

Capstone 1 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>

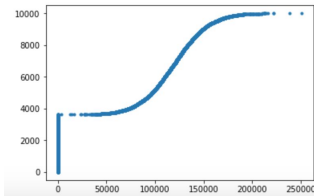
- No missing values but Balance column has 36.2% zero values.
- No outlier was detected either so we continued without removing or replacing any values.
- Some irrelevant columns such as Customer ID were dropped to simplify the analysis process.
- Then categorical columns such as Geography were transformed 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 |

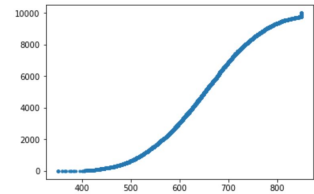
Initial Findings

- Basic summary statistics and histograms for each variable's distribution were computed.
- Empirical cumulative density functions of the variables and pairplots were plotted.
- Some of the variables seem to be normally distributed.

```
plt.plot(np.sort(data.Balance), np.arange(1, len(data.Balance)+1), marker='.', linestyle='none')
```

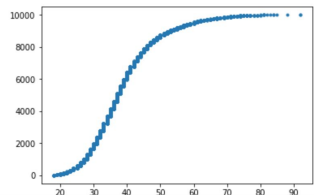


```
#ecdf
plt.plot(np.sort(data.CreditScore), np.arange(1, len(data.CreditScore)+1), marker='.', linestyle='none')
```



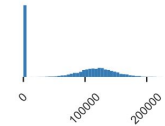
```
plt.plot(np.sort(data.Age), np.arange(1, len(data.Age)+1), marker='.', linestyle='none')
```

ck to expand output; double click to hide output



Balance
Numeric

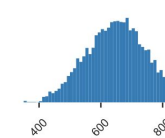
| | | | |
|----------------|-------|-----------|-------------|
| Distinct count | 6382 | Mean | 76485.88929 |
| Unique (%) | 63.8% | Minimum | 0 |
| Missing (%) | 0.0% | Maximum | 250898.09 |
| Missing (n) | 0 | Zeros (%) | 36.2% |
| Infinite (%) | 0.0% | | |
| Infinite (n) | 0 | | |



[Toggle details](#)

CreditScore
Numeric

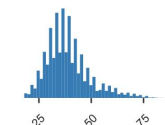
| | | | |
|----------------|------|-----------|----------|
| Distinct count | 460 | Mean | 650.5288 |
| Unique (%) | 4.6% | Minimum | 350 |
| Missing (%) | 0.0% | Maximum | 850 |
| Missing (n) | 0 | Zeros (%) | 0.0% |
| Infinite (%) | 0.0% | | |
| Infinite (n) | 0 | | |



[Toggle details](#)

Age
Numeric

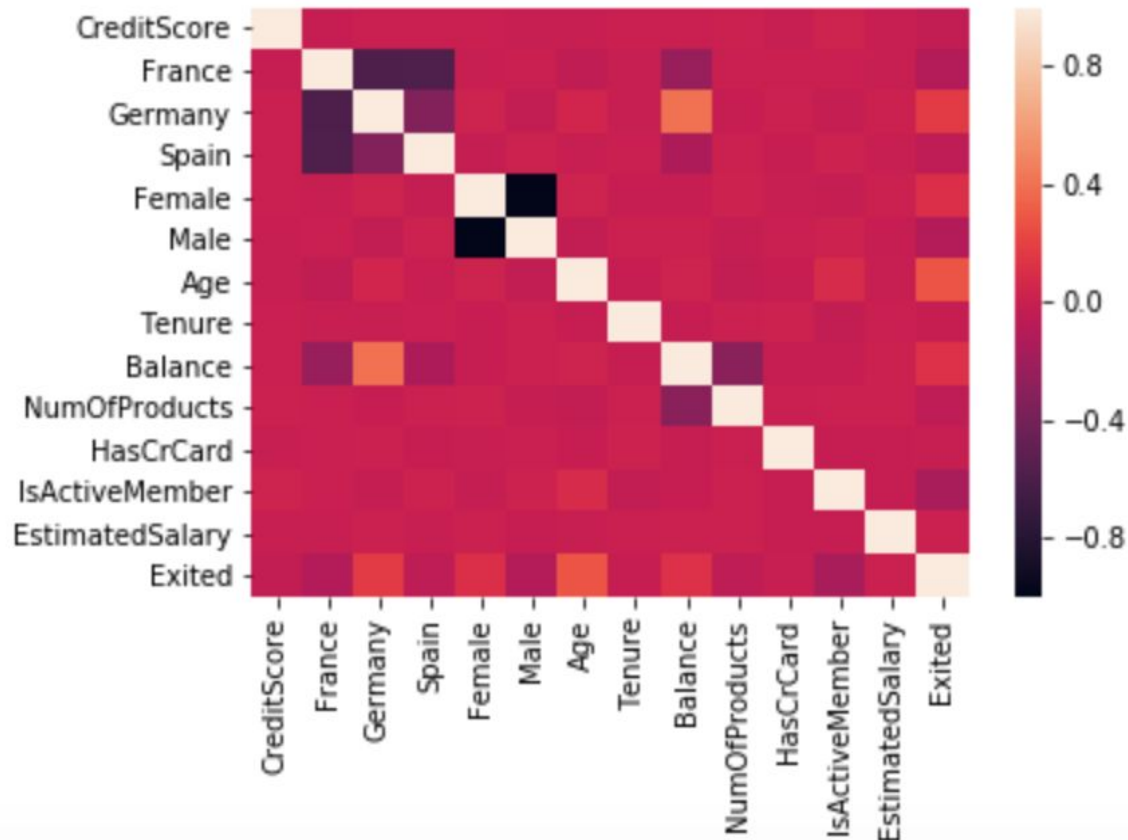
| | | | |
|----------------|------|-----------|---------|
| Distinct count | 70 | Mean | 38.9218 |
| Unique (%) | 0.7% | Minimum | 18 |
| Missing (%) | 0.0% | Maximum | 92 |
| Missing (n) | 0 | Zeros (%) | 0.0% |
| Infinite (%) | 0.0% | | |
| Infinite (n) | 0 | | |



[Toggle details](#)

Initial Findings

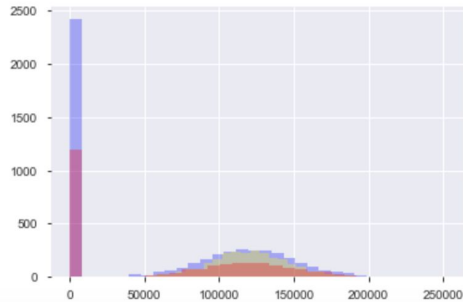
Pearson's correlation matrix heatmap



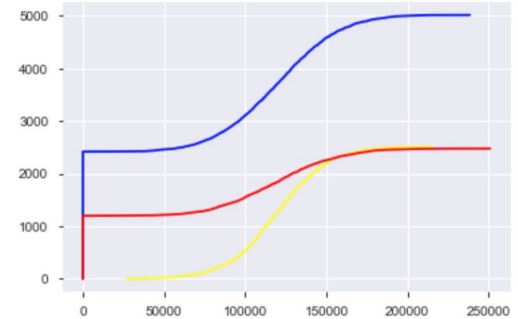
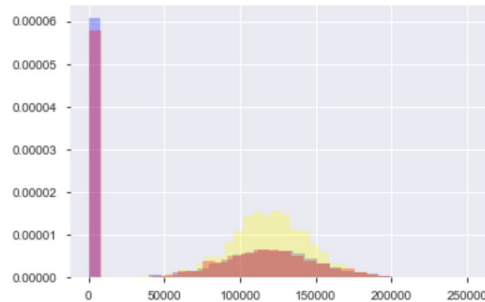
Initial Findings: Geography

- FR, GR and SP are similar in number of products and credit card holding rate.
- GR has no zero balance customers but a higher rate of exit and tend to have more centered distribution of balance.

```
_ = 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)
```

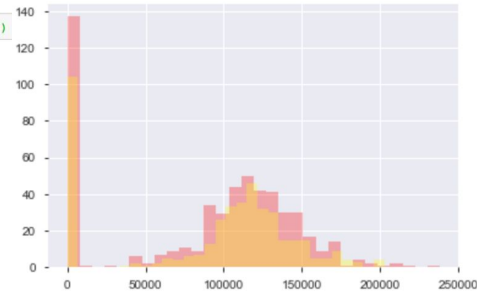
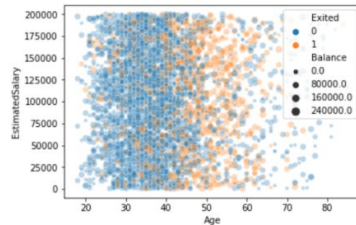


Initial Findings: Gender

- Both genders especially female customers tend to leave after a certain age (45). (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).
- For the exited customers that are over age 45, the distribution of balance tends to be narrower and there are more females having balance around a certain range (102500 - 150000).

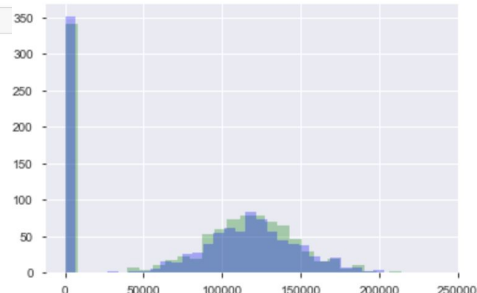
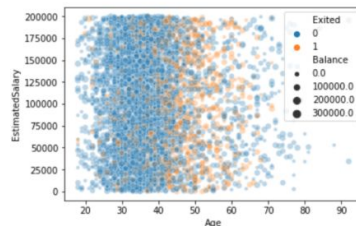
```
sns.scatterplot(x='Age', y='EstimatedSalary', data=Female, hue='Exited', size='Balance', alpha=0.3)
```

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

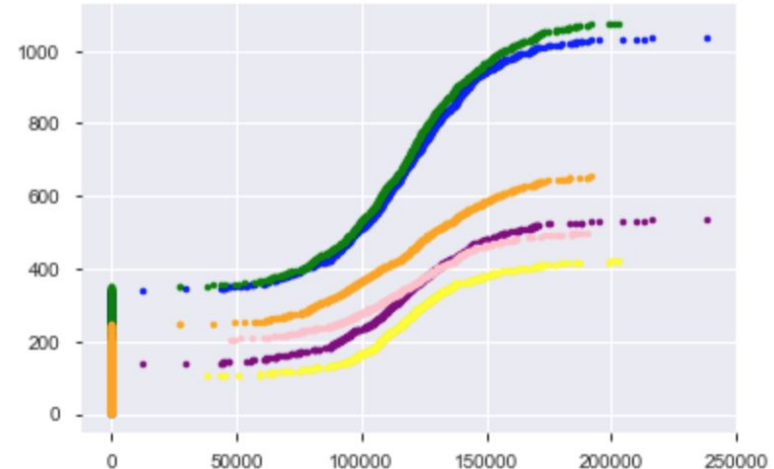
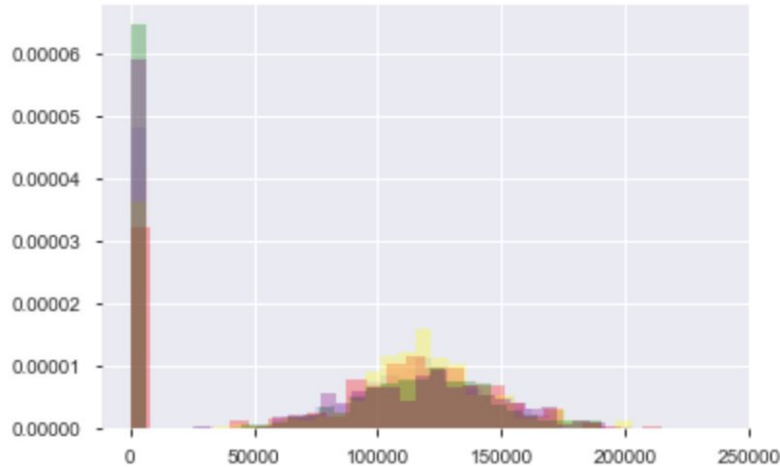


```
sns.scatterplot(x='Age', y='EstimatedSalary', data=Male, hue='Exited', size='Balance', alpha=0.3)
```

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



- 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.

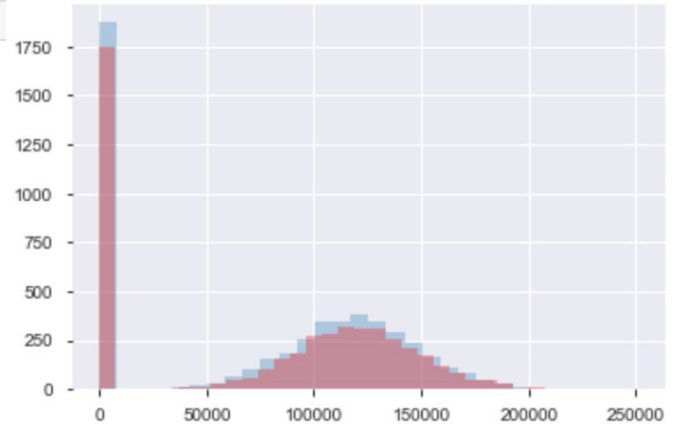
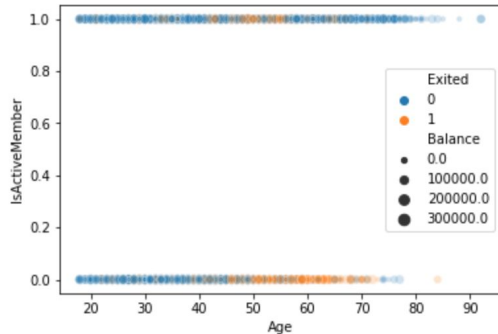


Initial Findings: Active member

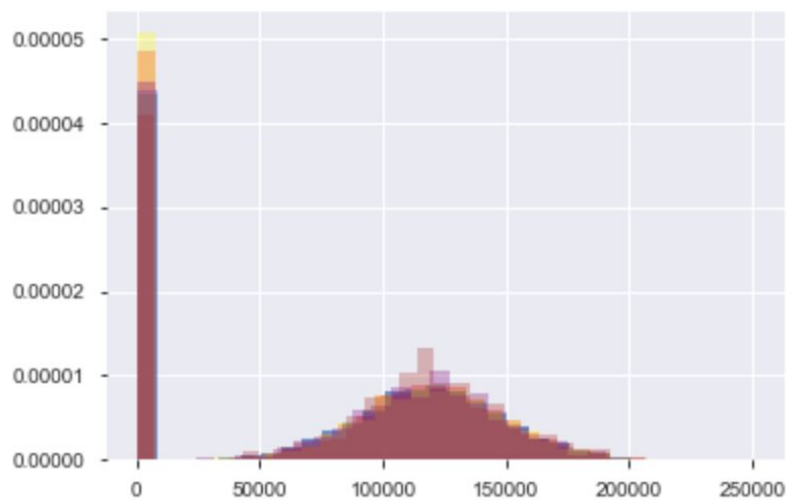
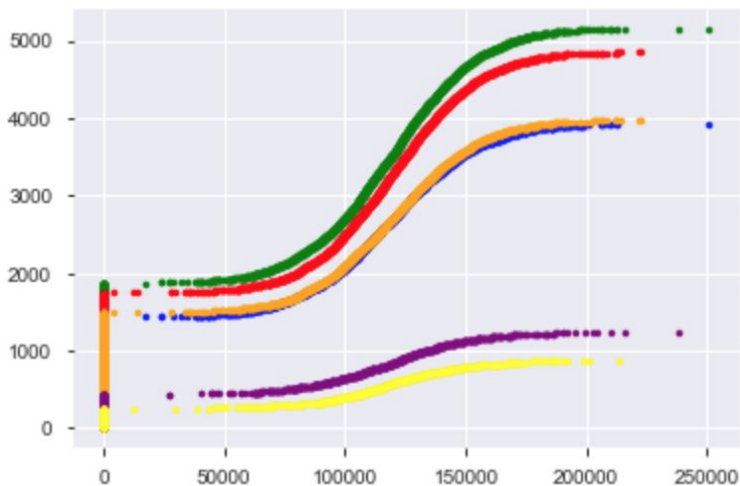
- Before 45 active members have higher retention rate (retention rate for active and not active members: 0.901, 0.826) and after 45 not active members have very low retention rate (retention rate for active and not active members: 0.720, 0.303)
- Active and not active members have quite similar patterns on balance.
- 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>
```

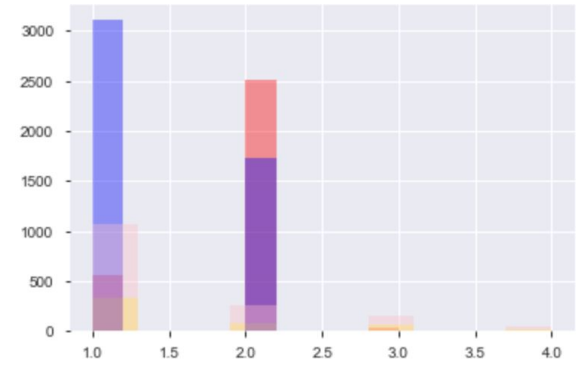
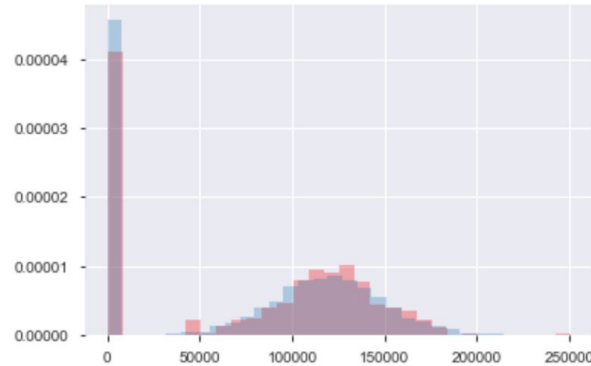
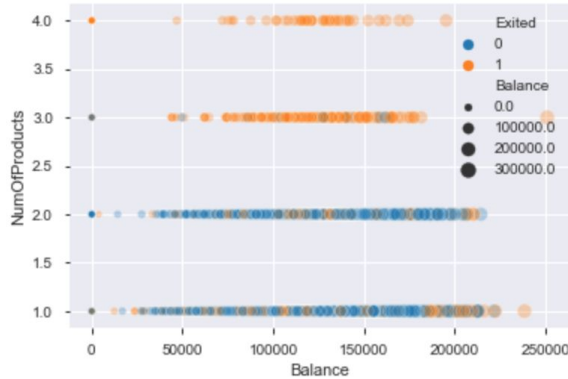


- I grouped active and inactive members by age and plotted a normalized histogram and their ECDFs. We can see that they have very similar distributions especially of non zero balance parts. Age > 45 not active group has less members with zero balance and slightly more of balance within the range just 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.



Initial Findings: Number of products





- Most customers tend to purchase 1-2 products but for those who purchased 3-4 products only 14.11% of them chose to stay. 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 1. But overall more customers tend to buy less products.
- Balance of customers of 1-2 and 3-4 products have no significant difference in the distribution patterns.
-



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

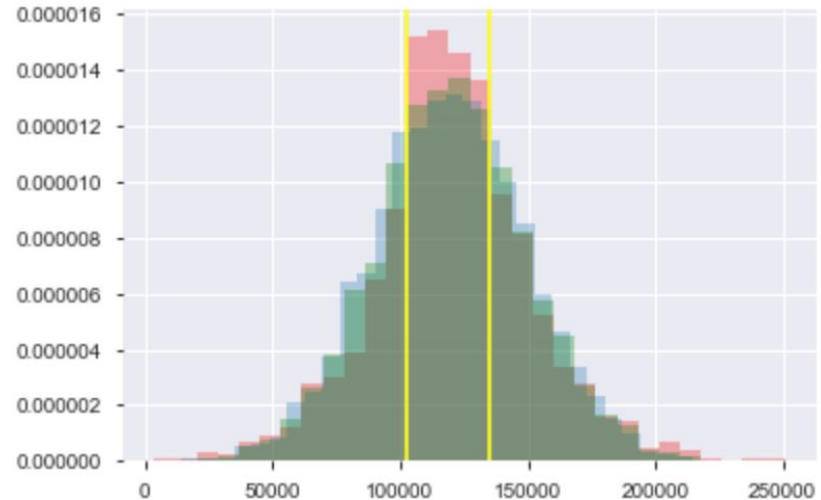
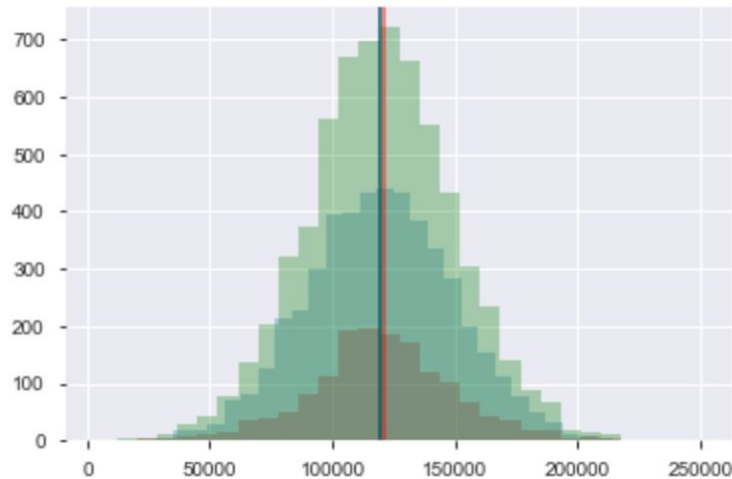
Initial Findings: Balance

- 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 (retention rate for zero balance and non zero balance customers: 0.862, 0.759)
- No significant correlation between a non zero balance and exited.
- Zero balance customers only show different distribution on number of products.

| Value | Count | Frequency (%) | |
|-------|-------|---------------|---|
| 1 | 5084 | 50.8% |  |
| 2 | 4590 | 45.9% |  |
| 3 | 266 | 2.7% |  |
| 4 | 60 | 0.6% |  |

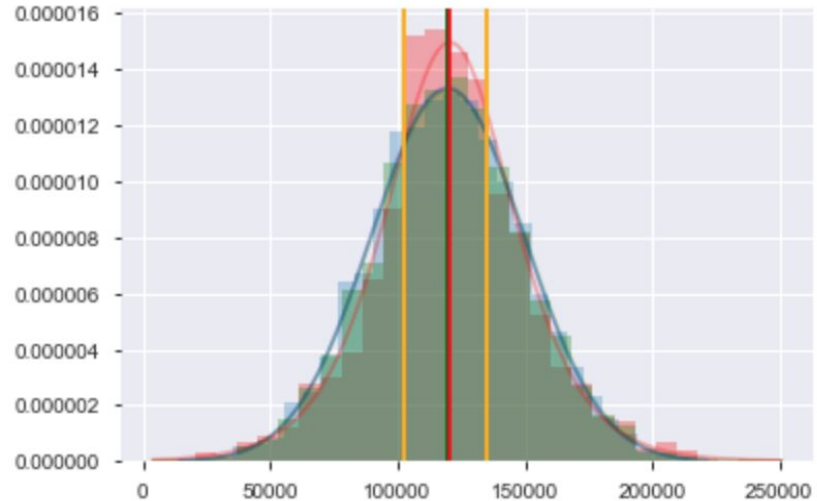
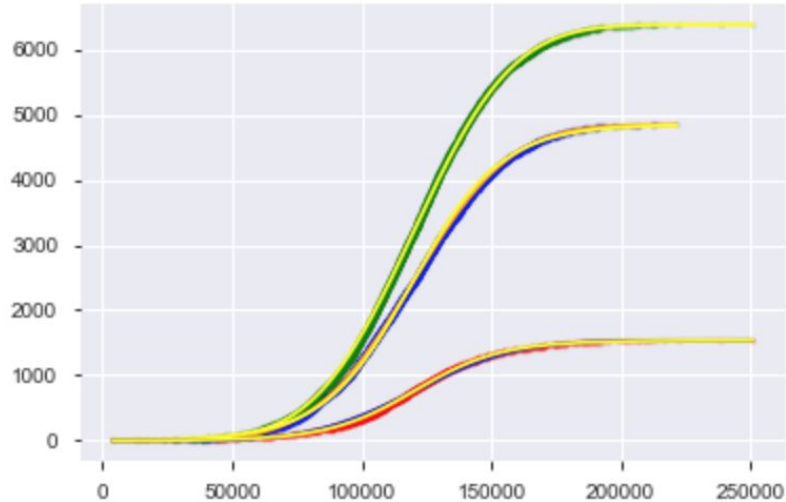
Initial Findings: Balance

- Balance of non zero balance exited and stayed customers have very close means and might all be normally distributed.
- Normalized histogram shows that within the balance range marked by the yellow lines customers tend to exit. (around [25, 70] percentile of exited group)



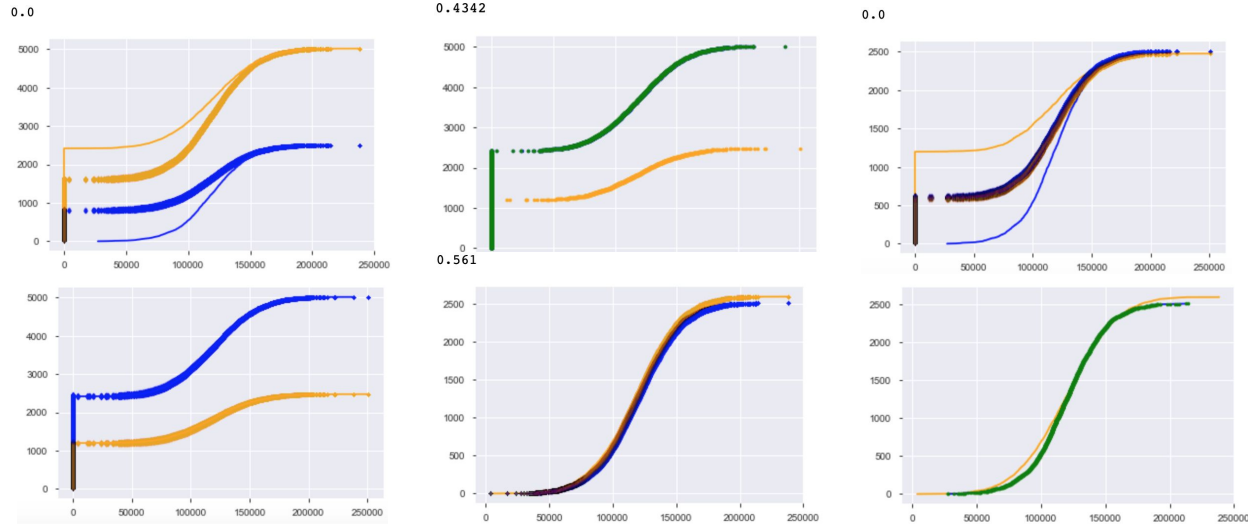
Initial Findings: Balance

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.



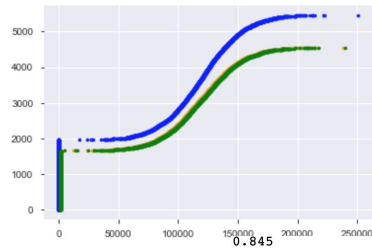
Summary of Findings

- FR and SP are more similar in balance distribution, and FR's non zero balance distribution is similar to GR.
- FR and SP could have the same mean and they are likely two parts of a same distribution.
(bootstrap test (n = 10000) p value of 0.4342)
- it is possible non zero balance FR and GR have the same mean and distribution.(permutation test p value 0.561, bootstrap test p value 0.4117)
- SP and GR shows that the two countries are not likely from the same distribution.

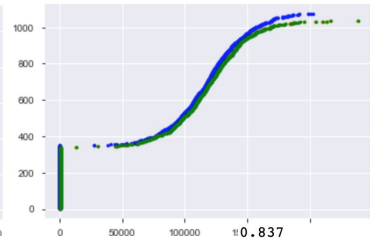


- It's possible age >45 female and male groups have the same balance mean(bootstrap sampling (n = 10000) p value 0.3859). two groups are from the same distribution it's possible they have the same means (p value 0.397).
- No significant difference in female and male members mean was found, and they are most likely from the same distribution. (p value for bootstrap test is 0.115, and 0.118 for permutation test)
-
- Active and inactive groups are likely to have the same mean (bootstrap p value 0.845), and are likely to have the same mean and distribution(p value 0.837).
- The non zero balance members of two groups are very likely of the same mean (p value 0.9492).

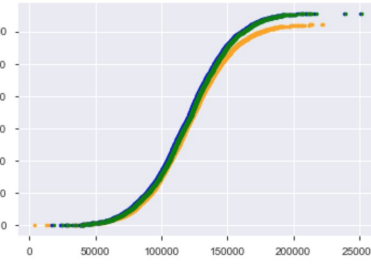
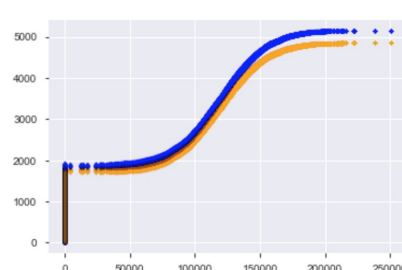
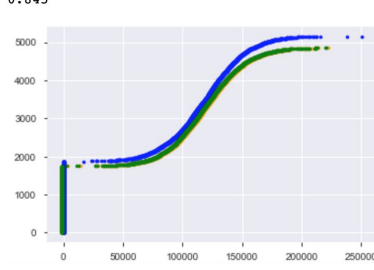
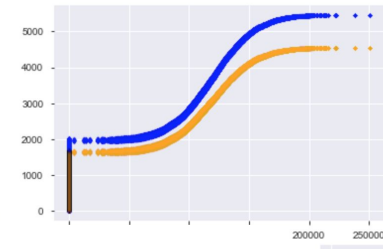
0.115



0.118

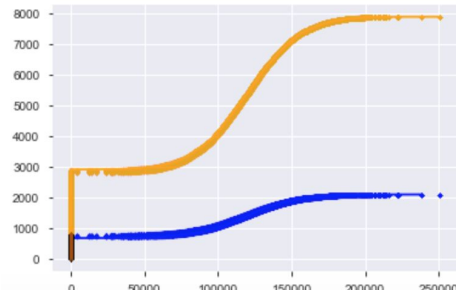


0.837

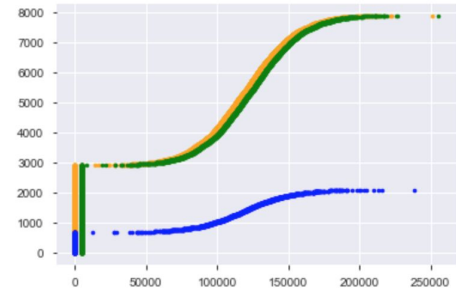


- Permutation ($n = 1000$) shows that two age groups divided by 45 are two parts of the same distribution, a indicates that two groups are not likely to have the same mean (bootstrap sampling p value 0.001, permutation test p value 0.0008).
- By performing bootstrap sampling and regressions ($n=1000$) around [30, 70] percentile of balance of exited group is where customers tend to leave.
- 95% confidence intervals for means and stds of the groups ($n=10000$):

0.001



0.0008

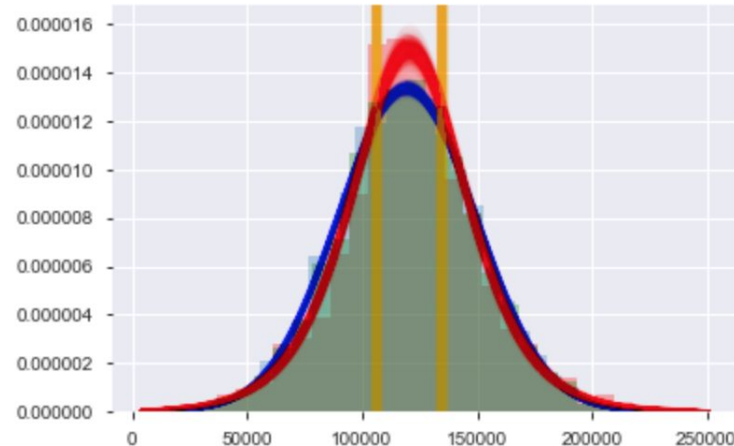


Non Zero Balance Exited Group Mean: [119204.61233393, 122283.12203969]

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]



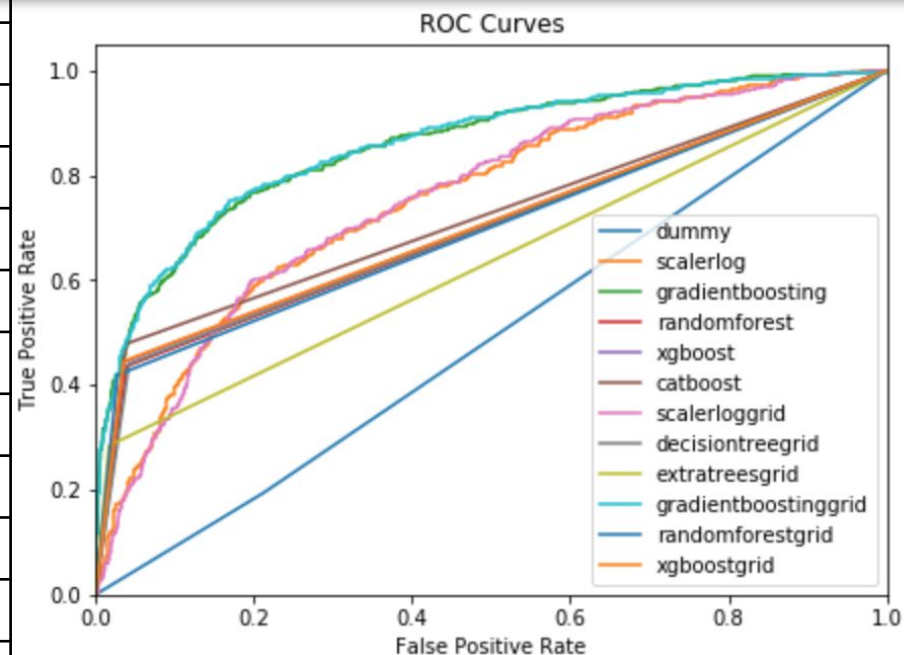
- In conclusion, customers of a certain range of balance tend to have a higher exit rate;
- Zero balance customers only show different distribution when purchasing products
- Most customers tend to purchase 1-2 products, those who have 3-4 products tend to leave though two groups (1-2 and 3-4 products) 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
- Germany and France without zero balance members can have the same mean and distribution.

In-Depth Analysis

- Split raining data (80%) and testing data (20%)
- Dummy classifier as baseline
- logistic regression with/without chi square select k best/standard scaler, decision tree, extra tree, extra trees, gradient boosting, random forest, svc, catboost, xgboost
- Validation: 5 fold
- Metrics: accuracy score, roc auc score
- Feature engineering
- Hyperparameter tuning

Prediction performance

| 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 |



- Gradient boost with tuned hyperparameters seems to be the best model for this dataset
- Catboost, xgboost, logistic regression with standard scaler are also good choices
- All model performed better than dummy model
- Improvement: feature engineering (new features), new classifiers, combine classifiers