Honouring the policy of KBC, I felt the quickest way to do the analysis of the data to find the variables and important models is through:

* Make Three different jupyter notebooks for each product
* Join the dataset using sales and revenue column for each three categories
* Separate the datasets into training and test datasets using the client id from sales revenue
* look for any correlation and relationship between variables
* Check for missing values
* Look for outliers
* Select the important variables of the three datasets.
* Try different algorithms to check which gives the most accuracy.
* Use the scripts from past projects.

Data analysis was done using Python’s library: Numpy, Pandas, Statsmodel and Scikit Learn.

**Analysis of Soc-dem**

We have 1615 values with Sex as categorical variable. There is no duplicate client id. We can get an idea of a possible skew in the data by comparing the mean to the median, i.e. the 50% figure. It looks like there is no skewness. The need to explore further observing from above:

* Sex has three missing values. Filled these using mode
* Age has a minimum of zero (looks like parents are opening accounts for their children)
* maximum of 97?? (It could be because the people are living longer in Ireland)
* Tenure has some values as 0 ( newly opened accounts)

**Analysis of Products\_ActBalance**

There are 1615 distinct values for products. There are a lot of missing values. We can assume the nan is because of the clients not having certain type of bank product. The following are all categorical variables.

* Count\_CA
* Count\_SA
* Count\_MF
* Count\_OVD
* Count\_CC
* Count\_CL

There is skewness in the following accounts.

* ActBal\_CA
* ActBal\_SA
* ActBal\_MF
* ActBal\_OVD
* ActBal\_CC
* ActBal\_CL

**Analysis of Inflow\_Outlow**

There are 1587 unique client values. We need to keep in mind if the missing values are from either the training or test set when compared to the sales\_revenue. There are no missing values.

There are 18 client's id missing from training set, and 10 from test set.

There is skewness in the following columns:

* VolumeCred
* VolumeCred\_CA
* VolumeDeb
* VolumeDeb\_CA
* VolumeDebCash\_Card
* VolumeDebCashless\_Card
* VolumeDeb\_PaymentOrder

**Analysis of Sales\_Revenues**

There are 969 values which is 60 percent of 1615 clients. There are three different marketing products :

* Mutual fund
* Credit Card
* Consumer Loan

The following are target variables for classification:

* Sale\_MF
* Sale\_CC
* Sale\_CL

The following are target variables for regression:

* Revenue\_MF
* Revenue\_CC
* Revenue\_CL

The maxima of the values are hugely out of proportion to the rest. The standard practice in these cases is to limit the upper limit using standardization. We will do it when we are doing modelling.

**Due to time constraint, it will be important to analyse a big dataset and do feature selection.**

The datasets are joined using LEFT JOIN. It makes more sense as it returns all records on the left hand side table whether it has not match on the other table.

**Continuous Variables**

* Age
* Tenure
* ActBal\_CA
* ActBal\_SA
* ActBal\_MF
* ActBal\_OVD
* ActBal\_CC
* ActBal\_CL
* VolumeCred
* VolumeCred\_CA
* VolumeDeb\_CA
* VolumeDebCash\_Card
* VolumeDebCashless\_Card
* VolumeDeb\_PaymentOrder
* Revenue\_MF
* Revenue\_CC
* Revenue\_CL

**Categorical Variables**

* Sex
* Count\_CA
* Count\_SA
* Count\_MF
* Count\_OVD
* Count\_CC
* Count\_CL
* TransactionsCred
* TransactionsCred\_CA
* TransactionsDeb
* TransactionsDeb\_CA
* TransactionsDebCash\_Card
* TransactionsDebCashless\_Card
* TransactionsDeb\_PaymentOrder

**Few Observations after visualization of data**

Tenure is not correlated with age. Age affect Revenue\_CL till 60 and hen falls. Outlier at 10-12. Most active between 20 and 45. For Revenue\_CC , there is an outlier at 70 and most of revenue comes from age group of 30-40. Revenue\_MF growa with age till 50 years old at 80,000 . There is an outlier of 225,000.

ActBal\_MF grows with age and max at 20-30. Same ActBal\_SA grows with age with max at 60 while ActBal\_CA is at 50. Sale for all three products is correlated with the revenue.

Revenue\_MF goes down as tenure increases with 20 months as most. Revenue\_CC consistently increases with tenure while there is no relationship between Revenue\_CL and tenure. ActBal\_CA increases Revenue\_CL and is linear with Revenue\_MF. ActBal\_SA and ActBal\_MF , both help in increasing Revenue\_CL and Revenue\_MF but in smaller amounts.

ActBal\_SA increase but is not very linear.

There are outliers in most of the columns. It is due to errors or due to scattered human behaviour is not clear through data.

There are atleast 4 clear outliers for VolumeDebCashless\_Card and VolumeDeb\_PaymentOrder. There is a linear relationship between VolumeDeb and VolumeDeb\_CA. VolumeDeb\_CA has also linear relationship with VolumeCred. ActBal\_CC is linear with ActBal\_OVD and ActBal\_CC with ActBal\_CL.

To detect outliers, median absolute deviation function is used with threshold of 3.5.

There were many values missing such as:

Count\_SA has 725 missing', 'Count\_MF has 777 missing', 'Count\_OVD has 719 missing', 'Count\_CC has 864 missing', 'Count\_CL has 884 missing', 'ActBal\_SA has 725 missing', 'ActBal\_MF has 777 missing', 'ActBal\_OVD has 719 missing', 'ActBal\_CC has 864 missing', 'ActBal\_CL has 884 missing', 'VolumeCred has 18 missing', 'VolumeCred\_CA has 18 missing', 'TransactionsCred has 18 missing', 'TransactionsCred\_CA has 18 missing', 'VolumeDeb has 18 missing', 'VolumeDeb\_CA has 18 missing', 'VolumeDebCash\_Card has 18 missing', 'VolumeDebCashless\_Card has 18 missing', 'VolumeDeb\_PaymentOrder has 18 missing', 'TransactionsDeb has 18 missing', 'TransactionsDeb\_CA has 18 missing', 'TransactionsDebCash\_Card has 18 missing', 'TransactionsDebCashless\_Card has 18 missing', 'TransactionsDeb\_PaymentOrder has 18 missing

**Handle Missing Data for continuous data**

* If any column contains more than 50 entries of missing data, drop the column
* If any column contains fewer that 50 entries of missing data, replace those missing values with the median for that column (the median imputation used on missing values is very crude.)
* Remove outliers using Median Absolute Deviation
* Calculate skewness for each variable and if greater than 0.75 transform it
* Apply the sklearn.Normalizer to each column

**Handle Missing Data for Categorical Data**

* If any column contains more than 50 entries of missing data, drop the column
* If any column contains fewer than 50 entries of missing data, replace those values with the 0

**Feature Selection was done using:**

* Variance Threshold
* Random Forest Classifier
* Recursive Feature Elimination

As some of the features are negative, I skipped Univariate feature selection using chi2chi2 test.

**Algorithms**

The best algorithms working for the classifier are Logistic Regression, Support Vector Classifier and Ensemble.

Logistic regression is more robust as the independent variables do not have to be normally distributed or have equal variance in the group. It may also handle non-linear effects. It does not require independents be interval or unbounded.

SVC maximizes margins and model is slightly robust for even non-linear relations. It has a regularisation parameter and helps in avoiding over fitting.

Neural Networks do not work because of the imbalanced dataset.

**Ensembles**

**Bagging (Bootstrap Aggregation)**

Involves taking multiple samples from the training dataset (with replacement) and training a model for each sample. The final output prediction is averaged across the predictions of all of the sub-models.

**Bagged Decision Trees**

Bagging performs best with algorithms that have high variance (i.e. decision trees without pruning).

**Random Forest**

An extension to bagged decision trees. Samples of the training dataset are taken with replacement, but the trees are constructed in a way that reduces the correlation between individual classifiers. Also the tree size is much slow due to max\_features.

Ensemble methods are good because they average out the biases, reduce the variance and unlikely to overfit.

**After Hyper parameter tuning, Logistic Regression, Linear Discriminant Analysis and Random Forest and creating a voting Ensemble gives an accuracy of 70% for Sales\_CL , 75% for Sales\_CC and 80% for Sales\_MF model.**