P a g e | **1**

**Data Science Project**

**Nushura Islam**

**Contents**

|  |  |
| --- | --- |
| **Business Understanding Data Understanding Data Preparation Modelling**  **Evaluation Best practices**  **References** | **03**  **04**  **11**  **16**  **17**  **30**  **31** |

# Business Understanding

The personal loan campaign, introduced last year, attracted fifteen per cent of Crédit Nationale Azur’s customers. No proper method was used to target appropriate clients, so a lower percentage of individuals got involved with the campaign and took loans. There are always costs associated with launching any campaign, and in addition, if it is unsuccessful, then in return, the price rises even more.

Therefore, the business problem that needs to be addressed is to lower the cost and increase the success rate of campaigns. As Crédit Nationale Azur intends to relaunch a similar campaign and thus, to solve the business issue, customer data from the previous campaign needs to be analysed. The data contains specific characteristics of clients, which consists of customer id, age, years of work experience, family size, education level, income level, mortgage amount if they have a credit card account, monthly credit card spending if they have a shared trading account, fixed deposit account, or online banking account and lastly, the result of if they took a personal loan during the last campaign.

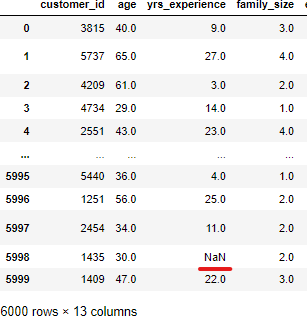
Hence to address the problem at hand, an appropriate machine learning algorithm will be used to create a model for predicting which customers are likely to accept personal loan offers. Analysing the data set from the previous campaign will assist in understanding the characteristics of the clients who are most likely to take personal loans. Thus, those individuals can be targeted for the campaign. Furthermore, having the information regarding the number of clients most likely to take loans will assist Crédit Nationale Azur in preparing and making a cost estimation required for the campaign. This will prevent the bank from making additional spending. Hence, Crédit Nationale Azur can efficiently utilise its cost and have a better personal loan campaign success percentage.

# Data Understanding

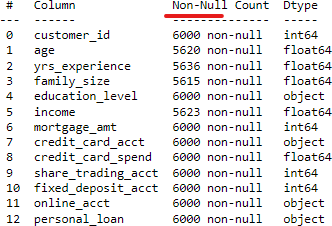
Data Issues and Solution

When loaded the dataset initially, missing values were present. Missing values can be considered when there is a presence of any of these characters within the dataset, '?', '--', ' ', 'NA', 'N/A', '-'. Hence, to address the missing value issue, a variable called ‘missing\_values’ was created where all the characters mentioned were stored in an array. When the dataset was being loaded through the ‘read\_csv’ method, the ‘missing\_values’ variable was passed

within the ‘read\_csv’. The action will go to all those places with missing values and label them as ‘NaN’(Not a Number). This is done to keep the dataset consistent regarding missing values and assist in further data analysis.



A data information table was also generated. From there, it could be observed that most attributes have six-thousand records. However, few features have less than six-thousand records.



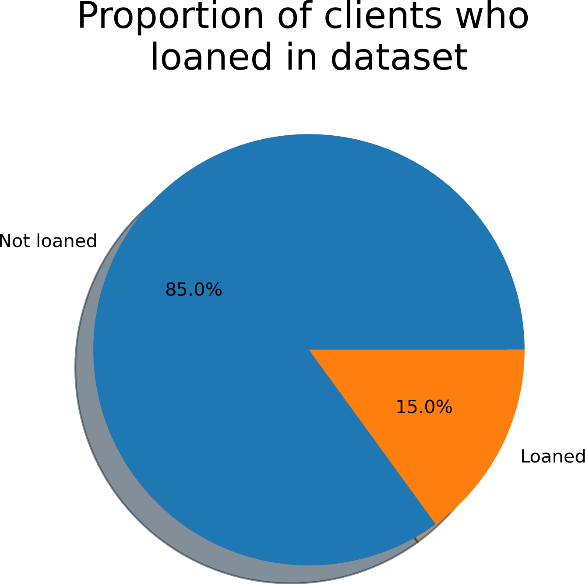
This proved that there are some missing values too. This issue is further addressed in the data cleaning please of the data analysis.

Data Exploration

A variety of graphs and chart were used to explore the data to identify relationships between the attributes and the personal\_loan target variable. The graphs and charts which were used are heatmap, pie chart, histogram, bar-graph and scatter-plots .

Pie-chart

To begin with, a pie-chart was used to visually demonstrate the portion of people within the dataset who took a loan and who did not. The percentage of people who took a loan is fifteen percent. However, the vast majority of clients (85%) did not take a personal loan during the last campaign.



Heatmap

The heatmap did not include the columns which were strings. Hence, the personal loan attribute was not present within the heatmap. After analysing the heatmap, it can be concluded that the attributes which have the strongest correlation with each other are:

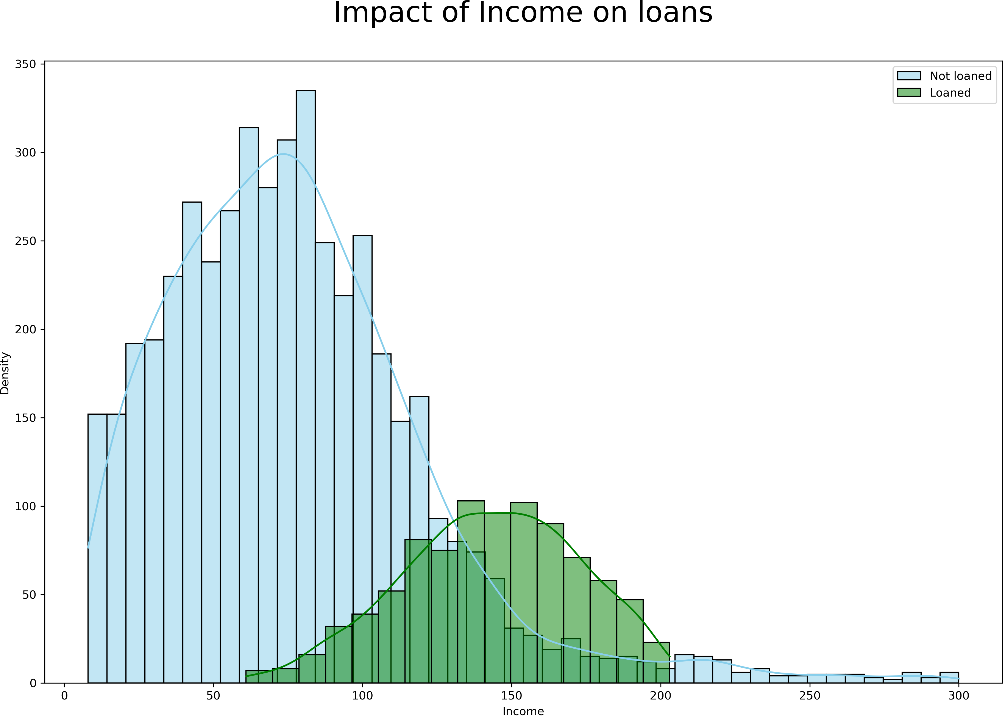
* Income and mortgage\_amt: 0.13
* Income and credit\_card\_spend: 0.15
* Income and fixed\_deposit\_acct: 0.19
* mortgage\_amt and fixed\_deposit\_acct: 0.1

These attributes, income, mortgage\_amt, credit\_card\_spend, and fixed\_deposit\_acct play a vital role in determining the probability of a particular individual taking a personal loan. For example, if an individual earns above a certain amount, they are most likely to take a loan as clients consider their chances of paying back before taking loans. Credit card spending will also increase the chances of someone taking a loan. This is because if their monthly

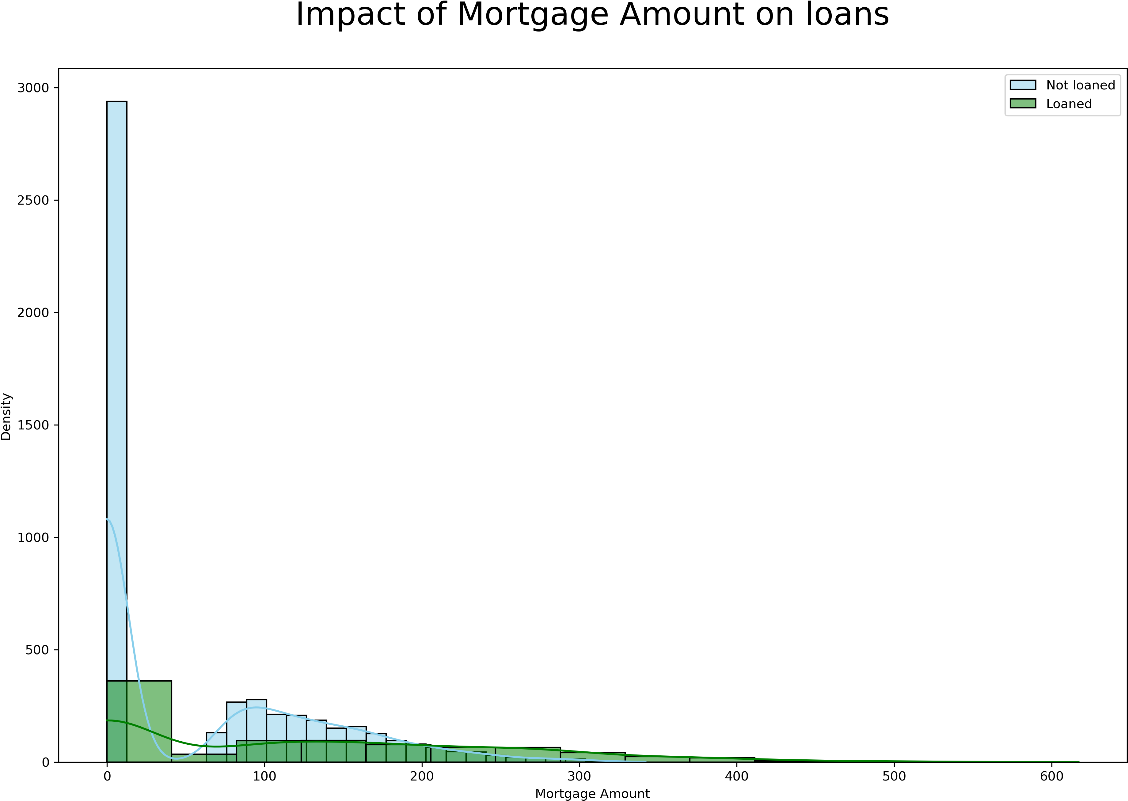
expenditure is above a specific limit, then they will not lean towards taking any loans. People with fixed deposit accounts are likely to save money and hence can take a loan.

Histograms

The histograms were used to analyse the impact of the strongest correlation attributes which were a continuous variable. (Income and Mortgage Amount)



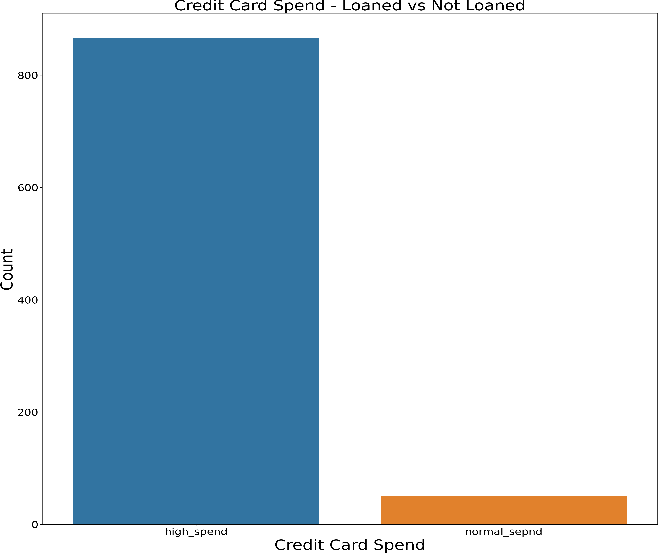
The horizontal x-axis represents ‘Income’ and the vertical y-axis represents the ‘Density’ of income. The impact of income on loans is visibly strong. Both groups, loaned (green) and not loaned (blue) show a normal distribution. However, there was a uniform distribution at the peak of loaned group. There is a noticeable difference in the distribution of income level in not loaned versus the loaned groups. The income of the clients who loaned are highest between 130 to 170, however, the income of clients who did not take a loan are less then the people who loaned. This justifies that income had a vital role in terms of Crédit Nationale Azur client’s taking a loan. Hence, income level would be a good variable to predict if a certain individual will take a personal loan or not.



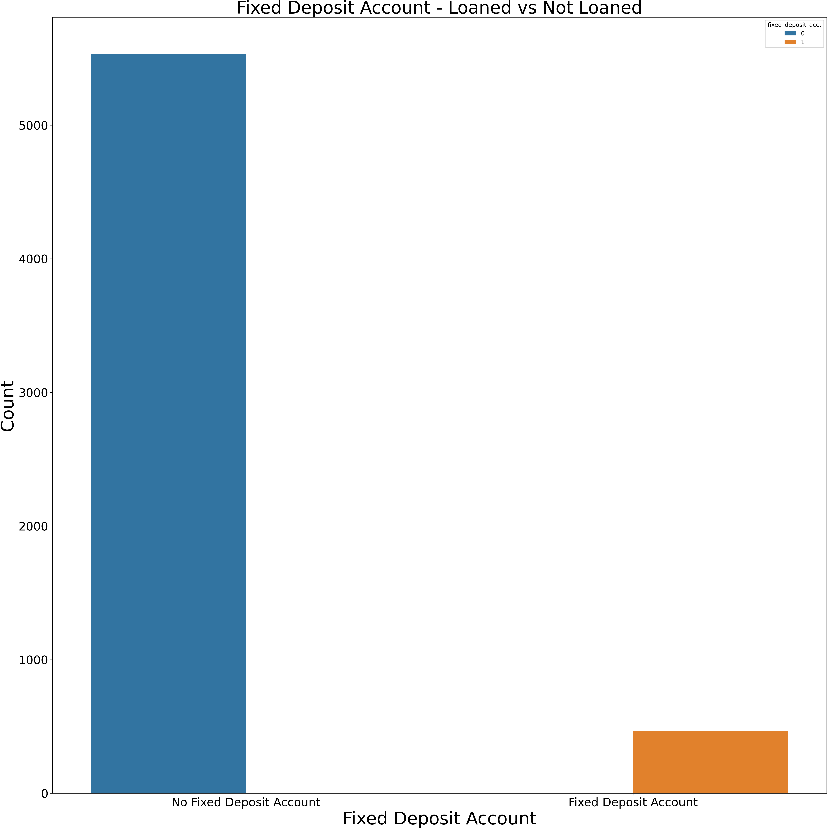
The horizontal x-axis represents ‘Mortgage amount and the vertical y-axis represents the ‘Density’ of mortgage amount. The impact of mortgage amount on loans is unclear from the histogram. However, it can be concluded that the clients who took a loan has a greater mortgage amount and the client’s who did not take a loan has lower mortgage amount.

Bar-graphs

The Bar-graphs were used to analyse the impact of the strongest correlation attributes which were a discrete variable. (Credit card spend and fixed deposit account)



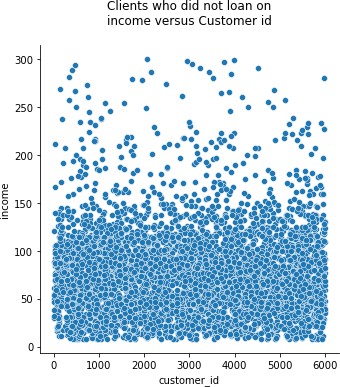
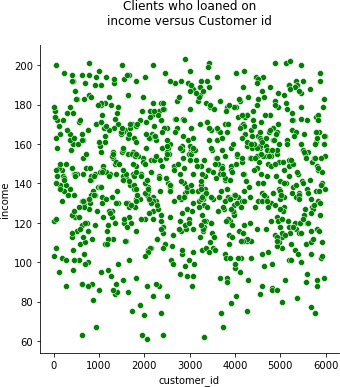
The horizontal x-axis represents ‘Credit Card Spend’ and the vertical y-axis represents the ‘Count’ of Credit Card Spend. The impact of credit card spend on loans is visibly strong. There is a noticeable difference in the distribution of credit card spend level in not loaned (blue) versus the loaned groups(orange). The count of credit card spend of the clients who loaned is below around 50, however, the credit card spend of clients who did not take a loan are above 800. This justifies that credit card spend had a vital role in terms of Crédit Nationale Azur client’s taking a loan because individuals who spend more did not take a personal loan but people who spend less took the loan. This can be because those people save more and thus, they consider that they will be able to pay back the loan. Hence, credit card spend level would be a good variable to predict if a certain individual will take a personal loan or not.



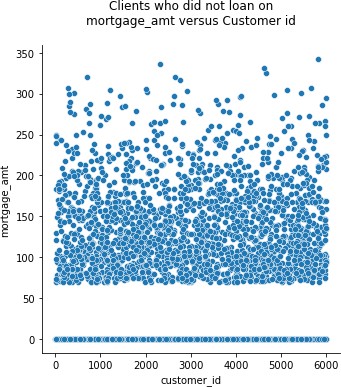
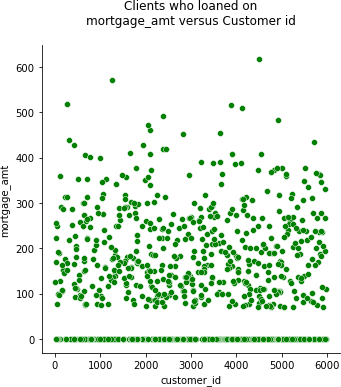
The horizontal x-axis represents ‘Fixed deposit account’ and the vertical y-axis represents the ‘Count’ of Fixed deposit account. The impact of fixed deposit account on loans is visibly strong. There is a noticeable difference in the distribution of fixed deposit account level in not loaned (blue) versus the loaned groups(orange). The count of fixed deposit account of the clients who loaned is below around 500, however, the fixed deposit account of clients who did not take a loan are above 5500. This justifies that having fixed deposit account had a vital role in terms of Crédit Nationale Azur client’s taking a loan because individuals who had a fixed deposit account meant that they were making an active effort to save money whereas others did not. Hence, fixed deposit account level would be a good variable to predict if a certain individual will take a personal loan or not.

Scatter-plots

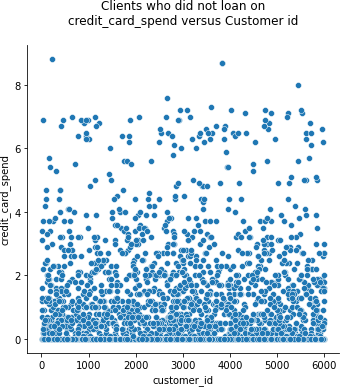
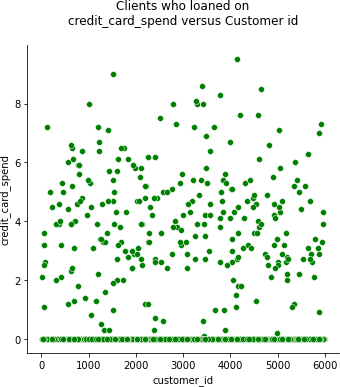
Income versus customer\_id: It can be observed that a high percentage of customers with income below 130 did not take loan and income above 130 took loans. Thus, emphasising on income as a good features to indicate who will take a personal loan.



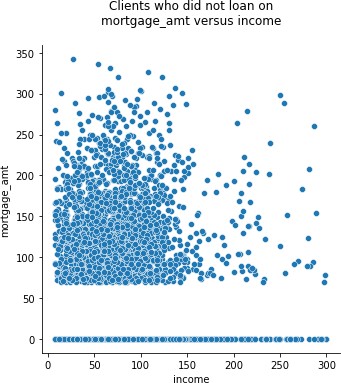
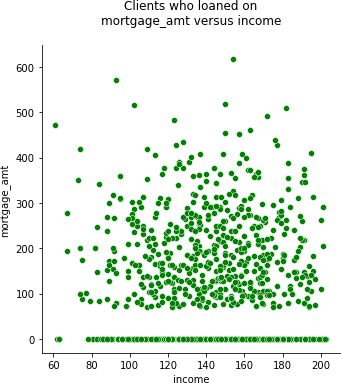
Mortgage amount versus customer\_id: By observation clients with more count of mortgage amount did not take the loan and clients with less number of mortgage amount took loan. Thus, a good feature to determine who will take a loan.



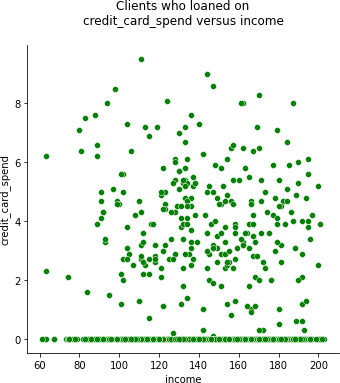
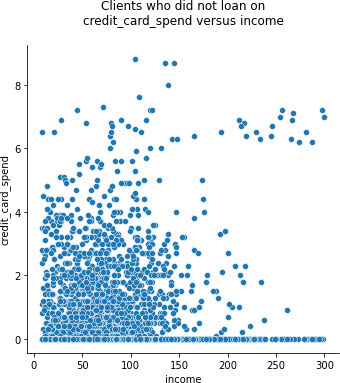
Credit card spend versus customer\_id: This proved again that it is a good measure for detecting if a client will take a personal loan or not. The plots demonstrated that the more a client spends, the less likely they are to take a personal loan and vise versa.



Mortage ammount versus income: It can be seen that more the income, the more the count of mortgage amount and vise versa. Thus, the two attributes are a good measure in determising if a customer to take a loan or not.



Credit card spend versus income: People who earn less, spend less and thus, they will not be able to pay back the loan even if they do take it. However, people who earn more, spend more comparatively and thus, if they take a personal loan then they will be able to pay back the loan. Hense, these two features are a good differentiator between people who may take a loan.



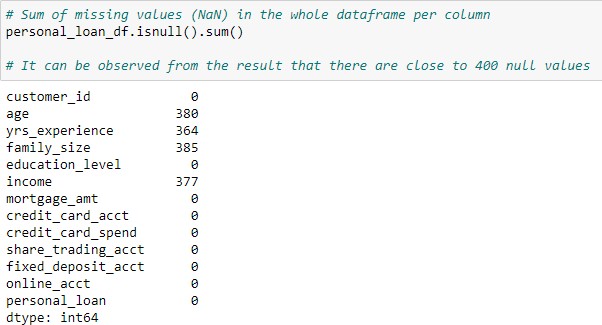
# Data Preparation

Duplicate rows

Data sets can contain duplicated rows. To make sure that there are no duplicated rows, drop\_duplicate() method was used. Before dropping the rows the size of the record was six thousand rows and thirteen columns. The result depicted that the number of rows and columns were same as before which means that there are no duplicated rows.

Missing values

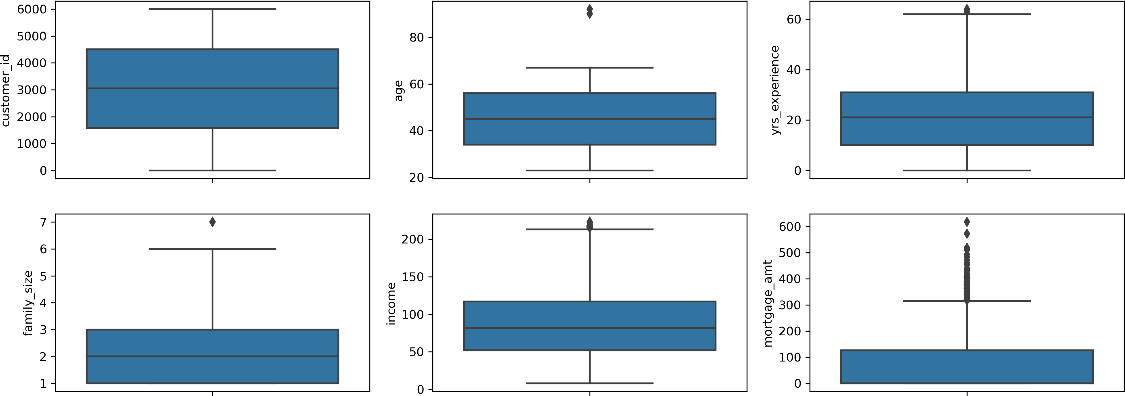
As previously mentioned, when loading the dataset, all the missing values were converted to NaN (not a number) to maintain consistency. To check the total number of NaN values per column this step was done in the data preparation step:



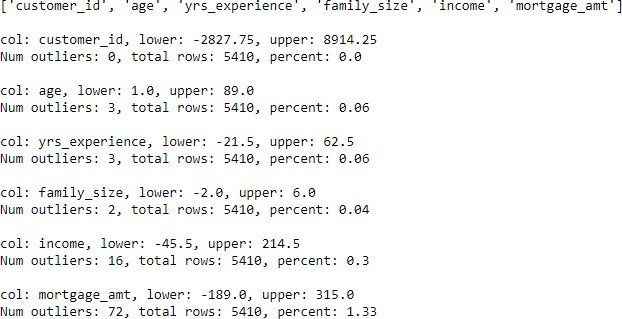
Isnull() method is addressing all the null/ NaN values within the dataframe called ‘personal\_loan\_df’. Then the sum() method is calculating the sum of NaN per column within the dataframe. It can be understood from the result that there were near 400 NaN values in four columns; age, yrs\_experience, family\_size, income. The percentage of missing data per columns is also calculated and it is found that there are six percent missing values present within those four columns. As the percentage of missing values is not high, all the rows where at least one NaN value is present, they are dropped. After the drop is done using the dropna() method, it can be observed that the number of rows drops from 6000 to 5410. This method of dropping rows is justified because the amount of rows dropped is not much and there is still a lot of data available to do the analysis on.

Outliers

All the numerical attributes were used to create box-plots to identify outliers. It is illustrated below.

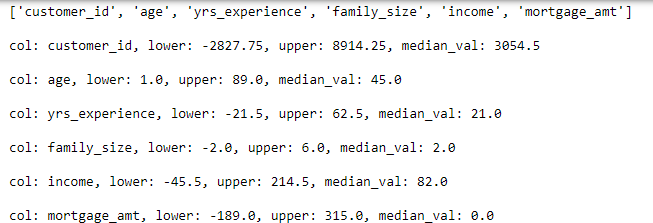


It can be observed that there are outliers present within all the plots expect for customer\_id. The exact percentage of outliers present within the five plots is calculated.



The column, customer\_id did not have any outliers. The highest number of outliers present are inside the credit\_card\_spend. The lowest number of outliers present are within family\_size.

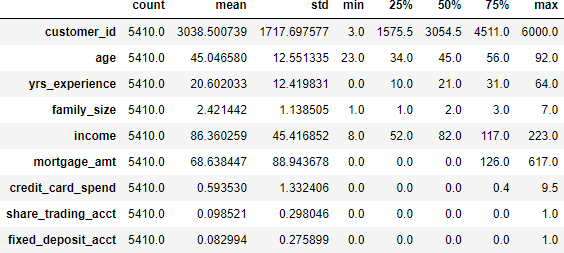
To address the outliers and remove them, the median was used to replace outliers.



By replacing the outliers with the median values it meant that all the values less than the

‘lower’ value and more than the ‘upper’ value in the screenshot above is replaced with the median. After the median replaced the outliers the maximum values of each attribute with outliers also changed. For example, the max value of age before the median replaced outliers was 92 which changed to 67 afterwards. It is illustrated below.

Statistics Before:



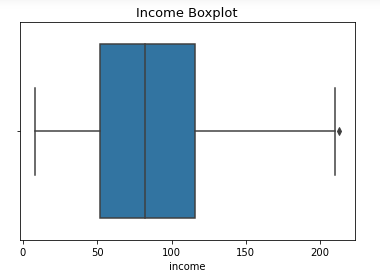
Statistic After:



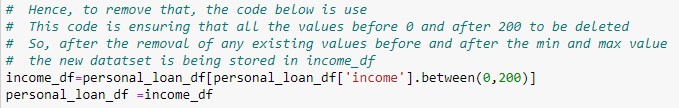
However, it could be observed that the income and mortgage\_amt columns had more outliers present even after replacing the outliers with median. More data cleaning was done to remove those outliers.

Income

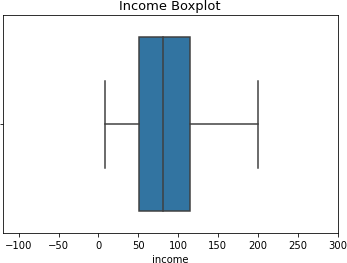
Initial box-plot (after being replaced with median):



It can be confirmed that there were not outliers before 0 and after 200 within income boxplot. Hence, to remove any more outliers before 0 and after 200, this method was implemented for both income and mortgage\_amt.



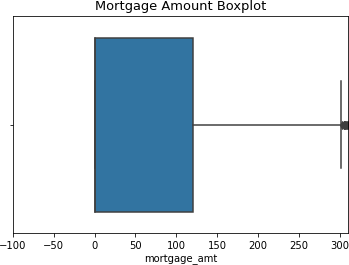
After implementing the above code, it could be observed that no more outliers were present within the income column. Hence, properly cleaned income attribute.



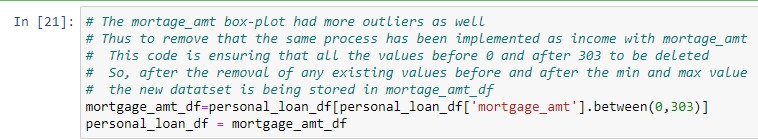
Mortgage\_amt

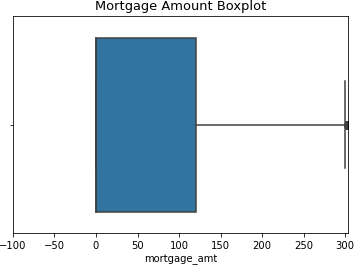
The same process was implemented for mortgage\_amt as well. However, it is observed that even after trying to remove the outliers with the code illustrated above, a new set of outliers appears. Hence, to achieve a completely cleaned mortgage\_amt set, the above method is repeated two times. The box plot for the two stages are demonstrated below.

Initially (after being replaced with median):



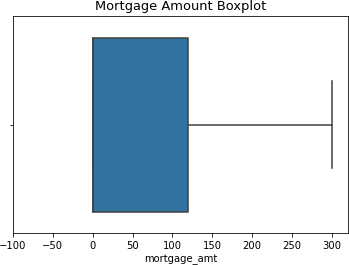
First time:





Second time:





In conclusion, the attributes used for data modelling are income, mortgage amount, credit card spending and fixed deposit account. The features are identified by plotting heatmaps, pie charts, histograms, bar charts and scatter plots. Furthermore, the data set is cleaned and organised.

In addition, for modelling, the data needs to go through data transformation, which will be done in the next step.

# Modelling

The classification models used were k-nearest neighbors (kNN) and decision tree (random forest):

Initial model (with all features)

As previously mentioned, before modelling, different kinds of transformation are done: scaling and discretising continuous variables, label encoding and one hot encoding.

* Scaling of continuous variables is used to scale attributes of continuous variable type. This type of transformation cannot be implemented on any other variable types.
* Discretization of continuous variables is used to transform continuous variables into discrete categorical variables.
* Label encoding is used for categorical columns that have two categories.
* OneHot encoding is used for categorical columns with more than two categories.

To measure the performance of each classification approach, confusion matrix, recall, precision and f1 scores were used. These methods were used to measure the performance matrix because the personal loan dataset is unbalanced. That is also why the accuracy score was not used to compute performance, or else the measurement would not be accurate with the accuracy score.

Improved model (with features selected)

This time, a few attributes were selected based on feature selection. The model was improved by tuning parameters and feature selection. All the other steps taken to build the model by transformation and computing performance matrix are the same as mentioned above.

# Evaluation

In general, the building of initial and improved model, hyper parameter tuning and feature selection, all began with three constant stages. The stages were import of libraries, loading of data and splitting of testing and training set.

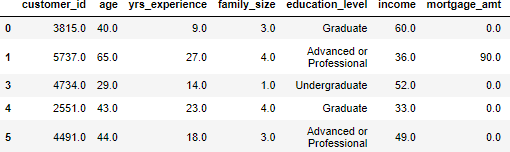
## kNN classification initial model (with all features)

Transformation

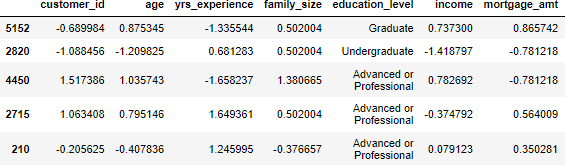
After the basic task were carried out, the modelling process started with carrying out data transformation. Transformation is carried out in both training and testing sets separately. The transformation tasks included scaling continuous variables, discretization of continuous values, label encoding and OneHot encoding.

The features which were considered for scaling of continuous variables are customer\_id, age, yrs\_experience, family\_size, income, and mortgage\_amt. It is observed that after the scaling is done, the values of the variables gets formatted in a consistent manner. For example, in the result of scaling illustrated below, it can be seen that customer\_id values changed from thousands to zeros and ones. The age values changed from tens to zeros and ones. Same goes for other columns mentioned.

Before scaling



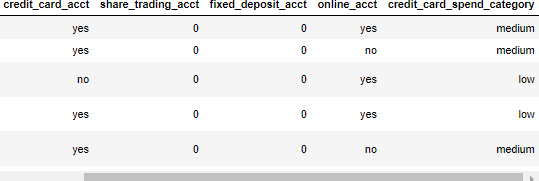
After Scaling



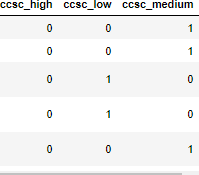
The features which were considered for Discretization of continuous value is credit\_card\_spend. It made more sense to carry out discretization on the column followed by OnHot encoding because the column refers to the amount of money each client spends.

Hence, the spending is categorized as low, medium and high first followed by discretization. It is observed that after the discretization is done, the values of the variables gets categorized.

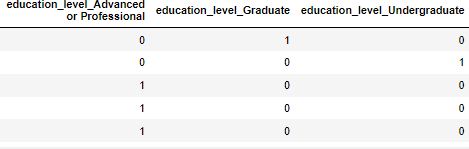
In the result of discretization illustrated below, it can be seen that credit\_card\_spend\_category column got created, values of credit\_card\_spend changed from numeric to low, medium and high categories which in turn gets stored in credit\_card\_spend\_category and the original credit\_card\_spend column is deleted.



The features which were considered for OnHot encoding are credit\_card\_spend\_category and edication level. Here, the categories are replaced with its own individual column so that it can be changed from string values to numeric zeros and ones. Credit card spend category is allocated three separate columns as illustrated below.



The same goes for education level. There are three columns allocated to education level as well. The result is illustrated below.



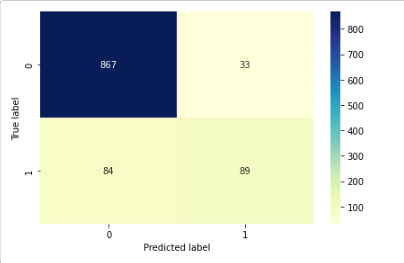
OneHot encoding was carried out for columns with multiple categories. To address columns with two categories label encoding is used. The features included are credit\_card\_acct and online\_acct. The categories replace strings as zeros and ones where zero if a client do not have an account and one if a client has an account.

To conclude, the reason why transformation is required before the actual modelling is because the data set is not consistently scaled or formatted across all columns. It is not formatted correctly, which means that columns have a different range of data (continuous and categorical data types) whose scale differs from each other. Furthermore, the computer only understand zeros and ones. Hence, for humans and computers to understand the data, it needs to be adequately formatted to zeros and ones by various transformations tasks.

Modelling

Then instantiating a k-nearest neighbors (kNN) model and fitting is done on training data set to train the model because the model needs to learn before it can predict outcomes by itself. Default parameters are used as it is the initial model.

Initial Model Evaluation Confusion matrix result:



The result of the confusion matrix states that based on the model created with kNN classification there are

* 89 true positive cases,
* 867 cases of true negative,

These true positive and negative cases determine, that the model is successfully predicting that a client will take a loan 89 times and they will not 867 times.

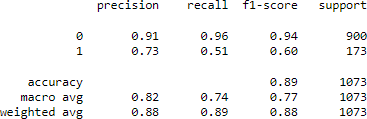
* 33 cases are classified as positive when they are not, meaning that the model is predicting 33 time that a client will take a personal loan whereas they will not.
* 84 cases are classified as negative when they are actually true, meaning that the model is predicting 74 times that a client will not take a personal loan but they would.

Recall result: It states that the initial model can successfully determine 51% of all people characteristics who would be interested in taking a personal loan.

Precision result: It sated that model can successfully predict the characteristics 73% times.

F1 score: It suggests that 60% of the times, the model can predict. The result means that the model contains low false positives and low fasle negatives. Hence, emphasising more on the result of confusion matrix result.

Classification report:

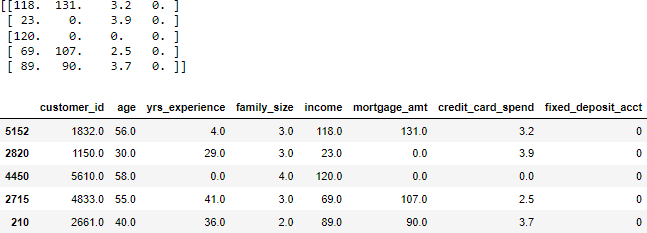


Hyper parameter tuning of kNN

Parameter tuning is normally done to figure out the best parameters that will assist in making the model perform better. It increases the success rate of the model predicting right. When optimising kNN model to find out the suitable parameters, it is found that having n\_neighbors of 6 and weight of ‘distance’ will increase the f1 score to be 62% whereas after building the initial model f1 score was 60%. The best parameter found in this step, is going to be used to improve the model.

Feature selection

To further improve the model, the column set is reduced from thirteen to eight. The four of the features within was discovered to have a strong correlation earlier. However, to test which features gets selected by the score function, eight features were selected. It was found during the fifth attempt, to reduce the features from eight to four, the features with the strongest correlation stayed. The function removed all the other four features. The result is illustrated below.



If the first row of the arrays and table is looked at then it can be observed that income (180), mortgage\_amt (131), credit\_card\_spend (3.2) and fixed\_deposit\_acct (0) stayed. Hence, these four columns are selected to improve the model along with the tuning of parameters mentioned earlier.

## kNN classification improved model (with selected features)

Transformation

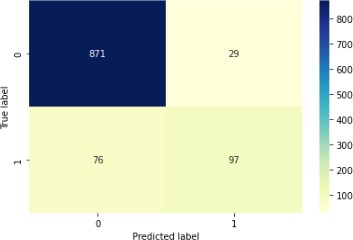
The process used to do the transformation is same as building the initial model, however, the columns used to do the splitting of data are reduced to the four columns found in feature selection stage (income, mortgage\_amt, credit\_card\_spend and fixed\_deposit\_acct). In addition as the features are recued, the transformation included different columns. For scaling continuous variables income and mortgage\_amt were scaled. For discretization of continuous values credit\_card\_spend was discretized. For OneHot encoding credit\_card\_spend\_category was encoded.

Modelling

Then instantiating a k-nearest neighbors (kNN) model and fitting is done on training data set to train the model because the model needs to learn before it can predict outcomes by itself. However, this time new parameters were included to improve the performance which are n\_neighbors of 6, algorithm of auto (even though this is the default value) and weights of

‘distance’.

Improved Model Evaluation Confusion matrix result:



The result of the confusion matrix states that based on the model created with kNN classification there are

* 97 true positive cases,
* 871 cases of true negative,

These true positive and negative cases determine, that the model is successfully predicting that a client will take a loan 97 times and they will not 871 times.

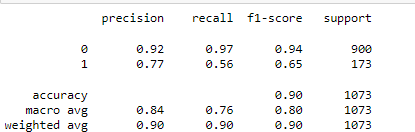
* 29 cases are classified as positive when they are not, meaning that the model is predicting 29 time that a client will take a personal loan whereas they will not.
* 76 cases are classified as negative when they are actually true, meaning that the model is predicting 76 times that a client will not take a personal loan but they would.

Recall result: It states that the initial model can successfully determine 56% of all people characteristics who would be interested in taking a personal loan.

Precision result: It sated that model can successfully predict the characteristics 78% times.

F1 score: It suggests that 65% of the times, the model can predict. The result means that the model contains low false positives and low false negatives. Hence, emphasising more on the result of confusion matrix result.

Classification report:



Initial VS Improved kNN Model

|  |  |  |
| --- | --- | --- |
|  | Initial model | Improved model |
| Confusion matrix | Model is successfully predicting that a client will take a loan 89 times and  they will not 867 times. | Model is successfully predicting that a client will take a loan 97 times and  they will not 871 times. |
| F1-score | 60% times can predict who  will take the loan | 65% times can predict who  will take the loan |
| Recall | 51% times can predict who  will take the loan | 56% times can predict who  will take the loan |
| Precision | 73% times can predict who  will take the loan | 77% times can predict who  will take the loan |

From the comparison table constructed above, it can be observed that the improved model performed better than the initial model. It is because the percentages in all the computation ways showed an increase. Confusion matrix increased from 89 to 97 times, f1-score increased from 60% to 65%, the recall increased from 51% to 56% and lastly, precision increased from 73% to 77% times. Hence, the model can successfully predict the chances of a client taking personal loan from Crédit Nationale Azur better from the improved kNN model.

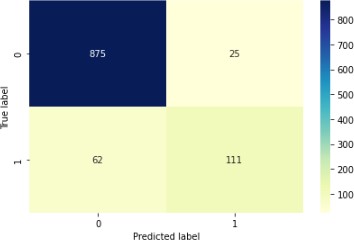
## Random forests classification initial model (with all features)

The random forests initial classification model building process is exact same as kNN initial classicisation until the transformation part.

Modelling

After transformation random forest classification is used instead of kNN. Default values of n\_estimators of 100 is considered for the initial model with all features.

Improved Model Evaluation Confusion matrix:



The result of the confusion matrix states that based on the model created with random forest classification there are

* 111 true positive cases,
* 875 cases of true negative,

These true positive and negative cases determine, that the model is successfully predicting that a client will take a loan 111 times and they will not 875 times.

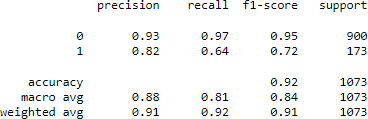
* 25 cases are classified as positive when they are not, meaning that the model is predicting 25 time that a client will take a personal loan whereas they will not.
* 62 cases are classified as negative when they are actually true, meaning that the model is predicting 62 times that a client will not take a personal loan but they would.

Recall result: It states that the initial model can successfully determine 64% of all people characteristics who would be interested in taking a personal loan.

Precision result: It sated that model can successfully predict the characteristics 82% times.

F1 score: It suggests that 72% of the times, the model can predict. The result means that the model contains low false positives and low fasle negatives. Hence, emphasising more on the result of confusion matrix result.

Classification report:



Hyper parameter tuning of Random forests

When optimising the random forest model to find the suitable parameters, it is found that having n\_estimators of 15, criterion of entropy and max\_features of 4 will increase the f1 score to 73%, whereas after building the initial model f1 score was 71%. The best parameters found in this step will be used to improve the model.

Feature selection

The feature selection for the random forest is the same as the knn classification. Hence the same set of four attributes is being used to improve the model further. The features are income, mortgage\_amt, credit\_card\_spend and fixed\_deposit\_acct stayed. Hence, along with these four columns the tuning of parameters are used to improve the model.

## Random forests classification improved model (with selected features)

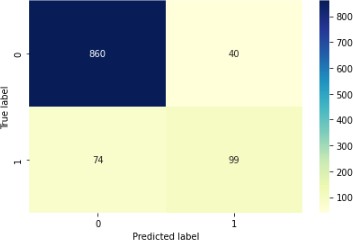
Transformation

The process used to transform the improved random forest model is the same as the improved knn model.

Modelling

Then instantiating random forest model and fitting is done on training data set to train the model because the model needs to learn before it can predict outcomes by itself. However, this time new parameters were included to improve the performance which are n\_estimators of 15, criterion of entropy and max\_features of 4.

Improved Model Evaluation Confusion matrix result:



The result of the confusion matrix states that based on the model created with kNN classification, there are

* 99 true positive cases,
* 860 cases of true negative,

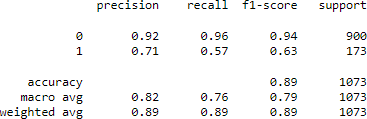
These true positive and negative cases determine that the model is successfully predicting that a client will take a loan 99 times and they will not 860 times.

* 40 cases are classified as positive when they are not, meaning that the model predicts 40 times that a client will take a personal loan whereas they will not.
* 74 cases are classified as negative when they are true, meaning that the model predicts 74 times that a client will not take a personal loan but would.

Recall result: It states that the initial model can successfully determine 57% of all people who would be interested in taking a personal loan.

Precision result: The model could successfully predict the characteristics 71% time.

F1 score: It suggests that 63% of the time, the model can predict. The result means that the model contains low false positives and low false negatives. Hence, emphasising more on the result of the confusion matrix result. Classification report:



Initial VS Improved random forest Model

|  |  |  |
| --- | --- | --- |
|  | Initial model | Improved model |
| Confusion matrix | Model is successfully predicting that a client will take a loan 111 times and  they will not 875 times | Model is successfully predicting that a client will take a loan 99 times and  they will not 860 times |
| F1-score | 72% times can predict who  will take the loan | 63% times can predict who  will take the loan |
| Recall | 64% times can predict who  will take the loan | 57% times can predict who  will take the loan |
| Precision | 82% times can predict who  will take the loan | 71% times can predict who  will take the loan |

From the comparison table constructed above, it can be observed that the improved model performance degraded more than the initial model. It is because the percentages in all the computation ways showed a decrease. The confusion matrix decreased from 111 to 99 times, f1-score decreased from 72% to 63%, recall decreased from 64% to 57%, and precision decreased from 82% to 71%. Hence, the model can better predict the chances of a client taking a personal loan from Crédit Nationale Azur from the initial random forest model.

## Comparison between kNN and Random Forest models

The comparison is being made to determine which model will be the most suitable for targeting the clients interested in taking a personal loan. The weighted average of precision and recall is f1-score. Hence, on the comparisons below, only the confusion matrix and f1- scores are considered to determine the best model. kNN VS Random Forest initial model performance

|  |  |  |
| --- | --- | --- |
|  | kNN Initial model | Random Forest  Initial model |
| Confusion matrix | Model is successfully predicting that a client will take a loan 89 times and  they will not 867 times. | Model is successfully predicting that a client will take a loan 111 times and  they will not 875 times |
| F1-score | 60% times can predict who  will take the loan | 72% times can predict who  will take the loan |

The above comparison table identifies that random forests initial model performs better in terms of determining who will take a loan as its prediction rate is hundred and eleven times. In addition, the f1-score for random forests initial model is great than kNN’s initial model as well.

kNN VS Random Forest improved model performance

|  |  |  |
| --- | --- | --- |
|  | kNN  Improved model | Random Forest  Improved model |
| Confusion matrix | Model is successfully predicting that a client will take a loan 97 times and  they will not 871 times | Model is successfully predicting that a client will take a loan 99 times and  they will not 860 times |
| F1-score | 65% times can predict who  will take the loan | 63% times can predict who  will take the loan |

The above comparison table identifies that random forests improved model performs better in terms of confusion matrix as it can determine who will take a loan as its prediction rate is ninety-nine times. However, the f1-score for random forests improved model is less than kNN’s improved model as well.

kNN initial VS Random Forest improved model performance

|  |  |  |
| --- | --- | --- |
|  | kNN Initial model | Random Forest  Improved model |
| Confusion matrix | Model is successfully predicting that a client will take a loan 89 times and  they will not 867 times. | Model is successfully predicting that a client will take a loan 99 times and  they will not 860 times |
| F1-score | 60% times can predict who  will take the loan | 63% times can predict who  will take the loan |

The above comparison table identifies that the random forests improved model performs better in terms of the confusion matrix and can determine who will take a loan as its prediction rate is ninety-nine times. In addition, the f1-score for the random forests improved model (63%) is also greater than kNN’s initial model (60%).

Random Forest initial VS kNN improved model performance

|  |  |  |
| --- | --- | --- |
|  | kNN  Improved model | Random Forest  Initial model |
| Confusion matrix | Model is successfully predicting that a client will take a loan 97 times and  they will not 871 times | Model is successfully predicting that a client will take a loan 111 times and  they will not 875 times |
| F1-score | 65% times can predict who  will take the loan | 72% times can predict who  will take the loan |

The above comparison table identifies that the random forests initial model performs better in the confusion matrix and can determine who will take a loan as its prediction rate is a hundred and eleven times. In addition, the f1-score for the random forests initial model

(72%) is more significant than kNN’s improved model (65%).

## Conclusion

To reduce the costs and increase the chances of success of the new personal loan campaign, Crédit Nationale Azur should use random forest's initial modal. The model can efficiently predict the characteristics of clients most likely to take a personal loan. The initial modal was chosen instead of the improved one because the initial modal had a better performance when compared to two random forest models. In addition, the random forest model is faster when compared to the kNN model optimisation as well. Hence, it will reduce optimisation

time and cost. Therefore, by using random forest’s machine learning algorithm, the model

can successfully predict which customers are likely to accept personal loan offers and identify those customers' characteristics. Hence, this leading Crédit Nationale Azur to use a more targeted approach, lower costs and enhance the bank's success rate.