Customer Churn Prediction Using Supervised Machine Learning Algorithms

Tamanna Manek (tm3734)

Ritik Lnu (rl4017)

Deepti Gouthaman (dg3781)

Guided by: Aline Bessa

1. Abstract:

In this project, we will be trying to predict the customers who are about to churn from a credit card company by using a supervised learning model and deriving with an accuracy of our predictions. Below are the three important aspects of the project:

- we will be Performing exploratory data analysis so that the bank can take enough insights from it to understand which kind of customers are more likely to leave.
- Building and tuning classification machine learning model to safely predict whether a customer will leave or not and the bank can target the leaving customers accordingly.
- Featuring importance, so that the bank can mitigate further churn by applying necessary remedial and actions.

2. Introduction:

Customer churn is one of the maximum critical and tough troubles confronted in corporations consisting of credit score card agencies, cable providers, SASS providers, and telecommunication agencies international because it entails excessive risks, excessive earnings, and revenue. Customer churn additionally recognized as "consumer attrition", is certainly considered one of the largest troubles confronted through the enterprise nowadays because it at once affects operating, marketing, and agency budgets. It normally represents the wide variety of clients who forestall shopping for out of your enterprise inside a fixed time frame. If we should discern out why a consumer leaves and once they depart with affordable accuracy, it'd immensely assist the company to strategize their retention tasks manifold. In this version, we can be predicting What is the probability of an energetic consumer leaving a company and are key signs of a consumer churn. These kinds of evaluation are in particular utilized in Business Analytics which analyzes the important thing factors of an enterprise (Finding the capabilities that at once affect the enterprise more) and uses those capabilities to give you giant enterprise solutions.

.

3. Models :

We will be using supervised learning to predict customer churn. A supervised machine learning algorithm generally analyzes the training data and produces an inferred function that in turn can be used for mapping new examples. Given that we have data on current and prior customer transactions in our dataset, this is a standardized supervised classification problem that tries to predict a binary outcome (Y/N).

- **Decision tree classifier:**

Here we use a Decision tree classifier as our base model which we further expand into the development of Random Forest and XGBoost models. All three models are focused on making Binary outcomes or solutions (Weather the customer will churn/ The customer will not churn).

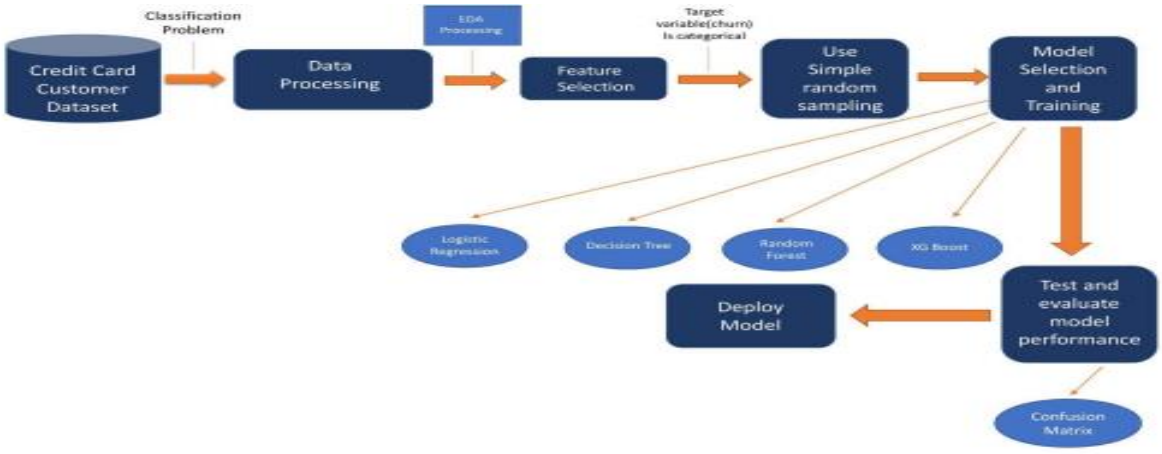
- **Random Forest Classifier**

Random Forest is used mainly for the classification of data and it is a supervised algorithm that implies that the data fed to the model for training is labeled. As we know that a forest is made up of trees and more trees means a more robust forest. Random forest algorithm implements a voting method wherein the decision trees are used on the data and thereafter selects the optimal solution by voting. Random forest performs better than a single decision tree as it averages the result and thus reduces over-fitting.

- **XGBoost**

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 (also known as GBDT, GBM) that solves many data science problems in a fast and accurate way. The same code runs on major distributed environments (Hadoop, SGE, MPI) and can solve problems beyond billions of examples.

4. Workflow Diagram:

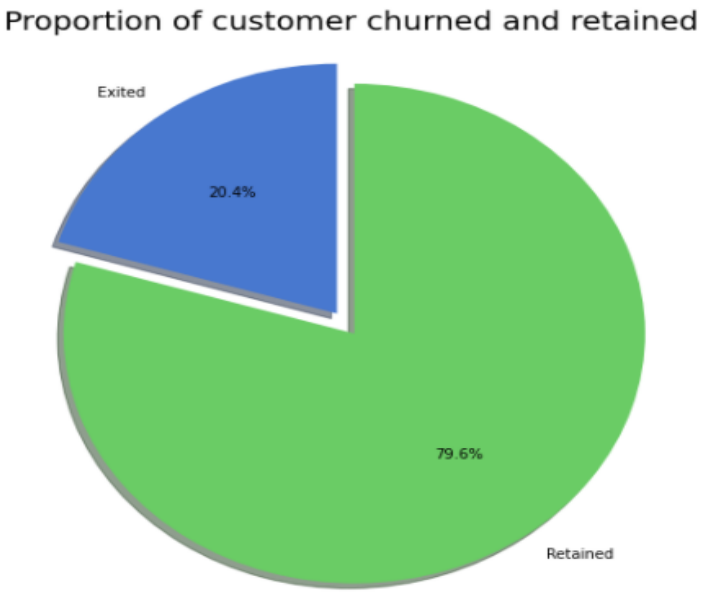


5. Data collection:

Our data set is from a credit card company, where we are able to review customer attributes such as geography (location), gender, age, tenure, balance, number of products they are subscribed to, their estimated salary, and if they stopped the subscription or not (Exited). Our data contains 10000 rows and 13 columns with 3 data types (int, object, and float), the attributes to our project that contains the maximum information gain/ key attributes are Geography, gender, and age. We are using open-source reliable datasets from Kaggle(https://www.kaggle.com/kmalit/bank-customer-churn-prediction/data). As shown in the figure below there are no null entries, which implies that we do not have any missing values in our data set. Hence no further data cleaning or preprocessing of data was required during the development of the project. The screenshot below also lists all the attributes contained in our dataset.

```
In [6]: df.info()
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 10000 entries, 0 to 9999
Data columns (total 14 columns):
 #   Column           Non-Null Count  Dtype
---  ------           --------------  -----
 0   RowNumber        10000 non-null  int64
 1   CustomerId       10000 non-null  int64
 2   Surname          10000 non-null  object
 3   CreditScore      10000 non-null  int64
 4   Geography        10000 non-null  object
 5   Gender           10000 non-null  object
 6   Age              10000 non-null  int64
 7   Tenure           10000 non-null  int64
 8   Balance          10000 non-null  float64
 9   NumOfProducts    10000 non-null  int64
 10  HasCrCard        10000 non-null  int64
 11  IsActiveMember   10000 non-null  int64
 12  EstimatedSalary  10000 non-null  float64
 13  Exited           10000 non-null  int64
dtypes: float64(2), int64(9), object(3)
```

80% of the customers in our data set are retained and 20% of them have decided to leave the credit card company. Below is the graphical representation for the same.

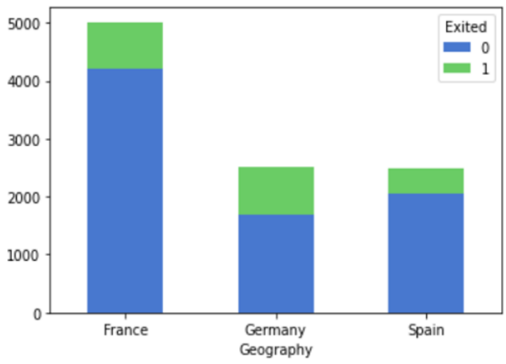


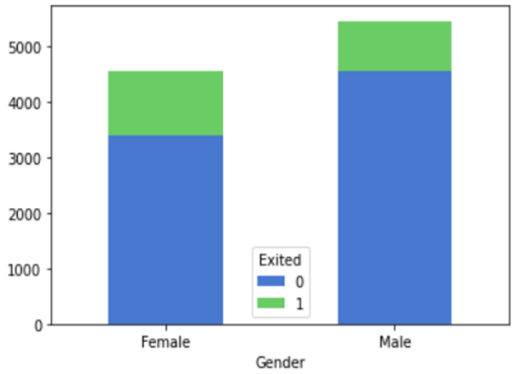
6. Explanatory Data Analysis

We have two important categorical variables in our project

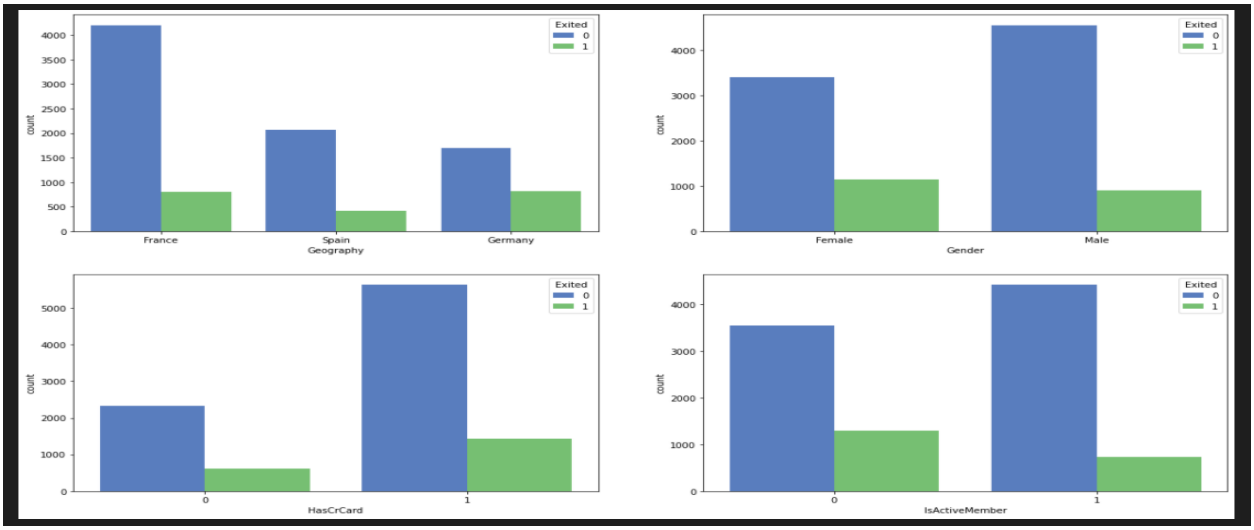
1) Geography
2) Gender

We are handling our categorical variables through the Label Encoding technique.
In label encoding, the labels are transformed into numeric forms in order to be readable by machines. By using machine learning algorithms, we can then decide more effectively how to operate the labels. Preprocessing a structured dataset is an important step in supervised learning.



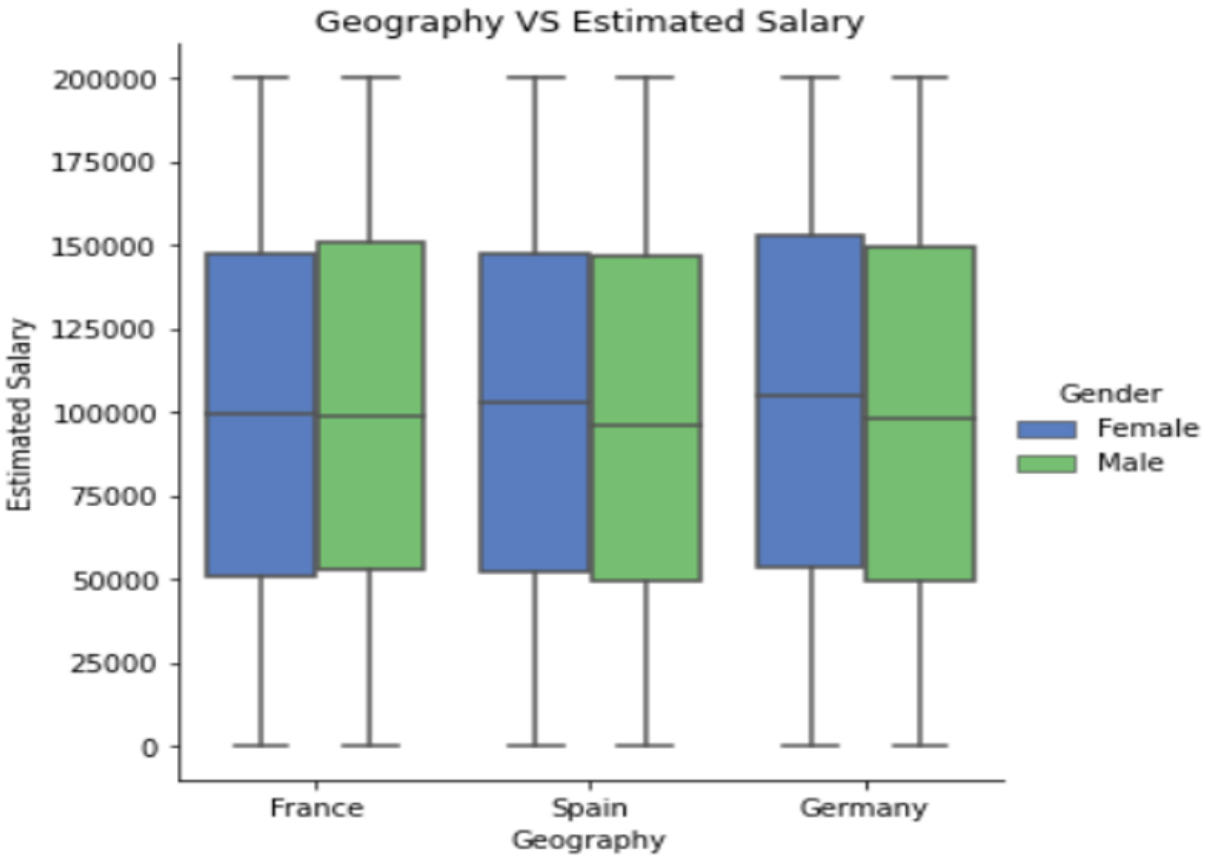


In the above two diagrams, we can see that the number of people leaving the bank is not dependent on the population of the people in that area. As we can see in the first figure, the churning rate in Germany and France is almost the same, despite Germany's population being substantially lower than France's. In the second figure, we see that women are more likely than men to leave the bank. Although the female population is smaller than that of males, the proportion of churning is higher in females.



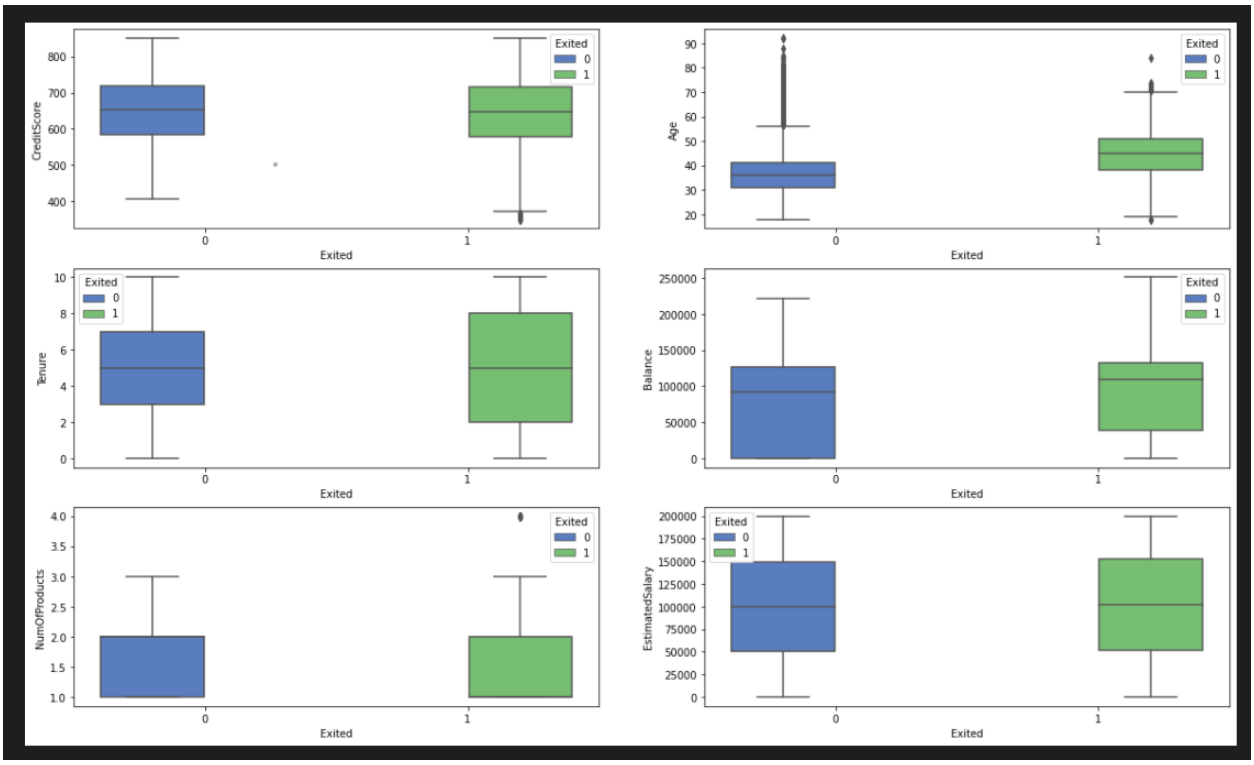
From the above figure, we note that:

- While most of the data comes from French persons, the churn rate is closely related to the population of customers, suggesting that the bank may have a problem (perhaps not enough customer service resources) in those areas where they have fewer customers.
- In addition, the percentage of female customers churning is higher than that of male customers. Therefore, the bank should look for some distinctive offers that are appealing to female customers in order to retain them.
- Interestingly, most of the churning customers are those with credit cards. It could be that having credit cards is purely coincidental
- The bank's inactive customers have a higher churn rate, but what's more concerning is that the overall proportion of inactive members is quite high, suggesting that the bank might need to implement a program to turn them into active customers, which will definitely have a positive effect on customer churn.



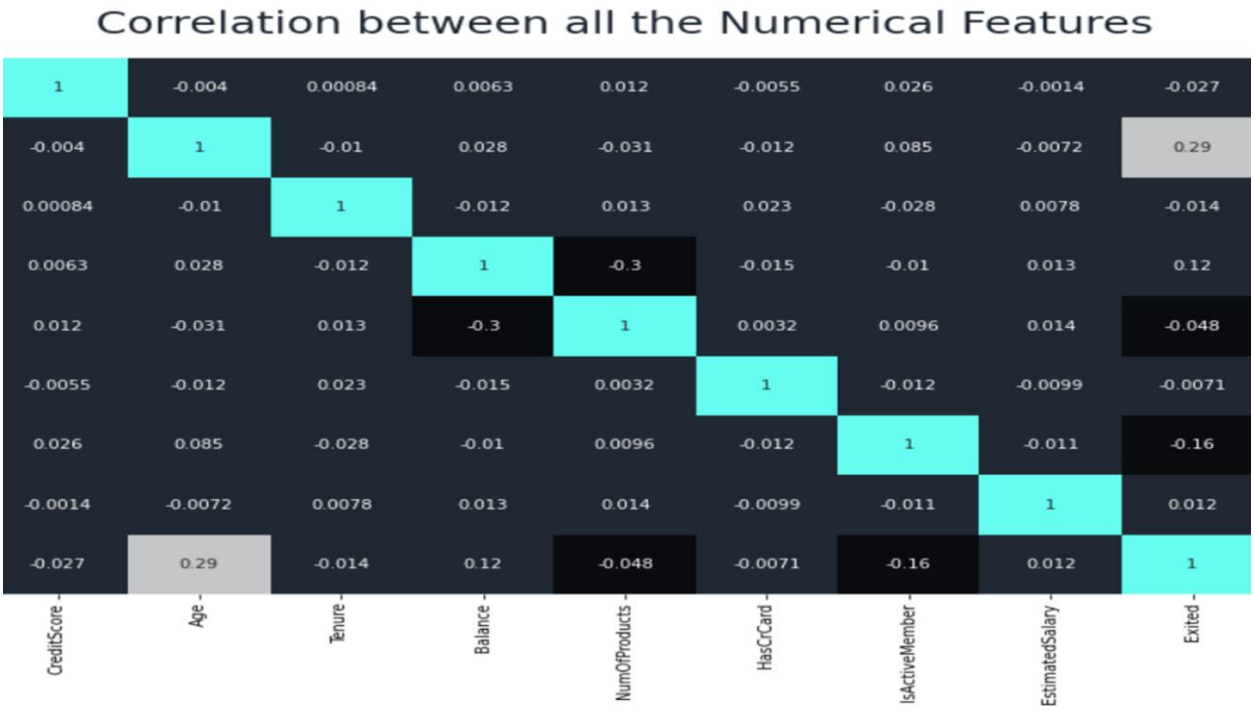
As shown in the above figure, the Estimated Salary of males and females in different locations is almost the same. Males earn slightly more than females in France, while females earn slightly more than males in Germany.

Relationship between exited and other variables



We note the following:

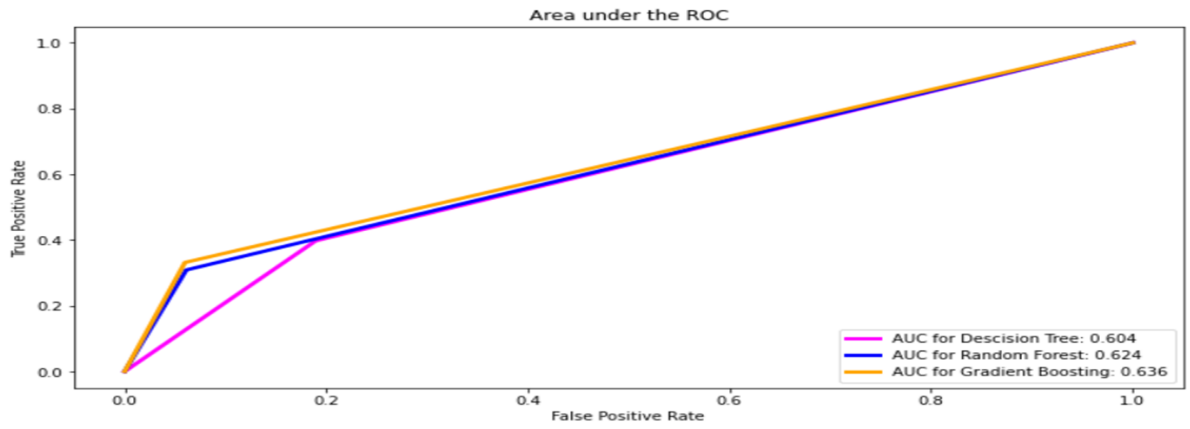
- A comparison of retained and churned customers shows no significant differences in credit scores.
- It appears that older customers churn at a higher rate than younger customers, indicating a difference in service preferences between the age groups. The bank may need to evaluate its target market or its retention strategy.
- Clients who have either spent little time with the bank or very much time with the bank are more prone to churning than those who have an average tenure.
- Sadly, the bank is losing customers with large bank balances, which could reduce its lending capacity.
- No significant relationship exists between the number of subscribers or the salary and the likelihood of churn.



Generally If the dataset has perfectly positive or negative attributes then there is a high chance that the performance of the model will be impacted by a problem called — "Multicollinearity". Luckily, decision trees and boosted trees algorithms are immune to multicollinearity by nature. because when they decide to split, the tree will choose only one of the perfectly correlated features. But just to make the job easier we were trying to just check for dimensionality reduction of obvious correlation. We found no highly correlated attributes and hence no attributes were to be removed.

7. Training and Evaluating Models:

Using 80:20 Training , Testing Split :



Using the base parameters for the Decision Tree we found AUC to be 0.604 and also found that it overfits the training dataset. We also found that the Decision Tree Algorithm to be very slow and in the real world where if the bank tries to use this algorithm on its dataset this algorithm would require high compute time and resources and also any addition to data points makes the algorithm run from the start. Exploring alternatives for Decision Tree we found Random Forest and Gradient Boosting algorithms to be more effective. Training these models on the same datasets using the same strategy and finding out the ROC curve we found out that they had very similar AUC as that can be seen in the figure. Just to get better insights and how to choose the better algorithm for our dataset we needed more evaluation metrics and also run the models on more test data.

Using Cross-Validation:

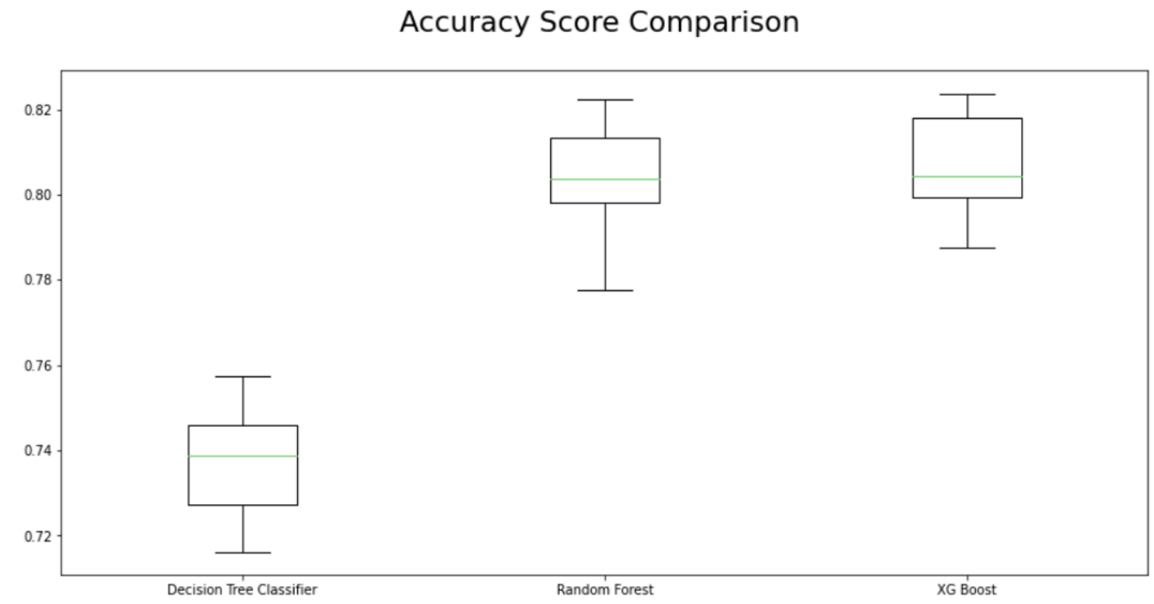
To find the best algorithm and to remove the doubt of an algorithm running with better accuracy only on a specific part of the dataset we used cross-validation with 10 folds. Gathering the mean and standard deviation of all algorithms the below figure shows the summary of how each model performed.

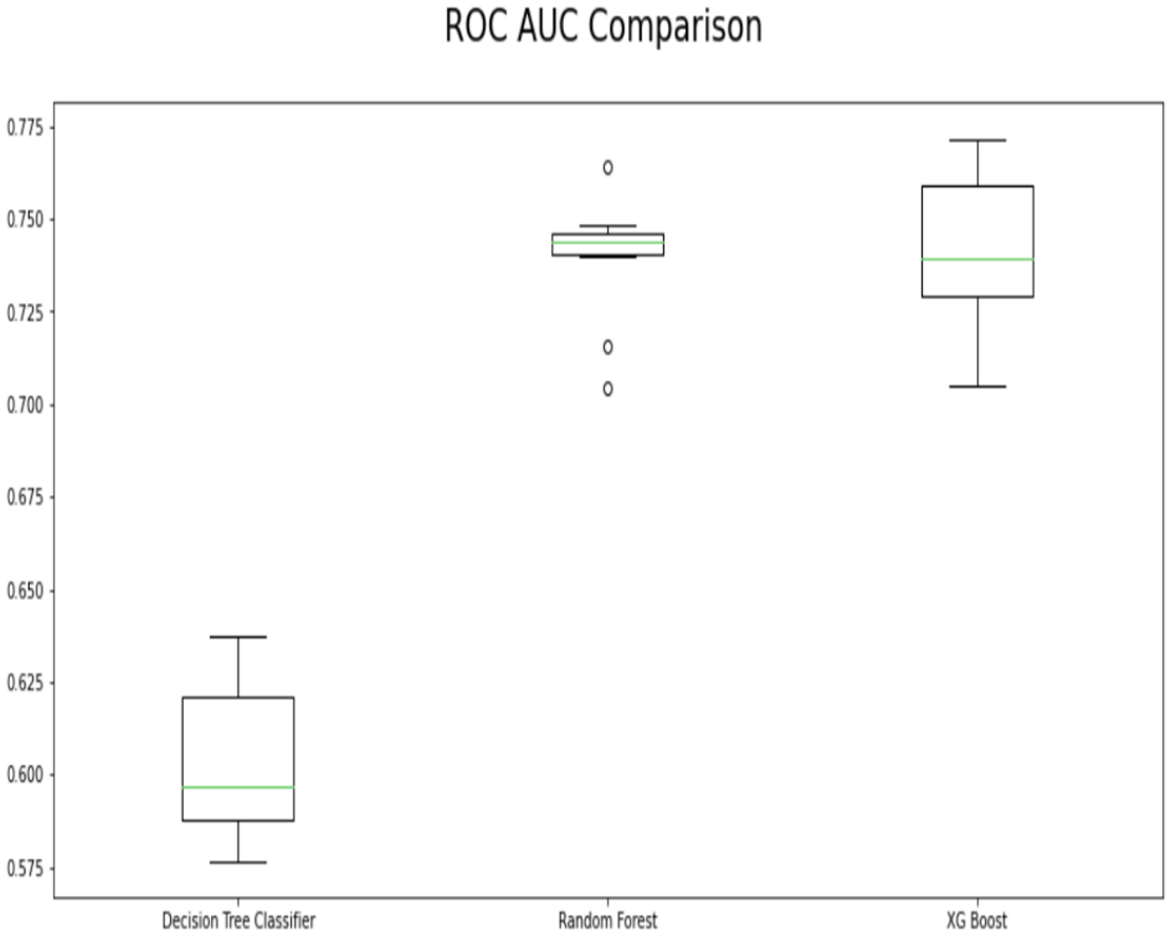
	Algorithm	ROC AUC Mean	ROC AUC STD	Accuracy Mean	Accuracy STD
2	XG Boost	74.17	1.93	80.71	1.11
1	Random Forest	73.94	1.62	80.50	1.29
0	Decision Tree Classifier	60.30	1.96	73.75	1.37

We still found nearly the same results for mean ROC AUC for XGBoost and Random Forest but both surpassed the Decision Tree Classifier with a huge range. We can see that the accuracy standard deviation of Random Forest was greater than XGBoost. To get a better idea we plotted the above summary on a box plot.

We find some outliers for the random forest algorithm in the ROC AUC comparison indicating that it could be not stable on various datasets.

Having an accuracy of 80% is an indication of a good model but there are other metrics that matter for any machine learning model evaluation. Below is the table which gives the summary of all model's evaluation metrics.

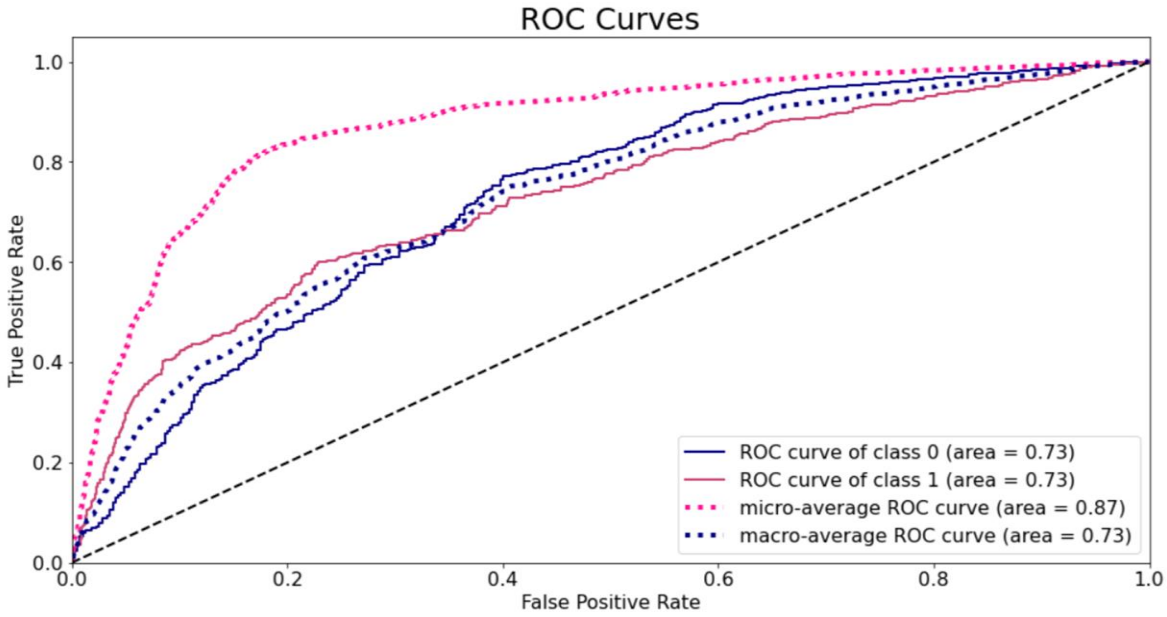

Accuracy Score Comparison
0.82
0.80
0.78
0.76
0.74
0.72
Decision Tree Classifier
Random Forest
XG Boost


ROC AUC Comparison
0.775
0.750
0.725
0.700
0.675
0.650
0.625
0.600
0.575
Decision Tree Classifier
Random Forest
XG Boost

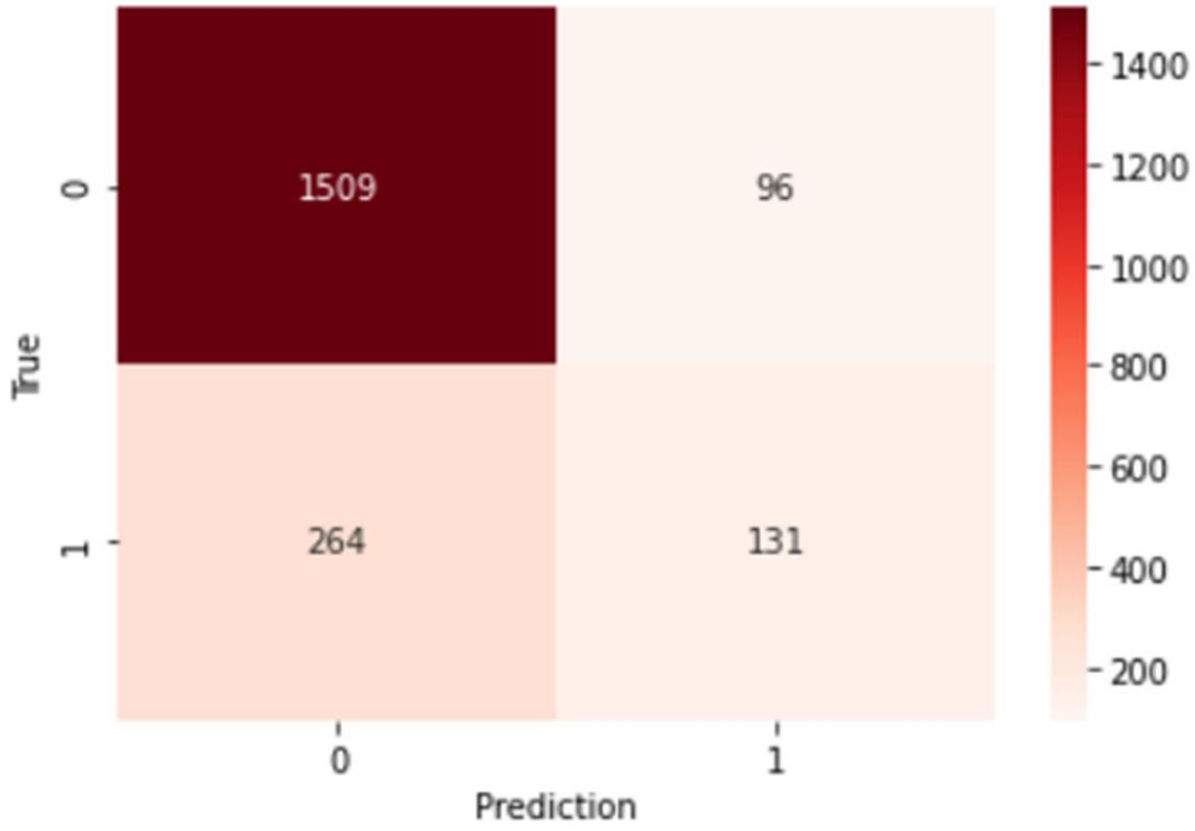
	Algorithm	Accuracy	Precision	Recall	F1 Score	F2 Score
2	XG Boost	0.8200	0.577093	0.331646	0.421222	0.362479
1	Random Forest	0.8135	0.552885	0.291139	0.381426	0.321588
0	Decision Tree Classifier	0.7390	0.353349	0.387342	0.369565	0.380030

We clearly see that XG Boost again beats Random Forest in terms of other model evaluation metrics like precision, Recall, and F-Scores. Based on the above findings, summary, and research on the above algorithms we found XGBoost to be the best suitable algorithm for our dataset.

To get more understanding now on how good the XGBoost algorithm is performing a classification we performed an AUC analysis on positive and negative class classification as seen in the figure.



The average ROC curve indicates that class 1 (in our dataset non-churned) is being classified very effectively and class 0(churned) is also good but not that great as compared to class 1. Precision metric supports this claim, to verify the class1 being classified correctly with that high rate we used a confusion matrix to support and understand this.



We found no surprise as our dataset has 80% of data labeled as non churned and we were only working with 20% of data labeled as churned. Hence the accuracy of predicting non churned customers will be higher than churned customers.

8. Model Hyperparameter Tuning:

Using GridSearchCV strategy we found the best parameters for XGBoosting to be as 'colsample_bytree': 0.8178138551591502,

'gamma': 1.236053761680807,

'max_depth': 15.0,

'min_child_weight': 5.0,

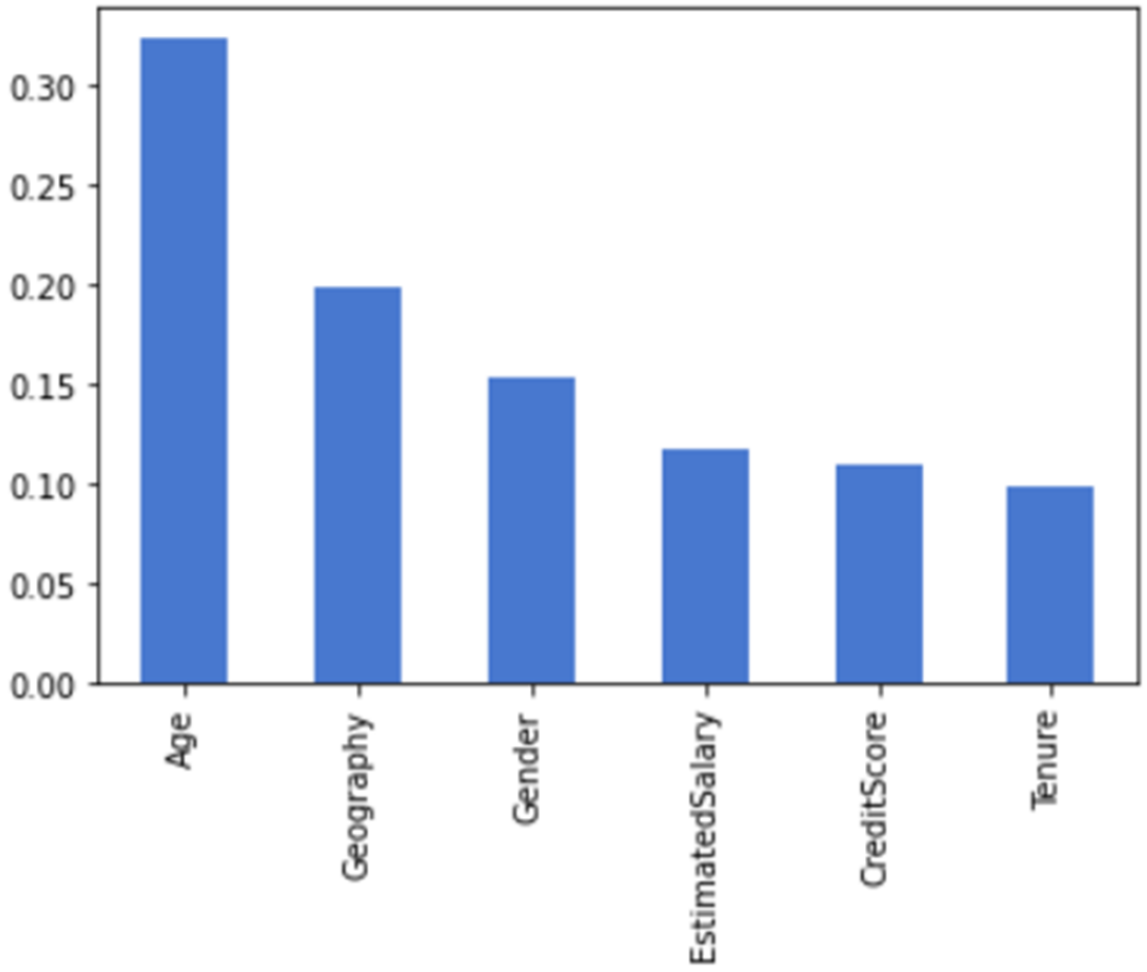
'reg_alpha': 71.0,

'reg_lambda': 0.4127493018835883 ,

Resulting with an highest accuracy of 0.82.

9. Feature Importance :

We find the AGE, Geography and Gender to be top important features which really makes an important decision if the customer will churn or not.



10. Conclusion :

In this project, we found XGBoost to be the best algorithm among Decision Tree and Random Forest for our Dataset and would be reliable for large dataset and parallel processing as well. The feature importance combined with EDA analysis helps the bank to understand which features are important and how they affect a customer to churn and come up with appropriate retaining strategies on predicted customers who are going to churn.