INTE2047 Information Systems Solutions and Design

Assessment 3

Technical Report

Date: 27/05/2022
Student ID: S3741327
Name: Juyeon Kim

Table of Contents

Business Understanding

Crédit Nationale Azur wants to conduct a more targeted approach to the potential customers and this practice would allow further cost savings as well as provide further business opportunities. This goal can be achieved by creating a machine-learning prediction model to predict who a customer will be able to repay his loan or who a customer will leave the company.

As it is of paramount importance to understand the historical data, we will perform it through descriptive analysis as well as visualization approach, then clean the data using multiple data transformation and handling missing data approach. With this cleaned data, in the modeling process, we will take the iterative approach with the two most common machine learning algorithms and will increase its accuracy by feature selection and parameter tuning.

This data science project should work out in terms of being a tech-savvy bank industry leader and cost savings. A significant number of companies in various industries have started to treat data as invaluable assets and as a future means of value creation. As time goes by, more data will be accumulated, which means it is no good to just hold in a hand without utilizing it. Machine learning skills will allow them to step ahead of other competitors, positioning themselves as a tech-savvy leader in the bank industry. On top of that, as profit maximization is the ultimate goal of the business, the more targeted approach with the latest technology will not only help Crédit Nationale Azur to save cost by eliminating redundant marketing approaches but also eliminating traditional manual prediction.

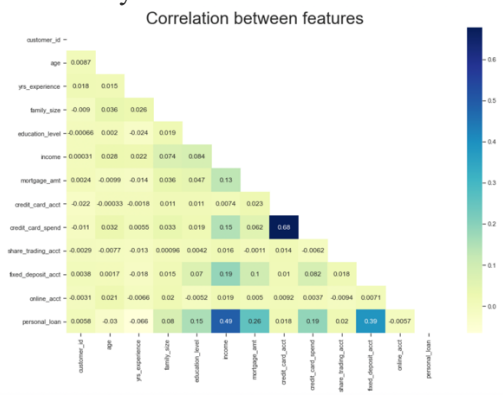
Data Understanding

(1) Missing data & object type of features identified

We identified the values of 4 features that contain NaN values, and 4 other features are encrypted as a string that should be converted into a numeric value for later model prediction. In the visualization section, we decided to delete the rows containing null values as many visualization techniques will not work with the null value as well as our purpose is not to manipulate it but to visualize only. However, it can be seen that 509 rows contain NaN values which account for approximately 10% out of the whole historical data, we decided to keep the rows in the modeling section and replace those values with non-null values for better prediction. For an object data type, we firstly extract the unique values of each column to confirm all variables are categorical data and to convert them into numeric value.

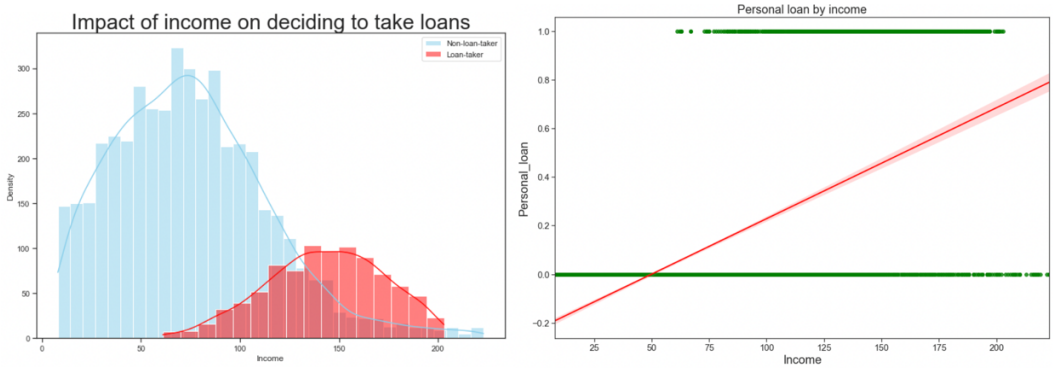
(2) Income, fixed deposit account, and mortgage amount are what matters

We have conducted 3 different approaches for correlation analysis. Values greater than 0.7/8 are considered as a significant relationship but none of the variables seems to have that much strong relationship with a personal loan, which implies if they are likely to take a loan or not. However, there seems relatively strong relationship between income (49%), fixed deposit account (39%), and mortgage amount (26%) which means a person with a higher income, fixed deposit account, and a lower mortgage is likely to take a loan next year. But it is also worth to notice that each variable alone does not let the user decide to take a loan, but other factors may have been involved.

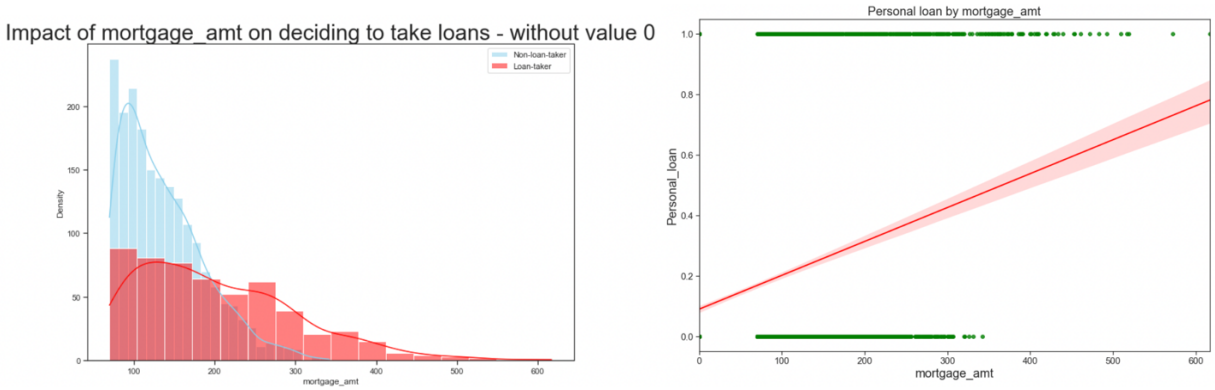


(3) Visualize correlation

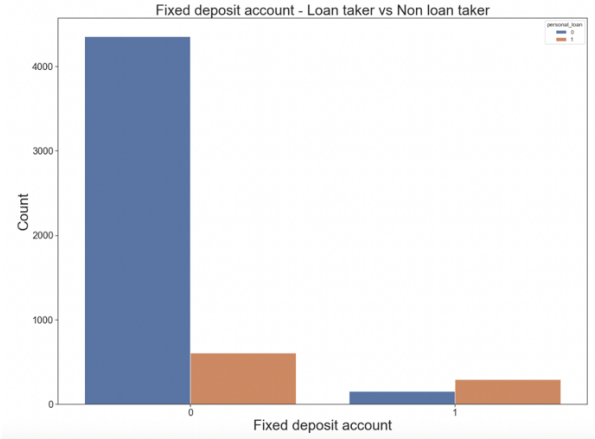
Different visualization was conducted by the type of variable to visualize the correlation, continuous variable with histogram & regression plot, and categorical variable with count plot. There seems to have a positive relationship between personal_loan and income. The more income, the more likely to take a loan. Many of non-loan taker took a loan when their income is around 60k. Income of loan-taker is higher than the income of non-loan-taker. Both histograms were normally distributed which means there is less need for parameter encoding.



Since the majority of people are not having a mortgage, we created a histogram without the value of 0 to see the clear trend of how the mortgage amount is correlated to the loan. There seems to have a positive relationship between loans and income. The more mortgage is, the more likely you are likely to take a loan. Since non-loan takers and loan-taker look to have the same peak around 100k of mortgage, it looks less strong relationship than with an income.

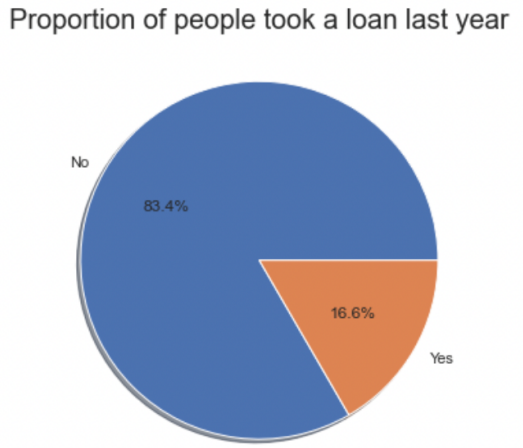


A group of people who have a fixed deposit account is more likely to take a personal loan. Looking at the people who took a loan, there does not seem to be a remarkable difference between the 2 categories, except for the fact that people without a fixed deposit account are slightly more likely to take a loan than people with a fixed deposit account. We might need further evidence to support the hypothesis that a fixed deposit account will be a good predictor of being a potential loan-taker.



(4) Proportion of Potential customers

Last year, the company launched a campaign to all customers and was only able to achieve 16.6% of them to take a loan. This can potentially mean that there is still 83.4% of a big chance where businesses can take advantage of the targeted approach with a machine learning prediction model.



Data Preparation

1. Data transformation

Firstly, we transformed feature values with “object” data type to numeric “categorical” data. We extract unique values and create the dictionary with the unique value as a key and the target value we want to transform it into.

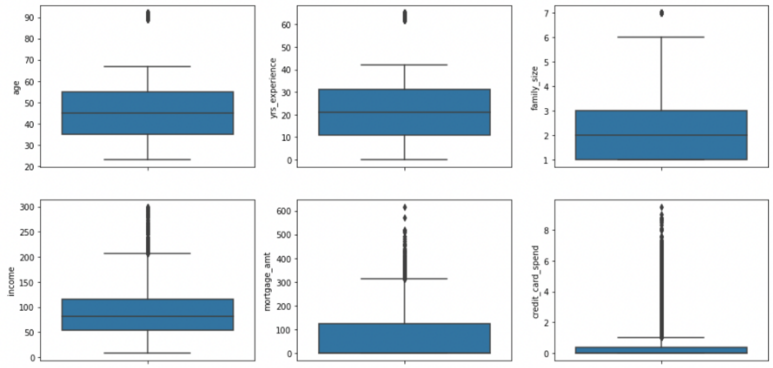
```
# k:v = string value of 'k' shall be converted to numeric value of 'v'.
edu_level_dict = {
    'Graduate': 2,
    'Advanced or Professional': 1,
    'Undergraduate': 0
}

yes_no_dict = {
    'yes': 1,
    'no': 0
}
```

2. Handle missing values

Generally, if less than 5% of values are missing then it is acceptable to ignore them. However, we have around 10% of records containing null values, and simply dropping all those records is not acceptable. To resolve this, we will make use of "SimpleImputer" to replace all missing values with media values.

3. Handle outliers

We use a box plot to visualize the outliers of continuous variables. All variables are right-skewed and have some outlier on the higher side which can be clipped. Credit card spending is mostly 0 and is right-skewed and has the most outliers (18.5%). By using Inter Quartile Range, we replace the detected outliers and any values greater than the upper cut-off point with the median value.



Feature selection

With those cleaned data, it is believed that removing non-informative features can reduce data noise and can increase the model prediction. By using scikit-learn feature selection module called SelectKBest, we tried multiple attempts by passing different k parameters to find out the best number of features required showing the highest accuracy score. For KNN model, the model performs the best with 7 features; that are; ['yrs_experience', 'family_size', 'education_level', 'income', 'mortgage_amt', 'credit_card_spend', 'fixed_deposit_acct'] by increasing accuracy from **0.8817** (rounded up in digit 4) to **0.8933.** Random forest model performs the best with 9 features; that are; ['age', 'yrs_experience', 'family_size', 'education_level', 'income', 'mortgage_amt', 'credit_card_spend', 'share_trading_acct', 'fixed_deposit_acct'] by increasing accuracy from 0**.8942** to **0.8992**.

Modelling

Two models of KNN and Random forest were selected for this project.

A. KNN

Scaling for continuous variables is expected to increase KNN performance. Thus, we perform scaling for continuous variables (age, yrs_experience, income, and credit_card_acct by using a standard scaler. As can be seen from EDA, the 0 value accounts for the majority of credit card spending and mortgage amount values. Under the hope to alleviate this skewness, we perform extra discretization to those values and then perform one-hot encoding to transform string-encoded value into a number. After the process of parameter tuning, we found {'algorithm': 'auto', 'n_neighbors': 6, 'weights': 'distance'} as the best

parameters. Looking at the accuracy table, KNN model after feature selections performs the best with default parameters (before tuning process).

B. Random Forest

As Discretization of continuous variables is highly recommended for decision tree model, discretization and one-hot encoding were applied on mortgage_amt and credit_card_spend feature. After 3 hours of executing parameter tuning, we found this model performs the best with the following params: {'criterion': 'gini', 'max_depth': 9, 'max_features': 'auto', 'max_leaf_nodes': 30, 'min_samples_leaf': 7}. Looking at the accuracy table, the Random Forest model performs the best after the feature selections & tuning. It can be seen that the **Random Forest model after the feature selection & tuning process** would be the best prediction model with the highest accuracy score of around **0.9033**.

Accuracy score	KNN	Random Forest
Before feature selection		
Before tuning	0.8816666666666667	0.8925
After tuning	0.8841666666666667	0.8941666666666667
After feature selection		
Before tuning	0.8983333333333333	0.8991666666666667
After tuning	0.8933333333333333	0.9033333333333333

Evaluation

Our model should target to choose the model that maximizes True Positive (TP) and minimize True Negative (TN) as we don't want to predict the potential loan-taker as a non-potential taker. We used various performance metrics to measure the performance of each model.

To align with our goal, we will first look at the Precision metric which is the % of the positive classes that were correct out of the total positive predicted classes. Random Forest model after feature selection has a precision of 90% which mean when predicting whether a customer is likely to take a loan, it is correct 90% of the time. With that being said, we can say it is a very competitive model.

Weighted average of Precision	KNN	Random Forest
Before feature selection	0.87	0.86
After feature selection	0.89	0.90

Recall metric is also considered to measure the % of TP out of actual positive classes. The higher recall is, the better the company can approach the targeted customer. Same as the Precision metric, the Random Forest model after feature selection has the highest precision of 90% that annotates the model can identify 90% of all potential-loan takers.

Weighted average of Recall	KNN	Random Forest
Before feature selection	0.88	0.89
After feature selection	0.89	0.90

The higher AUC, the better the performance of the model in terms of distinguishing between positive and negative classes. Random Forest after feature selection has the highest AUC.

AUC	KNN	Random Forest
Before feature selection	0.72	0.72
After feature selection	0.75	0.77

Throughout performance metrics analysis, we came to choose **Random Forest after the parameter tuning and feature selection** that performs the best in prediction in terms of accuracy score, precision, recall, and AUC, all the performance metrics.