The Analysis of Default in Loan

* The Analysis of default status in Source One Financial

Dojin, Kwon

August 12, 2019

Table of Contents

1. Definition of problem and Hypothesis statement 04

1.1. Definition of Problem 04

1.2. Hypothesis Statement 04

1.2.1. The 1st Hypothesis 04

1.2.2. The 2nd Hypothesis 04

1.3. Hypothesis Tests 04

1. Analysis 05

2.1. Data Analysis Frame 05

2.1.1. Data Preprocessing 05

2.1.1.1. Data Cleaning 06

2.1.1.2. Imputation 07

2.1.1.3. Transformation 07

2.1.2. Exploratory Data Analysis(EDA) 08

2.1.2.1. Statistical Analysis 08

a. Statistical Hypothesis Test 08

b. Analysis of Variance 09

c. Correlation Analysis 10

2.1.2.2. Visualization Analysis 11

a. Income Analysis 11

b. Asset Analysis 11

c. Default Analysis by Province 11

d. Further Extension 12

2.1.2.3. Feature Analysis 13

a. Age 13

b. Vehicle 13

c. Interest Rate and Risk Score 13

d. Others (Job and Dealer) 14

e. Others (Product Type) 14

2.1.3. Modeling for Prediction of Default 15

2.1.3.1. The Process of Modeling 15

2.1.3.2. Building up Model 15

a. Selecting Model 15

b. The Achievement of Model 16

c. Feature Importance 16

2.1.3.3. The Improvement Process of Model 17

a. The Factors in Consideration 17

b. Cross Validation and K-Fold 17

c. The Improvements 17

2.1.3.4. Parameter Tuning 18

2.1.3.5. Evaluation of Model 18

a. The Function of 1st Model 18

b. Improvement of Model 19

1. Conclusion 20

**The Process of Analysis**

1. Definition of problem and Hypothesis Statement
2. Analysis
   1. Data Extraction and Data preprocessing
   2. Data Exploration
   3. Data Modeling

* Evaluation of data model
* Improvements

1. Conclusion
2. **Definition of Problem and Hypothesis Statement**
   1. Definition of Problem

Figure out the main factors of ‘Default’ to identify criteria which is helpful to decide who can be clients as borrowers with lower risk of default. Build up machine learning to predict client’s default possibility.

* 1. Hypothesis Statement
     1. The 1st Hypothesis

The 1st Hypothesis : The annual income will be the most important factor in loan

The reason : To evaluate the borrower’s ability about repayment and interest charges, the regular amount of income is critical.

* + 1. The 2nd Hypothesis

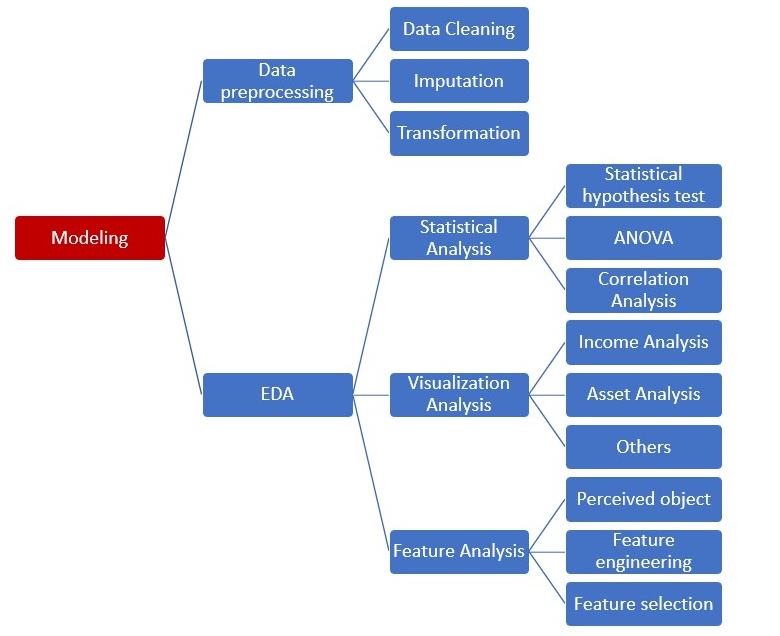
The 2nd Hypothesis : The asset would be the other main factor.

The reason : Borrowers with more assets would be able to write off the debts in the worst case.

* 1. Hypothesis Tests

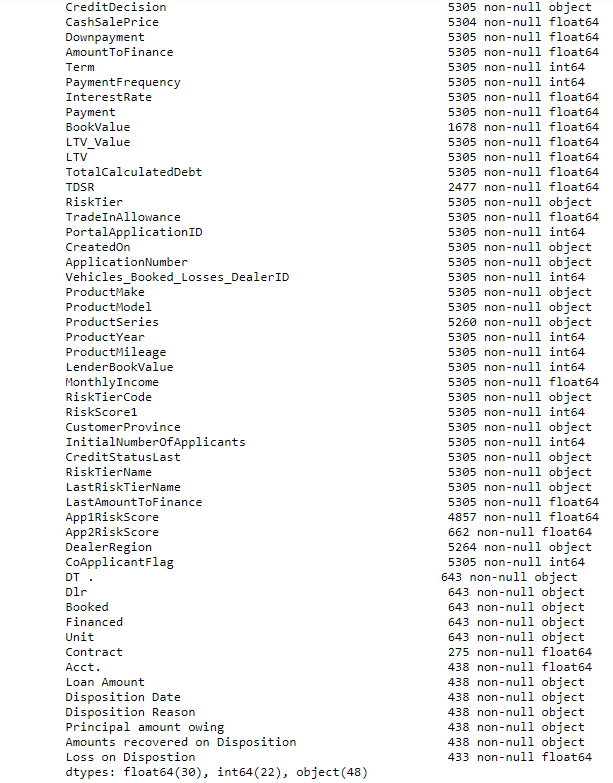
To recognize which group will repay the loan in the end, we need to distinguish their traits. For this, we will research based on three keywords 1) Income 2) Assets 3) Credit rating.

1. **Analysis**
   1. Data Analysis Frame



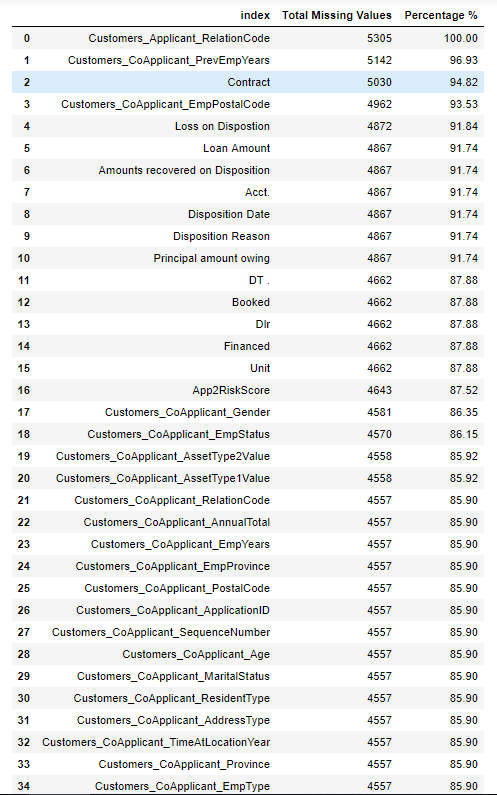
* + 1. Data Preprocessing

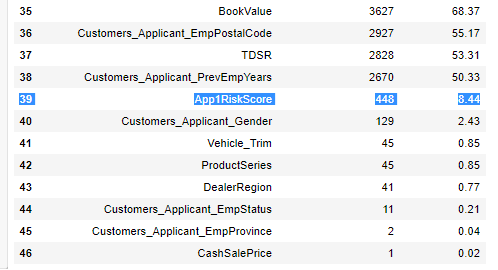
- The data set : 100 variables and 5,305 loan data



* + - 1. Data Cleaning

Cleaning missing data consists of variables that do not have 50% of values because it is hard to predict approximate value.

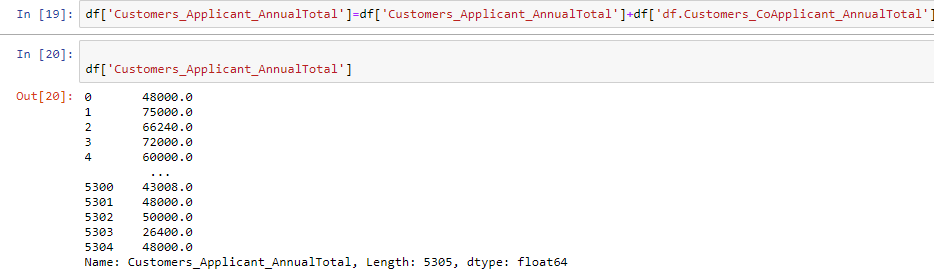




* + - 1. Imputation

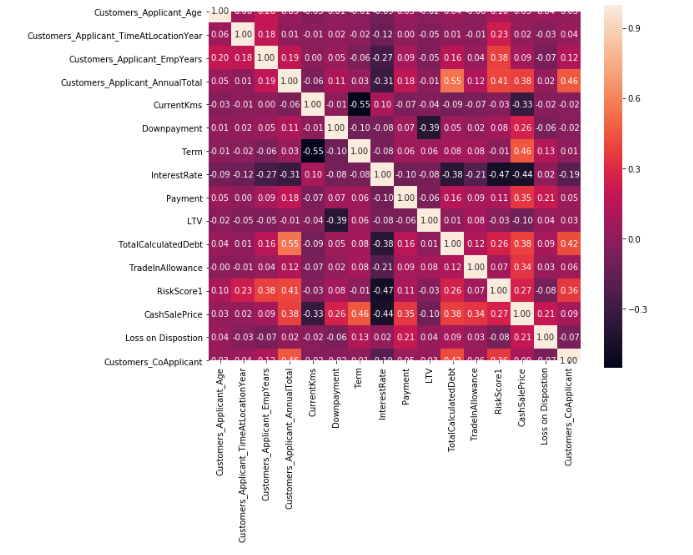
4,557 loan applications do not have information because they are co-applicants. It means 748 are co-applicants from 5,305 applicants. For comparing between co-applicants and applicants, we create a binary variable as shown below.





* + - 1. Transformation

Removal of Multicollinearity, a situation in which two or more explanatory variables in a multiple regression model are highly linearly related.



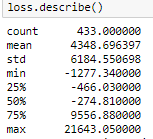
After the completion of data preprocessing, we increase the density in a qualitative

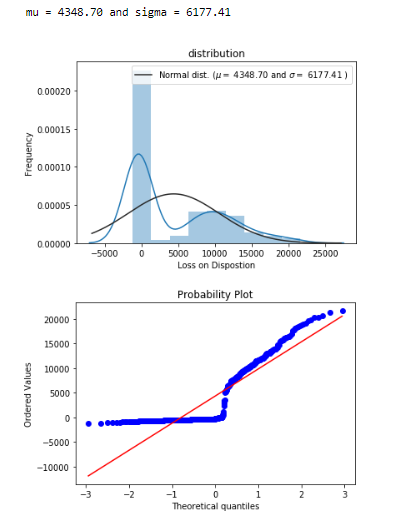
manner.

* + 1. Exploratory Data Analysis(EDA)
       1. Statistical Analysis

1. Statistical Hypothesis Test

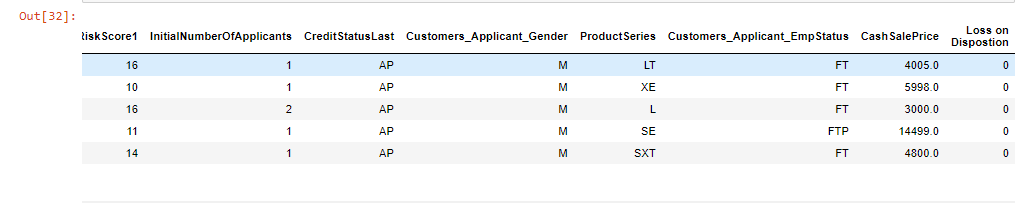
- The value of Loss on Disposition





Since we need to figure out the trait of default from the data, we created a distribution chart of defaults. After a vehicle is repossessed and disposed, there are some profits and the number of net loss is 189 out of 433 defaults. The average of defaults is 4,348. After disposition, the profit is not enough and most of defaults result in a huge loss.

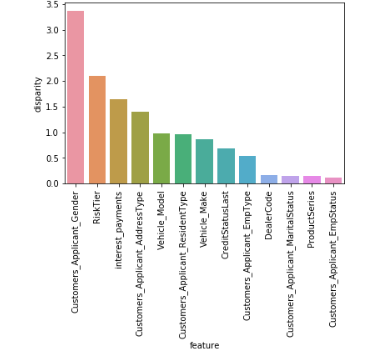
The resulting profit or loss does not change the risk of default. Therefore, the data from default can be used as predictor variables like below.



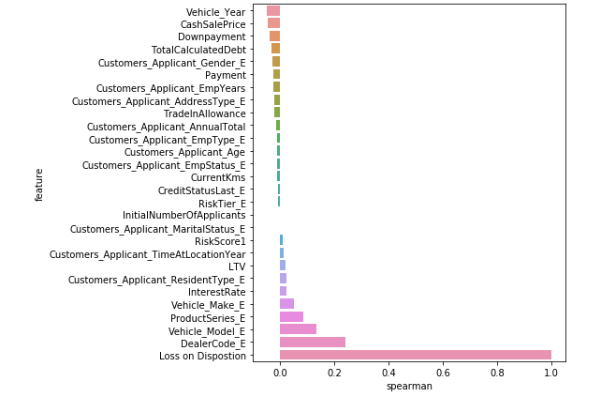
1. Analysis of Variance (ANOVA)

If independent variables depend on each other, it will cause an over-fitting problem in the model. To remove the dependent variable, figure out the dependency of specific variables.

Detect the unbalanced variance in each category and select the qualitative variable which are most relevant.



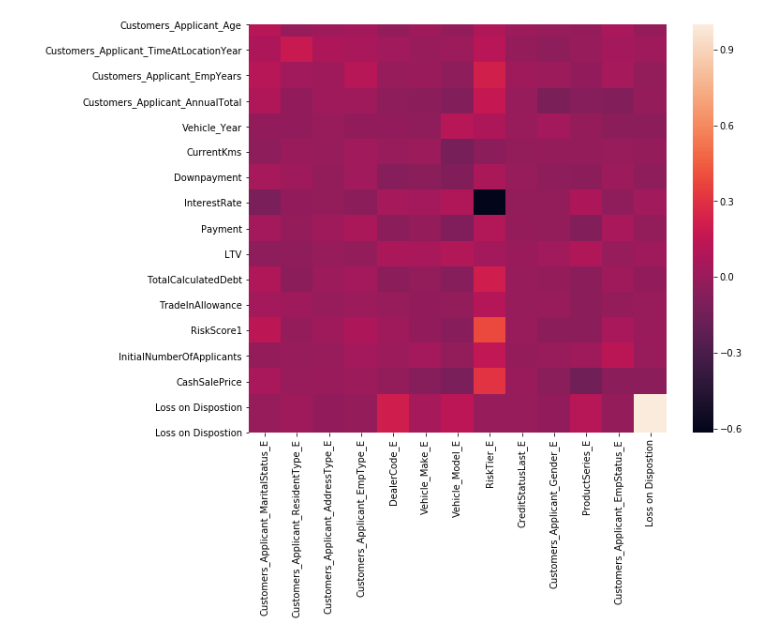
Look into the values in descending order from each category related to the possibility of default as follows.



The variables of Vehicle year, Cash Sale Price, and Down Payment show negative correlation. On the contrary, the variables; Interest rate, LTV, etc. have a positive correlation. It seems obvious that if the interest rate is high and credit rate is low, the default would be high. Unexpectedly, dealer and the model of vehicle affect largely comparing to other factors. This might imply that dealer’s ability is important whether default will be happened or not. However, it shows the low quality of data.

1. Correlation Analysis

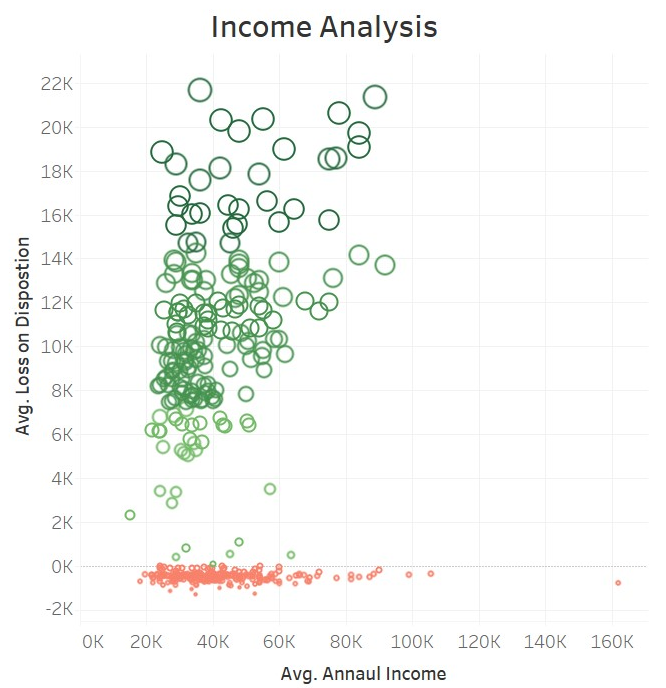
From this analysis, we can decide which variables are related closely. This analysis will be the foundation of Feature Analysis.



* + - 1. Visualization Analysis

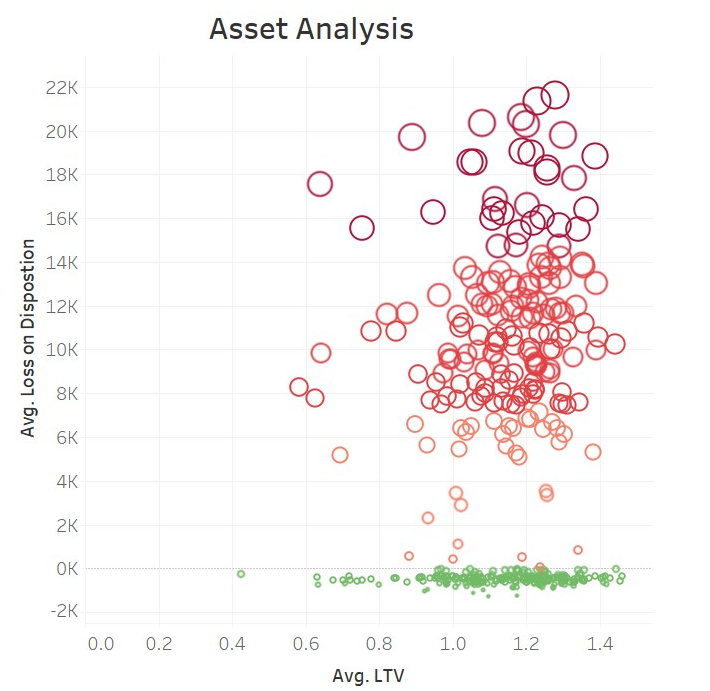
1. Income Analysis

According to the hypothesis, we look into the relationship between income and defaults. Most of defaults are linked to the borrowers’ income from 20k to 60k.



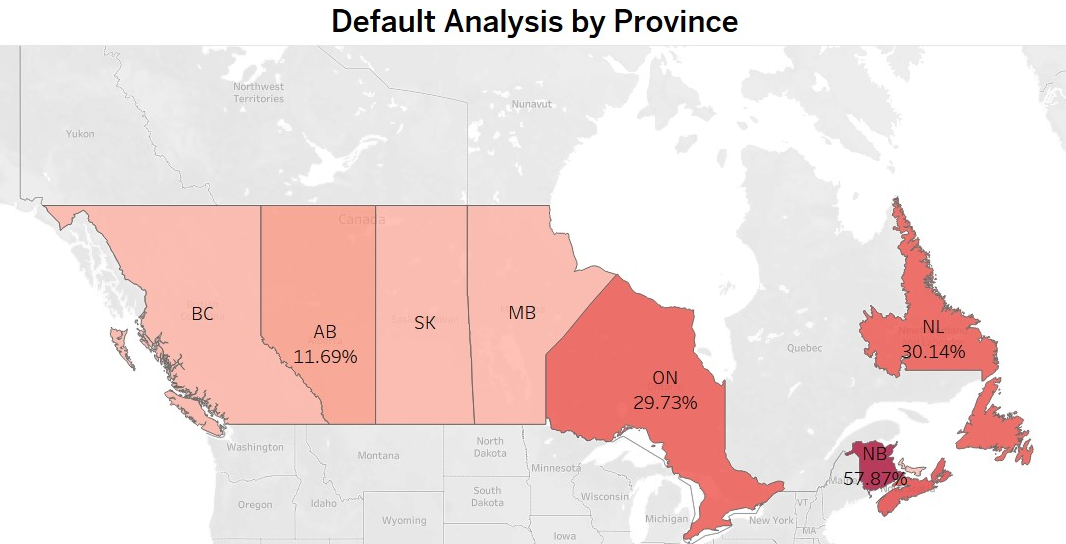
1. Asset Analysis

If the loan increased, the defaults tend to be high.



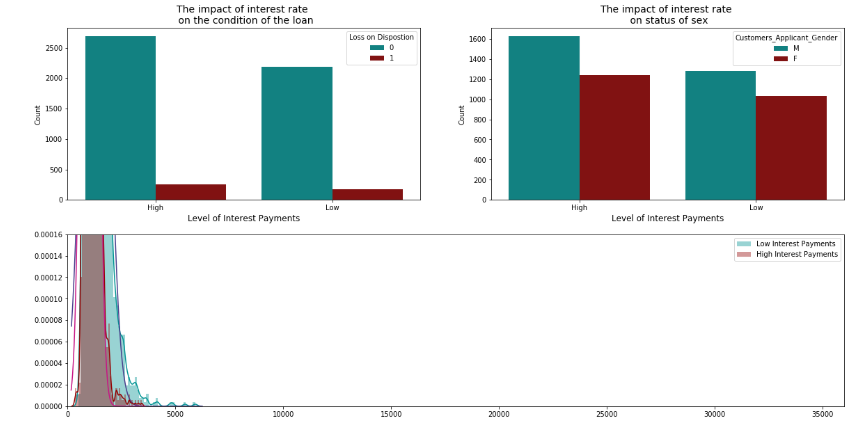
1. Default Analysis by Province

This is the chart which shows the default volume from total loan by province. New Brunswick has high percentage.



1. Further Extension



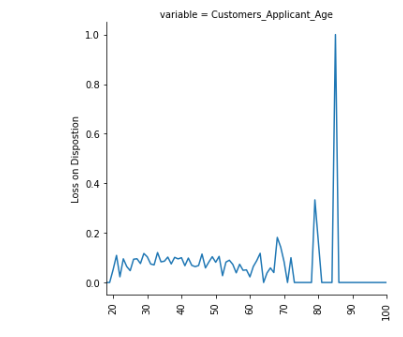


Compare the loan status by making two groups from the average interest rate. Compare the impact of interest rate on the condition of the loan and gender. However, the default rate is similar.

* + - 1. Feature Analysis

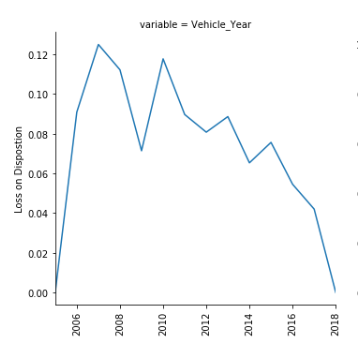
We need to look into the specific variables to figure out the traits of default.

1. Age

The default rate increase dramatically especially clients, over 70 years old. Most clients over the age of 85 are linked to defaults.

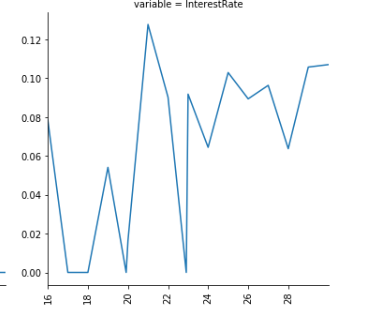
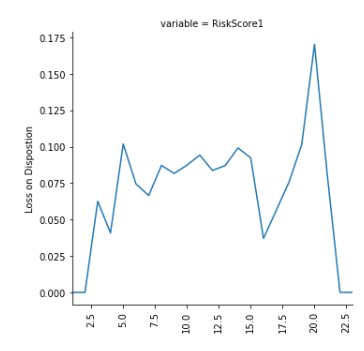
1. Vehicle

Older vehicles have a higher default rate. We assume that the people who live well prefer to buying recent model of vehicle.



1. Interest Rate and Risk Score

High interest rates and risk scores have a higher default.

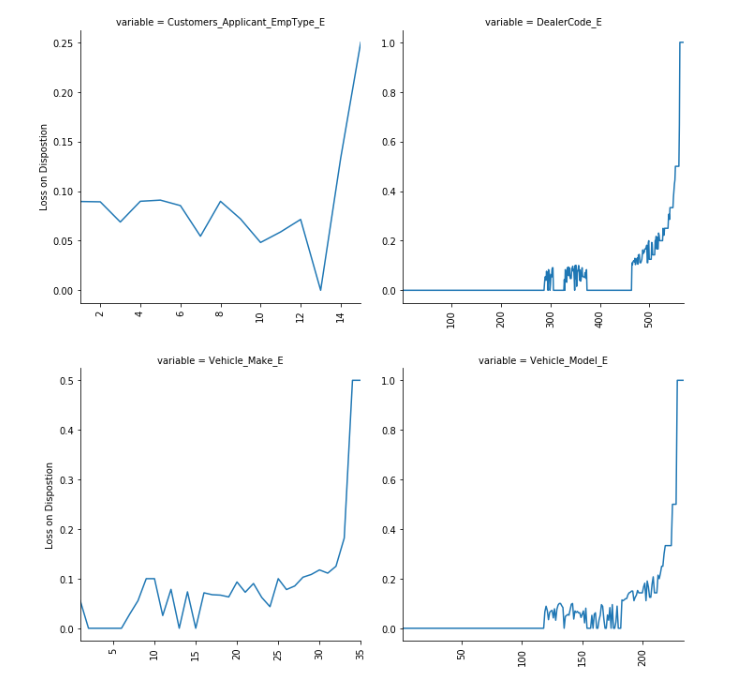
 

1. Others (Job and Dealer)

Based on chart below, the people who work at specific field show many defaults.

In the next graph, we see that a few dealerships have many defaults.

The last graph, show that defaults are occurring on specific vehicle models. The dealership and vehicle make results could be related if these dealerships are selling mostly vehicles of a specific company.

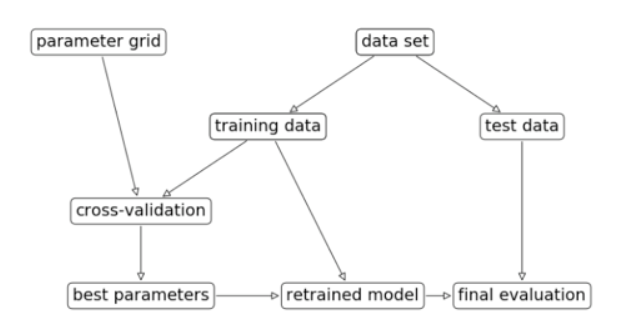


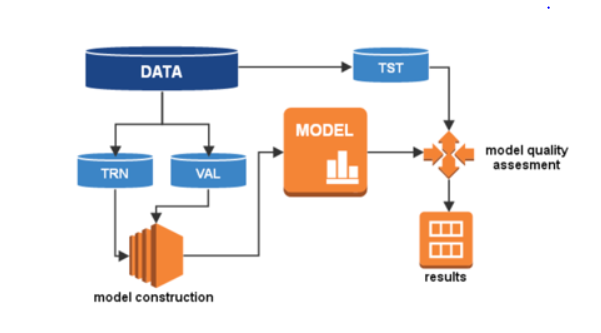
1. Others (Product Type)

The default happened from below product series.



* + 1. Modeling for Prediction of Default
       1. The Process of Modeling





* + - 1. Building up Model

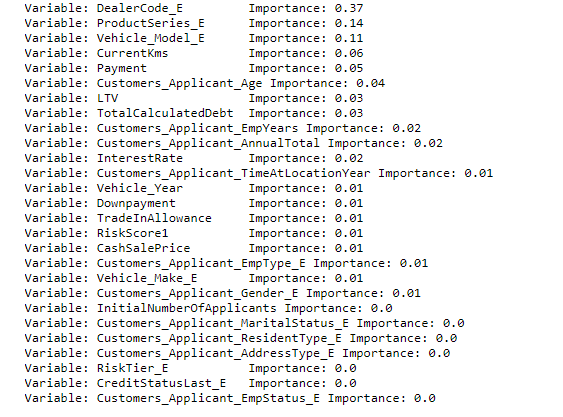
1. Selecting Model

* Logistic Regression
* Native Bayes
* K-Nearest Neighbor
* Tree-based Model
* SVM (Support Vector Machines)

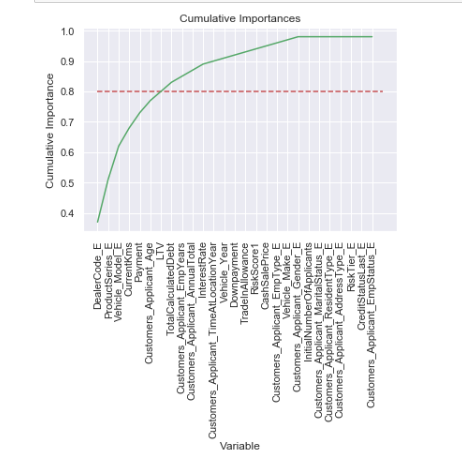
1. The achievement of Model



1. Feature Importance



The top 5 variables achieved 80%. Since the number of data is not enough, especially lack of the number of default, the function of prediction of default is weak.



* + - 1. The Improvement Process of Model

1. The factors in Consideration

* Which model is good for the purpose of analysis?
* How about generalization and the possibility of overfitting or underfitting?
* Which indicators will be used for performance measure? How can we use the Feature Engineering to improve performance?
* Which feature we should consider for language or the volume of calculation when adapting products and system?
* How will we set up the update cycle of products or system?

1. Cross Validation and K-fold

To increase the possibility for generalization, adapt the Cross Validation and K-FOLD. For this, we divide the whole dataset into three parts as follow.

* 50% : Training set
* 20% : Validation set(for grid search)
* 30% : Testing set (to be used just once at the last moment)

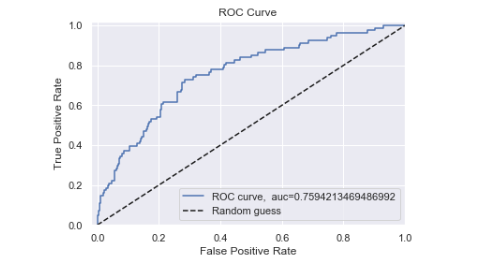


1. The Improvements

* After 10 times cross validation

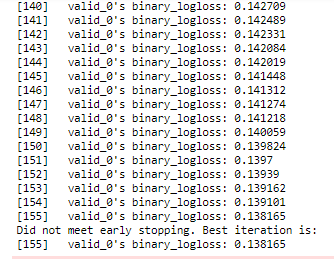


* The function for detecting default improve more than 50% compared to Random guess



* + - 1. Parameter Tuning

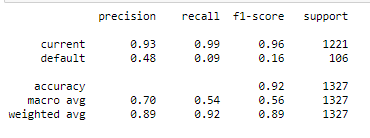
To optimize the parameters, we use K-fold and Grid search which easily increases the accuracy of the model.



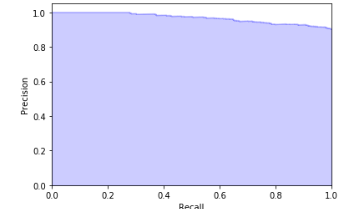
* + - 1. Evaluation of Model

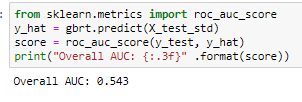
We used ‘Confusion Matrix’ as a model assessment.

1. The Function of 1st Model



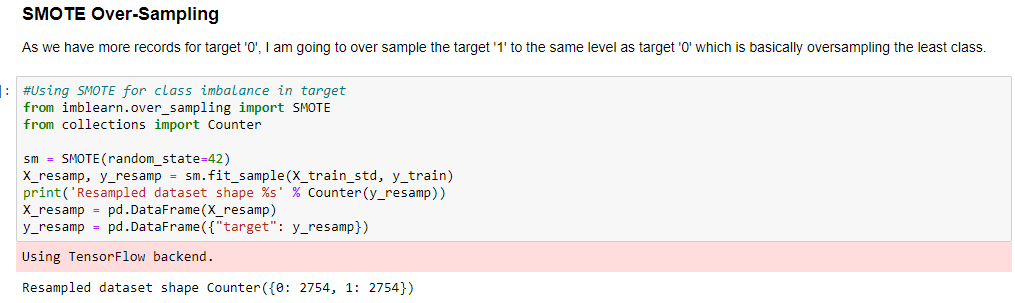
* The accuracy is high, but the default prediction is low.

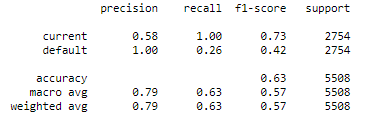


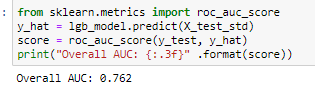


1. Improvement of Model:

Increased the prediction of defaults by 3 times compared to the previous one.



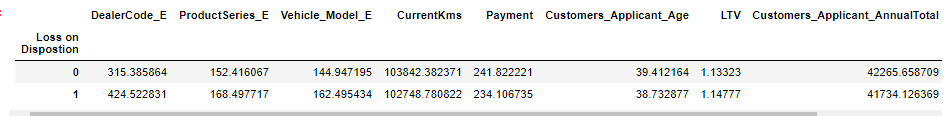




1. **Conclusion**

Divide group based on default and look into the important factor.

Following the table, the default group is younger and has low payment and annual income. However, it is not significant difference because most of clients are on the same level and the default data is not enough for analysis.



Also, the dealership appears to be an important factor of defaults, so we analyze as follow.



As going the bottom of table, the dealer has many default clients. This is not meaning that the dealer has poor performance because most of them has many clients. As long as the dealer has many borrowers, the number of default will increase.

The data set is not enough for getting exact result so we cannot evaluate their performance accurately. However, based on this table, we can check the dealer’s performance. If the dealer has too much defaults compared to the number of borrowers, we can consider training them to improve their ability.

I used oversampling techniques because data sets do not have enough duration to achieve our intended goals and the number of default data is too small.

Even though it increased the prediction of defaults by 3 times compared to the previous one, classification Performance is still not enough. But, this analysis is still meaningful to identify the important variables to improve Source One’s business model to address their risk associated with defaults.

The amount of effort in data engineering required at the initial stage of our project highlights the need for Source One to have internal resources to control and have a better understanding of their data. As the loans we analyzed come to the end of their terms over the next few months and years, the methodology and models described in this report will have a much better level of accuracy. This will give Source One the ability to better identify their high-risk clients and potentially reduce their rate of defaults and losses on those defaults.