Banking Customers Analysis

By:Christos Jovanovic

Introduction

- This is a data analysis for Bank customers and their churn status.
- Will be using a classification dataset with target variable if the customer has left the bank or not.
- https://www.kaggle.com/datasets/saurabhbadole/bank-customer-churn-prediction-dataset - here is a link to the kaggle dataset

Analysing Data

Features:

RowNumber: The sequential number assigned to each row in the dataset.

CustomerId: A unique identifier for each customer.

Surname: The surname of the customer.

CreditScore: The credit score of the customer.

Geography: The geographical location of the customer (e.g., country or region).

Gender: The gender of the customer.

Age: The age of the customer.

Tenure: The number of years the customer has been with the bank.

Balance: The account balance of the customer.

NumOfProducts: The number of bank products the customer has.

HasCrCard: Indicates whether the customer has a credit card (binary: yes/no).

IsActiveMember: Indicates whether the customer is an active member (binary: yes/no).

EstimatedSalary: The estimated salary of the customer.

Exited: Indicates whether the customer has exited the bank (binary: yes/no).

Analysis

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

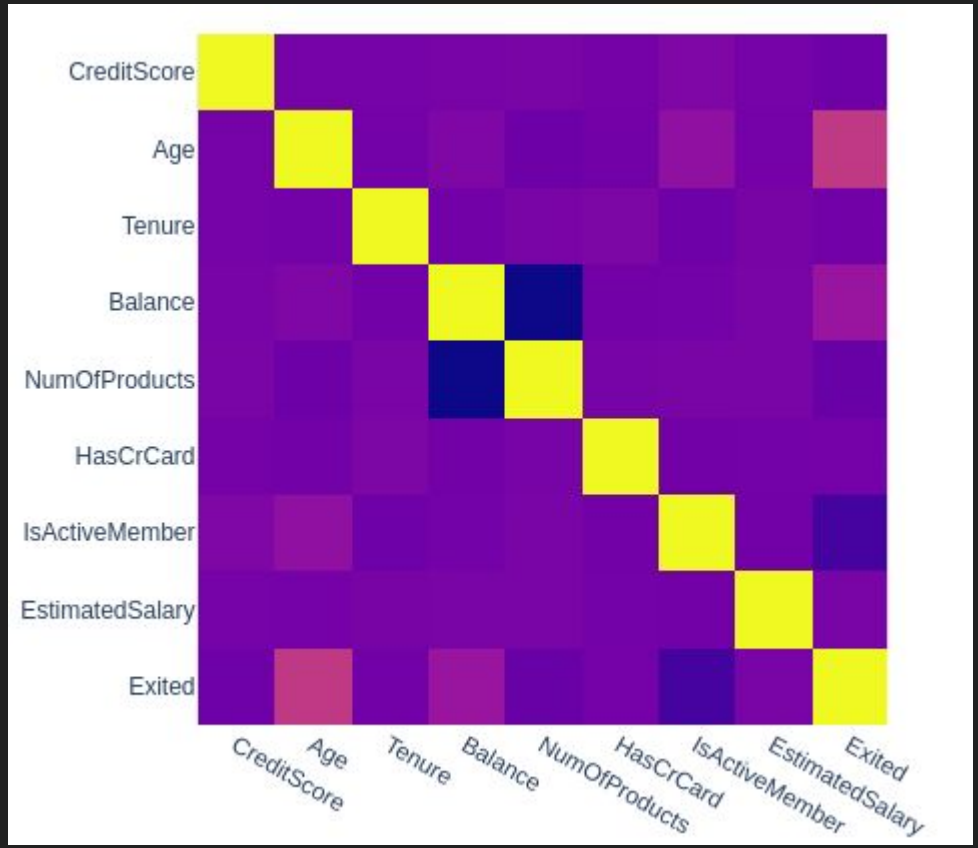
Here is an example of how the dataset looks like

There are about 10,000 rows

Graphing Data

For the graph part of the analysis, we are going to look for skewness in the data, outliers and how the overall data looks. Also the importance of each feature against the target variable.

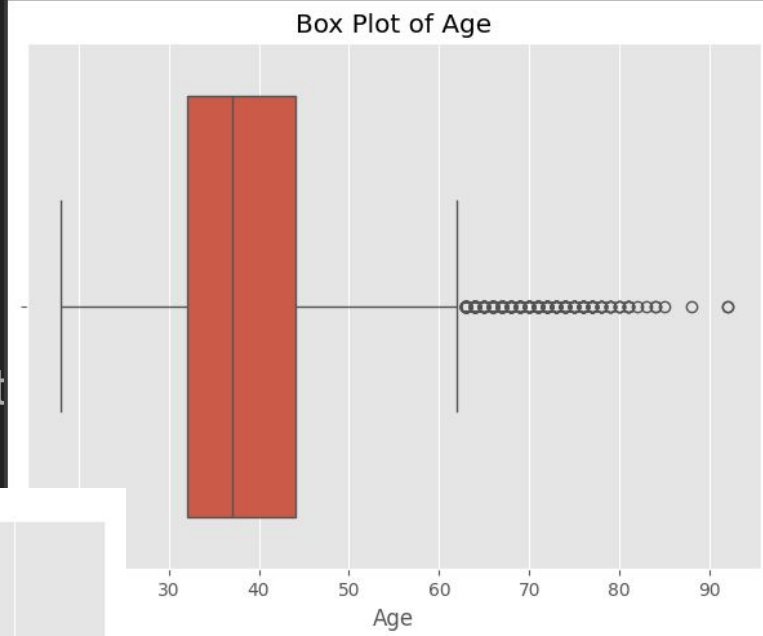
In this confusion matrix it shows us that age and balance have a bigger affect on the target variable.

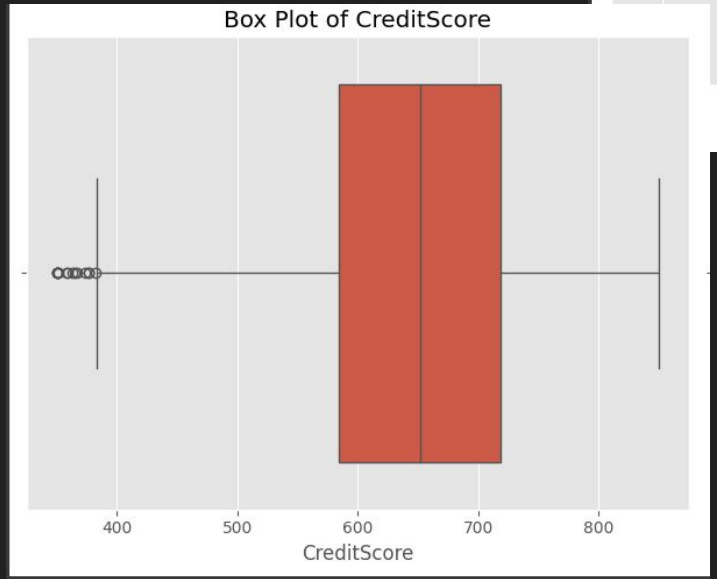


Graphs

Here we can see there are a few outliers on age

And credit score so we’re going to take care of that





Cleaning Data

- First dropped all the features that will not be need(RowNumber, CustomerID, Surname, Geography and Gender)
- Then checked for nulls in the data(which weren't any), the only ones was when balance was 0, which was not removed because 0 in this case matters
- Then the as seen above in the box plots, with the use of z-score we removed outliers in age and creditscore
- Lastly, we put age and creditScore in labeled bins to separate the data, and normalized Estimated salary and balance.

Metrics Used

The metrics that were used for the classifiers were:

- Accuracy
- Precision
- Recall
- F1
- Support

Tree Classifiers

Used two tree classifiers

Decision Tree

- Accuracy:0.83

Random Forest

- Accuracy:0.85

Non-Tree Classifiers

Non-Tree Classifiers that were used:

Logistic Regression:

Accuracy: 0.80

SVM:

Accuracy:DNF

SVM would not finish even after 2 minutes

Best one

Based on all the information Random Forest Performed the best.

Somethings I tried to so, is XGBoost which I think it would had performed the best, but I couldn’t get it to work so I just went with those options.

Application from class

- I used techniques we learned in class for data cleaning and preparation, for example z-score and normalization
- Decision trees that we learned in class, with the help of outside sources I was able to implement them in google colab
- As well as other classification algorithms like logistic regression

References

https://www.kaggle.com/code/durgancegaur/a-guide-to-any-classification-problem/notebook#Training-our-Machine-Learning-Model-:

https://www.datacamp.com/tutorial/decision-tree-classification-python

https://scikit-learn.org/stable/modules/model_evaluation.html