Departing Customer Prevention

By Pablo Nguema

Objectives

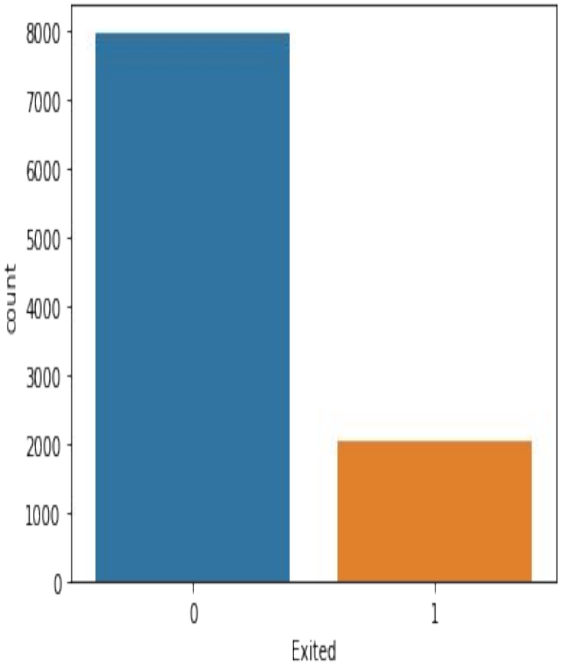
1) Do an Analysis of the Data

2) Build a Model that would show the possible Departing customers.(It was a Neural Network Model)

3) Have the results of the potential staying and leaving customers

Analysis I(Departed Customers)

RowNumber	10000
CustomerId	10000
Surname	2932
CreditScore	460
Geography	3
Gender	2
Age	70
Tenure	11
Balance	6382
NumOfProducts	4
HasCrCard	2
IsActiveMember	2
EstimatedSalary	9999
Exited	2

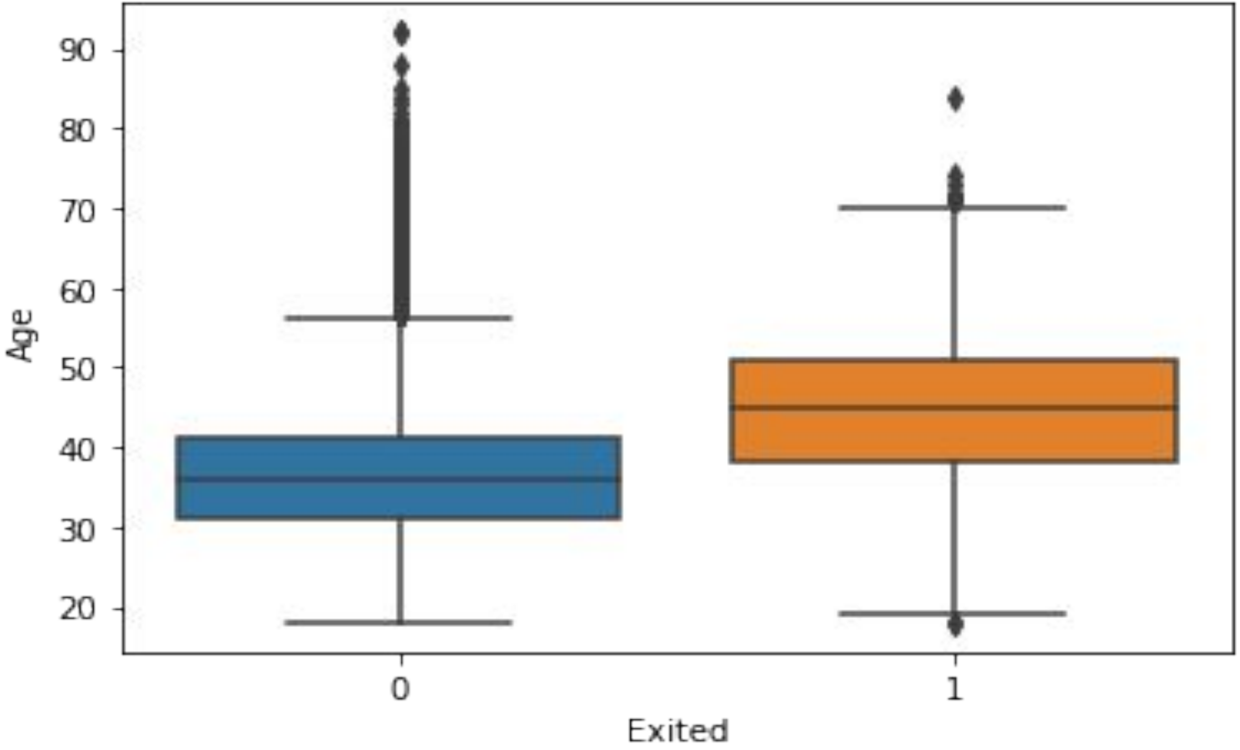


Customer_Id has as many unique values as there is data. Suggesting no repetition

20.37% of total customers Departed from the bank.

Lost of 1/5 of our customers

Analysis II(Departed In Age)

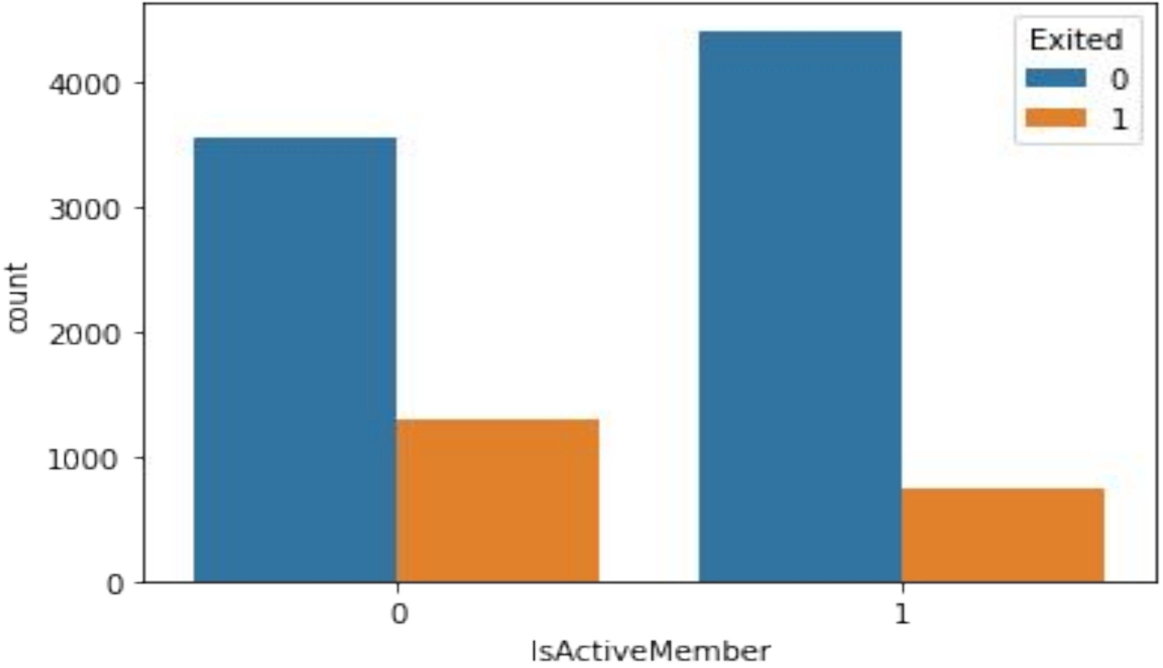


Comparison of People that left according to age.

I seems apparent that when the customers get around 40 years old they tend to leave.

For the older customers,it might be optimal to check living status.

Analysis III(Active Members)

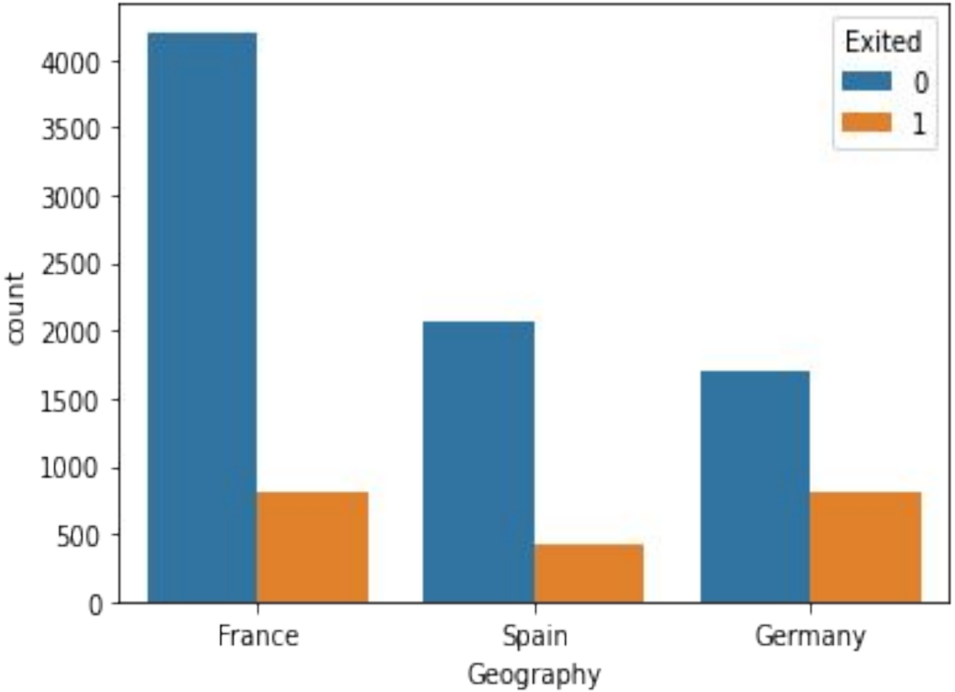


51.51% of members are actives,and 48.49%

Out of the deserters 63.9% are inactive members,while 36.1% are active.

26.85% of inactive members leave,while 14.27% of active members leave

Analysis IV(Geography)



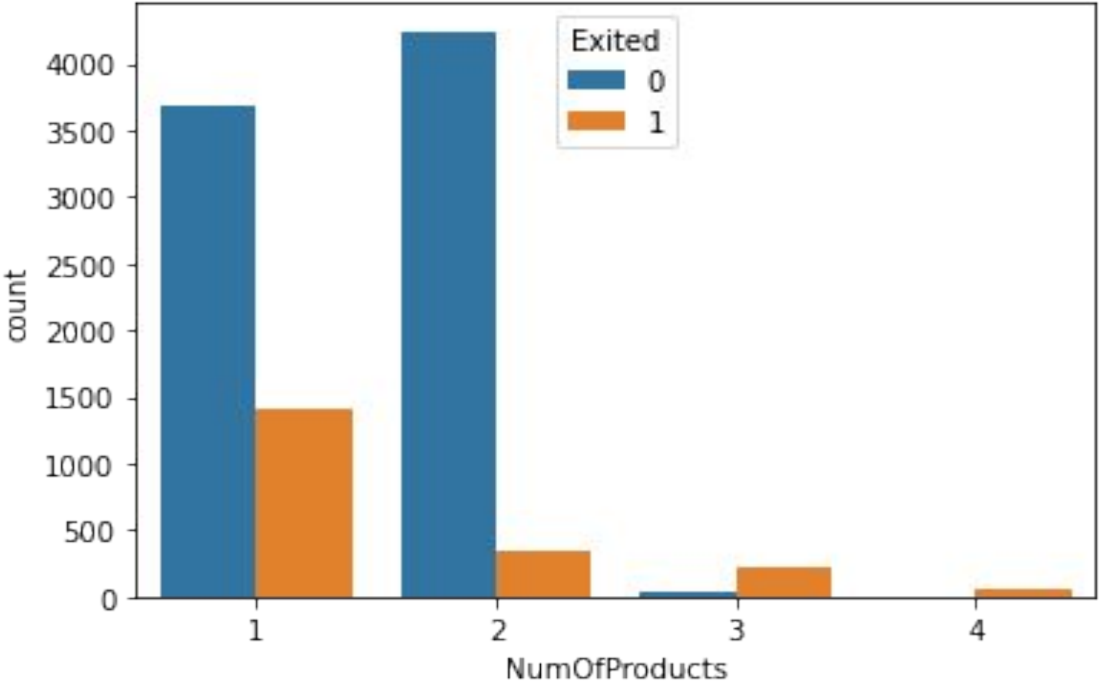
50.14% are in France,25.09% are in Germany ,and 24.77% are in Spain

Out of the customers that left 39.76% are in France,39.96% are in Germany and 20.27% are in Spain

32.44% of the customer in Germany leave,15,16% of France's customer leave,and 16.67% of Spain's customers leave

Germany might b e focal point for customers that leave. As it almost doubles the rest in percentages per region.(Germany have similar numbers as France when counting)

Analysis V(Number of Products)

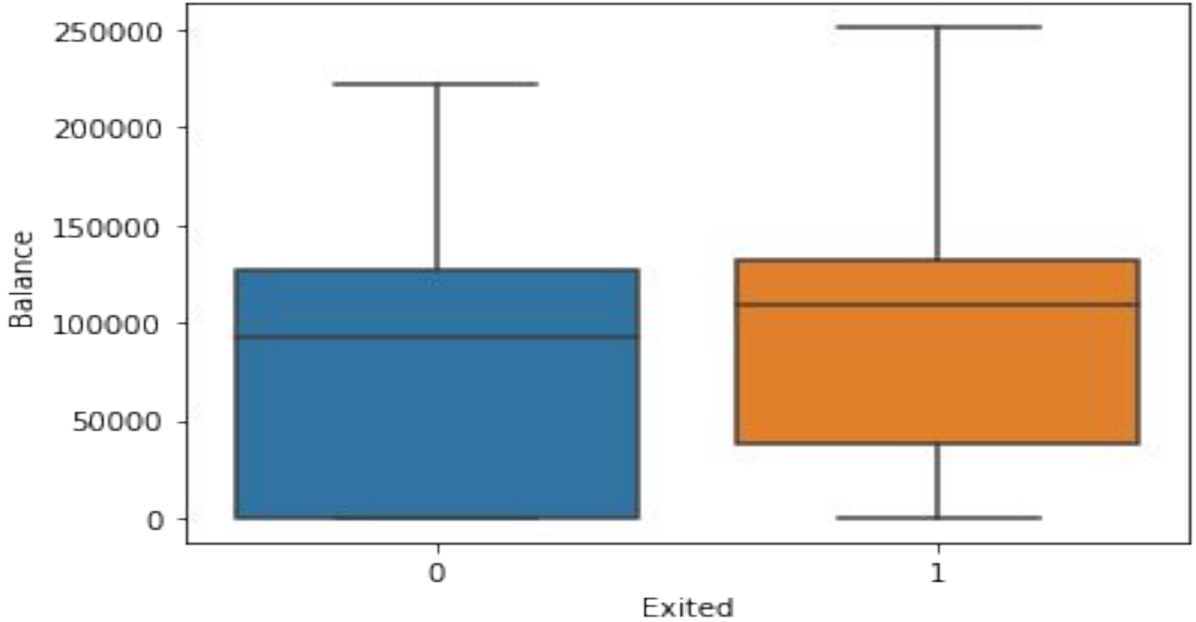


50.84% of customer took 1 product ,45.90% took 2 products, 2.66% took 3 products,and 0.6% took 4 products

% of leavers per products taken are: 1 product 69.17%,2 products 17.08%, 3 products 10.8%,and 4 products 4 2.95%.

% of leaver in amount of product taken: 1 product is 27.71%, 2 products is 7.58%,3 products 82.71%, and 4 products 100%

Analysis VI(Balance)



People start to leave in bulk after having 50K in their balance in their accounts

However the biggest increase is after 100K in their balance

Results

Real Values		Predictions: 0	Predictions: 1
	0	2243(75%)	130(4.3%)
	1	291(9.7%)	336(11%)

0 represent Staying, and
1 represents Departing.

75% of the customer stayed

11% of cumers that where considering leaving, have left

9.7% of customers that were considering leaving could be convinced to stay

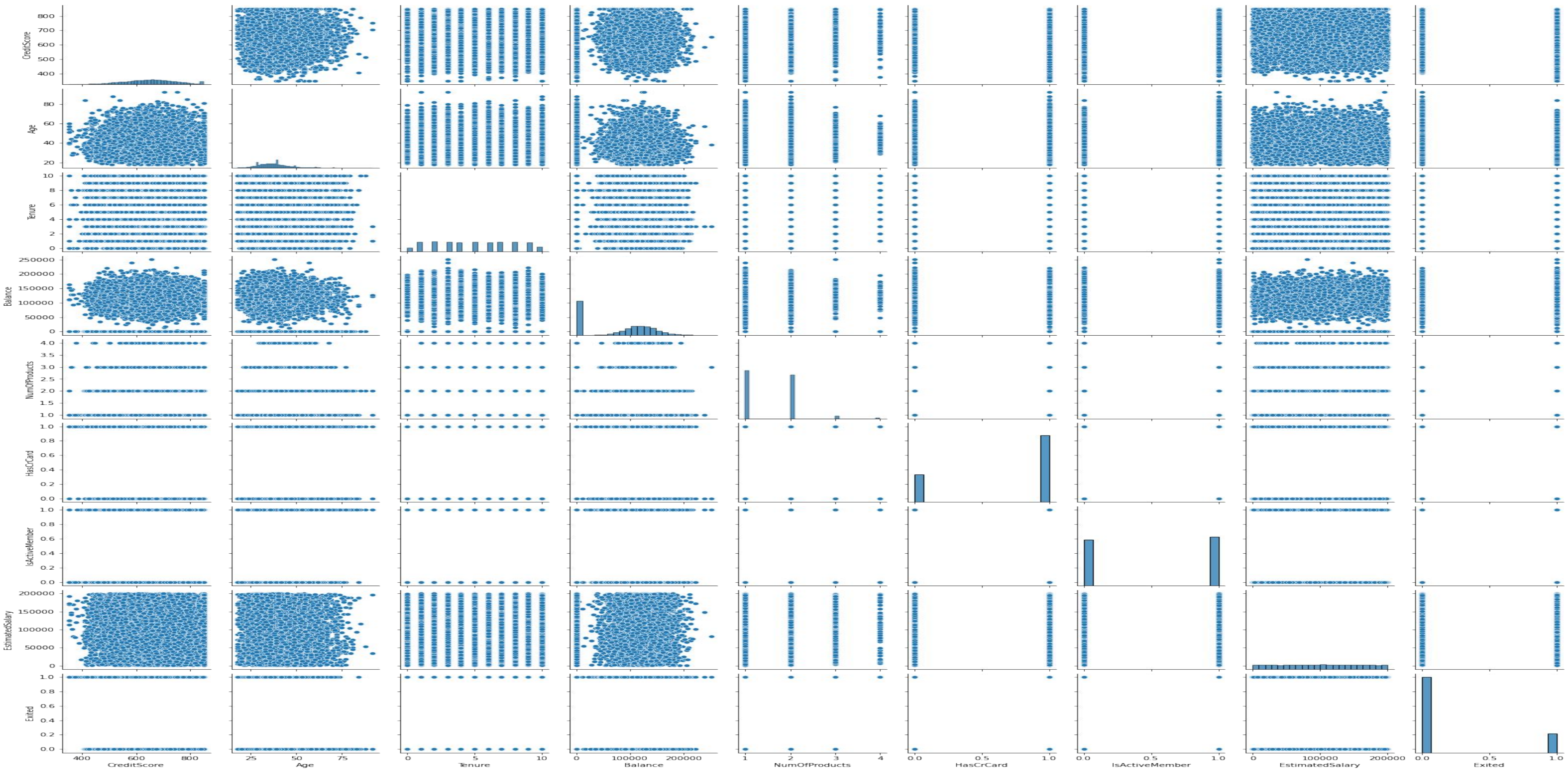
4.3% that are were considering staying, could leave

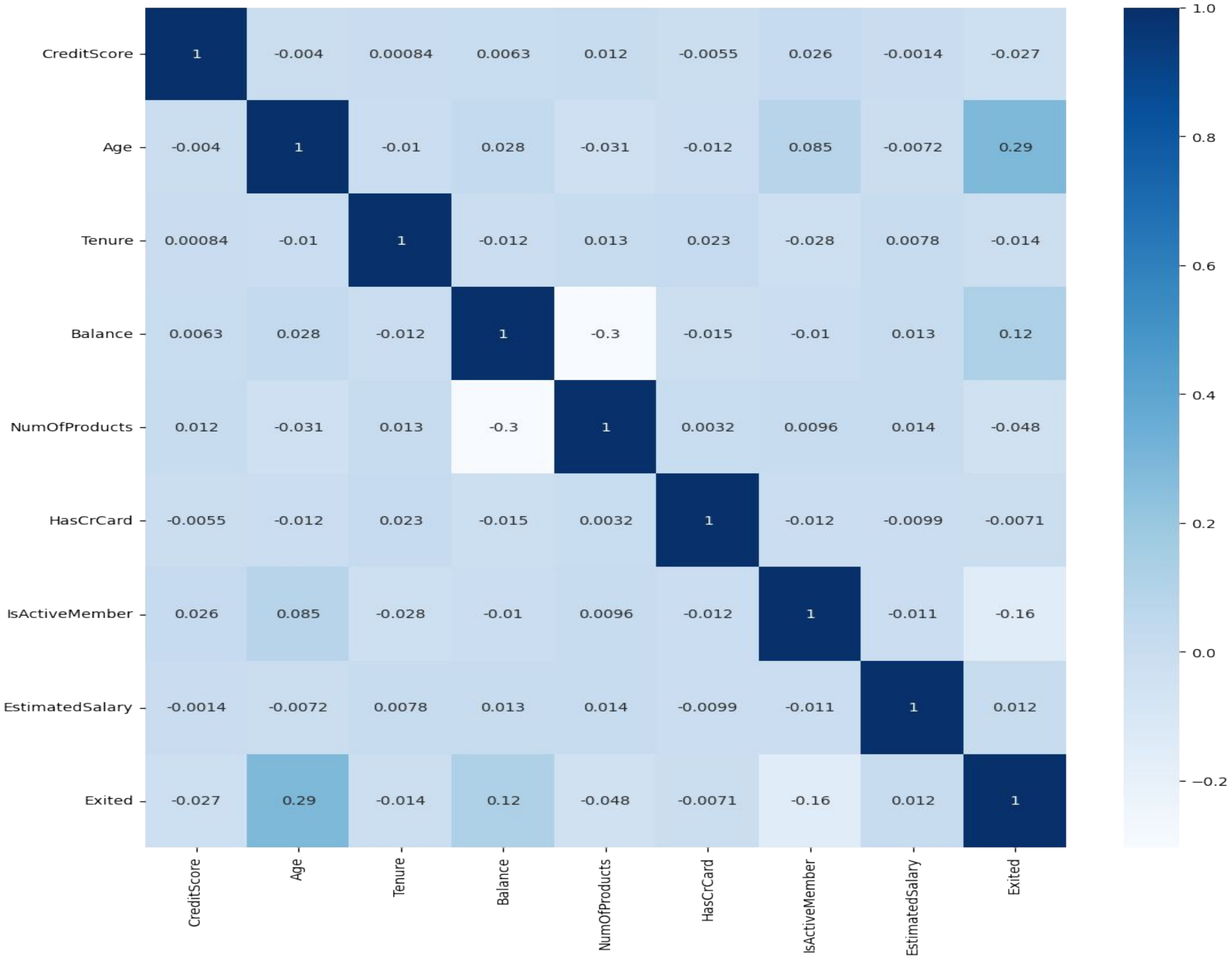
Conclusion

1) There is 4.3% of customers that we might think want to stay that might consider leaving. A possible solution is to check in, and see what could be done to keep them in enrolled with us.
2) There is an 9.7% of customer that might retract from their decision to leave, these customers will need most of the attention, as it might be ideal to see what we could offer to confirm their stay.
3) Around 86% of our customers have made their decision finale whether it is to stay(75%) or to leave(11%)
4) Best case Scenario we could keep 89% of our customer base, however the worst case we could lose 26% of them.


APPENDIX

Analysis I (Pairplots)

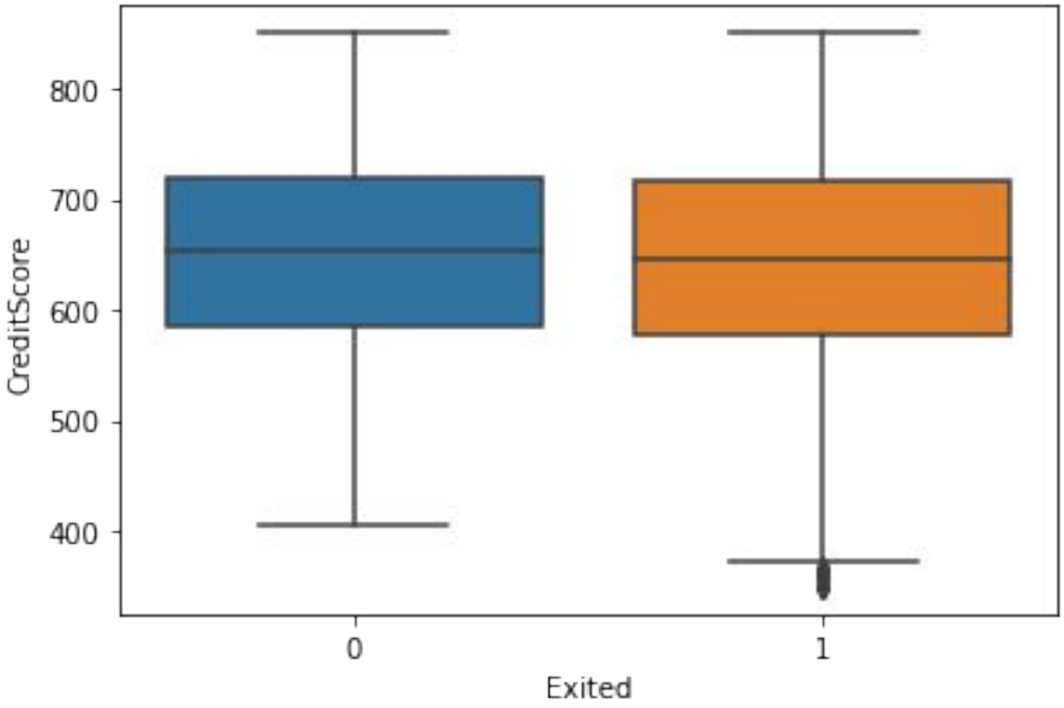




Correlation Matrix Heatmap

There is almost no Correlation within features

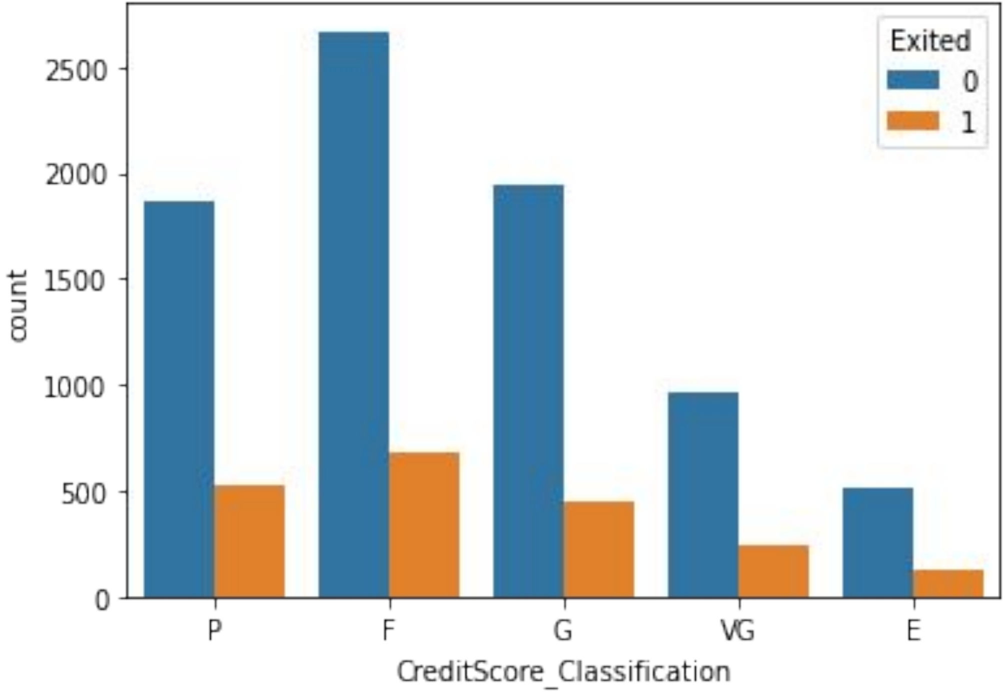
Analysis I(Credit Score)



The amount of customers that left vs those who stayed and distributed in the sme credit scores.

Maybe classifying them might provide better insights.

Analysis II(Credit Score Classification)



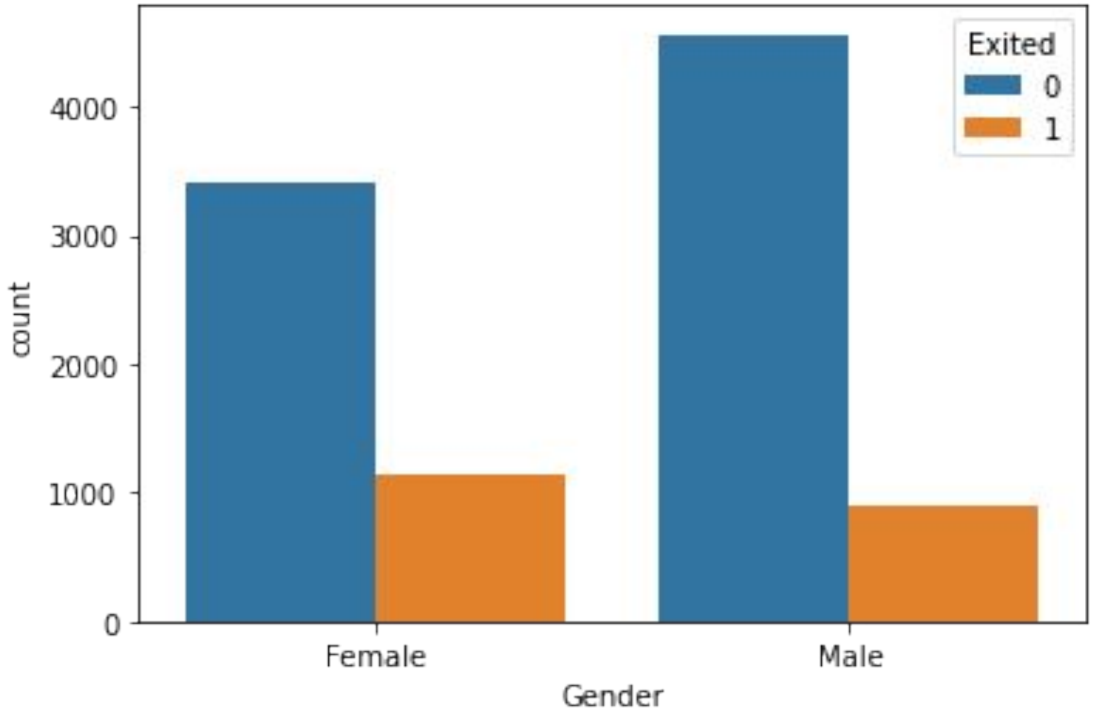
By %, the distribution is: F 33.50%, G 23.97%, P 23.93%, VG 12.15%, and E 6.45%

% distribution of the customers that leave.:
P 26.019%,F 33.53%,G 22.14%,VG 12.078%,and E 6.23%

% of Customers that leave per category:P 22.15%, F 20.39% ,G 18.82%, VG 20.25%, and E 19.69%

The distribution per class is very even. This categorizing verified that credit score can be ignored as a factor

Analysis III(Gender)



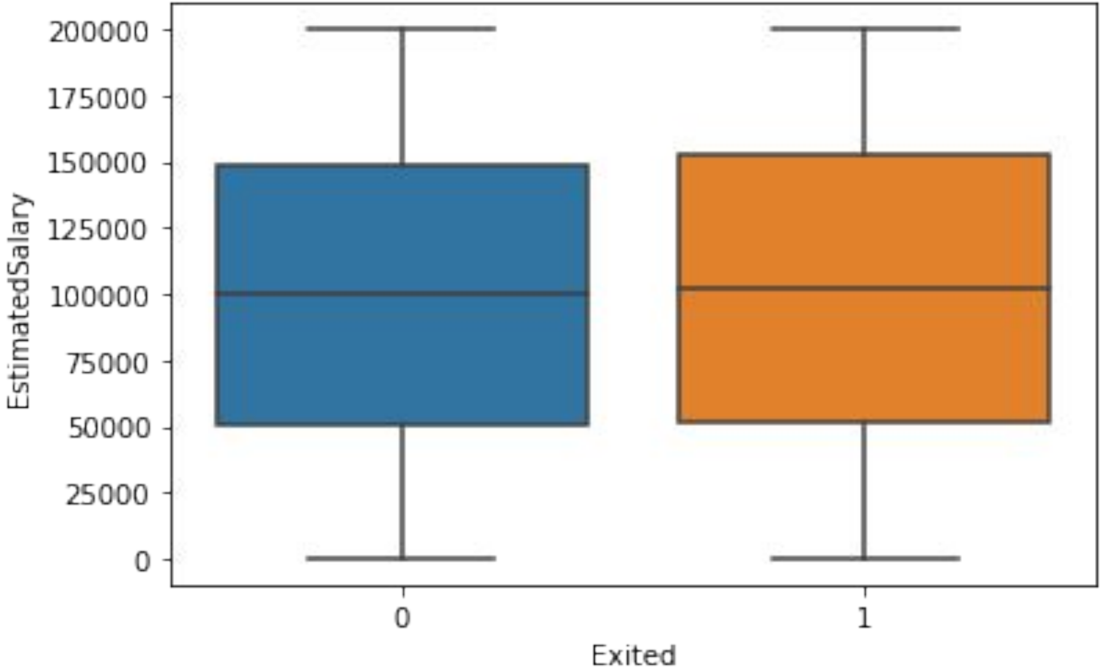
Overall distribution in % :Male 54.57% ,and Female 45.43%

Distribution of leavers overall in % :Female 55.92%, and Male 44.08%

% of leavers per Gender:Female 25.07%,and Male 16.46%

There is not much difference between categories,hence they could be ignored

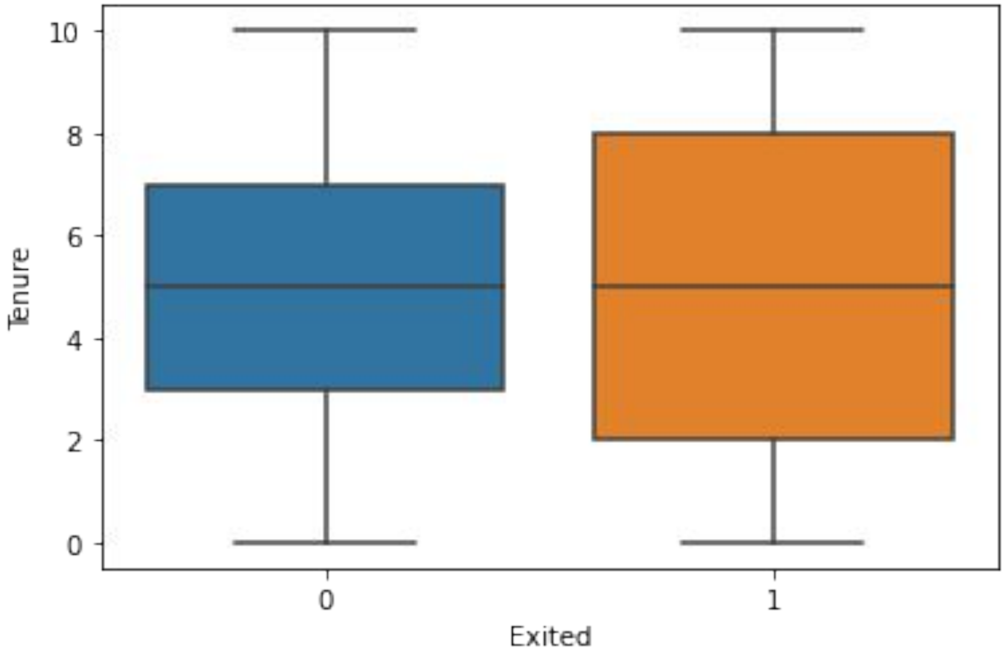
Analysis IV(Salary)



Similar distribution over salary

Could have caused noise for the model

Analysis V(Tenure)



Tenure does not appears to give any information in the distribution of those who leave and those who stay

Model Comparison

	Model	Accuracy	Recall	Precision	f_1_score
0	model_2	0.860	0.499	0.749	0.599
1	model_3	0.862	0.478	0.777	0.592
2	model_4	0.861	0.494	0.758	0.598
3	model_5	0.861	0.485	0.762	0.593
4	model_6	0.863	0.514	0.754	0.611
5	model_7	0.866	0.501	0.777	0.609
6	model_8	0.859	0.560	0.705	0.624
7	model_9	0.864	0.498	0.772	0.605
8	model_10	0.865	0.482	0.793	0.599
9	model_11	0.860	0.536	0.721	0.615
10	model_12	0.865	0.459	0.816	0.588

The best model to identify the costumes that might leave or might stay needed to have the highest Recall.
Model 11 was the best suited

Twelve models were trained with checkpoints set for the best weights

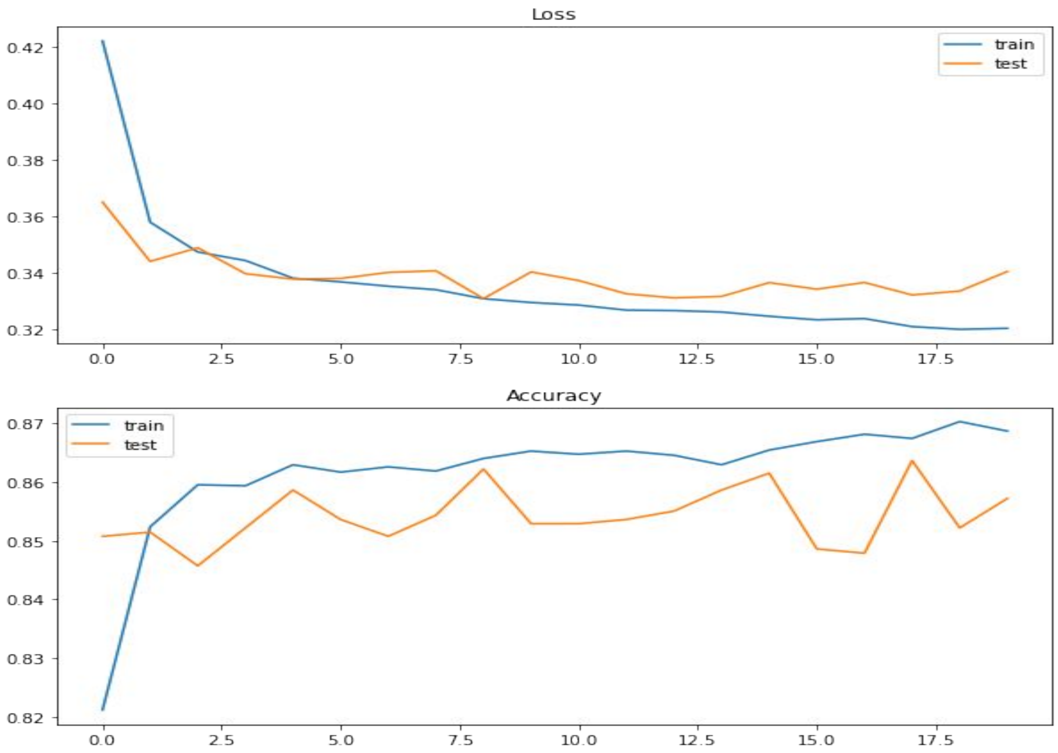
Summary of the best model

```
Model: "model_11"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 input_11 (InputLayer)       [(None, 10)]              0

 dense_48 (Dense)            (None, 100)               1100

 dense_49 (Dense)            (None, 50)                5050

 dense_50 (Dense)            (None, 25)                1275

 dense_51 (Dense)            (None, 12)                312

 dense_52 (Dense)            (None, 1)                 13

=================================================================
Total params: 7,750
Trainable params: 7,750
Non-trainable params: 0
_________________________________________________________________
```

Was made of 4 hidden layers

The activation Functions where a Relu/Tanh alternation.

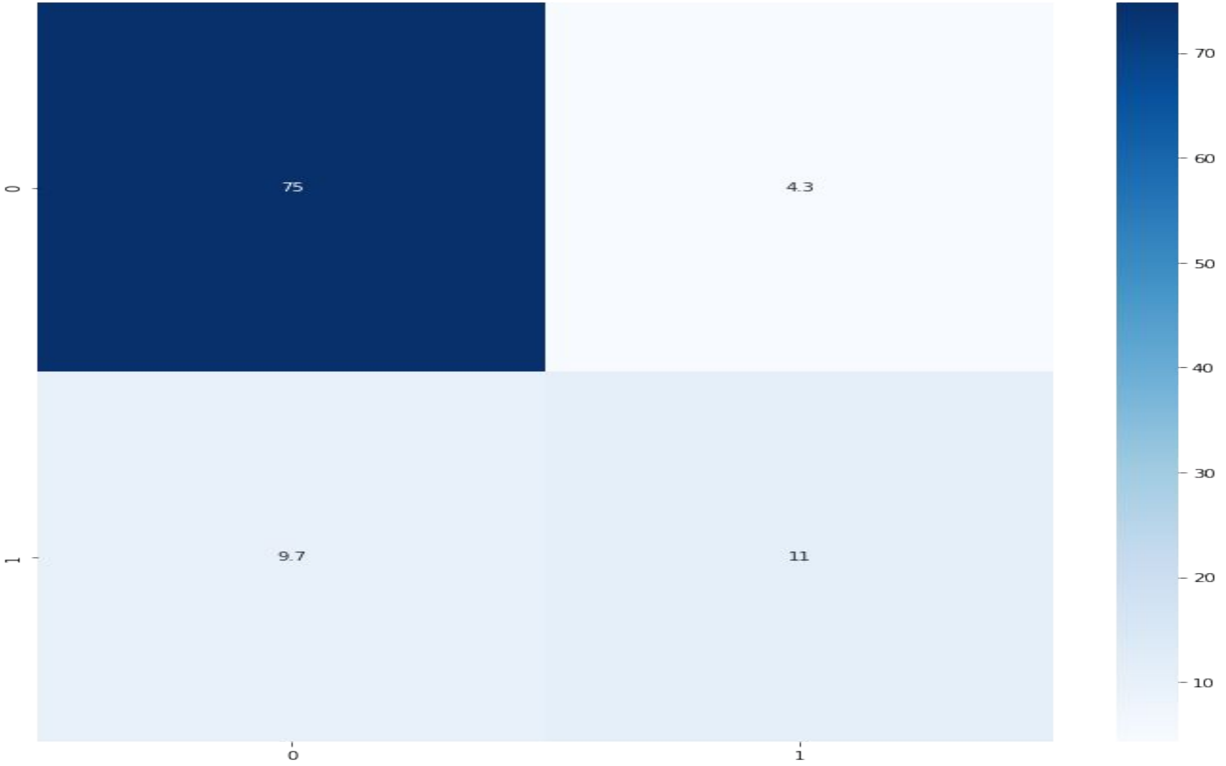
Histogram Of the Model



For the the test set. The best loss value was between 7-10 epochs

The highest accuracy was of 86%, around the same epoch area

Confusion Matrix in Percentage



My Thanks

Thank You for Your Attention

Further information can be found in the code, at the Jupyter Notebook

https://colab.research.google.com/drive/106rkMmTG-Xfsq9olmrwbEwDCpjw0syGC#scrollTo=XMAZZE5hK4_X