Technical Report

This report contains the technical background employed to maximise the revenue from marketing of three KBC products, as a part of the case study.

A. Data Cleaning:

1. The three tables, Soc_Dem, Products_ActBalance and Inflow_Outflow are joined on the 'Client' column containing the client ids.
2. There are a total of 13633 missing values in the resulting table. The missing values in the 'Sex' column are imputed with 'M' since missing values are few in number and the number of Male customers are more. The missing values in the Counts and Account balances of different products owned by the clients are filled with 0s since the missing values are a result of a particular client not possessing that particular product.
3. There are some outliers in the 'Age' and 'Tenure' columns. The 'Age' column is filtered such that it only contains clients between the age 18 and 100. Further, the data is filtered based on the difference between the Age and Tenure of a client. Any client with this difference less than 18 is removed. A new column is created by dividing the Age column into 4 groups, only for the sake of visualisation.
4. Further outliers in other columns, such as 'ActBal_SA', 'ActBal_CA', 'ActBal_CA', 'ActBal_MF', 'VolumeCred' and 'VolumeDeb' are removed based on the box plots of features. All values in the mentioned columns greater than 60000 are removed.

B. Data Exploration - Features only

1. Number of Male clients are higher than number of female clients.
2. The Age and Tenure distribution of Male and Female clients is more or less the same. Most of the clients are between the age group 30-50 years and the tenure is between 0-100 months. Also there are a high number of clients with tenure between 150-200 months.
3. The CA Balance of most clients is in the range 0-3000 €, SA Balance is between 0-2000€ and MF Balance is between 0-1000€.
4. Most clients have OVD, CC and CL balances in the range 0-1000€.
5. Most clients have their assets in the range 0-20000€ with people in the age group 18-40 years having highest number of assets and 80-100 years with lowest asset. Most clients have liabilities in the range 0-2000€ again with age group 18-40 years having largest liabilities and 80-100 having the lowest.

6. Most clients have credit transactions in the range 0-10000€ with age group 18-40 years having the highest credits. Also clients have debits within the range 0-10000€ with age group 18-40 years.
7. Clients in the age group 18-40 years have the highest number of transactions in the range 0-75.

C Feature Engineering

1. Correlation between the features was checked using Pairplots in seaborn and the corresponding Pearsons correlation coefficient. The Products owned and their balances are checked together for correlation and the transaction counts and the volumes are separately checked.
2. TransactionsDeb is correlated with TransactionsDeb_CA and TransactionsDebCashless_Card. Two of them can be removed.
3. VolumeDeb and VolumeDeb_CA are correlated. One of them can be removed
4. VolumeDeb and VolumeDeb_PaymentOrder are correlated. One of them can be removed.
5. VolumeDebCashless_Card and TransactionsDebCashless_Card are correlated. One of them can be removed.
6. The overall observation is that the number of debit transactions is proportional to volume of debit transactions only for cashless cards.
7. The complete debit volume mostly consists of VolumeDeb_CA, VolumeDebPaymentOrder since they have higher correlation with it. Most number of debit transactions consist of TransactionsDeb_CA and TransactionsDebCashlessCard.
8. The initial assumption that the more number of products a client owns, the more debit transactions they will have is proved wrong since none of the attributes related to products owned and the debit volumes.
9. The initial assumption that the larger credit volumes for client means that they have higher spending potentials is cross checked by observing correlation between volume of credit transactions to that of debit transactions. The actual observation is that except for Cash and Cashless debit volumes, the correlation of the rest of the attributes show that the assumption is correct.
10. After removing the features as mentioned above, a heatmap is used to check for any residual correlations.
11. Eventually, 27 features are left after within-feature correlation analysis.

D. Feature-Target correlation analysis

1. Number of sales and no sales for MF, CC and CL are checked and an imbalance

is found with no sales being majority class in all cases.
2. The Sale counts are checked for all the three products and the sales of CL is highest, followed by CC and MF.
3. In the case of CL, there are 25% clients with a sale and 75% with no sales
4. The mean balances of clients with a CL sale are checked and below observations are made;
	a. Average Tenure of clients who took CL is higher, hence loyal customers are more likely to respond to these offers.
	b. Average MF balance of clients who purchased CL is higher, probably because bank offers loan over MF balance.
	c. Average debit transactions of clients who purchased CL is higher, hence people with high spendings are likely to respond to CL offers.
	d. Average age of clients who took loan is less and probably in the age group of 18-50.
5. Number of Sales of MF is 20% against 80% with no sales of MF.
6. The mean balances of clients with MF sale are checked and below observations are made;
	a. Clients who purchased MF have a lesser average Age and Tenure.
	b. Clients having higher average asset value are more likely to purchase MF.
	c. Clients with higher liabilities are unlikely to purchase MF.
	d. Clients with higher credit transactions, and probably higher income, are more likely to invest in MF.
7. Number of Sales of CC is 25% against 75% with no sales of CC.
8. The mean balances of clients with CC sale are checked and below observations are made;
	a. Clients with a higher average Age and Tenure are more likely to purchase CC.
	b. Clients whose average overdrafts are more likely to purchase CC.
	c. Clients who already have a CC, are less likely to purchase CC.
	d. Clients with higher average asset values are likely to buy CC.
	e. Clients with higher liabilities are less unlikely to buy a CC.
	f. Clients with higher credit transactions are more likely to purchase CC.
9. Eventually, the correlation of all features with the categorical target variable, again using Pearsons coefficient since there are only two categorical values.

E. Propensity Modelling

1. The class imbalance is handled using SMOTE so that the count of datapoint eventually becomes 50%-50% after splitting the dataset into train and test with 20% validation data. The resulting train dataset is shuffled to avoid any bias.

2. Logistic regression is chosen as the model since it predicts probabilities for a given datapoint belonging to a given class using Maximum-likelihood estimation to estimate the coefficients for input features. The coefficients are estimated such that the predicted probabilities have minimum error.
3. A small Grid search experiment is carried out to tune the hyper parameters like 'solver' and 'C' with an L2 penalty and class weights are balanced.
4. The results are visualised through a confusion matrix and a precision recall curve, since ROC is not a good metric for class imbalance.
5. The propensities are given using the predict_proab function of the Logistic regression model. The propensities for the entire training and validation data are predicted as well as for the 40% data for which there are no labels available.
6. Achieve accuracies are as follows;
   a. CL model: 68%, b. MF model: 84% c. CC model: 67%

F. Revenue model

1. A Linear regression model is used to predict revenues for the datapoint with no labels.
2. Same data cleaning steps are carried out as the ones in Propensity modelling. Below results are obtained;
   a. MSE of CL revenue prediction model is 32.8 b. MSE of MF revenue prediction model : 17.7 c. MSE of CC revenue prediction model : 32.28