Capstone 1 Milestone Report

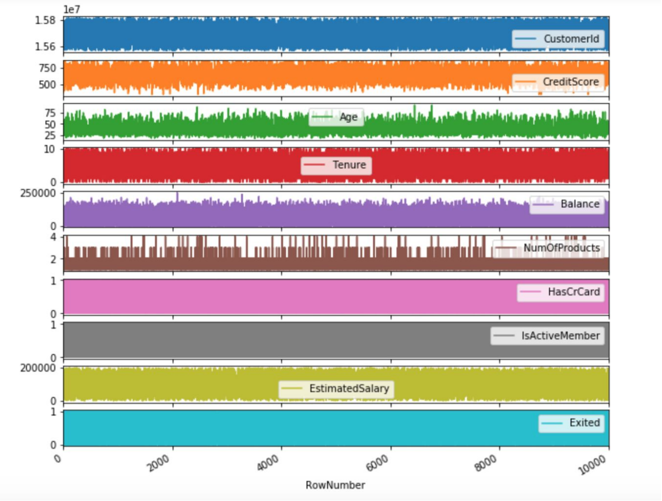
Problem Statement

Customer retention has always been the key to sustainable growth of a business. With the use of data, marketing campaigns can be more cost-effective not only in attracting new customers but retaining existing customers to generate more growth as well. In this case I'll look into a bank's customer data to discover more about how customers make their choice to stay or leave.

Dataset Description

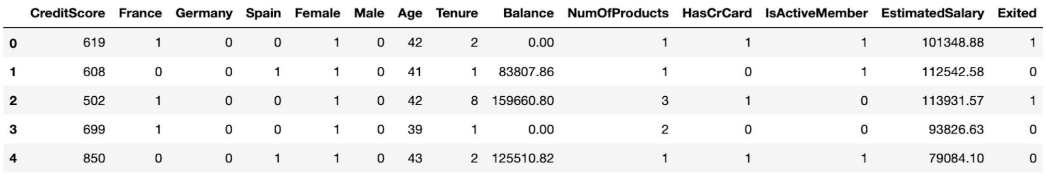
Data source:
https://www.kaggle.com/shrutimechlearn/churn-modelling

After importing the data as a Pandas dataframe, the data was checked for null or missing values, variable types and length of the dataframe. There is no missing values but Balance column has 36.2% zero values.
No outlier was detected either so we continued without removing or replacing any values.



Some irrelevant columns such as Customer ID were dropped to simplify the analysis process.

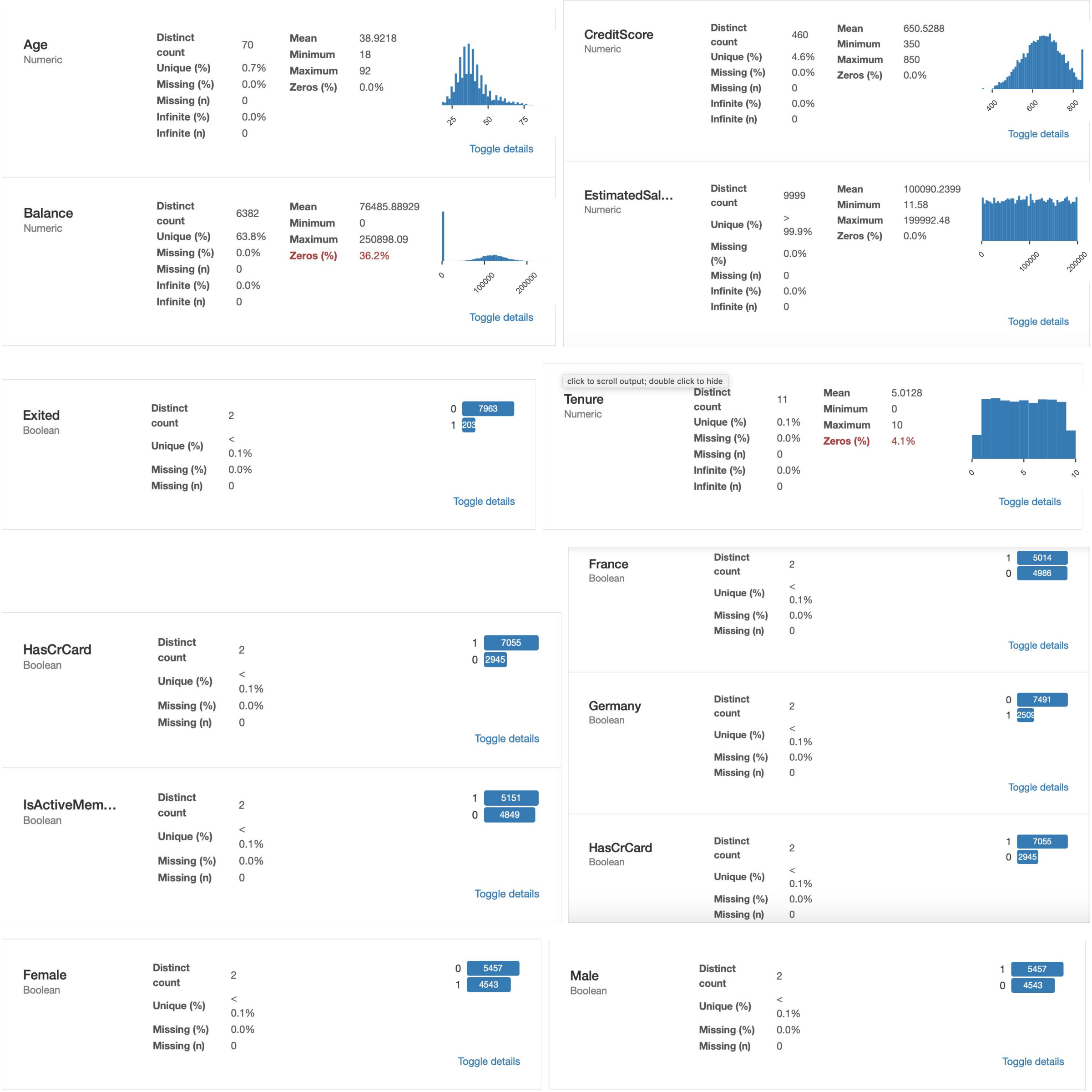
Then categorical columns such as Geography were transformed into numerical values using one hot encoding to perform further analysis. The cleaned dataset look like this:



	CreditScore	France	Germany	Spain	Female	Male	Age	Tenure	Balance	NumOfProducts	HasCrCard	IsActiveMember	EstimatedSalary	Exited
0	619	1	0	0	1	0	42	2	0.00	1	1	1	101348.88	1
1	608	0	0	1	1	0	41	1	83807.86	1	0	1	112542.58	0
2	502	1	0	0	1	0	42	8	159660.80	3	1	0	113931.57	1
3	699	1	0	0	1	0	39	1	0.00	2	0	0	93826.63	0
4	850	0	0	1	1	0	43	2	125510.82	1	1	1	79084.10	0

Initial Findings

For the exploratory analysis, basic summary statistics and histograms for each variable’s distribution were computed. Then empirical cumulative density functions of the variables and pairplots were plotted to see if there’s any obvious trends.



Age
Numeric

Distinct count	70	Mean	38.9218
Unique (%)	0.7%	Minimum	18
Missing (%)	0.0%	Maximum	92
Missing (n)	0	Zeros (%)	0.0%
Infinite (%)	0.0%		
Infinite (n)	0		

Toggle details

CreditScore
Numeric

Distinct count	460	Mean	650.5288
Unique (%)	4.6%	Minimum	350
Missing (%)	0.0%	Maximum	850
Missing (n)	0	Zeros (%)	0.0%
Infinite (%)	0.0%		
Infinite (n)	0		

Toggle details

Balance
Numeric

Distinct count	6382	Mean	76485.88929
Unique (%)	63.8%	Minimum	0
Missing (%)	0.0%	Maximum	250898.09
Missing (n)	0	Zeros (%)	36.2%
Infinite (%)	0.0%		
Infinite (n)	0		

Toggle details

EstimatedSal...
Numeric

Distinct count	9999	Mean	100090.2399
Unique (%)	> 99.9%	Minimum	11.58
Missing (%)	0.0%	Maximum	199992.48
Missing (n)	0	Zeros (%)	0.0%
Infinite (%)	0.0%		
Infinite (n)	0		

Toggle details

Exited
Boolean

Distinct count	2
Unique (%)	< 0.1%
Missing (%)	0.0%
Missing (n)	0

Toggle details

Tenure
Numeric

Distinct count	11	Mean	5.0128
Unique (%)	0.1%	Minimum	0
Missing (%)	0.0%	Maximum	10
Missing (n)	0	Zeros (%)	4.1%
Infinite (%)	0.0%		
Infinite (n)	0		

Toggle details

France
Boolean

Distinct count	2
Unique (%)	< 0.1%
Missing (%)	0.0%
Missing (n)	0

Toggle details

HasCrCard
Boolean

Distinct count	2
Unique (%)	< 0.1%
Missing (%)	0.0%
Missing (n)	0

Toggle details

Germany
Boolean

Distinct count	2
Unique (%)	< 0.1%
Missing (%)	0.0%
Missing (n)	0

Toggle details

IsActiveMem...
Boolean

Distinct count	2
Unique (%)	< 0.1%
Missing (%)	0.0%
Missing (n)	0

Toggle details

HasCrCard
Boolean

Distinct count	2
Unique (%)	< 0.1%
Missing (%)	0.0%
Missing (n)	0

Female
Boolean

Distinct count	2
Unique (%)	< 0.1%
Missing (%)	0.0%
Missing (n)	0

Toggle details

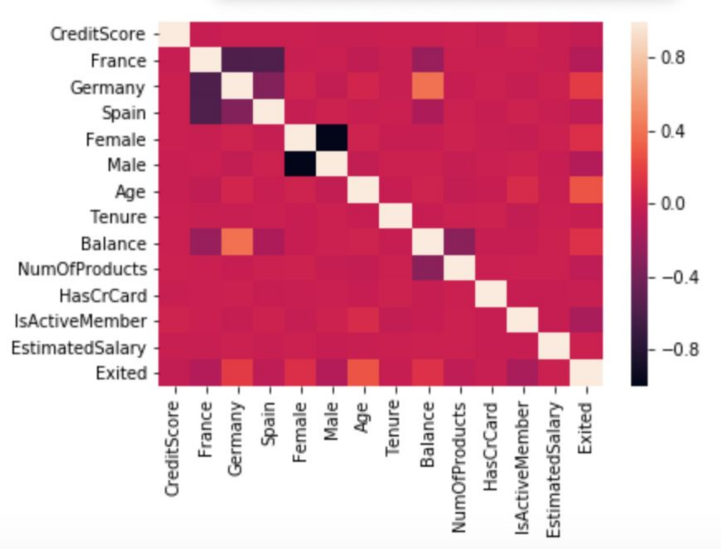
Male
Boolean

Distinct count	2
Unique (%)	< 0.1%
Missing (%)	0.0%
Missing (n)	0

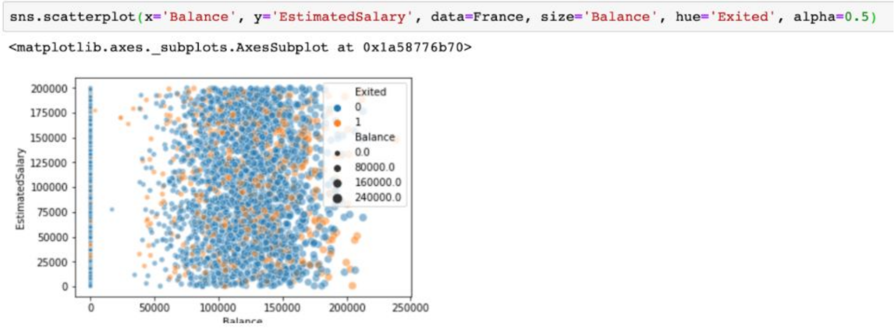
Toggle details

We can see that some of the variables seem to be normally distributed.

Then a Pearson's correlation matrix and a heatmap showing the correlations between variables were computed.

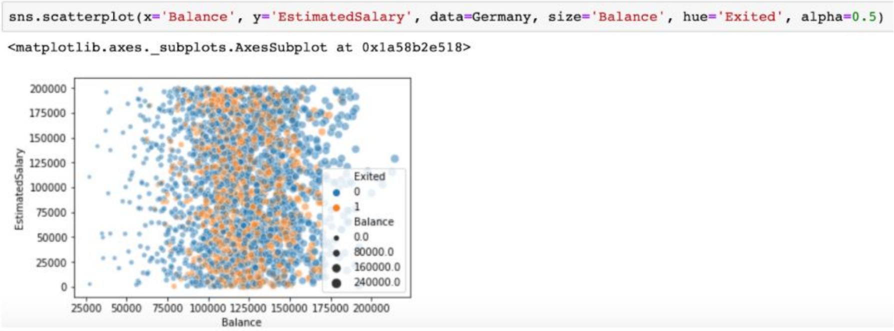


Countries seem to be correlated with balance and exited, so I plotted each country's balance and exited status in scatter plot and histogram. And retention rate (stayed customers/total customers) for each country was calculated.



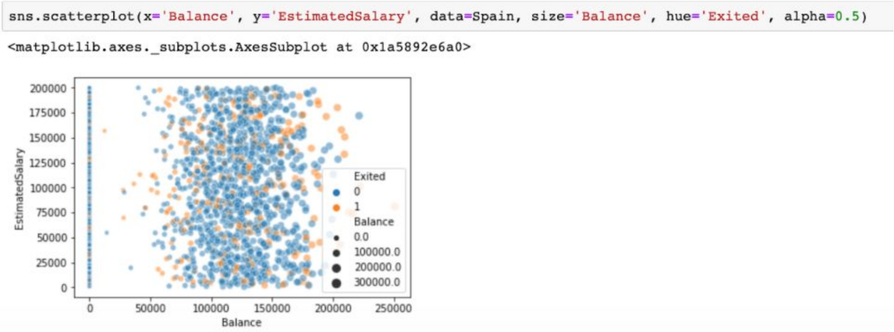
```
sns.scatterplot(x='Balance', y='EstimatedSalary', data=France, size='Balance', hue='Exited', alpha=0.5)
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x1a58776b70>
```



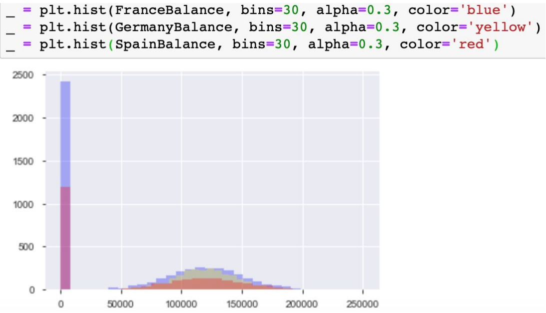
```
sns.scatterplot(x='Balance', y='EstimatedSalary', data=Germany, size='Balance', hue='Exited', alpha=0.5)
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x1a58b2e518>
```

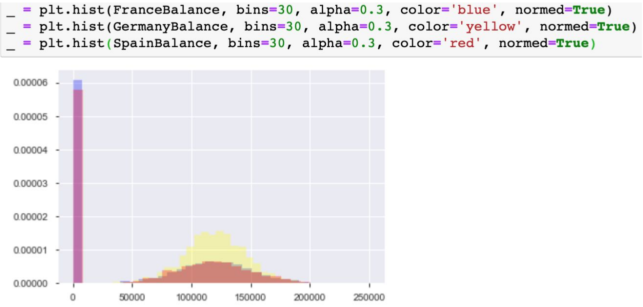


```
sns.scatterplot(x='Balance', y='EstimatedSalary', data=Spain, size='Balance', hue='Exited', alpha=0.5)
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x1a5892e6a0>
```



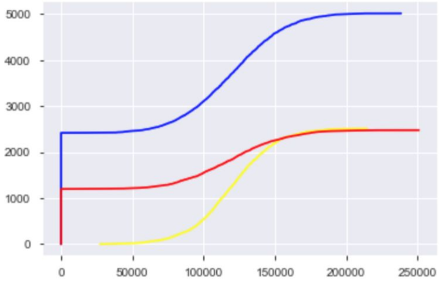
```
_ = plt.hist(FranceBalance, bins=30, alpha=0.3, color='blue')
_ = plt.hist(GermanyBalance, bins=30, alpha=0.3, color='yellow')
_ = plt.hist(SpainBalance, bins=30, alpha=0.3, color='red')
```



```
_ = plt.hist(FranceBalance, bins=30, alpha=0.3, color='blue', normed=True)
_ = plt.hist(GermanyBalance, bins=30, alpha=0.3, color='yellow', normed=True)
_ = plt.hist(SpainBalance, bins=30, alpha=0.3, color='red', normed=True)
```

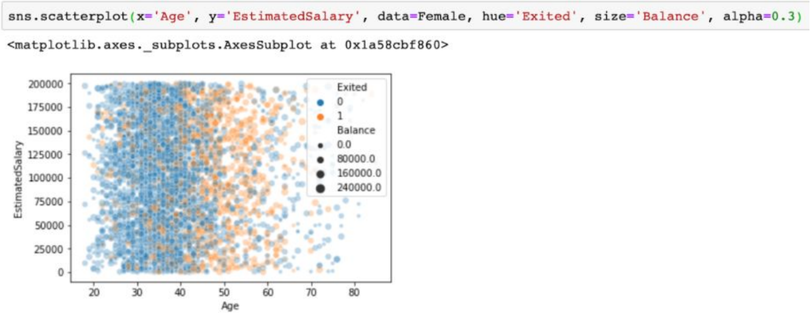
After calculation I found that three countries are similar in number of products and credit card holding rate.

Germany has no zero balance customers but a higher rate of exit and tend to have more centered distribution of balance. (Retention rate for France, Germany, Spain, respectively: 0.838, 0.676, 0.833)



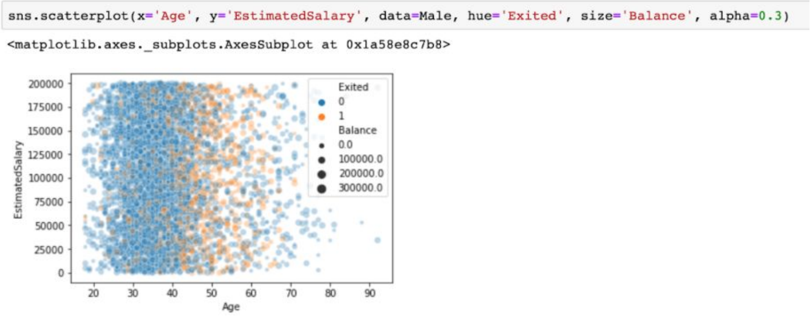
From the ECDFs of three countries' balance we can see France and Spain have similar patterns.

Next I looked into genders and made scatter plots showing the two gender groups' distribution in age, estimated salary, exited, and balance.



```
sns.scatterplot(x='Age', y='EstimatedSalary', data=Female, hue='Exited', size='Balance', alpha=0.3)
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x1a58cbf860>
```



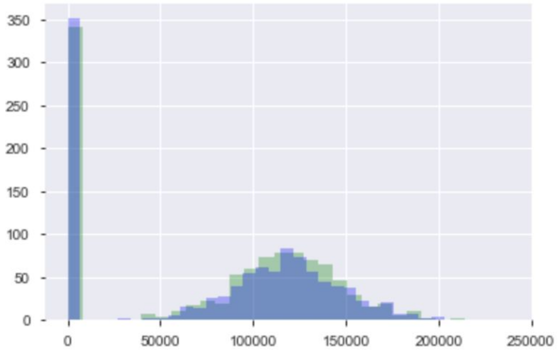
```
sns.scatterplot(x='Age', y='EstimatedSalary', data=Male, hue='Exited', size='Balance', alpha=0.3)
```

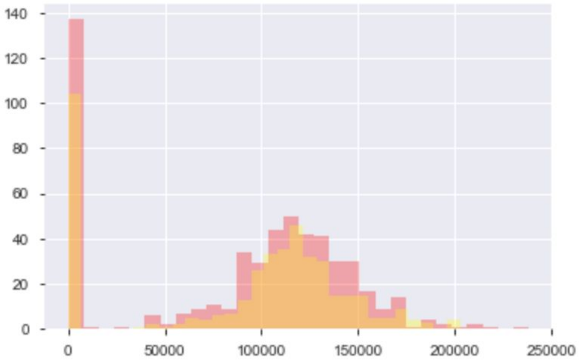
```
<matplotlib.axes._subplots.AxesSubplot at 0x1a58e8c7b8>
```

From the plots it seems like more both genders especially female customers tend to leave after a certain age (around 45). In general, male customers have a higher retention rate. (Retention rate for females and males: 0.749, 0.835)

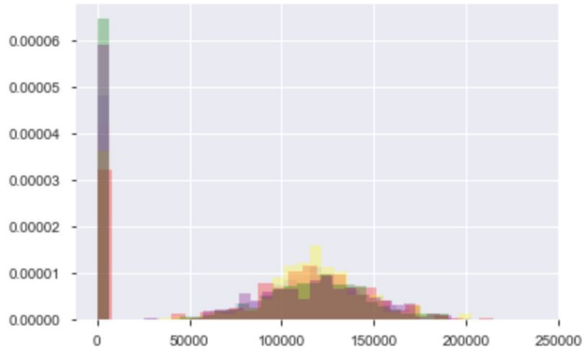
Female customers tend to be less active (active rate for males and females after 45: 0.614, 0.555) and leave after 45 (retention rate for females and males age>45: 0.483, 0.608).

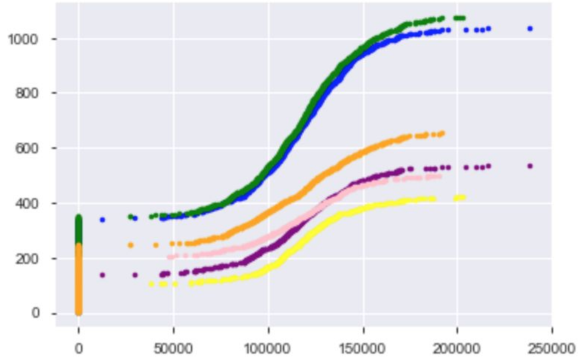
I divided customers by gender and age, exited to plot histograms of their balance and found that for the exited customers that are over age 45, the distribution of balance tend to be narrower and there are more females having balance around a certain range (102500 - 150000).





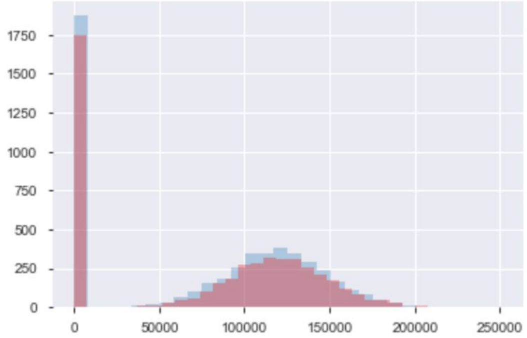
And from the normalized histogram and ECDFs we can see the other than the percentage of zero balance members, groups (divided by gender and age, exited status) seem to have similar distribution of balance, the exited groups seem to have more balance in the range mentioned previously. Gender doesn't seem to be affecting the shapes of the distributions.



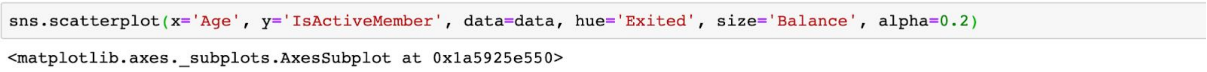


I then dug into age and active status and found that before 45 active members have higher retention rate (retention rate for active and not active members: 0.901, 0.826) and after 45 not active members have very low retention rate (retention rate for active and not active members: 0.720, 0.303)

We can see from the histogram that the balance distributions of active and not active members have quite similar patterns.

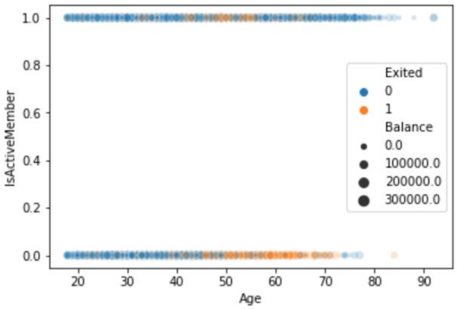


From the scatter plot it seems like for not active members, there is a tendency to leave after a certain age (around 45 as we previously noticed in gender groups).

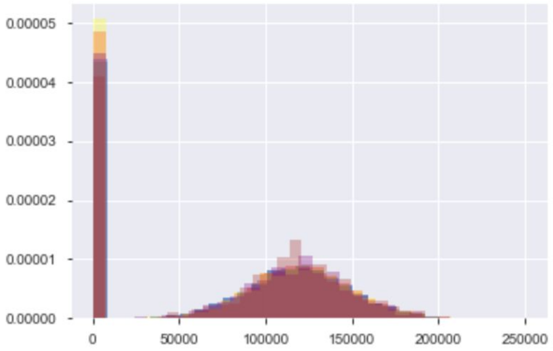


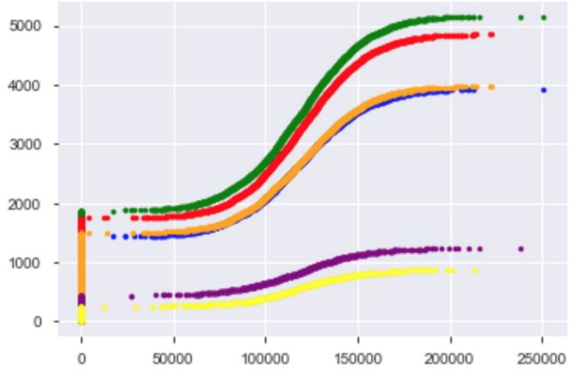
```
sns.scatterplot(x='Age', y='IsActiveMember', data=data, hue='Exited', size='Balance', alpha=0.2)
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x1a5925e550>
```

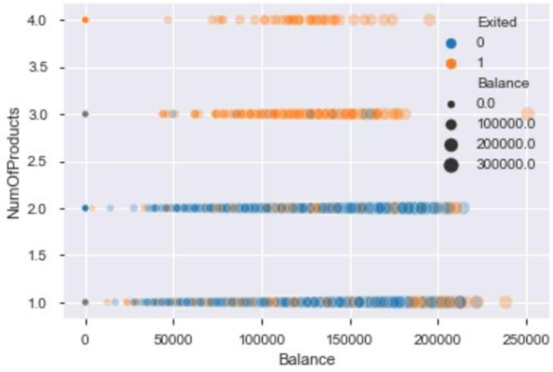


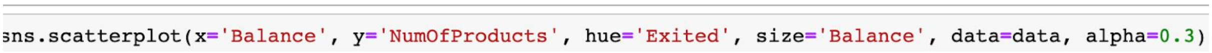
I grouped active and inactive members by age and plotted a normalized histogram and their ECDFs. We can see that they have very similar distributions especially of non zero balance parts. Age > 45 not active group has less members with zero balance and slightly more of balance within the range just mentioned. Not active groups tend to have more zero balance members. Active status doesn't seem to affect the shapes of the distributions but age seems to be an important factor.





Next I looked into balance and number of products. With the scatter plot here we can see that customers who purchased 3-4 products tend to leave.

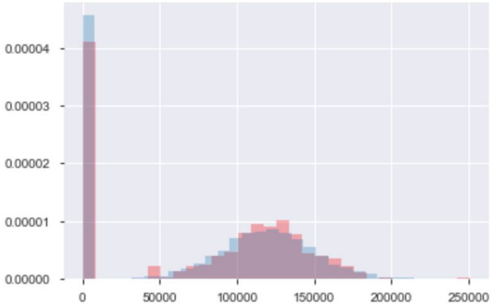




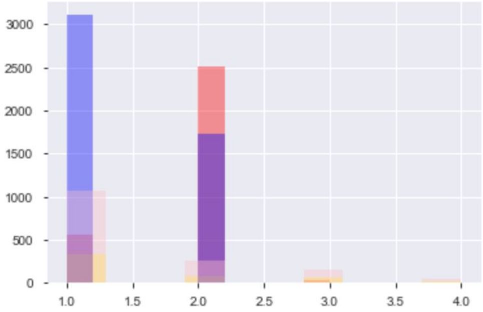
```
sns.scatterplot(x='Balance', y='NumOfProducts', hue='Exited', size='Balance', data=data, alpha=0.3)
```

And from counting values of the two groups of customers (1-2 products, 3-4 products since all customers purchased at least 1 product) we find that most customers tend to purchase 1-2 products but for those who purchased 3-4 products only 14.11% of them chose to stay. (Retention rate for customers with 3-4 products and 1-2 products: 0.141, 0.818)

From this normalized histogram of balance of customers with 1-2 and 3-4 products we see no significant difference in the distribution patterns.

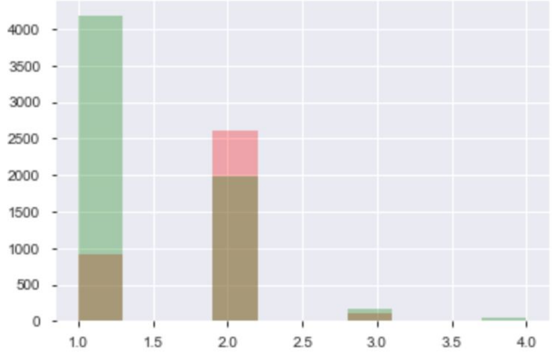


By dividing customers by balance (zero and non zero) and exited (exited and stayed) and we can see how many products they purchased. It shows that stayed customers tend to buy only 1-2 products and stayed with zero balance group tends to buy 2 while stayed with positive balance group tends to buy buy 1. But overall more customers tend to buy less products.



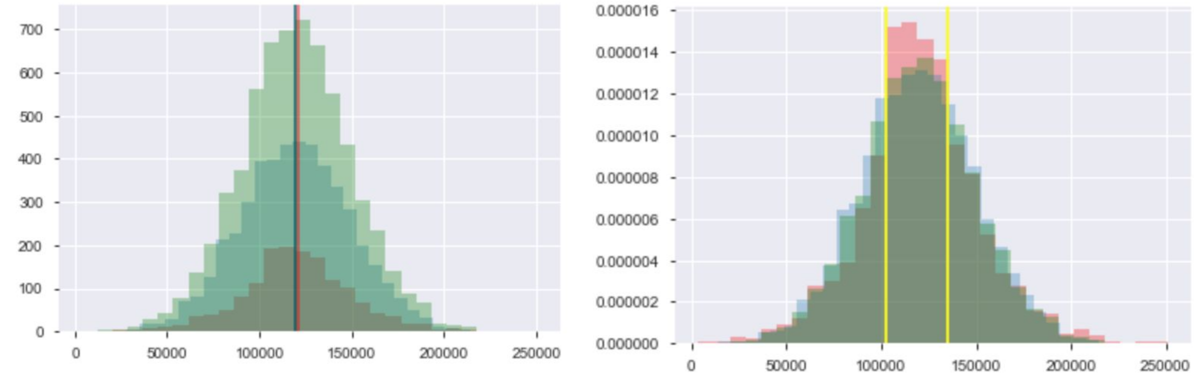
From the data we can see there are more customers with non zero balance but customers with zero balance have a higher retention rate compared to those with non zero balance, and there’s no significant correlation between a non zero balance and exited. (Retention rate for zero balance and non zero balance customers: 0.862, 0.759)

And by grouping customers into has balance and zero balance group and comparing distributions on age, credit score, number of products, tenure, estimated salary, credit card and active status, we find zero balance customers only show different distribution on number of products.

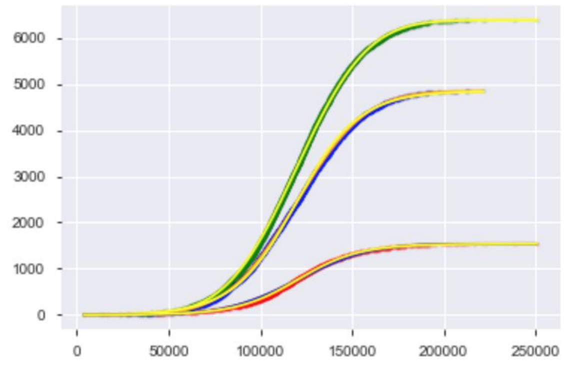


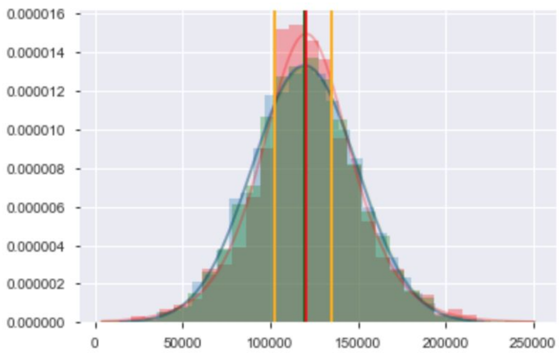
Then I looked into customers with a non zero balance to see if there’s any correlation.

I divided customers with non zero balance by exited and plotted the histogram on the left which shows they have very close means and might all be normally distributed. And the normalized histogram on the right shows that within the balance range marked by the yellow lines customers tend to exit. (around [25, 70] percentile of exited group)



Then by plotting the ECDFs and trying different regressions (normal, t, logistic) we find non zero exited customers' balance seems to be logistically distributed while the other two groups are normally distributed. And the regression functions fit well with the empirical data.





Tenure, credit score, and estimated salary doesn't seem to have strong correlations with other variables.

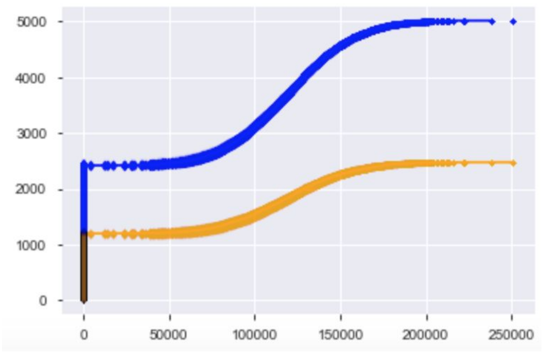
Summary of findings

Based on the initial findings, I tested some hypotheses using bootstrap and permutation simulations.

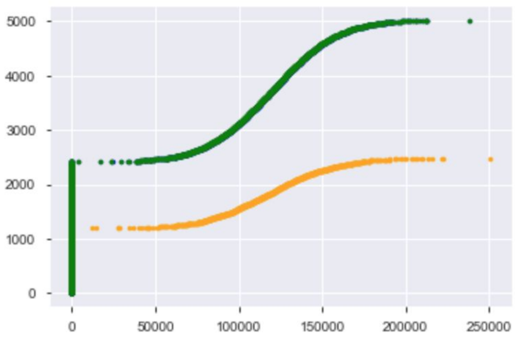
As the three countries' balance ECDFs showed, France and Spain are a bit similar, and France's non zero part looks similar in shape to Germany.

To test if France and Spain have the same mean and distribution, permutation test (n = 1000) was performed and a p value of 0.424 was generated, indicating that the two countries could have the same mean and they are likely two parts of a same distribution. A bootstrap test (n = 10000) then generated a p value of 0.4342, showing the two groups could have the same mean.

```
0.424
```

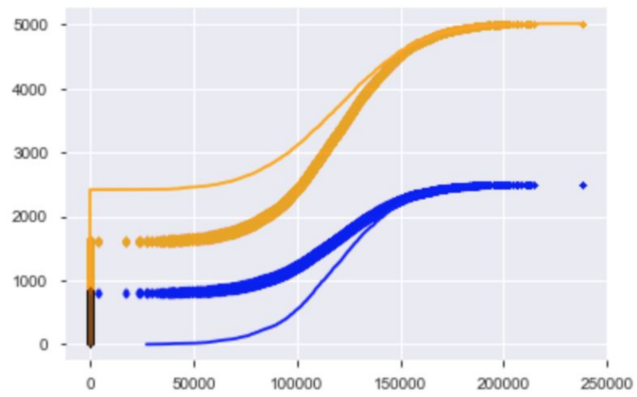


```
0.4342
```



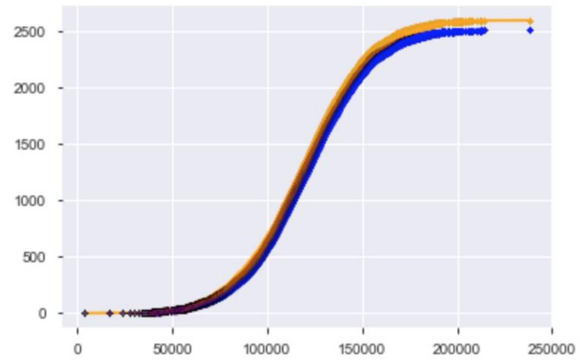
To compare France and Germany, we include the zero balance members first. And the permutation test (n = 1000) shows that the two groups are not likely to have the same mean or distribution (p value is 0.0).

```
0.0
```

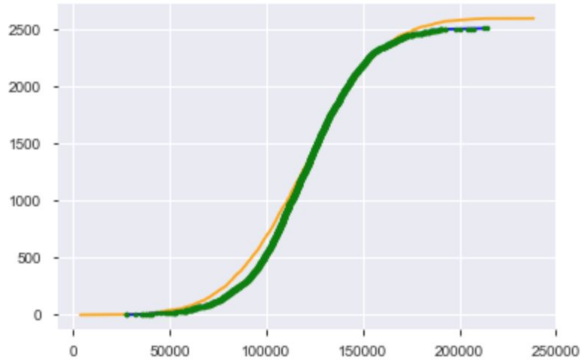


Then after removing the zero balance members in France and running the same permutation test again, we get a p value of 0.561 and they seem to be from the same distribution. And the bootstrap test generated a p value of 0.4117, indicating it is possible non zero balance France and Germany have the same mean and distribution.

```
0.561
```

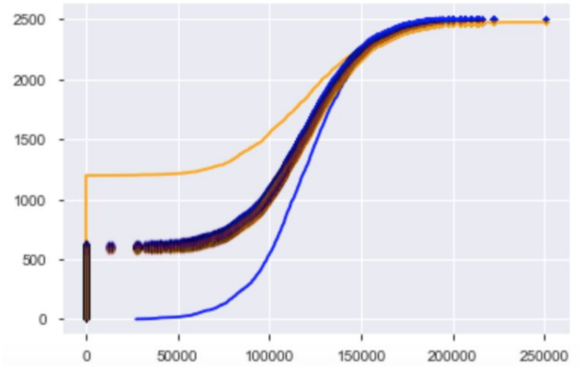


```
0.4117
```



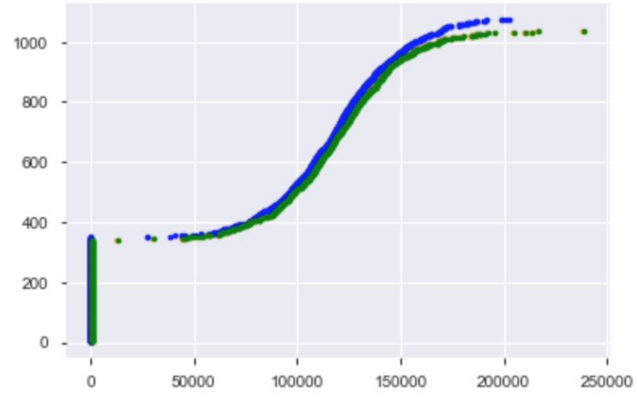
The permutation test on Spain and Germany shows that the two countries are not likely from the same distribution.

```
0.0
```

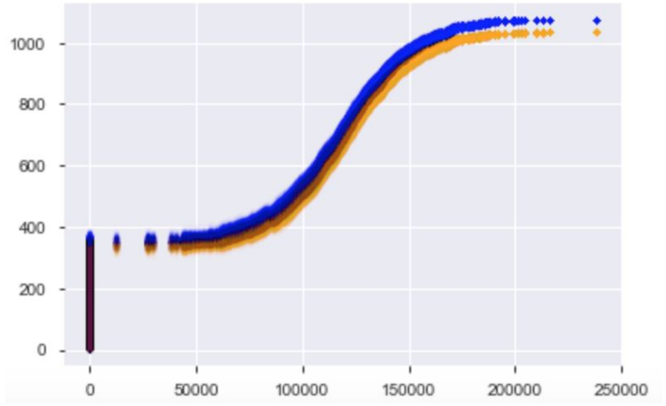


To test if age >45 female and male groups have the same mean and distribution, bootstrap and permutation tests were used.

Same as the bootstrap test before, the mean of one group was shifted to the same as the other group, and after bootstrap sampling (n = 10000), a p value of 0.3859 shows it’s possible the two groups have the same mean.

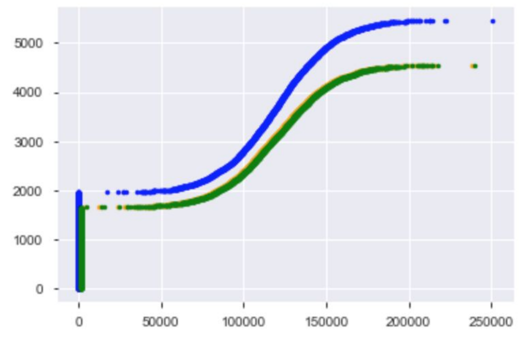


Then from the permutation (n = 1000) test we can see the two groups are from the same distribution and a p value of 0.397 shows that it’s possible they have the same means.

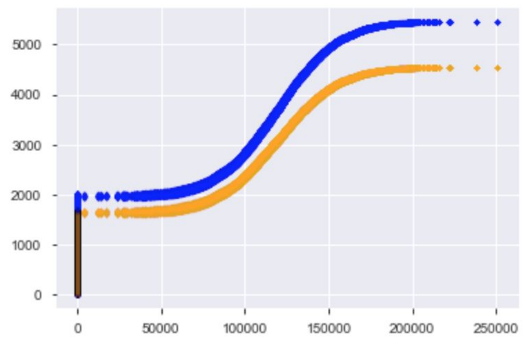


Then after performing the same tests on female and male members, no significant difference in their mean was found, and they are most likely from the same distribution. (p value for bootstrap test is 0.115, and 0.118 for permutation test)

```
0.115
```

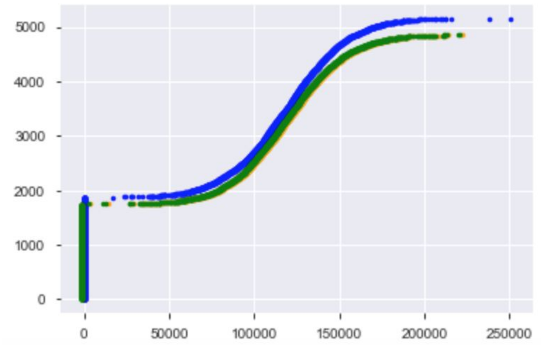


```
0.118
```

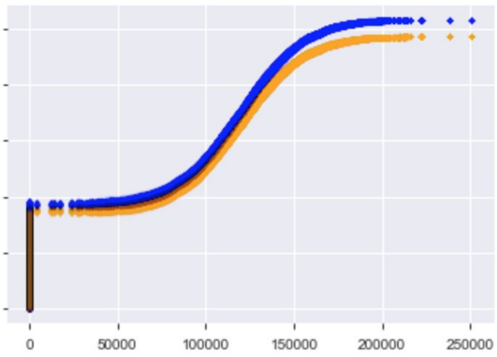


Then to test the effect of active status on balance, same bootstrap and permutation tests were performed on active and inactive groups. A p value of 0.845 from the bootstrap test shows that two groups are likely to have the same mean. And a p value of 0.837 indicates that they are likely to have the same mean and distribution.

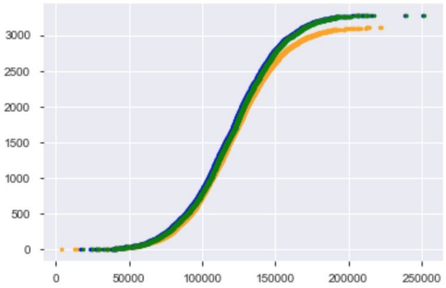
```
0.845
```



```
0.837
```



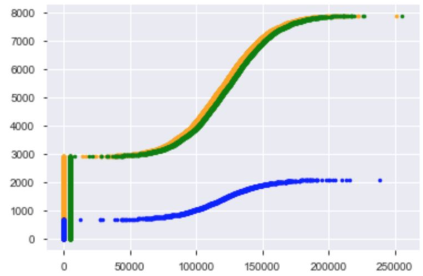
And after removing the zero balance members from the two groups and running the same bootstrap test again, a p value of 0.9492 was computed showing that the non zero balance members of two groups are very likely of the same mean.



To test if there’s a significant difference in balance between two age groups divided by 45, bootstrap and permutation tests were used.

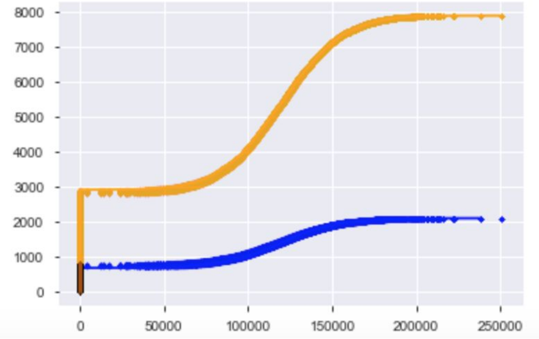
By shifting the mean of one group to the same as the other group and running bootstrap sampling (n = 10000) and comparing the simulated difference of means and empirical difference of means, a p value of 0.0008 was computed, indicating it’s not likely that the two age groups have the same mean.

```
0.0008
```

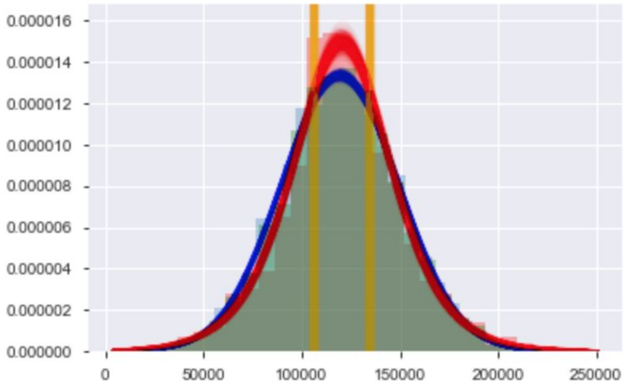


Permutation (n = 1000) shows that two groups are two parts of the same distribution, a p value of 0.001 indicates that two groups are not likely to have the same mean.

```
0.001
```



By performing bootstrap sampling and regressions (n=1000) we can get confidence intervals of the regression functions and we can tell around [30, 70] percentile of the exited group is where customers tend to leave.



Using bootstrap sampling we can get 95% confidence intervals for means and stds of the groups (n=10000):
Non Zero Balance Exited Group Mean: [119204.61233393, 122283.12203969]
Non Zero Balance Exited Group Standard Deviation: [29248.69979372, 31874.10350709]
Non Zero Balance Stayed Group Mean: [118684.88234797, 120385.14285602]
Non Zero Balance Stayed Group Standard Deviation: [29364.8433063 , 30524.31485055]

In conclusion, customers of a certain range of balance tend to have a higher exit rate; zero balance customers only show different distribution when purchasing products and most customers tend to purchase 1-2 products, those who have 3-4 products tend to leave though two groups (1-2 and 3-4 products) have very similar distribution of balance; active status and gender do not significantly affect balance, female and not active customers over 45 might be leaving for reasons other than balance; age > 45 group tend to have a higher exit rate and a lower active rate; It’s possible that France and Spain can have the same mean and come from the same distribution, and Germany and France without zero balance members can have the same mean and distribution.

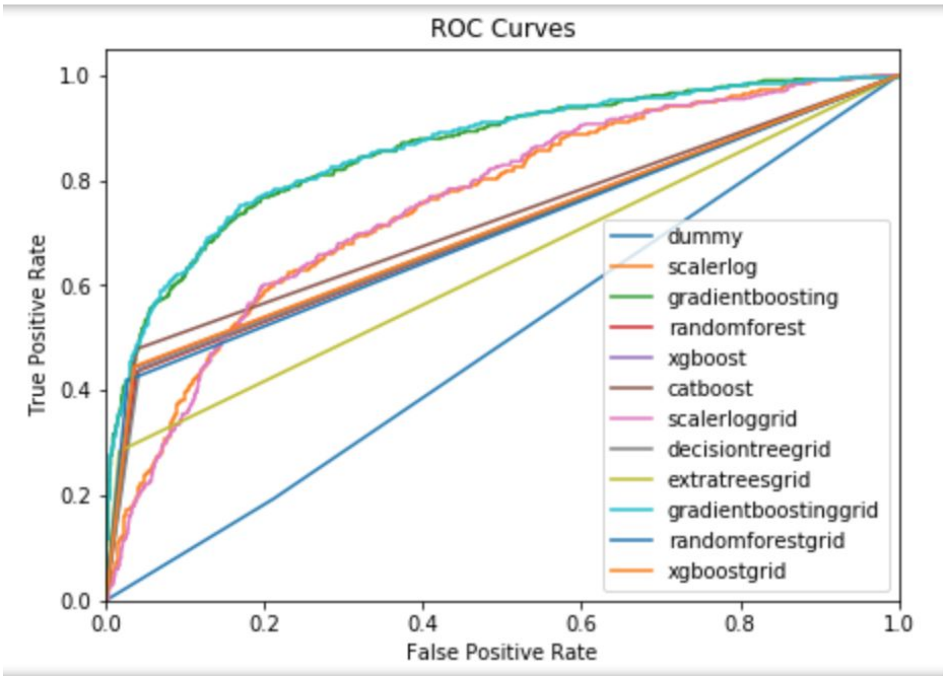
In-depth analysis summary

In this section some machine learning techniques were applied to the dataset to build a model that is able to generate good predictions.

The dataset was split into two to leave out a group for testing (20% of entire dataset), and a dummy classifier was then used to set up a baseline for model performance. For preprocessing and feature selection, I tested standard scaler and K best with chi square with logistic regression classifier, and new features created by combining Age and Balance as these two features seemed to be related to Exited. For classifiers, I tested logistic regression, decision tree, extra tree, extra trees, gradient boosting, random forest, svc, catboost and xgboost. For metrics, I used both accuracy score and roc auc score to compare the models and tune hyperparameters. After running cross validation scoring with training data on the models, a few less accurate models were eliminated (logistic regression, logistic regression with chi square k best, logistic regression with chi square k best and scaler, extra tree, svc) and the testing data was fitted into the rest models to get predictions and scores.

model	prediction accuracy score	prediction roc auc score
'dummy'	0.49100937655800425	0.6625
'scalerlog'	0.7534118417337923	0.7995
'gradientboosting'	0.8565802593720528	0.858
'randomforest'	0.6983643582674091	0.85
'xgboost'	0.7041368581022232	0.855
'catboost'	0.7200082292661536	0.8595
'scalerloggrid'	0.754238522576421	0.788
'decisiontreegrid'	0.699149742610179	0.8485
'extratreesgrid'	0.633624961706882	0.8355
'gradientboostinggrid'	0.8585812625015766	0.8605
'randomforestgrid'	0.6957739414580817	0.8555
'xgboostgrid'	0.7050048354447107	0.855

From the prediction scores and roc curves gradient boosting seems to be the best performing classifier and hyperparameter tuning only slightly enhanced the performance of the models. All models performed better than the dummy model which was set to be a baseline.



I was not able to create new features that significantly enhance the prediction's accuracy. I experimented with featuretools as well as creating new feature manually by dividing Balance by Age and Balance percentile by Age percentile, however the results were not affected. I also had to reduce the grids of hyperparameter tuning due to overly long computation time, and I didn't use random search which might have been helpful for some models' performance. If I had more time and computation power I would do more feature engineering and try out more classifiers or combine multiple classifiers with weights.