Credit Risk Modelling

Project Report Submitted For the Partial Fulfilment of Requirement for the Award of the Degree of

Master of Science

In

Statistics & Computing

By

Prakash Chandra

M.Sc. Statistics & Computing (Sem IV)

UNDER THE SUPERVISION OF

Prof. Umesh Singh

Department of Statistics

Banaras Hindu University, Varanasi

ENROLMENT NO-384381 SESSION: 2017-2018

Contents

CERTIFICATE

This is to certify that Mr. Prakash Chandra, a student of M.Sc. IV semester, DST CIMS has satisfactorily completed his project report entitled “**Credit Risk Modelling”** in our department for the partial fulfilment of the requirements for the award of M.Sc. in Statistics and Computing. This Project report represents independent work carried out by the candidate.

Prof. Umesh Singh
Department of Statistics
Institute of Science
B.H.U

1. ACKNOWLEDGEMENT

It is a matter of great Privilege for me to submit this project report entitled *"Credit Risk Modeling"* based on the secondary data available on the kaggle dataset.

I heartily extend my gratitude to Prof. Umesh Singh the supervisor, for his valuable support and encouragement. I wish to express my deep sense of gratitude to my supervisor; his valuable suggestions helped me throughout the course of my work.

Last, but not least, I am thankful to my classmates and research scholars for their co-operation and support.

Prakash Chandra

2. Summary

This report summarizes machine learning technique which is used for classification problem. With the enhancement in the banking sector lots of people are applying for bank loans but the bank has its limited assets which it has to grant to limited people only, so finding out to whom the loan can be granted which will be a safer option for the bank is a typical process. So in this paper I try to reduce this risk factor behind selecting the safe person whether he is applicable for loan or not. This is done by mining the Big Data of the previous records of the people to whom the loan was granted before and on the basis of these records/experiences the machine was trained using the machine learning model which give the most accurate result.

3. Introduction

Since, distribution of the loans is the core business part of almost every banks. The main portion the bank's assets is directly came from the profit earned from the loans distributed by the banks. The prime objective in banking environment is to invest their assets in safe hands where it is. Today many banks/financial companies approves loan after a regress process of verification and validation but still there is no surety whether the chosen applicant is the deserving right applicant out of all applicants. Through this system we can predict whether that particular applicant is safe or not and the whole process of validation of features is automated by machine learning technique The aim of this Paper is to provide quick, and easy way to choose the deserving applicants applicable for loan. Since, Loan Prediction is very helpful for employee of banks as well as for the applicant also. There are so many features on which loan prediction is depend like income of person, occupation, age, extra. In this dataset response variable is 0 & 1, 0 means an applicant who is not applicable for loan and 1 means an applicant who is applicable for loan i.e. response variable is binary variable therefore I used following two methods to tackle this problem:

1. *Binary Logistic Regression* &
2. *Random forest technique*

4. Objective of the study

A “Fintech” company wants to build a scorecard that would help automate lending decision. Since it is a new company that has no historical data, it wants to use publicly available data from different banks that has data for who was granted loan and who was not for a given year. Thus, objective of this study is to build a scorecard using the attributes to automate the lending decision (who should be given loan and who should be not)

5. Dataset

There are 45 variables and 500K observations in my dataset.

5.1 Data Dictionary

Variable definitions are shows in following figure 1.

F1 = As_of_Year	F2=Respondent_Id	F3=Agency_Id
F4=Loan_type	F5= Property_Type	F6=Loan_Purpose
F7=Occupancy	F8=Loan_Amount	F9=Proposal
F10=Action_Type	F11=MSA_MD	F12=State_Code
F13=Country_Code	F14=Census_Tract_n	F15=Applicant_Ethinicity
F16=Co_Applicant_Ethnicity	F27=Applicant_Sex	F28=Co_Applicant_Sex
F29=Applicant_Income	F30=Purchase_Type	F31=Denial_Reason_1
F32=Denial_Reason_2	F33=Denial_Reason_3	F34=Rate_Spread
F35=HOEPA_Status	F36=Lien_Status	F37=Edit_Status
F38=Sequence_Number	F39=Population	F39=Population
F40=Minoroty_population	F41=HUD_Median_Family_Income	

Table 1.

5.2 Data Visualization

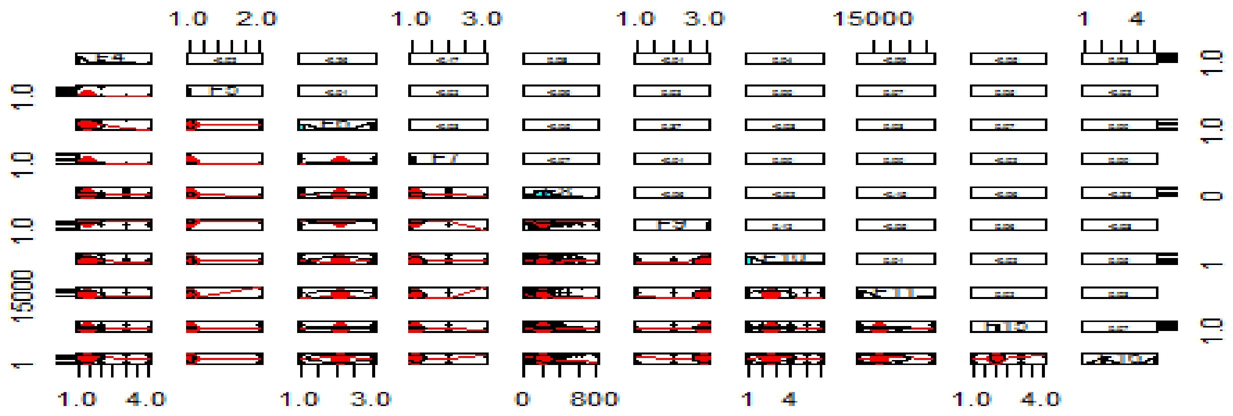


Figure1.

5.3 Data Preparation

In the given dataset, there are so many variables about which information is not available like as

F17, F18, F19, F20, F21, F22, F23, F24, F25, F26, F42, F43, F44, F45.

After removing those variable, there are two variable F34 & F37 contain large NA values. So, I also removed these from the dataset.

In this project for data preparing we use cross validation technique. In cross validation technique we split our data into two part in a ratio where 70% of data in train and 30% in test. We use cross validation technique for accuracy of our model.

We need to create an indicator variable which indicate whether an application is approved or not. We have three variables F31, F32 & F33 for this with the following information:

“There are three columns that has the data for denial reason (reason for denial of loan). If any of these columns is not null or has some value then the customer was denied a loan else he was granted loan. This way create an indicator variable.”

I created a variable Applicable using this information.

If   Applicable=0: not applicable for loan
     Applicable=1: applicable for loan

I have done some feature engineering and created a new variable

F46: Debt to income ratio using following two variables

F8:   Loan Amount

F29: Income

Debt to income ratio:

$$\mathbf{F46} = \frac{\boldsymbol{Loan\ Amount\ (F8)}}{\boldsymbol{Income\ (F29)}}$$

6. Methodology

Our aim is to classify a person as applicable for loan or not. It means we have two categories which are:

1- Person who are applicable for loan (1)

2- Person who are not applicable for loan (0)

It is a classification problem which can be easily counter with the help of following methods:

6.1 Binary Logistic Regression

Binary logistic regression is a regression technique which can be applicable for binary response variable.

It estimates the probability that a characteristic is present (e.g. estimate probability of "success") given the values of explanatory variables, in this case a single categorical variable; i.e., $\boldsymbol{\pi} = P(Y = 1|X = x)$.

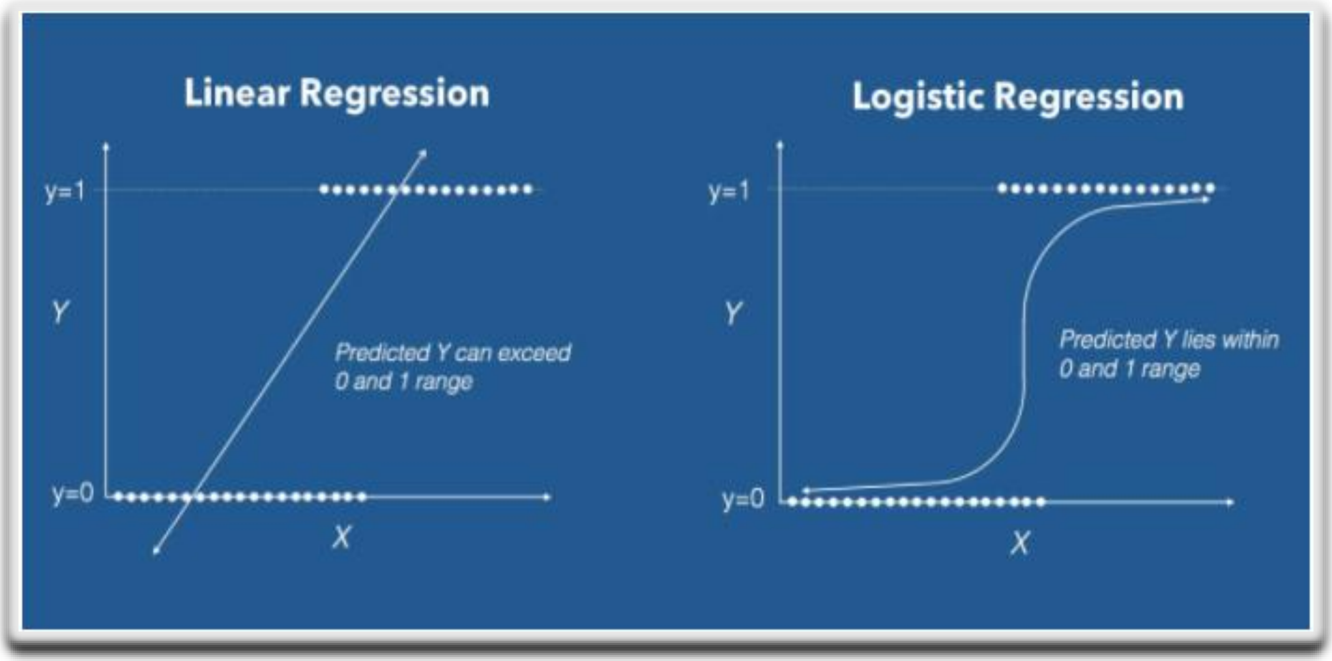


Figure 2.

6.2 Random Forest

Random Forest is a flexible, easy to use machine learning algorithm that produces, even without hyper-parameter tuning, a great result most of the time. It is also one of the most used algorithms, because its simplicity and the fact that it can be used for both classification and regression tasks.

Random forest builds multiple decision trees and merges them together to get a more accurate and stable prediction.

Below you can see how a random forest would look like with two trees:

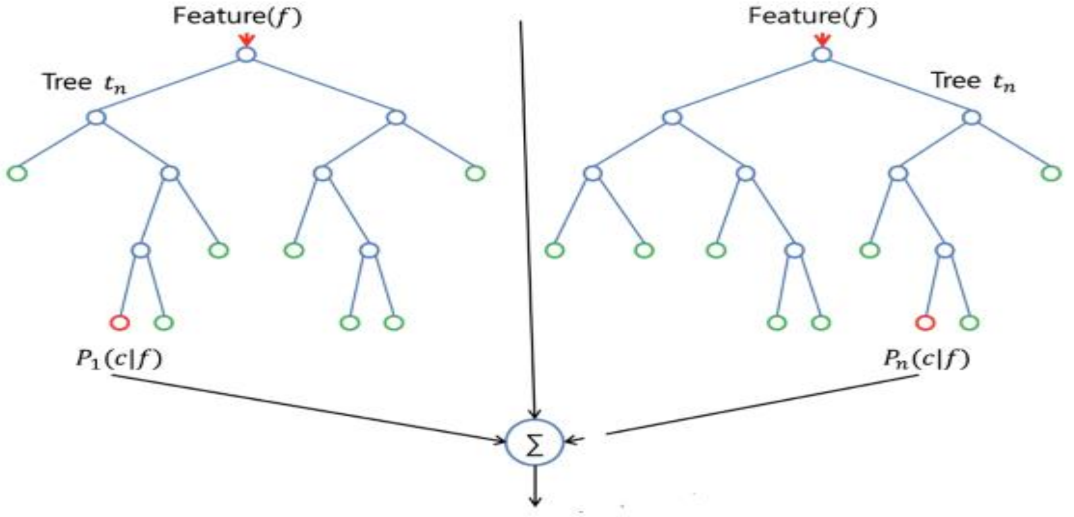


Figure 3.

Where

$$P(c|f) = \sum_{1}^{n} P_n(c|f)$$

7. Feature Selection

First of all, we have to select important feature for model building. This can be done by Feature Selection:

To find that which variable I have to use to build my model I use following two tests:

7.1 Chi-square Test of Independence - Categorical Data

Chi Square Test of Independence is used to determine if there is an association between two categorical variables.

We would like to statistically test if different variables of the applicants have any association to applicable status.

Null Hypothesis:

Different variables and applicable status are Independent of each other.

7.2 Correlation Test - Continuous Data

We would like to test if there is any significant correlation between variables and application status.

Null Hypothesis:

Applicant’s status is not correlated with the interest rate charged.

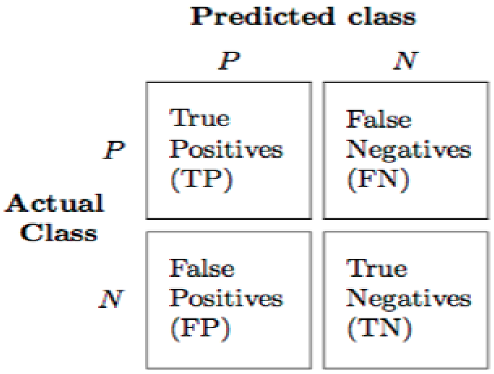
Using P-value we will find whether a variable is related with Application status.

8. Validation

We will validate our model using Confusion Matrix. For this we can use “***caret package”*** in R.

Confusion Matrix:

A confusion matrix is a table that is often used to describe the performance of a classification model (or "classifier") on a set of test data for which the true values are known.



		Predicted class	
		P	*N*
Actual Class	*P*	True Positives (TP)	False Negatives (FN)
	N	False Positives (FP)	True Negatives (TN)

Figure 4.

We can check following things after building a model.

- **Accuracy:** Overall, how often is the classifier correct

  (TP+TN)/total

- **Misclassification Rate:** Overall, how often is it wrong

  (FP+FN)/total

- **False Positive Rate:** When it's actually no, how often does it predict yes

FP/actual no

- **Sensitivity:** When it's actually yes, how often does it predict yes

TP/actual yes

- **Specificity:** When it's actually no, how often does it predict no

TN/actual no

9. Model Building

After feature selection we have to fit a model which classify better whether an applicant is applicable for loan or not. Dataset is divided into two sets namely training & testing. This can be done in R, by "*CaTools package in R"* We build a model on train dataset and then validate the model on test dataset this can be done by following model:

9.1 By Logistic Regression

I did run many models with various variables and after that I figured out that following variables are the most important variables in the dataset:

F10: Action Type

F27: Applicant Sex

F28: Co-Applicant Sex

F46: Debt to income Ratio

Our model is:

Applicable ~ F10 + F27 + F28 + F46

This model gives us the following accuracy result:

Confusion Matrix:

```
                Predicted
                   0        1
 Actual    0   42639   1439
           1   62601 158933
```

accuracy.logistic

```
[1] 0.8691477
```

9.2 Using Random Forest:

I fitted two Random Forest Classification Model one with various variables available in our dataset and the other one with the variables selected in Logistic Regression Model. To work with randomForest in R, we use “*randomForest Package”* in R.

Model 1: With various variable available in our dataset

Applicable ~ F4 + F5 + F7 + F8 + F10 + F27 + F28 + F39 + F40 + F41 + F46

Confusion Matrix:

```
                  Predicted
                     0        1
  Actual       0   41821   2257
               1   24080 197454
```

accuracy.rf1

```
[1] 0.9200493
```

In R, we select feature for random Forest technique by varImPlot function in R. Following figure shows the importance of features.

Model 2: With tested variables

Applicable ~ F10 + F27 + F28 + F46

```
Confusion Matrix:
          Predicted
Actual      0          1
    0     43515    563
    1     25566 195968
```

accuracy.rf2

[1] 0.9359114

Logistic Regression Vs Random Forest Model 1 Vs Random Forest Model 2

- accuracy.logistic

[1] 0.8691477

- accuracy.rf1

[1] 0.9200493

- accuracy.rf2

[1] 0.9359114

10. Conclusion

From a proper analysis of various features, it can be safely concluded that the some factors highly effective for loan prediction. This application is working properly and meeting to all Banker requirements. After applying various techniques, from the confusion matrix in various model, we conclude that “Random Forest” is quite a good technique for Classification problem in comparison with Logistic Regression. And our dataset can be classified into two categories using Random Forest technique with the following Model:

Applicable ~ F10 + F27 + F28 + F46

Model Accuracy**: 0.9359114**

I.e. our model predicts with 93% accuracy.

Where F10=Action_Type,   F27=Applicant_Sex

F28=Co_Applicant_Sex, F46= Debt to income ratio

11. References

1) Source of data from https://www.kaggle.com/datasets

2) https://www.datacamp.com/community/tutorials/logistic-regression-R

3) https://towardsdatascience.com/the-random-forest-algorithm-d457d499ffcd

4) https://www.youtube.com/watch?v=gmmV4drPTS4

5) https://www.youtube.com/watch?v=Z5WKQr4H4Xk

6) https://perso.math.univ-toulouse.fr/motimo/files/2013/07/random-forest.pdf

7) https://mregresion.files.wordpress.com/2011/04/logistic-regression-a-self-learning-text.pdf