Bank Customer Churn
Prediction

BAN 5573 Project

Bank Customer Churn Prediction

Authors:
Maryam Ahmedi
Dipendra Mainali

Table of Contents

Bank Customer Churn Prediction using supervised and unsupervised classification methods

Project Proposal for BAN5573- Visual Analytics & Business Intelligence
Team members: Maryam Ahmadi- Dipendra Mainali
Data: https://www.kaggle.com/datasets/mathchi/churn-for-bank-customers

1. Introduction

Churn customers are those customers who have stopped being customers during certain periods. Having churn customers is a difficult situation for a bank. They want to reduce the churn of customers as increasing the number of customers would put the business at a loss. Banks usually apply various customer retention strategies in order to reduce churn. They try to make customers renew their subscriptions and continue using their products or services.

2. Problem Definition

There are many reasons for customer churn. Credit score, geography, gender, and age play a certain role in churn. Balance in customers' account, number of products he/she uses, whether they have a credit card from that particular bank if they are active members or not and whether their estimated salaries match their actual salaries are other factors that could be deciding factors for churn in customers. It is usually more expensive to get new customers than keep existing ones. It is very important for banks to know the reason that leads their customers to leave the company. Reducing churns helps banks to increase their profits [1]. In this project, we will analyze, predict and visualize the different factors that affect the churn of customers.

Business/Research Questions

- Predict whether a customer will exit the bank or not, based on his/her characteristics.
- Which factors have the most effect on customer churn?
- What are the values of churn of customers associated with different factors?

3. Literature Review

Customer Churn

Kim et al. (2006, p 104) describe churn as 'the number or percentage of regular customers who abandon a relationship with a service provided. Customer lifetime value (CLV) (Jain and Singh, 2002) represents the present value of the expected benefits minus the costs of initializing, maintaining, and developing the customer relationship. It involves elements such as costs, revenue, a discount rate, and a time horizon. Another important aspect is retention rate or, conversely, probability of churn. As emphasized by Kim et al. (2006), it is necessary to consider the churn probability in the lifetime evaluation of a customer instead of only looking at past profit history. Lima, E., C. Mues, and B. Baesens. [2]

Supervised Learning

Supervised Learning is a machine learning paradigm for acquiring the input-output relationship information of a system based on a given set of paired input-output training samples. As the output is regarded as the label of the input data or the supervision, an input-output training sample is also called labeled training data or supervised data. Supervised Learning has been successfully used in areas such as Information Retrieval, Data Mining, Computer Vision, Speech Recognition, Spam Detection, Bioinformatics, Cheminformatics, and Market Analysis (Wikipedia, 2010). [3]

Unsupervised Learning

Unsupervised Learning refers to algorithms to identify patterns in data sets containing data points that are neither classified nor labeled. The algorithms are thus allowed to classify, label and group the data points within the data sets without having any external guidance in performing that task. The main goal of unsupervised Learning is to discover hidden and interesting patterns in unlabeled data. There are four types of unsupervised tasks: Clustering, Principal component analysis, Anomaly detection, and Autoencoders. [4]

Previous projects in Banking customer churn.

Isabelle Tandan & Erika Goteman from Uppsala University -Department Of Statistics did a similar project regarding bank customer churn in June 2020. They used a supervised statistical learning method; random forest, logistic regression, or K-nearest neighbor to predict customer churn. They also worked on finding a better cross-validation approach among k-Fold cross-validation and leave-one-out cross-validation. [5]
They worked on a research question:

- How successful are the statistical learning methods, logistic regression, KNN and random forest, in predicting customer churn and how do the methods compare to each other?
- Furthermore, regarding which method performs the best, which one of the cross-validation approaches, Leave-one-out and k-Fold, yields the most reliable predictions?

Similarly, Wong Hui Yeok, Ng Sin Yu, Lee Xin Yang, Lim Hong Chuan and Chin Chee Teng, have submitted a project on Rpubs by RStudio, regarding Bank customer churn prediction on June 2020. [6]
The objective of their project was to:
 To predict if a bank's customers will churn or stay with the bank.
 To predict how many years a customer would stay with the bank.
They used linear regression, decision tree, and support vector machine in this project.

Gaps in literature reviews

The writers of the journal and article talk about their research on the specific subject matter. They do not share border ideas on dealing with different nature of data or industries for prediction and analysis. More importantly, they do not discuss how their data set is trained for further prediction. It is difficult to figure out how exactly the method is implemented in the model as we can just see the report of the project but not the procedure of the projects.

Our project is based on a similar topic but our research question is quite different than these projects. We will be looking at the factors that affects customer churn.

4. Dataset main information

The dataset used in the project is the "Churn for Bank Customers" from the Kaggle repository.

Data Set Characteristics:	Multivariate	**Number of Instances:**	10000
Attribute Characteristics:	Categorical, Numerical, Boolean, String	**Number of Attributes:**	14
Associated Tasks:	Classification	**Missing Values?**	No

This database contains 14 attributes. The data set is publicly available on the Kaggle website. The classification goal is to find machine learning algorithms that are able to predict if a customer will leave the bank or not based on the customer's characteristics. Each attribute is a potential factor.

4.1. Dataset features description

The dataset contains the following features:

- RowNumber—corresponds to the record (row) number and has no effect on the output.
- CustomerId—contains random values and has no effect on customers leaving the bank.
- Surname—the surname of a customer has no impact on their decision to leave the bank.
- CreditScore—can have an effect on customer churn, since a customer with a higher credit score is less likely to leave the bank.
- Geography—a customer's location can affect their decision to leave the bank.
- Gender—it's interesting to explore whether gender plays a role in a customer leaving the bank.
- Age—this is certainly relevant, since older customers are less likely to leave their bank than younger ones.

- Tenure—refers to the number of years that the customer has been a client of the bank. Normally, older clients are more loyal and less likely to leave a bank.
- Balance—also a very good indicator of customer churn, as people with a higher balance in their accounts are less likely to leave the bank compared to those with lower balances.
- NumOfProducts—refers to the number of products that a customer has purchased through the bank.
- HasCrCard—denotes whether or not a customer has a credit card. This column is also relevant, since people with a credit card are less likely to leave the bank.
- IsActiveMember—active customers are less likely to leave the bank.
- EstimatedSalary—as with balance, people with lower salaries are more likely to leave the bank compared to those with higher salaries.
- Exited—whether or not the customer left the bank. [7]

5. Analysis and modeling methods

In this project, we are going to design the best classification model to detect if a customer will leave the bank or not by using the dataset, which can be found at https://www.kaggle.com/datasets/mathchi/churn-for-bank-customers. All the data are contained in multiple comma-separated values (csv) files.

In this dataset we have the target variable labels, we may use supervised learning techniques on it. Unsupervised algorithms, such as clustering, can be used to discover customers with similar traits and build specialized marketing tactics for each cluster. As a result, various machine learning methods, including supervised and unsupervised models, will be employed in this project.

To this aim, algorithms such as KNN, Decision Tree, and Logistic Regression as supervised models, and K-Mean clustering as an unsupervised model will be applied to the dataset, and their performance will be compared.

Different evaluation metrics are computed to assess the results with minimal error. Moreover, a cross-validation method will be used to improve the models. Also, since our dataset is imbalanced and the number of observations in the Not Exited class is 3.9 times of Exited class, it is required to apply oversampling method in order to balance our dataset.

The following table shows a summary of different algorithms as well as performance metrics that we are going to use in this project.

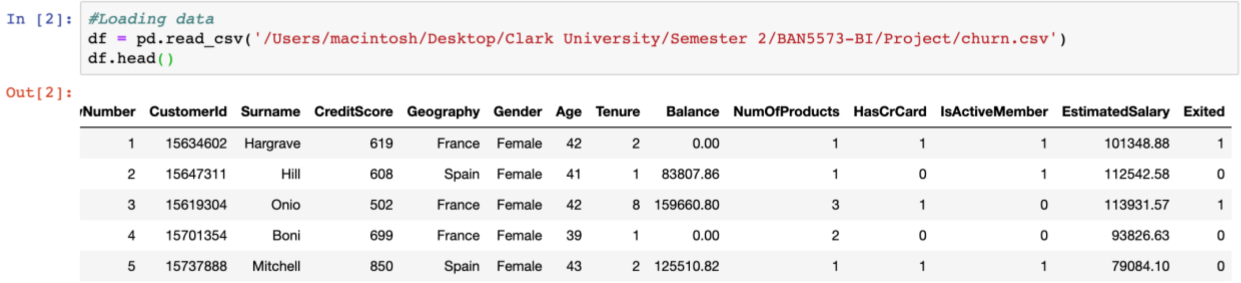
Type of models	Algorithms	Performance metrics
Supervised Learning	KNN	Accuracy, Precision, Recall, F1-Score, ROC Curve
	Decision Tree	
	Logistic Regression	
	Random Forest	
	Voting Classifier	

6. Software

In this project, we will use Power BI, and Tableau for visualization, and Python 3 on Jupyter Notebook to explore the dataset and build classification models. To this aim, the following libraries in Python will be implemented for data preprocessing, visualization, and fitting machine-learning models on the dataset.
List of Libraries: Numpy, Pandas, Scikit-learn (sklearn), Matplotlib, Seaborn, and researchpy.

7. Summary and Descriptive Statistics as Support for the Plan

In this section, we explore the dataset and implement descriptive analysis to discover the properties of the data and gain a better understanding of it.



```
In [2]: #Loading data
        df = pd.read_csv('/Users/macintosh/Desktop/Clark University/Semester 2/BAN5573-BI/Project/churn.csv')
        df.head()

Out[2]:
```

Number	CustomerId	Surname	CreditScore	Geography	Gender	Age	Tenure	Balance	NumOfProducts	HasCrCard	IsActiveMember	EstimatedSalary	Exited
1	15634602	Hargrave	619	France	Female	42	2	0.00	1	1	1	101348.88	1
2	15647311	Hill	608	Spain	Female	41	1	83807.86	1	0	1	112542.58	0
3	15619304	Onio	502	France	Female	42	8	159660.80	3	1	0	113931.57	1
4	15701354	Boni	699	France	Female	39	1	0.00	2	0	0	93826.63	0
5	15737888	Mitchell	850	Spain	Female	43	2	125510.82	1	1	1	79084.10	0

7.1. Number of Observations

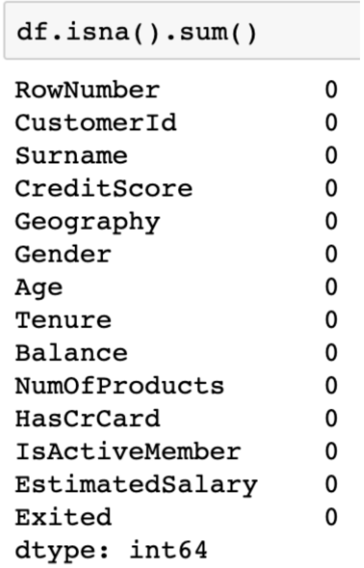
This data set includes data for 10000 rows. There are 7963 observations for the Not-Exited class and 2037 observations for the Exited class. The dataset is imbalanced, so oversampling or undersampling methods should be implemented to balance the training data.

7.2. Type of features:

Variable	Data Type
RowNumber	String
CustomerId	String
Surname	String
CreditScore	Numerical
Geography	Categorical
Gender	Categorical
Age	Numerical
Tenure	Numerical
Balance	Numerical
NumOfProducts	Numerical
HasCrCard	Boolean
IsActiveMember	Boolean
EstimatedSalary	Numerical
Exited	Boolean

7.3. Missing Values:

Now, let’s look at null/missing values in the dataset:



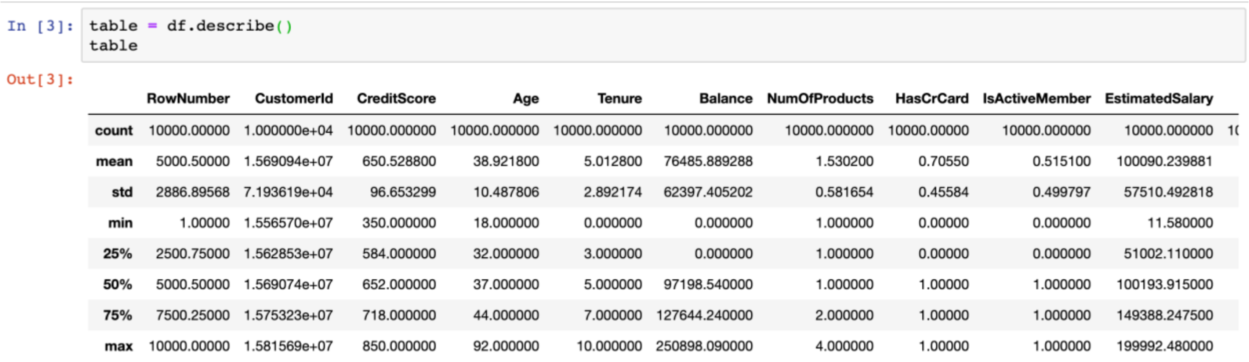
```
df.isna().sum()
```

```
RowNumber          0
CustomerId         0
Surname            0
CreditScore        0
Geography          0
Gender             0
Age                0
Tenure             0
Balance            0
NumOfProducts      0
HasCrCard          0
IsActiveMember     0
EstimatedSalary    0
Exited             0
dtype: int64
```

The dataset has no missing values.

7.4. Features statistical view

Every feature has specific statistical attributes that are necessary for further analysis and making a clear prediction model.



```
In [3]: table = df.describe()
        table
```

Out[3]:

	RowNumber	CustomerId	CreditScore	Age	Tenure	Balance	NumOfProducts	HasCrCard	IsActiveMember	EstimatedSalary	
count	10000.00000	1.000000e+04	10000.000000	10000.000000	10000.000000	10000.000000	10000.000000	10000.00000	10000.000000	10000.000000	1(
mean	5000.50000	1.569094e+07	650.528800	38.921800	5.012800	76485.889288	1.530200	0.70550	0.515100	100090.239881	
std	2886.89568	7.193619e+04	96.653299	10.487806	2.892174	62397.405202	0.581654	0.45584	0.499797	57510.492818	
min	1.00000	1.556570e+07	350.000000	18.000000	0.000000	0.000000	1.000000	0.00000	0.000000	11.580000	
25%	2500.75000	1.562853e+07	584.000000	32.000000	3.000000	0.000000	1.000000	0.00000	0.000000	51002.110000	
50%	5000.50000	1.569074e+07	652.000000	37.000000	5.000000	97198.540000	1.000000	1.00000	1.000000	100193.915000	
75%	7500.25000	1.575323e+07	718.000000	44.000000	7.000000	127644.240000	2.000000	1.00000	1.000000	149388.247500	
max	10000.00000	1.581569e+07	850.000000	92.000000	10.000000	250898.090000	4.000000	1.00000	1.000000	199992.480000	

Histograms

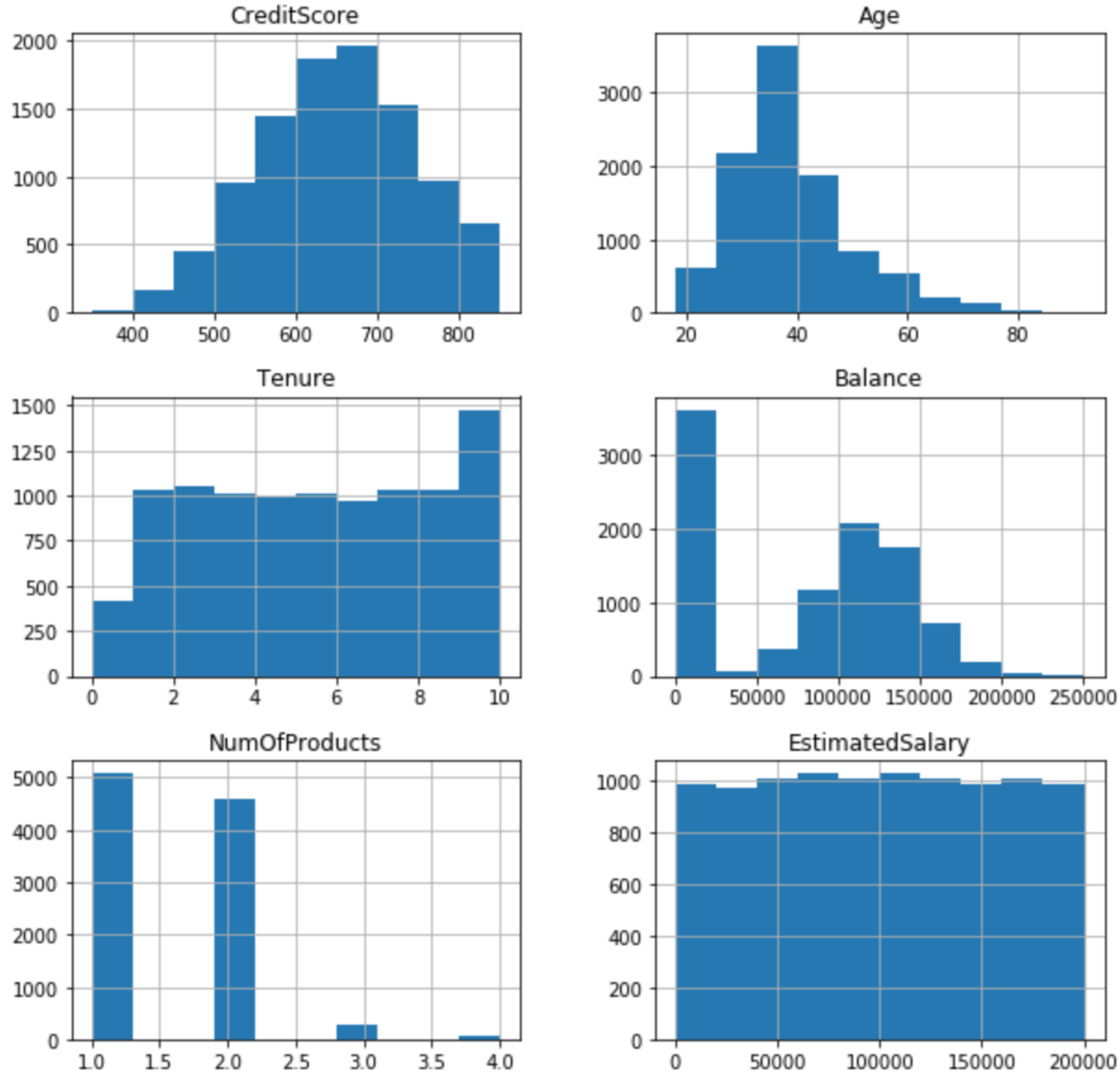


Fig: *Multiple histograms for different variables*

We can analyze the distribution of the numerical variable. For example, CreditScore and Age seem to have a bell-shape distribution while CreditScore is skewed to the left, but Age is skewed to the right. The variable "Balance" distribution attracts attention. It seems to have two separate distribution, one for people who have balance less than 25000, and another normal distribution for the rest, which are customers with Balance more than 25000. It's an important point that we should consider in our analysis.

Variables correlation

Using matplotlib and seaborn libraries, we calculate the correlation of the variables.

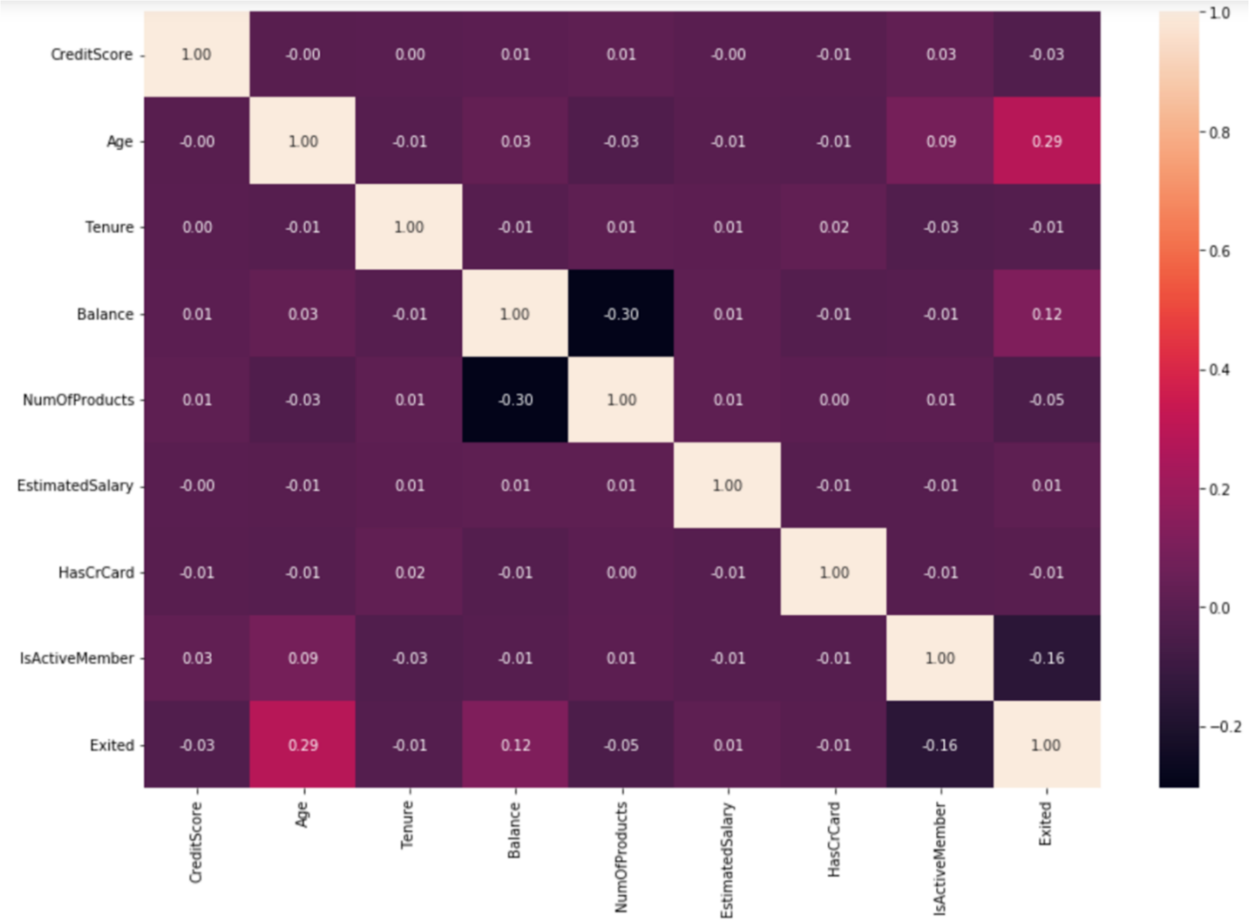
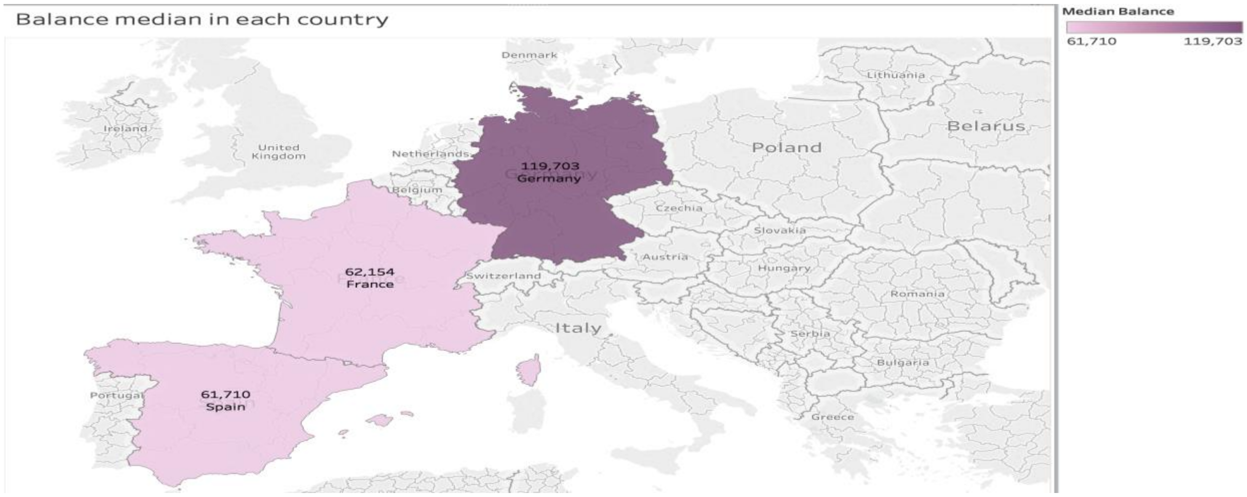


Fig: Correlation heat map

Based on the correlation heat map result, there is not a strong correlation between the target variable (Existed) and other variables. This means that more complicated machine learning methods are required to predict the target variables using those features.

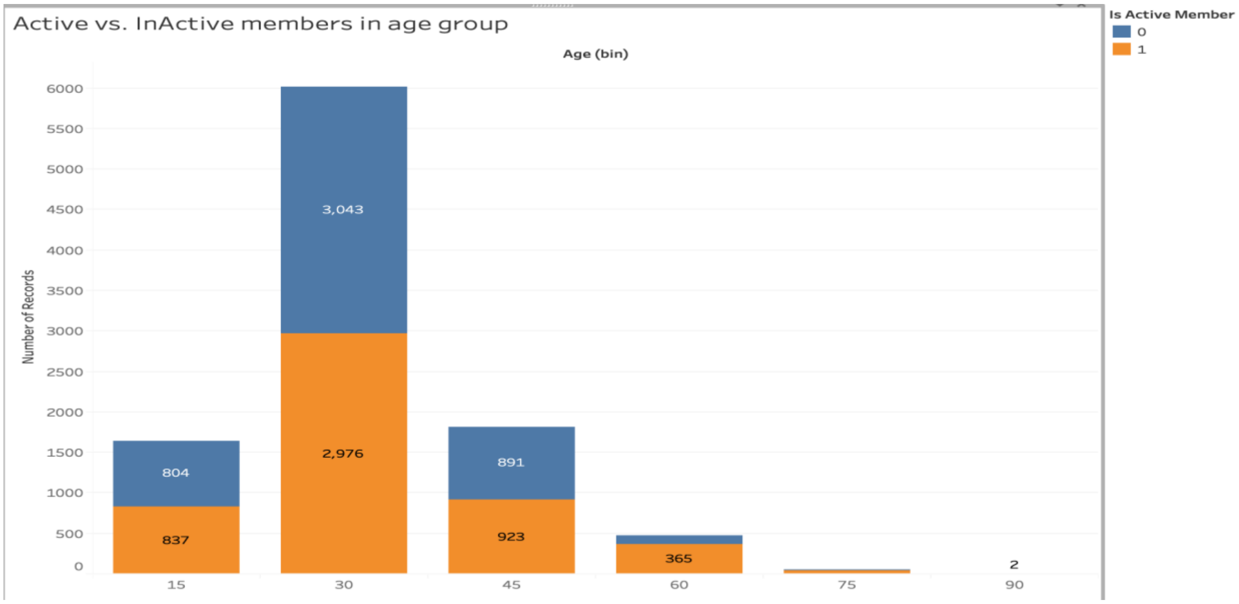
7.5. Data Visualization

In this section, we have a graphical representation of information hidden in the dataset.
One of the variables that seems to have a strong effect on customer churn is balance. Hence, it would be useful to investigate this variable. Using Tableau features, we compare median of customers' balance in each country. The reason we select median instead of mean is that median is not sensitive to outliers and variable balance includes outliers.

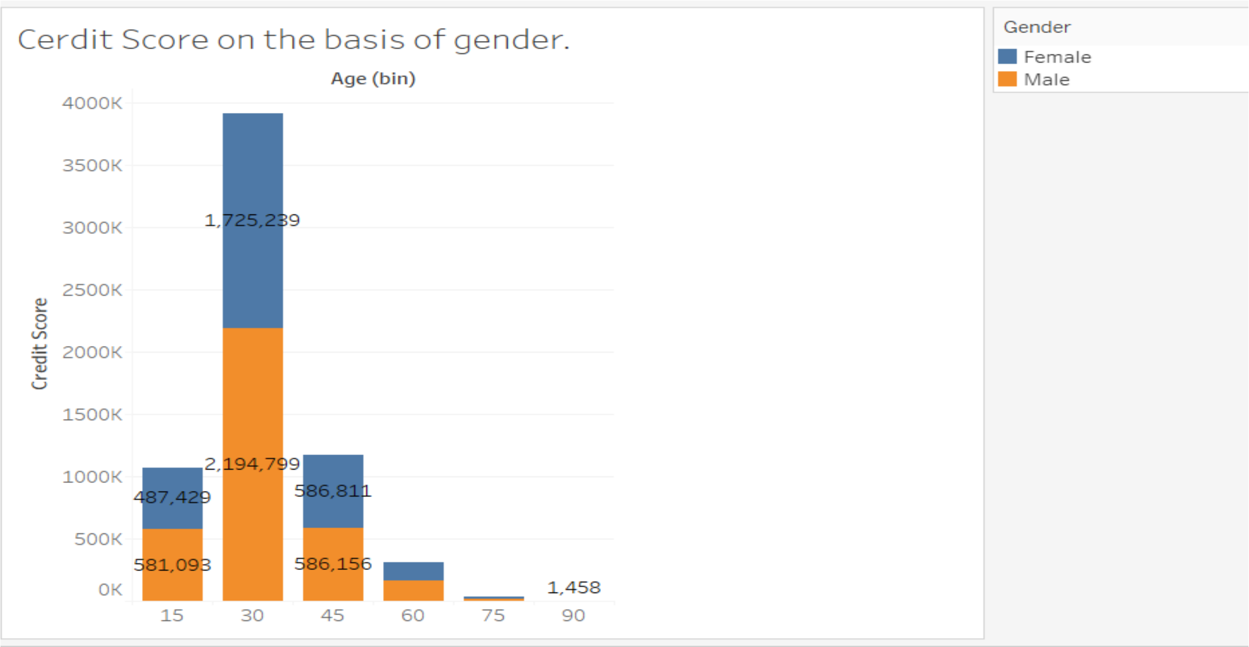


The map shows the gap between balance median of Germany and those two other countries, France and Spain. It can give us a clue that this feature will be effective in the model.

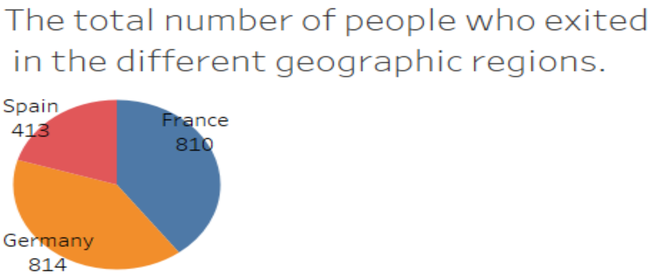
In the next step we are interested to investigate the number of active or inactive members in each age group. To this aim, we have grouped costumers based on their age in 6 group. 0-20, 20-40, 40-60, 60-80, and 80-100 are the age ranges.



As expected, there are few customers older than 70 years, while majority of the customers are between 20 to 40 years old. This is a insightful fact for managers to design specific marketing strategies for this group. In all age groups almost half of the customers are inactive. If our final model shows that this feature, being active or inactive, significantly affects the target variable, then we should design strategies to active costumers.



In the figure we have the sum of credit score of male and female between different age group. We have grouped costumers based on their age in 6 group. 0-20, 20-40, 40-60, 60-80, and 80-100 are the age ranges. The sum of the credit score of male is higher than female in age group 0-20 and age group 20-40. However the age group of 40-60 share similar sum of credit score. As the age increases female have a better sum of credit scores from age 60-80. And over 80 years of age the credit score of male is greater than that of female.



Here, we can see the total number of user that exited the bank. Each country is represented with a different color in a Pie chart. 413 people in Spain, 810 in France and 814 people in Germany have exited the bank. The total user those who exited from the bank were 2037.

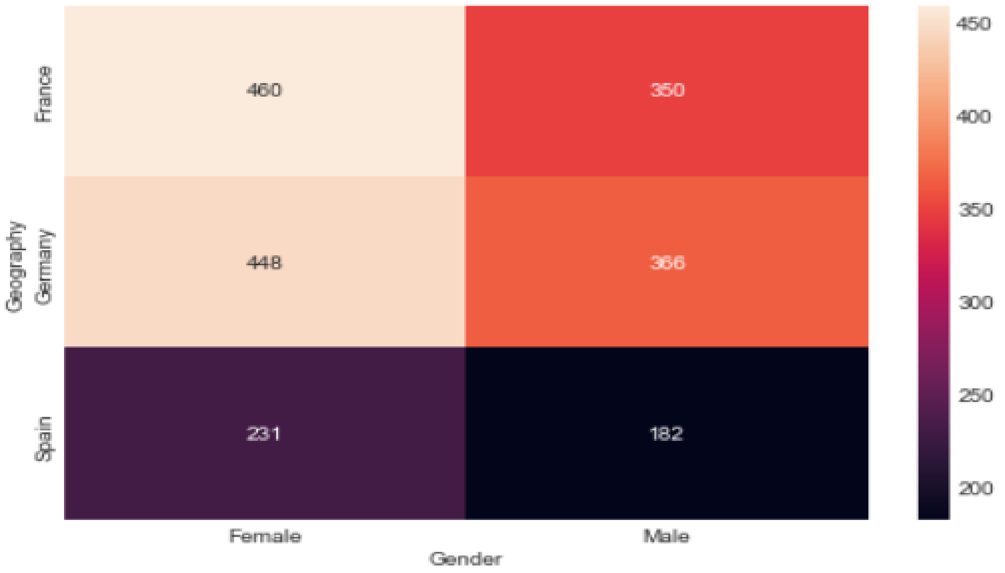
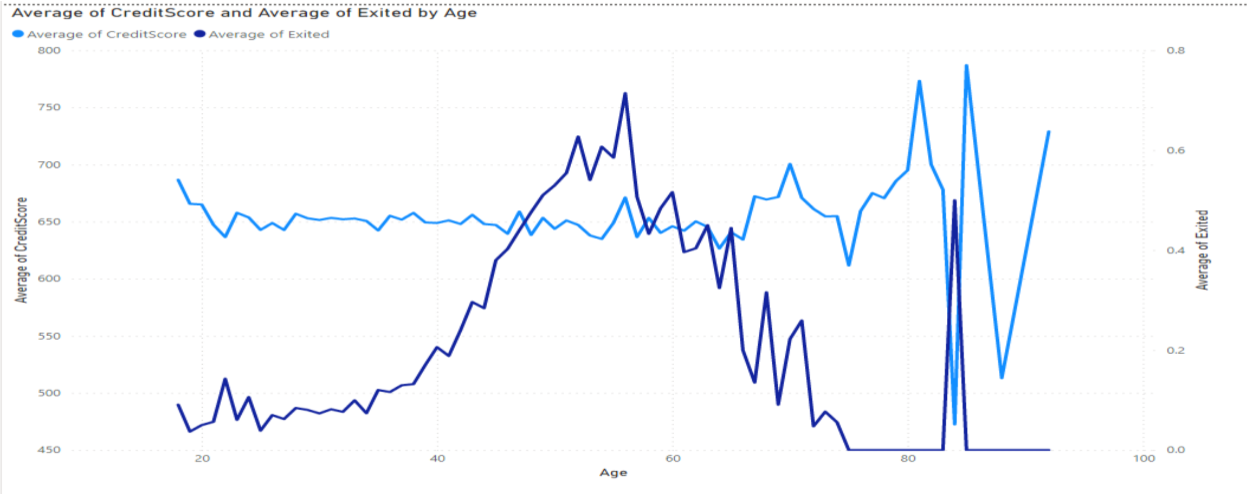


Figure: Heat map for exited number of male and female from different geographical region.

In the Heat map, we can see the number of male and female that exited the bank. In France, 460 female and 350 male exited where as in Germany 448 female and 366 male exited the bank. In Spain 231 male and 182 female exited the bank.



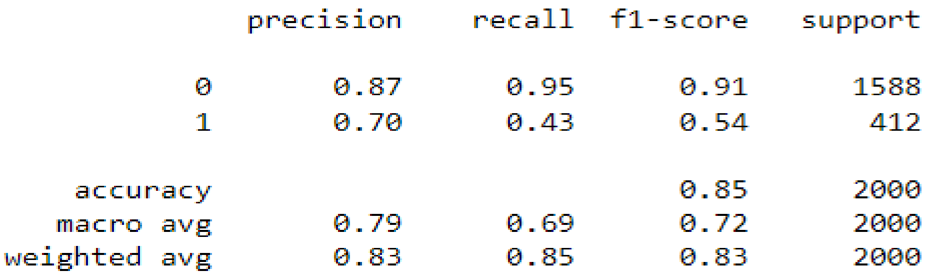
The above figure illustrate the line graphs of average credit score and average of exited by age. The average credit score is around 600-700 for age group of 20-60. After the age of 60 there is a fluctuation in credit scores. The user existing the bank is generally in the age group of 45-60.

8. Methods implemented:

After we cleaned our data we tested our data using various methods. The models used and the finding of the models are as follows.

8.1. Decision Tree

We first splitted the data set into training and testing dataset. We have a test size of 0.2, which means 20 % of data is taken a test data. We than performed the decision tree and got good accuracy.



```
              precision    recall  f1-score   support

           0       0.87      0.95      0.91      1588
           1       0.70      0.43      0.54       412

    accuracy                           0.85      2000
   macro avg       0.79      0.69      0.72      2000
weighted avg       0.83      0.85      0.83      2000
```

Figure: Classification report of decision tree.

We can see that the accuracy of our decision tree is 85%, which is a good accuracy and we and accept our decision tree. We also have the feature importance of decision tree.

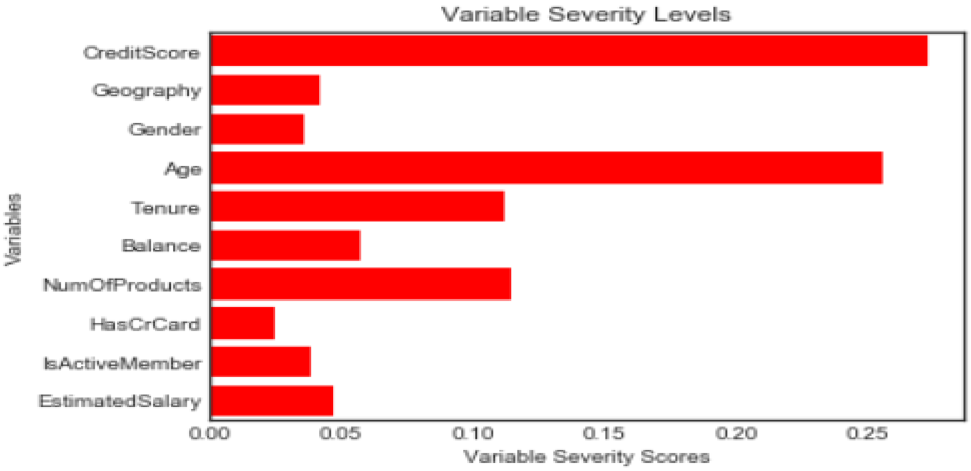
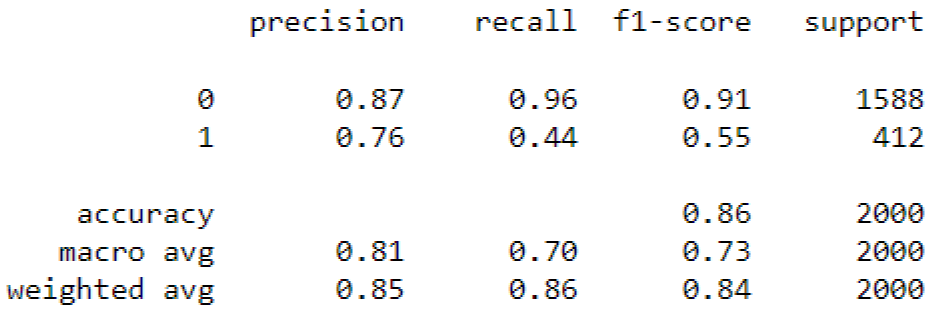


Figure: Feature importance of decision tree.

Here, in the feature importance we can see that the credit score is the most important and age is another important features among all other features.

8.2. Random forest

We performed Random forest after decision tree and got a little better result in accuracy than our decision tree.



```
              precision    recall  f1-score   support

           0       0.87      0.96      0.91      1588
           1       0.76      0.44      0.55       412

    accuracy                           0.86      2000
   macro avg       0.81      0.70      0.73      2000
weighted avg       0.85      0.86      0.84      2000
```

Figure: Classification report of random forest.

The accuracy of random forest is 86% which is a bit better than the accuracy that we got from decision tree. We have a different feature importance in random forest than the decision tree. The feature importance of random forest is shown below:

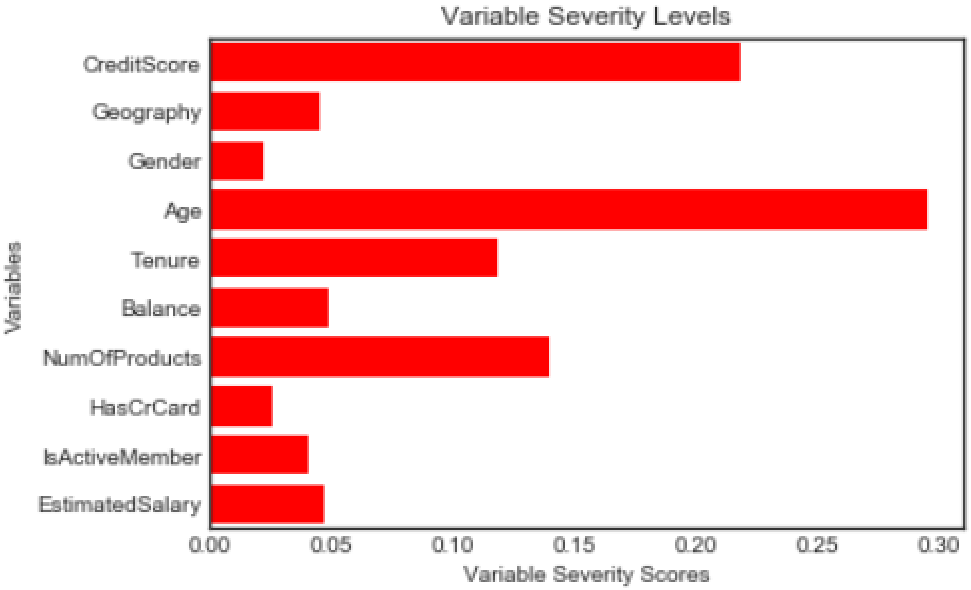
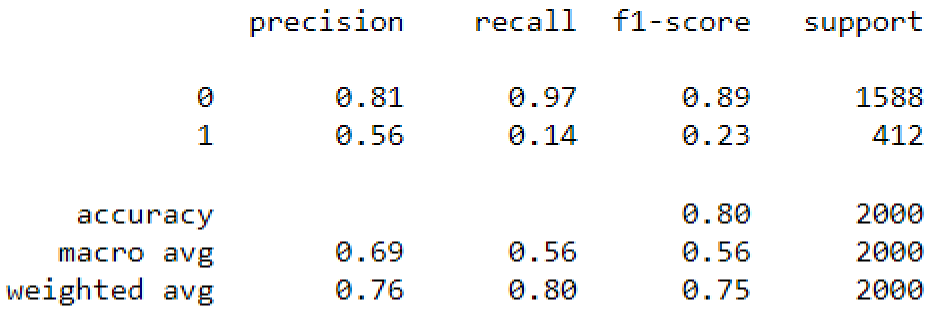


Figure: Feature importance of random forest.

Unlike decision tree the important feature is age for random forest followed by credit score. Number of product is third important feature followed by tenure.

8.3. Logistic Regression

We have also performed logistic regression but the accuracy was quite lower than the decision tree and random forest. The classification report of logistic regression is shown below:



```
              precision    recall  f1-score   support

           0       0.81      0.97      0.89      1588
           1       0.56      0.14      0.23       412

    accuracy                           0.80      2000
   macro avg       0.69      0.56      0.56      2000
weighted avg       0.76      0.80      0.75      2000
```

Figure: Classification report of logistic regression.

From the classification report we can see that the accuracy of logistic regression is 80 % which is lower than our previous two models. We do not have a feature importance for the logistic regression but we do have a correlation regression of the variable which is shown below in the figure.

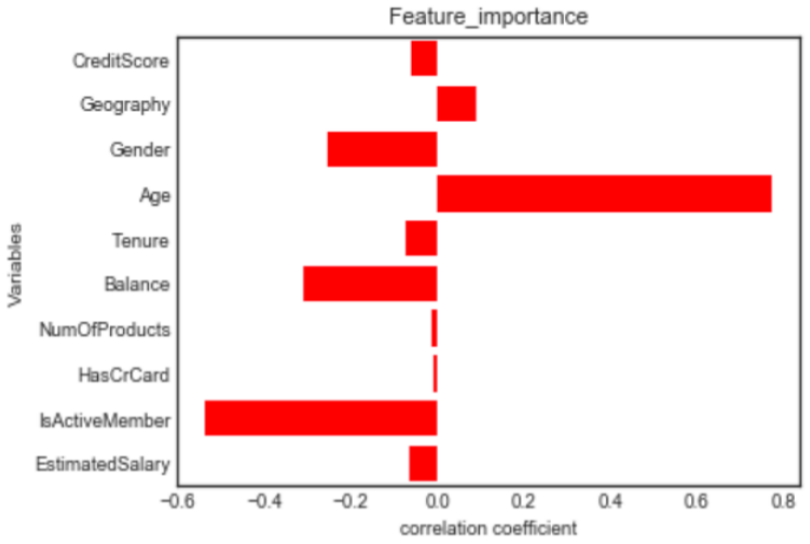
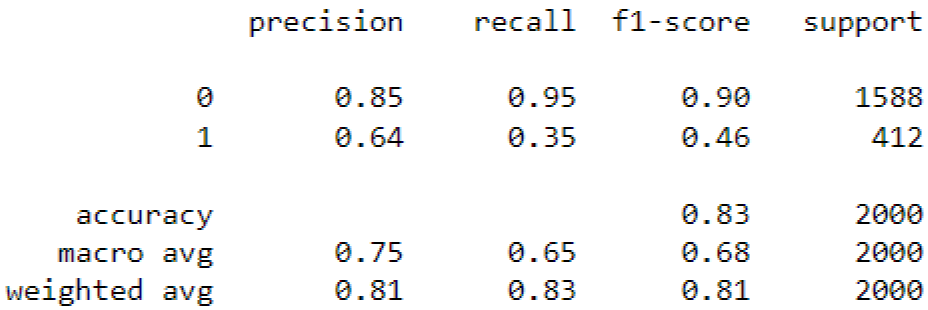


Figure: Correlation of features

In the figure, we can see that the age is highly correlated followed by is active member. Balance has third highest correlation followed by balance.

8.4. KNN (K-Nearest Neighbor)

We also performed the KNN algorithm and figures out that we have a good accuracy in this model too. The classification matrix of KNN is shown below:



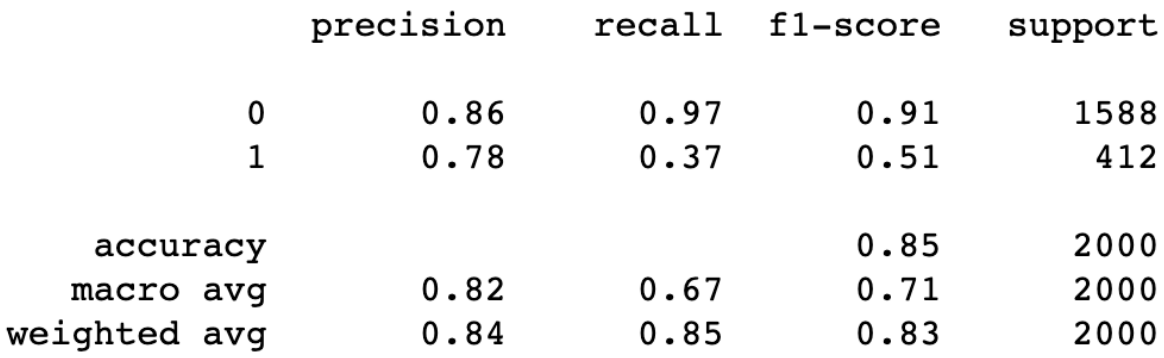
```
              precision    recall  f1-score   support

           0       0.85      0.95      0.90      1588
           1       0.64      0.35      0.46       412

    accuracy                           0.83      2000
   macro avg       0.75      0.65      0.68      2000
weighted avg       0.81      0.83      0.81      2000
```

Figure: Classification report of KNN.

From the classification report we can see that the accuracy of KNN is 80%. This is a decent accuracy but still less than the random forest and decision tree.

8.5. Voting Classifier

After performing all these model we also performed the voting classifier to get accuracy to our result. After we performed this classification we got a impressive result in accuracy. The classification report of voting classifier is shown below.



```
              precision    recall  f1-score   support

           0       0.86      0.97      0.91      1588
           1       0.78      0.37      0.51       412

    accuracy                           0.85      2000
   macro avg       0.82      0.67      0.71      2000
weighted avg       0.84      0.85      0.83      2000
```

Figure: Classification report of voting classifier.

From the report we can see the accuracy of voting classifier is 85%. Which is a very good accuracy according to our report so far.

8.6. AUC Curve

We used different models to test our data and calculated the accuracy of each model. The accuracy of all the models were close. Here we have generated the AUC-curve for all the models.

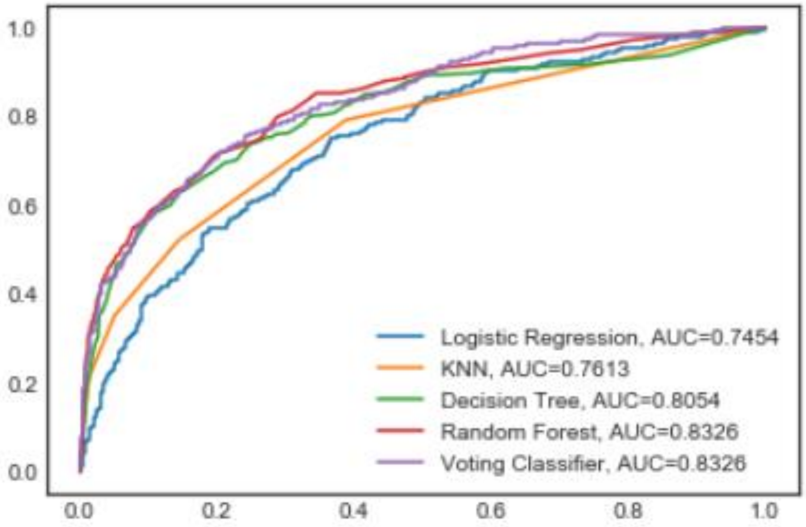


Figure: AUC-curve for all the models

As we can see in the figure, Random Forest, Decision Tree and Voting Classifier performed much better than KNN and Logistic Regression. The performance is for Random Forest and Voting Model with AUC = 0.8326. Therefore, instead of using all models, we can use Random Forest to predict customer churn with high accuracy.

9. Conclusion:

- Based on the implemented models, customers' age, credit score, and the number of products they use are the most important features for customer churn prediction.
- We implemented four classification models and one voting classifier to predict which customer is going to leave the bank. Among them, Random Forest and voting models have the best performance based on accuracy, precision, recall, f-1 score, and area under the curve measurements.
- Based on visualization, we conclude that women are more likely to leave the bank in comparison to Men. Moreover, almost half of the customers from Germany leave the bank. Therefore, we should investigate these two groups and find the reason so we can design proper strategies to prevent them from exiting.
- Using machine learning algorithms, we were able to predict 70% of exited customers. By offering promotions to them, it is possible to encourage them to stay.

10. Comments / Concerns:

- There are 7963 observations for the Not Exited class and 2037 observations for the Exited class. The dataset is imbalanced, so under sampling method is implemented to balance the training data.

- Based on the correlation heat map result, there is not a strong correlation between the target variable (Existed) and other variables. This means that more complicated machine learning methods are required to predict the target variables using those features. This fact justifies the weak performance of Logistic Regression model, since this model, looks for linear relationships among dependent and independent variables.

11. References

[1]. Dr. U. Devi Prasad, S. Madhavi(2012) ‘Prediction of Churn Behavior of Bank Customers Using Data Mining Tools’, Business Intelligence Journal, Volume (5), 96

[2]. "Domain Knowledge Integration in Data Mining Using Decision Tables: Case Studies in Churn Prediction." The Journal of the Operational Research Society 60, no. 8 (2009): 1096–1106. http://www.jstor.org/stable/40206835.

[3]. Liu, Qiong & Wu, Ying. (2012). "Supervised Learning," Researchgate, Qiong Liu, January 15, https://www.researchgate.net/publication/229031588_Supervised_Learning.

[4]. Dridi, Salim. (2021). "Unsupervised Learning-A Systematic Literature Review," Researchgate, Salim Dridi.
https://www.researchgate.net/publication/357380639_Unsupervised_Learning_-_A_Systematic_Literature_Review

[5]. Bank Customer Churn Prediction, A comparison between classification Prediction and evaluation methods. Isabelle Tandan & Erika Goteman Supervisor: Patrik Andersson UPPSALA UNIVERSITY DEPARTMENT OF STATISTICS, June 4th , 2020

[6]. https://rpubs.com/leexinyang/bankcustomerchurnprediction

[7]. https://www.kaggle.com/datasets/mathchi/churn-for-bank-customers

—--