DATA SCIENTIST CANDIDATE

REPORT ON ANALYSIS

Eugene Ambe Ndamukong

January 20, 2020

1 ANALYTICAL DATASET

The original dataset which was received contained 4 excel sheets : "Soc_Dem" "Products_Actbalance", "Inflow_Outflow" and "Sales_Revenues". So, these sheets were merged into a single dataset based on clients with available sales & revenue information. This dataset resulted in 969 clients along with 31 features and 6 targets. The client IDs in the "Sales_Revenues" sheet was used in merging all these excel sheets into a single dataset. This data was saved as **Analytical_data.csv**. The display of the first few observations are found in the **Analysis for KBC.ipynb**.

2 ANALYSIS OF THE CREATED DATASET

Documented python codes along with output can be found in the **Analysis for KBC.ipynb** file.

2.1 PRE-PROCESSING

Prior to creating the propensity models, certain features were pre-processed because of irregularities in the data such as missing data.

The first thing noted as abnormal from the data was that some clients had an age of zero. Logically, it is impossible to have an individual with zero age. This case was handled as missing information and their respective zero ages were imputed with the mean age of all the clients.

The second observation noted was missing data. 12 features were found to have missing data which are: *number of live current accounts, number of live saving accounts, number of live mutual accounts, number of live overdrafts, number credit cards, number of live consumer loans, actual current accounts balance, actual saving accounts balance, actual mutual fund balance, actual overdrafts balance, actual credit cards balance, actual consumer loans balance* and *sex*. The missing data in the quantitative features were imputed with their respective means. Also, the missing data in the qualitative feature, *Sex*, was imputed by the mode category(Males).

The third observation noted was outliers. Outliers were identified using boxplots for each feature except *Tenure, Age, number of live credit cards,number of live overdraft balance*. A truncation function was created, ***truncate***, which imputes the identified outliers with the upper limit value of the distribution if the outlier was above it or lower limit value if the outlier was below it. Treatment of the outliers was successful as the boxplot of the truncated features showed insignificant number of outliers left.

Lastly, it was observed that certain features had very low variance or information. These features had a variance of zero so they could not be used in building the propensity models. The features with zero variance happen to be : *number of live current accounts, number of live saving accounts, number of live mutual accounts, number of live overdrafts, number credit cards, number of live consumer loans, actual saving account balance, actual mutual fund balance, actual overdrafts balance, actual credit cards balance, actual consumer loans balance.* A function, ***feature_sel***, to perform this feature selection was created.

2.2 Propensity models

Prior to building the models, the number clients with and without ownership of either of the products (consumer loan, credit card and mutual fund) was noted. The data shows huge class imbalance i.e much larger clients with a product than without a product. Building the propensity model with this type of distribution will result in it always predicting the majority class (0) even if it is false. Therefore, synthetic minority over-sampling (SMOTE) was applied on the train dataset to ensure that the number clients were same in each class. Because there were already defined classes of clients (o & 1), supervised learning methods were used and their performances can be seen in table 2.1.

Models	Consumer loan	Credit card	Mutual fund
Logistic model	0.541	0.560	**0.586**
Random forest classifier	**0.584**	0.530	0.516
7-nearest neighbour classifier	0.479	0.578	0.536
Ensemble classifier	0.533	**0.611**	0.481

Table 2.1: Performance of propensity models

Validation set approach was used as the sampling technique to build the models where 80% of the data was used to train the model and 20% used to evaluate the model. Models were built for each product i.e consumer loan, credit card and mutual fund. The models built for each product were logistic regression, random forest classifier and ensemble classifier (combination of logistic regression, random forest classifier and K-nearest neighbour classifier). The AUC values for each product can be found in table 2.1 with those of the best model in bold.

3 Target clients

The models with highlighted AUC values (table 2.1) were used to predict the probability to own the respective product i.e Random forest classifier for consumer loan; ensemble classifier for credit card and logistic model for mutual fund. After that, the clients with >60% chance of owning or buying a respective product were obtained. These clients along with their respective revenues and IDs per product can be found in **contact_clients.csv** revenues. The clients have been ordered from clients with the highest propensity to lowest propensity based on their probability to own a product. The summary of the selected clients can be found in the table 3.1. These were obtained on the test set.

	Consumer loan	Credit card	Mutual fund	Total
Number of clients	14	22	42	**78**
Sum of revenue	89.49	50.213	152.593	**292.296**

Table 3.1: summary of highly propensity clients