**Credit Card Approval Prediction Using Sklearn**

**Furkan Solmaz**

**October 4, 2020**

1. **Introduction**
   1. **Business Problem**

Credit score cards are a common risk control method in the financial industry. It uses personal information and data submitted by credit card applicants to predict the probability of future defaults and credit card borrowings. The bank is able to decide whether to issue a credit card to the applicant. Credit scores can objectively quantify the magnitude of risk.

Generally speaking, credit score cards are based on historical data. Once encountering large economic fluctuations. Past models may lose their original predictive power. Logistic model is a common method for credit scoring. Because Logistic is suitable for binary classification tasks and can calculate the coefficients of each feature. In order to facilitate understanding and operation, the score card will multiply the logistic regression coefficient by a certain value (such as 100) and round it.

* 1. **Interest**

At present, with the development of machine learning algorithms more predictive methods such as Boosting, Random Forest, and Support Vector Machines have been introduced into credit card scoring. However, these methods often do not have good transparency. It may be difficult to provide customers and regulators with a reason for rejection or acceptance.

1. **Data**

Data set: <https://www.kaggle.com/rikdifos/credit-card-approval-prediction>

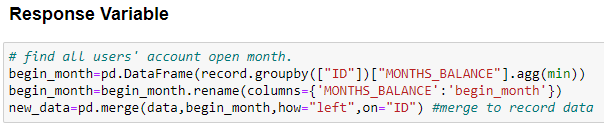
Data downloaded or scraped from multiple sources were combined into one table.

There're two tables could be merged by ID:

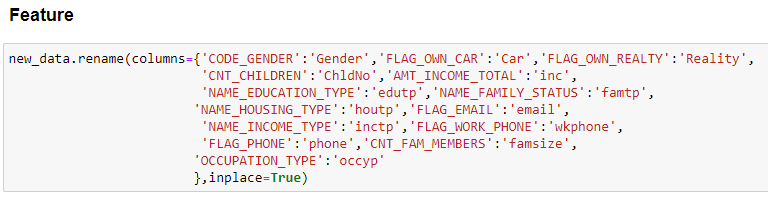
| application\_record.csv |  |  |
| --- | --- | --- |
| **Feature name** | **Explanation** | **Remarks** |
| ID | Client number |  |
| CODE\_GENDER | Gender |  |
| FLAG\_OWN\_CAR | Is there a car |  |
| FLAG\_OWN\_REALTY | Is there a property |  |
| CNT\_CHILDREN | Number of children |  |
| AMT\_INCOME\_TOTAL | Annual income |  |
| NAME\_INCOME\_TYPE | Income category |  |
| NAME\_EDUCATION\_TYPE | Education level |  |
| NAME\_FAMILY\_STATUS | Marital status |  |
| NAME\_HOUSING\_TYPE | Way of living |  |
| DAYS\_BIRTH | Birthday | Count backwards from current day (0), -1 means yesterday |
| DAYS\_EMPLOYED | Start date of employment | Count backwards from current day(0). If positive, it means the person currently unemployed. |
| FLAG\_MOBIL | Is there a mobile phone |  |
| FLAG\_WORK\_PHONE | Is there a work phone |  |
| FLAG\_PHONE | Is there a phone |  |
| FLAG\_EMAIL | Is there an email |  |
| OCCUPATION\_TYPE | Occupation |  |
| CNT\_FAM\_MEMBERS | Family size |  |

| credit\_record.csv |  |  |
| --- | --- | --- |
| **Feature name** | **Explanation** | **Remarks** |
| ID | Client number |  |
| MONTHS\_BALANCE | Record month | The month of the extracted data is the starting point, backwards, 0 is the current month, -1 is the previous month, and so on |
| STATUS | Status | 0: 1-29 days past due  1: 30-59 days past due  2: 60-89 days overdue  3: 90-119 days overdue  4: 120-149 days overdue  5: Overdue or bad debts, write-offs for more than 150 days  C: paid off that month  X: No loan for the month |

1. **Data Analysis**



Generally, users in risk should be in 3%, thus I choose users who overdue for more than 60 days as target risk users. Those samples are marked as '1', else are '0'.

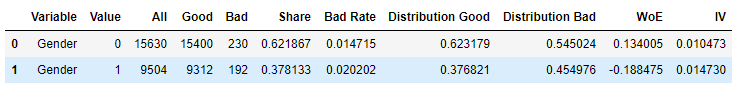


After renaming cloumns, I analyze relationship between features such as Relationships Gender, Having a car or not etc.

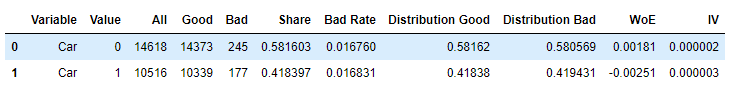
These are binary features:

**3.1 Binary Features**

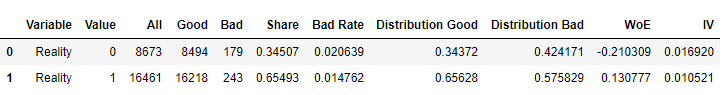
**Gender**



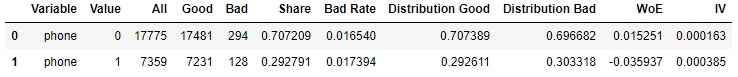
**Having a car or not**



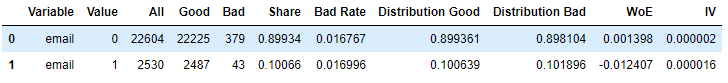
**Having house reality or not**



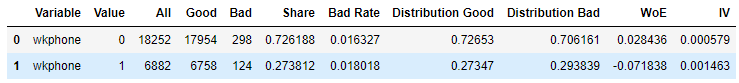
**Having a phone or not**



**Having an email or not**



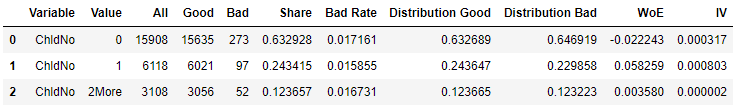
**Having a Work Phone or not**



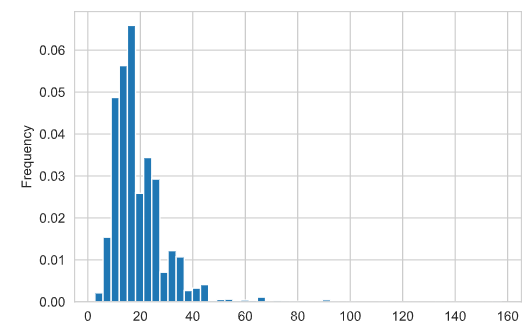
Then I analyzed other features like continuous variables.

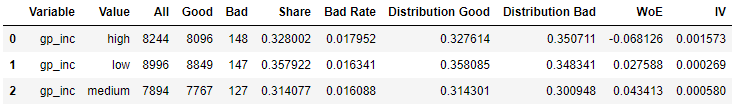
**3.2 Continuous Variables**

**Children Numbers**

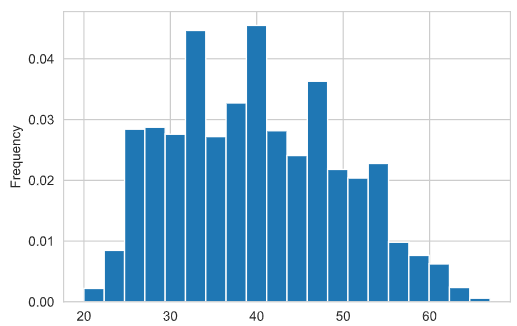


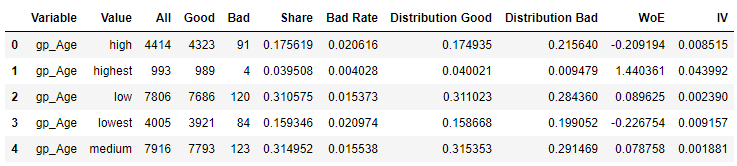
**Annual Income**



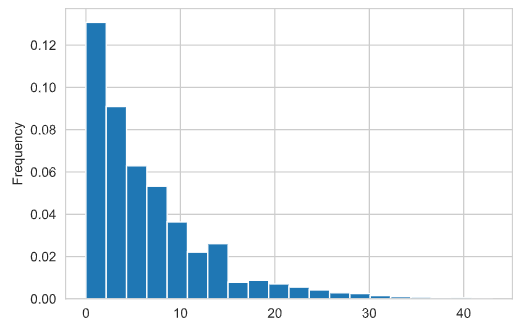


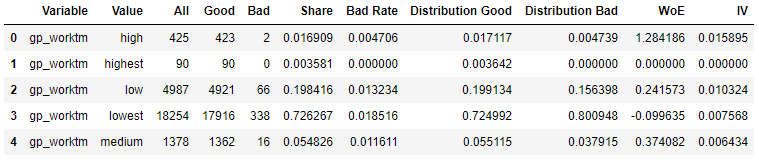
**Age**



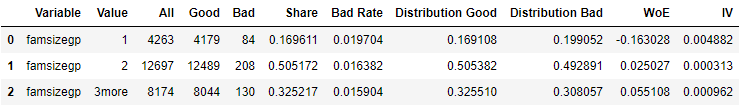


**Working Years**





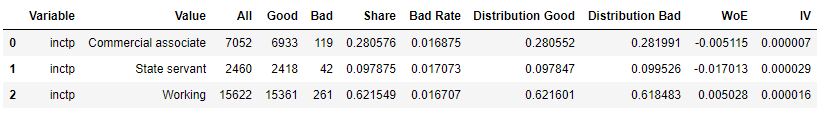
**Family Size**



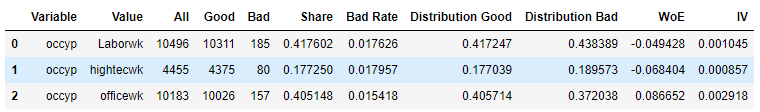
Then I analyzed other features such as categorical features.

**3.3 Categorical Features**

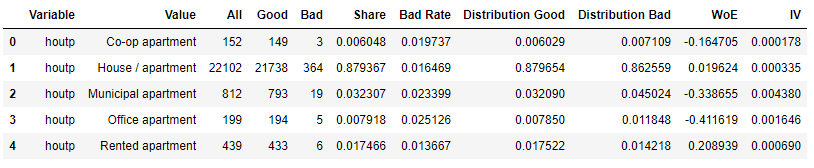
**Income Type**



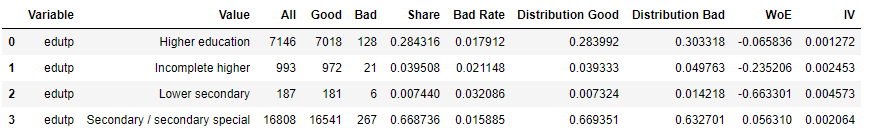
**Occupation Type**



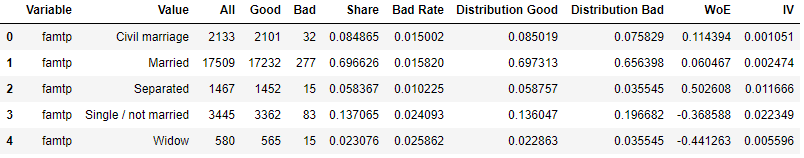
**House Type**



**Education**



**Marriage Condition**



1. **Modelling**

There are four types of algorithms, logistic regression, decision tree, random forest and SVM.

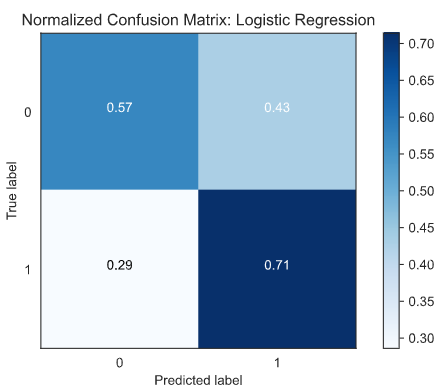
**Algorithms**

* Logistic Regression
* Decision Tree
* Random Forest
* SVM

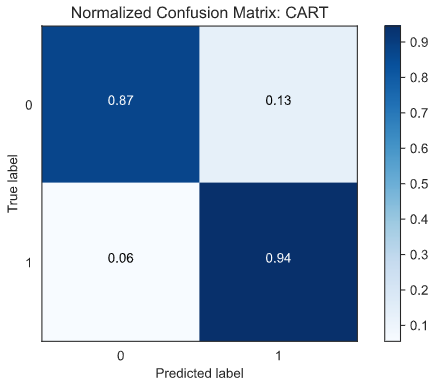
Using Synthetic Minority Over-Sampling Technique(SMOTE) to overcome sample imbalance problem.

After over sampling, the number between 1 and 0 is balanced. It can be seen from the confusion matrix.

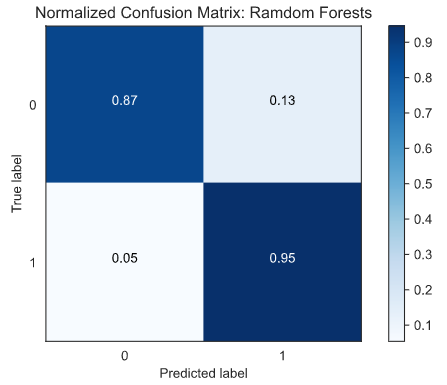
* 1. **Logistic Regression**



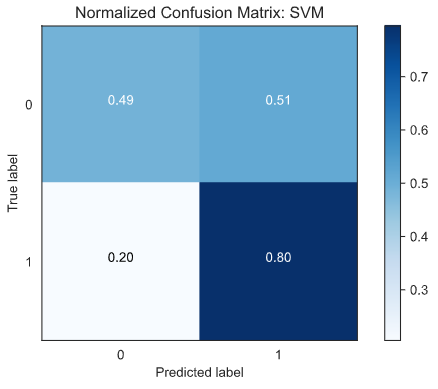
* 1. **Decision Tree**



* 1. **Random Forest**



* 1. **SVM**



1. **Conclusions**

In this study, I analyzed the relationship between gender, age, annual income etc. I build a machine learning model to predict if an applicant is 'good' or 'bad' client, different from other tasks, the definition of 'good' or 'bad' is not given. I use some techique, such as logistic regression or SVM using sckit learn to predict Credit Card Approval.

As a result, the plot shows how classification accuracy value varies across every epochs of training. The validation and train accuracy have no big difference.