions wrt. Churn



Number of Products Distributions - A large majority of people had either one or two products offered by the bank, with very few customers having four



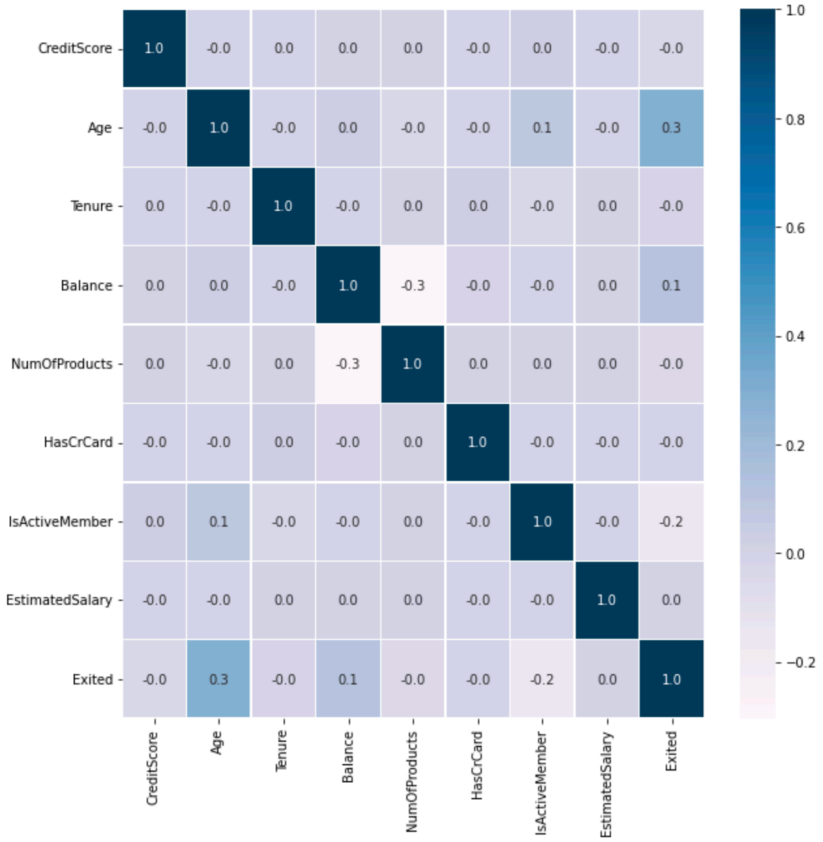
Furthermore, we see its effect on churn - All customers who owned 4 products exited the banking service, most that had 3 products exited as well (with a select few who stayed).
On comparing this to their balance, it is seen that customers who exited had a similar balance across customers owing 1/2/3/4 products



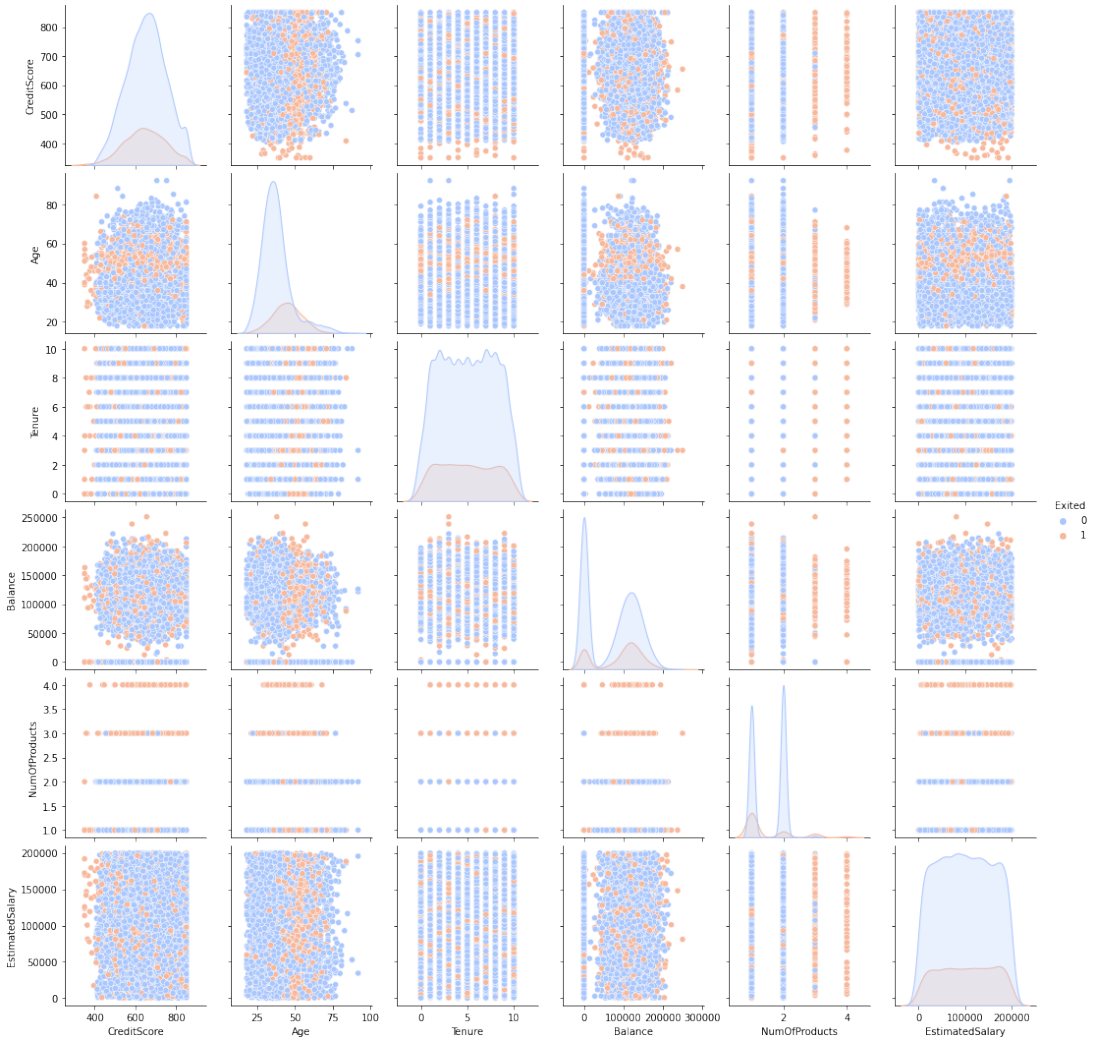
Heatmap - Finally, the correlation between the variables are calculated (geography and gender are one hot encoded and left out for the purpose of this heat map).

IsActiveMember and Exited have a correlation of -0.2, this corresponds with the previous findings; Age and Exited have a correlation of 0.3

These correlations are the ones ones of significance, however, the correlation is considerably low. Hence, we can conclude that there is no multicollinearity between these variables.



Scatter Plot Matrix - Finally, a scatter plot matrix is constructed to visualizes the relationships between all pairs of variables.



Model Exploration:

The problem that is being addressed is a supervised classification problem, there exist several models that can be implemented. For the purpose of this project we are will be exploring the following models -

Logistic Regression, Decision Tree, SVC, Random Forest, KNN, Gradient Boost, XGBoost Classifier, LGBM Classifier, and Neural Networks

Most of these models can be implemented using sklearn. XGB (used to check how it outperforms GradientBoost), LGBM (used since it is said to concentrate on accuracy and speed) have their own packages in python.

Model Implementation and Performance Evaluation:

The original dataset contains 7963 number of records for customers who have stayed with the bank and 2037 number of records for customers who have exited. The minority class (Exited =1) is the class that we want to place more importance on, so to correct this imbalance we will test out a synthetic data generation methods for over-sampling called SMOTE (Synthetic Minority Over-sampling Technique). A comparison between the a few models' performances on the original dataset and the over-sampled dataset will also be used to pick the final model. Accuracy, precision, recall, f1-score, and confusion matrixes will be used to check model performance, in addition to this ROC curve will be plotted for every classifier along with checking for the AUC value.

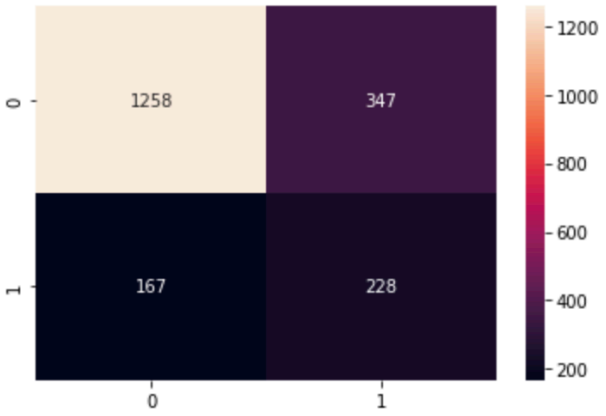
Upon implementing SMOTE we have 6358 number of records with exited value = 0 as well as for exited value = 1. It is important to note that in the case of our dataset problem, the goal is to find customers who are churning which means additional importance will be placed on the Recall value (minimize the chances of missing customer who might actually exit).

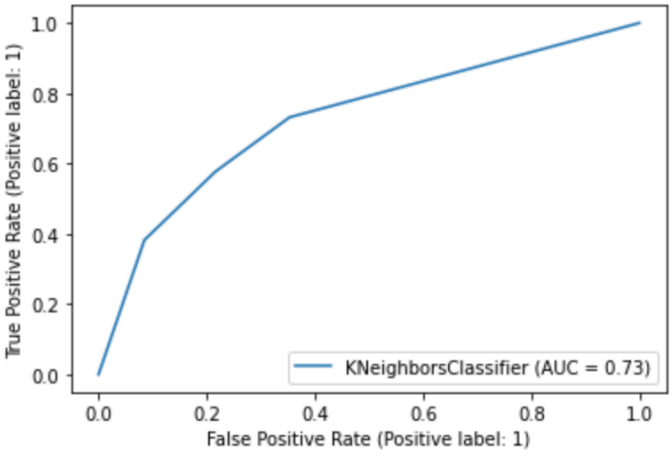
KNN:

K-Nearest Neighbors takes k nearest neighbors whose distances form that point are minimum and computes the average of those values. Although an accuracy of 74% is good enough we see poor performance in terms of recall, precision, and f1 score for the class of interest (exited = 1)

	precision	recall	f1-score	support
0	0.88	0.78	0.83	1605
1	0.40	0.58	0.47	395
accuracy			0.74	2000
macro avg	0.64	0.68	0.65	2000
weighted avg	0.79	0.74	0.76	2000

Classification matrix and ROC curve is plotted to evaluate performance and we see the AUC value = 0.73





Logistic Regression:

Logistic Regression is being used as it is another model that is simple and easy to implement, it is less prone to overfitting lower dimensional classes. On implementing the model we find the classification report for -

the imbalance data:

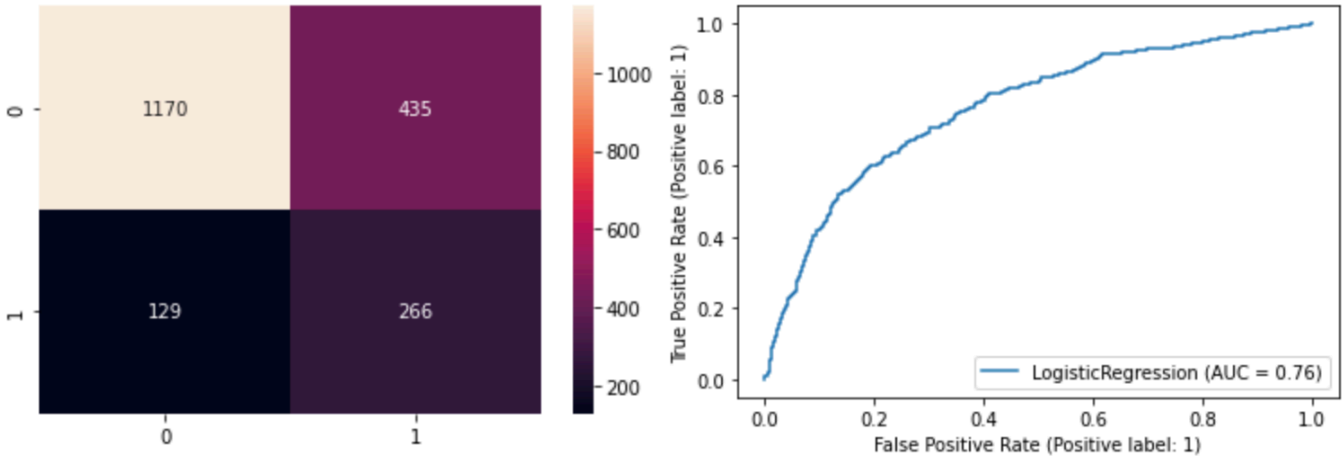
	precision	recall	f1-score	support
0	0.83	0.96	0.89	1605
1	0.57	0.23	0.32	395
accuracy			0.81	2000
macro avg	0.70	0.59	0.61	2000
weighted avg	0.78	0.81	0.78	2000

the balanced data:

	precision	recall	f1-score	support
0	0.90	0.73	0.81	1605
1	0.38	0.67	0.49	395
accuracy			0.72	2000
macro avg	0.64	0.70	0.65	2000
weighted avg	0.80	0.72	0.74	2000

Here is is important to note that although accuracy decreased, the value of recall for class 1 changed from 0.23 to 0.67 (and a higher f1 score), hence indicating how it is necessary to have balanced the dataset for the purpose of finding more accurate predictions for customers who might have exited.

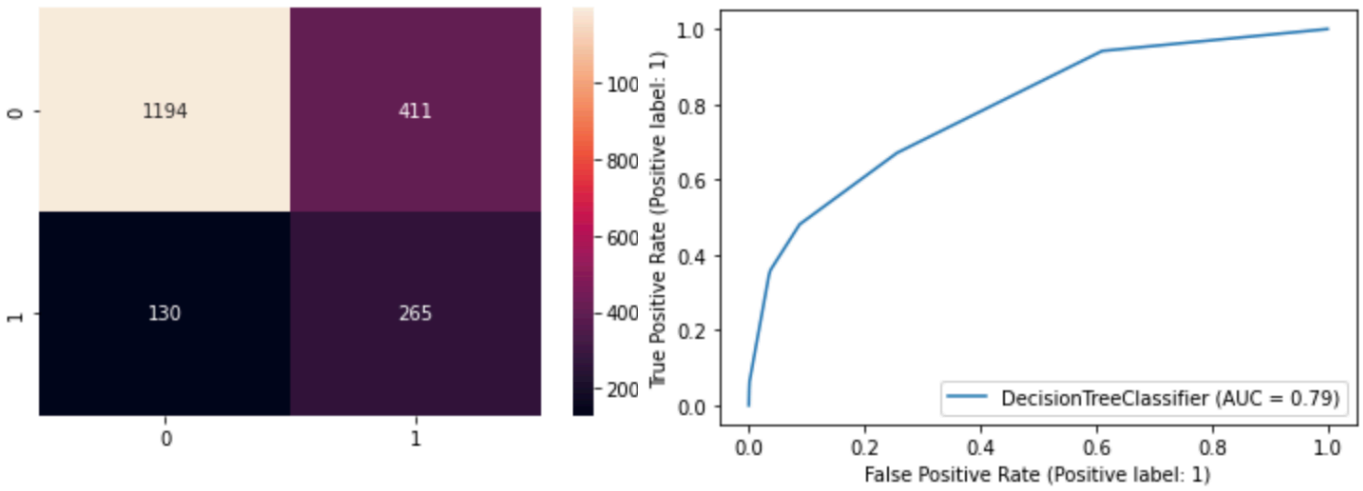
Comparing its performance evaluation with that of the KNN model we can say that it is pretty much on par with KNN in terms of how well the model performed for our dataset.



Decision Tree:

The decision rules are generally in form of if-then-else statements. The deeper the tree, the more complex the rules and fitter the model. Although decision tree has an accuracy that is comparable to the models used above, the AUC value is 0.79 which makes it better at separating the two classes.

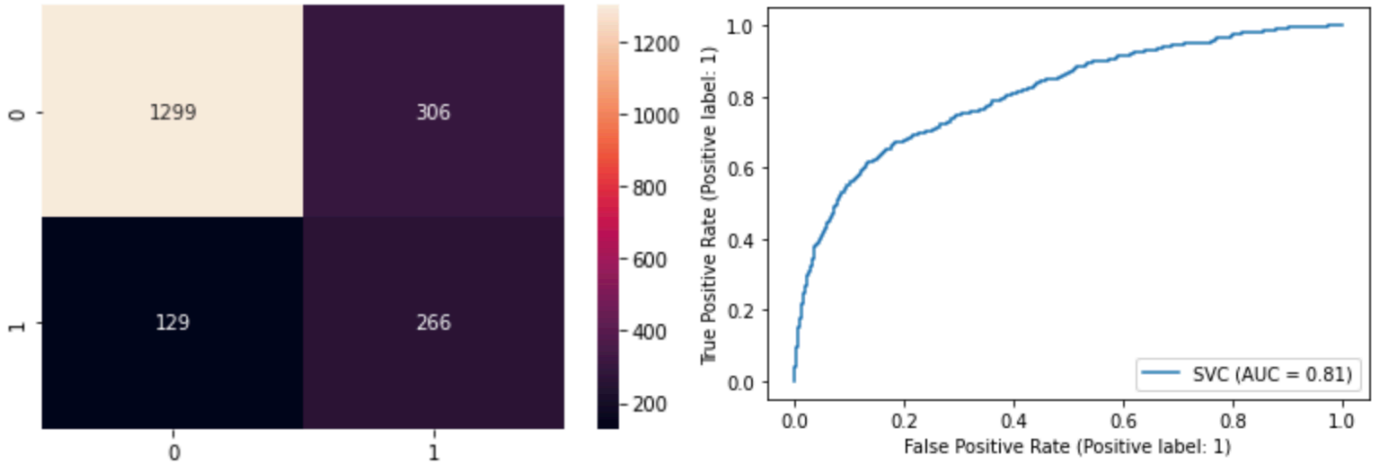
	precision	recall	f1-score	support
0	0.90	0.74	0.82	1605
1	0.39	0.67	0.49	395
accuracy			0.73	2000
macro avg	0.65	0.71	0.66	2000
weighted avg	0.80	0.73	0.75	2000



SVC:

SVC is a common data mining algorithm and gives us an accuracy of 78%, the AUC value is 0.81 which is a good performance value. Although SVC takes much longer to complete execution due to its high time complexity.

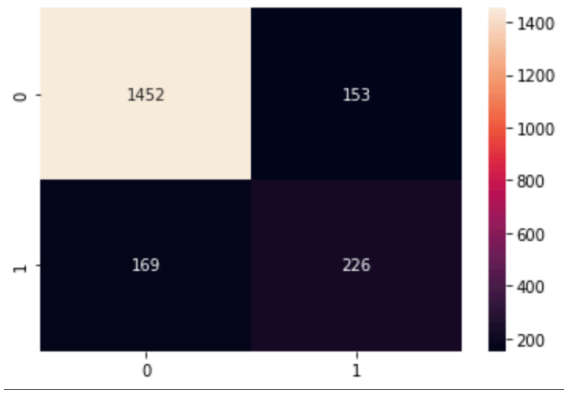
	precision	recall	f1-score	support
0	0.91	0.81	0.86	1605
1	0.47	0.67	0.55	395
accuracy			0.78	2000
macro avg	0.69	0.74	0.70	2000
weighted avg	0.82	0.78	0.80	2000

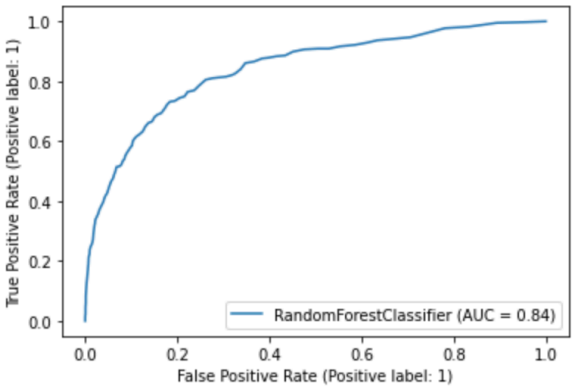


Random Forest:

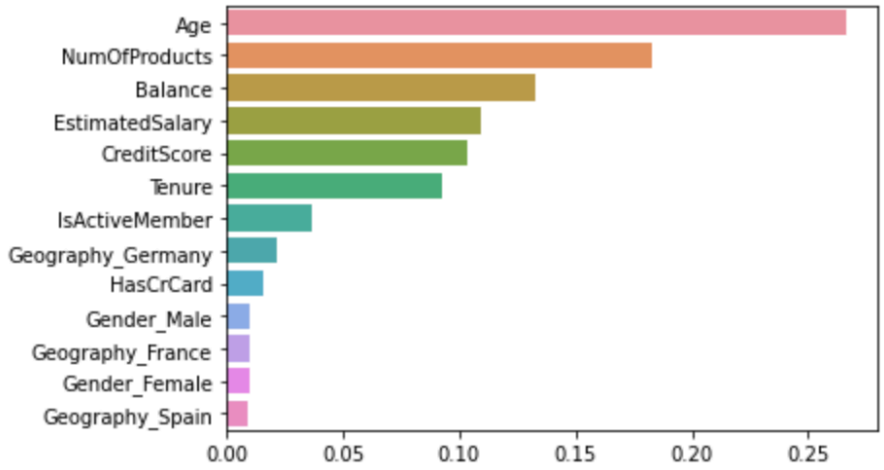
The Random Forest algorithm establishes the outcome based on the predictions of the decision trees. It predicts by taking the average or mean of the output from various trees. It utilizes ensemble learning, which is a technique that combines many classifiers to provide solutions to complex problems. Due to the added advantage provided by the many classifiers the accuracy of random forest is considerably higher than the previous models (84%) and an AUC value of 0.84.

	precision	recall	f1-score	support
0	0.90	0.90	0.90	1605
1	0.60	0.57	0.58	395
accuracy			0.84	2000
macro avg	0.75	0.74	0.74	2000
weighted avg	0.84	0.84	0.84	2000





Since this model performs considerably well, we look into it further by examining what features it places more importance on by using a barplot:



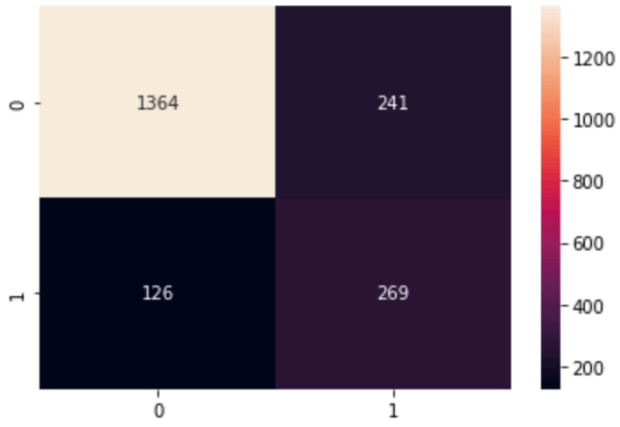
Age and No. Of products seem to be the features of most importance, we also see that among the dummy variables for geography Germany is the only one with any relevant significance.

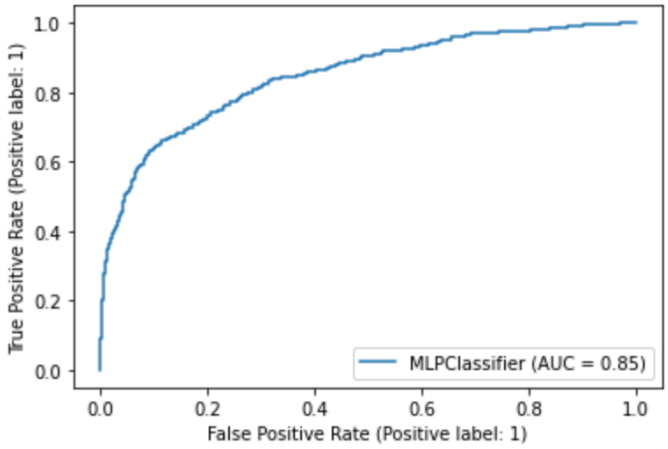
Neural Network:

The main reason for using neural network would be to leverage their ability to identify non linear relationships among the variables. Manually trying out different values for number of nodes (10,15,20,25) and different number of hidden layers (3,4,5) we find that using three layers with 15 nodes gives us the best value.

	precision	recall	f1-score	support
0	0.92	0.85	0.88	1605
1	0.53	0.68	0.59	395
accuracy			0.82	2000
macro avg	0.72	0.77	0.74	2000
weighted avg	0.84	0.82	0.82	2000

Using neural networks also gives us the highest Recall, F1 Score, and AUC value among all the previously tested datasets. Due to the number of iterations it takes an understandably long amount of time to complete execution.

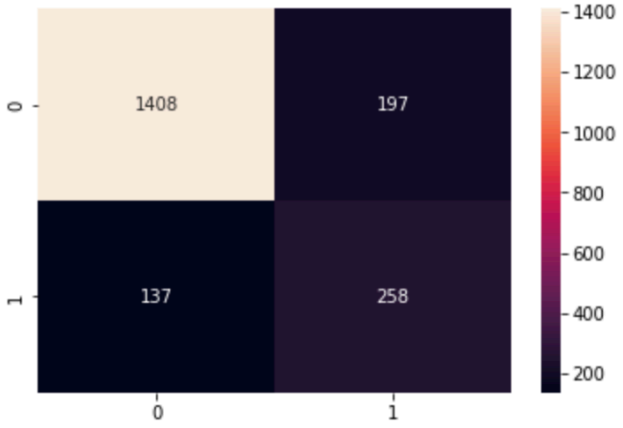


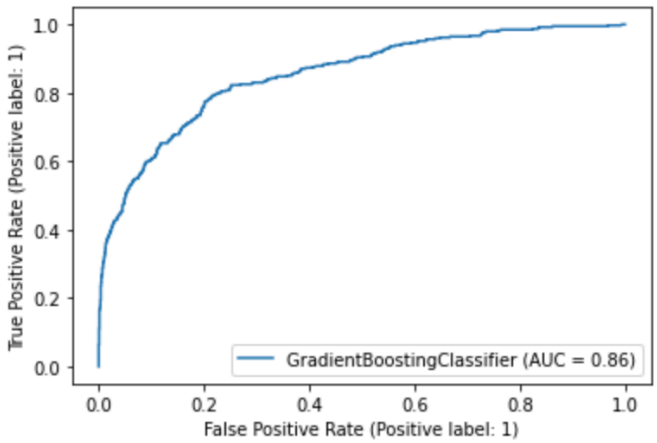


Gradient Boost:

Gradient boost focus on minimizing overall prediction error, it works on doing so with every subsequent model generated. Light Gradient Boosting Method and eXtreme Gradient Boost are also used to improve upon the Gradient Boost model. All performance evaluation metrics used indicate good performance.

	precision	recall	f1-score	support
0	0.91	0.88	0.89	1605
1	0.57	0.65	0.61	395
accuracy			0.83	2000
macro avg	0.74	0.77	0.75	2000
weighted avg	0.84	0.83	0.84	2000

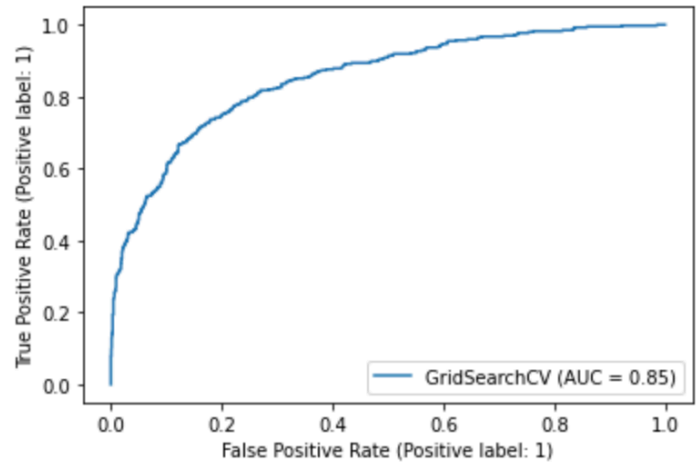




LGBM Classifier:

The LGBM classifier is essentially gradient boosted decision trees. A range of parameters are specified to try and improve the model, such as specifying the number of leaves, learning rate, lambda and alpha (L2 and L1 regularization) values - to prevent overfitting, number of estimators, and fraction of features to consider. This is done along with using grid search cross validation to find the best fit model parameters for lgbm. The recall value improved from 0.65 in Gradient boost to a 0.68

	precision	recall	f1-score	support
0	0.92	0.86	0.89	1605
1	0.55	0.68	0.61	395
accuracy			0.83	2000
macro avg	0.73	0.77	0.75	2000
weighted avg	0.84	0.83	0.83	2000



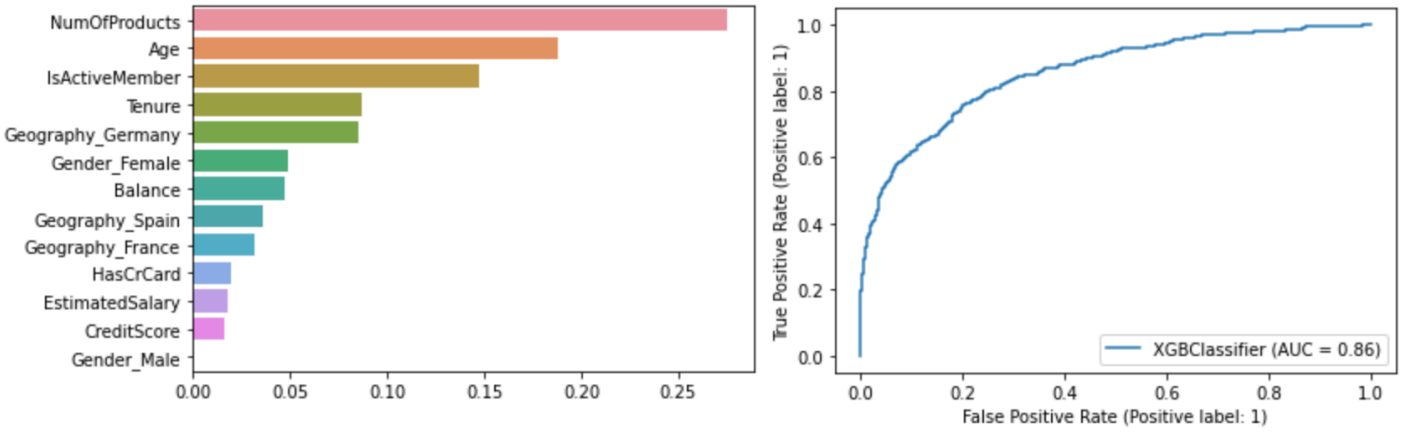
XGBoost Classifier:

XGBoost is an optimized distributed gradient boosting library designed to be highly efficient, flexible, and portable. It implements machine learning algorithms under the Gradient Boosting framework. XGBoost provides a parallel tree boosting that solves many data science problems in a fast and accurate way. It is also much faster than other boosting algorithms.

Parameters specified for XGB include, base score - which is a minimum initial prediction score of all instances, booster - tree or linear models, gamma - minimum loss reduction required to make next partition, learning rate, max depth, and number of estimators

	precision	recall	f1-score	support
0	0.90	0.93	0.91	1605
1	0.66	0.58	0.62	395
accuracy			0.86	2000
macro avg	0.78	0.76	0.77	2000
weighted avg	0.85	0.86	0.85	2000

Since this gives us the highest accuracy of all the models used as well as the highest AUC value, we examine the feature importance and find similarities to that of Random Forest as it also places most importance of Age and Number of Products (however, in this case the feature importance of number of products is more than that of age).



Project Results

Insights drawn from the model implementation include:

- Location played a minimal role in determining whether a customer would leave the banking service except that of Germany, being female played a greater role in determining churn. Calculating the percentage we see that female customers are more likely to leave the bank at 25%, compared to 16% of males.
- If we were to group the variables ‘Estimated Salary’ and ‘Credit Score’ to be closely related based upon the assumption that customers with higher salary are more likely to have a higher credit score, we can see a great deal of difference place on the importance of these features in the Random Forest model (ranked among top 5) and the XGB Classifier (ranked among bottom 5)

- LGBM and Neural Network gives the highest recall value which is the general goal for customer churn tasks hence making them great models to use as well, however due to the sheer number of iterations - 200 for LGBM and 300 for Neural Networks (even for this relatively small dataset) it takes about 25 and 49 seconds respectively.
- XGB gives the highest accuracy with the highest f1 score, which is why it would be one of the most useful models to use for this problem. It take noticeably less time to execute even with the number of estimators being set to a 100. Time taken to execute this model was 2.2 seconds while that of gradient boost was 3.2 seconds and random forest was 1.7 seconds

Impact of the Project Outcomes

Although having high accuracy for any given a model is a great way to asses the probability of successfully achieving our goal, the purpose of this project was not entirely focused on model accuracy. We wanted to analyze the model performance on factors like recall and f1 score, see how changing certain aspects of our dataset, like scaling and correcting class imbalance, would affect the results, as well as record how different model parameters would change our results. At the core of this project is finding out reasons why a customer would leave the banks' services, a few important observation and solutions are as follows:

- 1 in 3 people in Germany exited while also having a higher balance (lucrative), hence necessitating a look into competition in Germany - (what other bank are the customers a part of?) and a look into whether services in Germany are somehow subpar
- Customers who have 1 or 2 products have mostly stayed as opposed to customers who use 3 or 4 products, this could be because these products are not as good and require improvement and better customer service
- We see that active members have lesser chances of leaving hence it would be important to keep the customers engaged, as well as maintain the interest of female customers

Customer retention is a factor that has many facets and does not improve overnight, however if we keep up the interest of existing customers and find out reasons as to why customers are leaving they can be reeled back in with a few solid strategies and improve customer loyalty.