International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 05 Issue: 10 | Oct - 2021 ISSN: 2582-3930

Credit Risk Modelling with Logistic Regression & Random Forest Algorithms

Rohan Chaphekar[1], Anjali A. Shejul[2], Rajendra Pawar[3]

[1]Department of Mechanical Engineering, Dr. Vishwanath Karad MIT World Peace University, Pune
[2]Department of Computer Science & Technology, Dr. Vishwanath Karad MIT World Peace University, Pune
[3]Department of Computer Science & Technology, Dr. Vishwanath Karad MIT World Peace University, Pune

---

Abstract - Post the Global Financial crisis, a consequence of multiple factors including unsupervised lending, complex debt products, credit risk management has been given utmost importance with stricter regulations and capital requirements. Credit risk can be broadly defined as the risk of losing capital that the creditor or lender provides to a counterparty. Credit derivatives like credit default swaps, collateralised loan obligations (C.L.O's) are used to mitigate this risk and thus credit risk analysis forms an integral part of financial institutions. The purpose of this paper is to model & predict probability of default, exposure at default and loss given default of any given loan data based on a historical dataset using two different Machine Learning algorithms and perform a comparative analysis on the two methods. Additionally, we compare and analyse our outputs with Moody's historical database for debt structures.

Key Words: Logistic Regression, Random Forest, Exposure at default, Loss given Default, Probability of default, Machine learning, Credit risk modelling**.**

1. INTRODUCTION

Banks, NBFC's engage in capital lending to individuals, corporates and are important for the growth of an economy. The primary business model of banks is to capture the spread between interest rates on deposits and extended loans. The banks provide a small interest to the depositors while it lends the money at higher interest rates pocketing a 600 – 700 basis point spread. But how do the banks decide who to lend? Afterall in an event of default the bank is losing the depositor's money, which with sufficient numbers can become a huge liability to the bank. Banks can rely on external ratings agency to get credit scores like FICO scores to assess a potential borrower or could develop an inhouse system that predicts loan defaults based on past information. Basel accords provide standardised approach to model these parameters for internal as well as externally sourced services. It also provides regulatory Tier I and II capital requirements, adequate ratios, and important limits on lending and NPA's. Using Historical data for a loan dataset (2007-2014) containing different independent variables pertaining to an individual loan like months since last delinquency, funded amount, Loan status etc, we aim to find a relationship between these variables and an event of default. This relationship will in turn help us predict the default probability of newly sanctioned loan dataset (2015) with help of machine learning.

2. LITERATURE REVIEW

There are four basic types of machine learning algorithms namely supervised machine learning like Linear regression, unsupervised machine learning like k-means clustering, semi-supervised machine learning like uclassify and reinforcement machine learning like Q-learning [3]. We would be focusing on supervised machine learning algorithms especially – logistic regression, linear regression, and random forest. Supervised machine learning algorithms involves the developer to select the outputs to feed the model as well as select the nature of the desired output. It is in most cases used for predicting values based on the data fed to the algorithm [3]. To fit a selected model the data is split in train and test datasets. This is done to avoid any kind of bias in the dataset. This also prevents overfitting of data. Essentially the data is split into two parts by selecting appropriate random state and split factor. We fit the model to the train dataset and test the model using the test dataset by predicting values. We also select the inputs and target variables of the data to be fitted. Thus, we have net four sets of historical data vis-à-vis inputs-train, targets-train, inputs-test, targets – test. We superimpose the predictions of inputs-test with targets-test to find the accuracy of the model. There are many validation tests like Gini coefficient, area under the curve etc. that provide us with the quality of the model fit.

Linear regression is a widely used tool to establish a linear relationship between random variables. A multivariate model has 'k' explanatory variables that form a relationship and give the following equation for the dependent variable [4]

$$Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \ldots\ldots \beta_k X_k + \varepsilon \quad (1)$$

Ordinary Least squares method (OLS) measures the unique contribution of each explanatory variable. It can be interpreted that a unit change in X_k implies a marginal increment in Y by β_k. An important assumption of OLS is that the residual is normally distributed. The alpha and betas of all explanatory variables are estimated to give us the model equation.
The statistical significance of the explanatory variables can be assessed by obtaining p-values of the estimators. Insignificant variables can be left out and the model refitted depending on the scenario.

Logistic regression can be given by [5]

$$\text{Log}\left(\frac{Y}{1-Y}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots \beta_k X_k \quad (2)$$

Upon close observation it can be observed that if Y represents a probability of a certain event the logarithm of the ratio of the probability of occurrence to the probability of non-occurrence is a linear regression. Thus, logistic regressions are often used for prediction of probabilities of occurrence of binary event, in our case probability of default(Y)
and probability of no default(1-Y).

Random forest algorithm is a method of constructing multiple decision trees for training and it returns a class or instance selected by the majority of the trees. It has better

accuracy than single decision trees as it returns the average value of prediction of each component tree [3].

Probability of default can be defined as the probability of an event wherein the borrower fails to make payments. The time elapsed between a missed loan payment and loan default is called delinquency period. For this paper we would consider a payment more than 30 days delinquent or late as a default.

Loss given default can be described as the loss expected on the underlying extended credit after it defaults at a certain point in time. While giving out loans banks usually require the borrower to post collateral. The percentage of the funded amount that can be recovered by the bank in form of sale of collaterals can be defined as the recovery rate of the underlying loan. Thus, loss given default at a certain point in time post default would be [2]

(1 – recovery rate) x Funded amount (3)

Exposure at default can be defined as the net amount yet to be repaid by the borrower after accounting for all past payments made, which is vulnerable to loss in an event of default. The expected loss is the product of all these three variables [1].

Expected loss = PD x LGD x EAD (4)

For regulatory and economic capital allocations the loss distribution is plotted and value at risk at the 99.9% confidence level is calculated [1][7]. In times of financial stress, the correlation between defaults tends to 1. Thus, banks require to calculate worst case default rates in these scenarios and maintain ample liquidity. Usually, a default rate in normal scenarios can be considered acceptable when between 3-6 %, but it ultimately depends on the business nature of the bank. A conservative bank would limit its default rates to 2.5 - 3% while an aggressive approach would accept default rates up to 5.5 - 6% thereby increasing risk for the bank.

We try to establish a relationship between loan data variables and an event of default from the 2007-14 loan dataset. Using this relationship, we try to predict the future status of each loan in the 2015 loan dataset and ultimately decide whether the particular loan should be sanctioned or not.

3. METHODOLOGY

The first step would be to get a brief outlook on the data we have. The 2007-14 loan dataset has 466285 loans with 78 different variables including funded amount, grade, months since last delinquency, interest rate, recoveries, loan status-charged off, doesn't meet credit policy etc.

We will be using Visual studio code and python for the credit risk model. We first import the entire data in python and store it in a data frame. We then proceed to pre-process the variables. We first convert the data into a single datatype that is convert string data variables to numpy.float64 data type variables. We fill in missing values in the dataset. We then get dummies for variables like grade, subgrade, home ownership, verification status etc. We then take the loan status column, a discrete variable that is our target variable containing status of each loan. We classify charged off, doesn't meet credit quality, late (30-120) days as defaults assigning value 0 while other values like fully paid, late 16-30 days, current, in grace period as no default and assign value 1. We add the values to the original data frame so that we keep the original values as well as the processed values. Whenever we would require working on the data later, we would call them by only selecting the required processed columns from the data frame. We save this pre-processed data frame into a csv file for further use.

We now proceed to processing variables for the probability of default model (PD). We split the data into train and test sets with all variables as inputs except loan status which is our target variable. We calculate weight of evidence of each variable given by the logarithm of proportion of good loans to proportion of bad loans and plot it which gives the explanatory power of the variable. For example, the grade variable contains the values A, B, C, D, E, F, G. A has the highest weight of evidence while G has the lowest weight of evidence indicating A grade has the least amount of credit risk while G grade has highest amount of credit risk. We create a list of all variables and highlight the ones that have the highest credit risk and would thus have net zero value addition to probability of non-default. For continuous variables we class into categories that have a similar weight of evidence. We save these processed values as inputs processed. We repeat this same procedure for the test data as the shape of train and test data sets must be same and have the same fine-classed variables.

We now proceed to model fitting. We fit the model with both random forests and Logistic regression and elect the model with a better explanatory power by testing both models with the test data. While fitting the model we drop the statistically insignificant variables by assessing the p-values and refit the model. We use validation tests like Gini coefficient, Kolmogorov Smirnov coefficient, area under the curve. We develop a self-assigned credit score system based on the probabilities of non-default to provide easy readability of the result data of the model fitting. We save the model and our credit score analysis. We determine a threshold value for acceptance of loan in terms of probability of non-default and credit score by assessing false positive rates and expected value of defaults. For this paper we self-impose a restriction of having expected default rate less than or equal to 6.0 % and find the maximum threshold value for the same.

We now import the 2015 loan data and our saved model. We repeat the processing of variables for the 2015 data in the same way to get the same shape and variable classing. We then compute population stability index to detect any changes between the population data between past and present data. We find the PSI value for each variable as the difference between proportion contribution of population of the variable in both datasets. A higher difference would require investigating the variable in depth. We then proceed to predict the probabilities by inputting values in the saved model. We then find all loans greater than the threshold acceptance value and assign them as 'accepted loans' we also calculate the expected number of loans that can default out of the accepted values.

For the loss given default (LGD) model we first find recovery rates from the recovery column in the dataset. We also calculate the credit conversion factor for the exposure at default model (EAD) as the ratio of payments received till date divided by the funded amount. Thus, exposure would be 1- CCF. We find the distribution of recovery rates and credit conversion rates. Based on the distribution and our analysis of both the approaches from the PD model we either elect to form a single algorithm model for both or combination of the selected approach & a linear regression algorithm. For a distribution with majority of the values falling on the mean value it makes

sense to model with a combined approach which will give us the answer to the question- is the value a mean value or not? While the linear regression following it would give us the value if it were not the mean value. This is done to avoid mean-bias in predictions that is to avoid high concentration of values around mean to affect the tail values. We then fit the models for LGD and EAD separately as per the approach selected. We then test the models in a similar fashion to the PD model and consequently use the model to predict LGD and EAD of the 2015 dataset.

4. RESULTS

PD model

- Logistic regression seemed to be a better fit for the loan data than the random forest method. In detail comparative analysis can be found in the next section.
- 0.901 was found to be an ideal threshold acceptance value for the loan. The credit score threshold was found to be 600.

Table 1: PD model – Confusion matrix

Predicted	**0**	**1**
Actual		
0	7208	2834
1	35820	47395

The net expected default rate is thus given by 2834/47395

= 5.97191 %

- False negative error seemed to be slightly higher than expected but a net acceptance rate of 50.779 % seems to be a satisfactory result.
- Population stability index concluded that the population was in line with the past data except for the variable months since issue date. Upon closer examination it is obvious that months since issue date cannot be more than 12 months in the 2015 dataset and thus is bound to be significantly different than the historical data spanning over 7 years. Also, all data of 2015 would fall into only a single class i.e., 'months since issue date<= 38 'in the model. We don't want to completely drop this data, but we can modify the coefficient in such a way that it reduces its explanatory power on the model. Thus, we elect to set it to the least explanatory class of the months since issue date variable.
- The 2015 new data accepted 196,348 loans with a net acceptance rate of 46.62 % which is in line with our model. With an expected default rate of 5.97191 %, 11,720 of these loans may default.

LGD model

- The recovery rate was highly concentrated around 0. Thus, we used a combined approach to model LGD. A logistic regression to compute if recovery rate is zero or not and a linear model to compute value if not zero. Residuals were normally distributed in the regression.

Table 2: LGD – Historical dataset

mean	0.072793
std	0.052877
min	0.000000
max	0.236973

- We then fit the LGD model for the accepted loans.

Table 3: LGD – 2015 Dataset.

mean	0.019522
std	0.051602
min	0.000000
max	0.761581

EAD model

- The CCF distribution was significantly spread out and showed significant linearity and thus only linear regression was enough to model EAD. Residuals were normally distributed in the linear regression.
- We then proceed to fit the EAD model.

Table 4: EAD model – Historical dataset

mean	0.735992
std	0.105127
min	0.384774
25%	0.661553
50%	0.731750
75%	0.810625
max	1.000000

Table 5: EAD model – 2015 dataset

mean	0.814892
std	0.063040
min	0.670342
25%	0.770562
50%	0.798891
75%	0.839875
max	1.000000

5. INFERRENCE AND DISCUSSION

5.1 Comparative Analysis: Logistic Regression vs Random Forests (PD Model)

Table 6: Logistic Regression confusion matrix

Predicted	**0**	**1**
Actual		
0	7208	2834
1	35820	47395

Table 7: Random Forests confusion matrix

Predicted	**0**	**1**
Actual		
0	25	10017
1	75	83140

The logistic regression method showcased higher false negative rate than the random forest method. But the false positive rate of random forests is extremely poor as compared to the Logistic regression method. In an industry where a small increment in default rates can increase losses exponentially a 11 % false positive rate of a model can be catastrophic. The random forests method implies an expected default rate of 12.048 % which is way beyond the thresholds set by the Basel committee [7]. Moreover, the net default rate in the historical set is 12.067 % which is close to the expected rate implying this model is no different than a blind bet. As seen below in the graphs the validation tests of model showcase the random forest model is a poor fit. The area under the curve for the Logistic regression model is 0.7007 or a 70.07 % which is considered a good fit. On the other hand, the area under the curve for random forests is close to 50% indicating as stated above its closeness to a blind bet. Gini coefficient comes out as 0.4147 for the logistic regression implying a very good fit while it comes out to be 0.0018 for random forests model. Likewise, the Smirnov coefficient is 0.29 and 0.0063 for logistic regression and random forests respectively. These results imply logistic regression to be a good fit and definitely is superior to model credit risk more efficiently as compared to random forests. Hence, we proceed to model the probability of default using logistic regression only.

5.2 LGD & EAD outputs.

The recovery rates for LGD model in the 2015 dataset were significantly higher than the historic rates. Accordingly, the historic data had a mean LGD of 7.2% while the maximum LGD for the 2015 data was at 7.6% and hence needs to be examined further. Analysis and comparison to Moody's database of recoveries can be found in the next section.
As for the EAD model, the predictions indicated slightly higher mean exposures at 81.4 % than the historical dataset which stood at 73.59 %. Upon examination it can be inferred that exposure depends on the loan repayments. As repayments increase with time exposures will decline. Hence exposure at default is inversely related to time. More the time elapses since issue, the lesser will be the exposure. As the 2015 dataset contains loans no older than 12 months the consequent exposures at default will be higher.

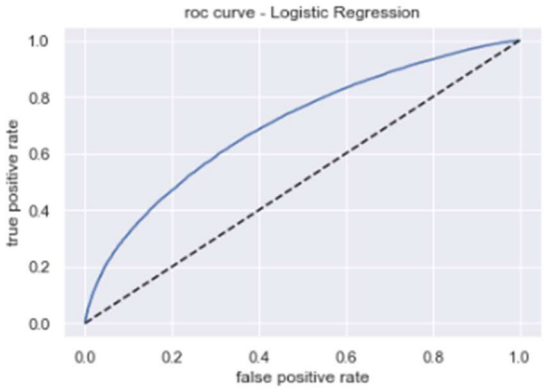

Fig 1: ROC curve – Logistic Regression

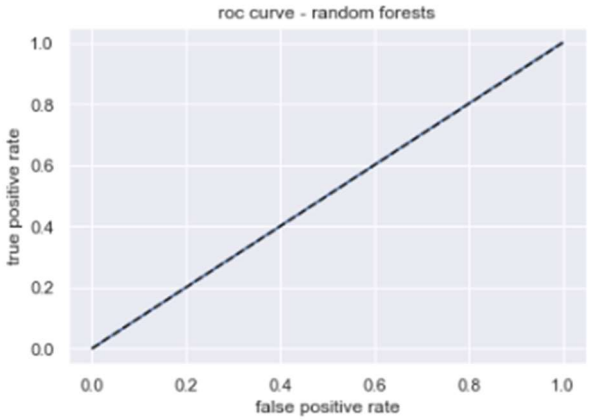

Fig 2 : ROC curve – Random Forests

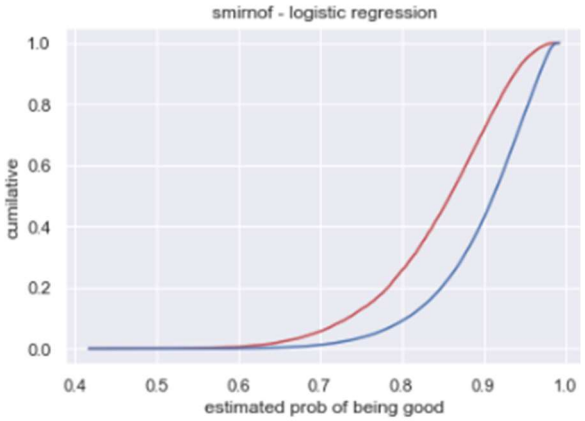

Fig 3: Smirnoff coefficient – Logistic Regression

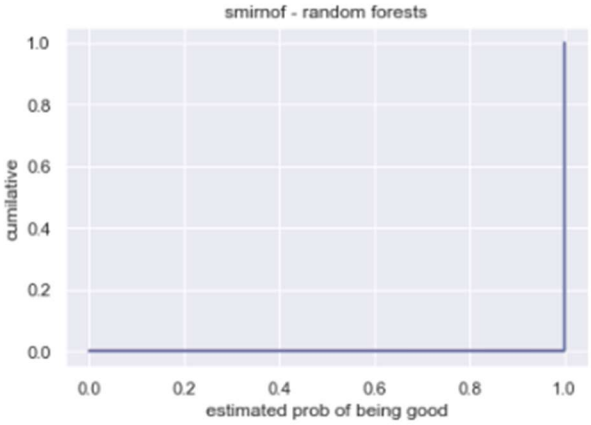

Fig 4: Smirnoff coefficient – Random Forests

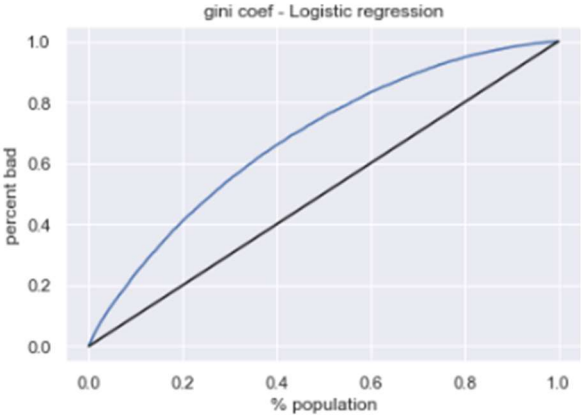

Fig 5: Gini coefficient- Logistic Regression

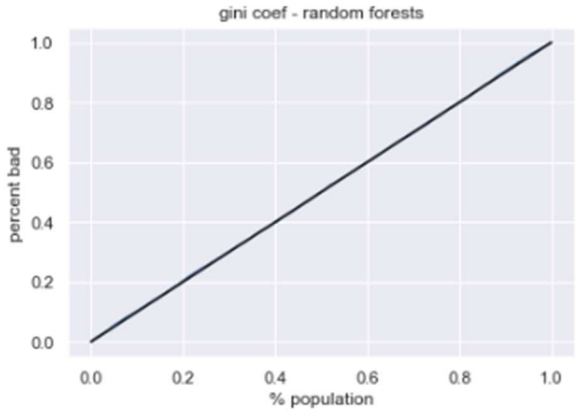

Fig 6: Gini coefficient- Random Forests

6. MOODY'S LGD DATA

Moody's recovery rate data suggests a mean 82% recovery and median 100% recovery for bank loans. Senior bonds have a 65 % mean and 67 % median recovery rate while junior and mezzanine tranche debt structure have even lower recovery rates at around 20 % for both [6]. According to our results the historical dataset has a mean LGD of 7.2 % giving a mean recovery rate of 92.8 % which is higher but considerably close to the mean data from Moody's. On the other hand, the LGD estimates for the 2015 data gave a mean 1.9 % implying a recovery rate mean of 98.1% which is significantly higher and requires investigation. Firstly the 2015 data contains newly sanctioned loans. On the other hand, the historical data contains loans spanning over 7 years while the Moody's database contains data since 1987.

Collaterals can be in form of cash, cash equivalents, property, gold or other assets that are exposed to depreciation over time. Moreover, macroeconomic factors like inflation can devalue cash equivalents over time. Secondly collateral valuations can decline significantly in times of stress as correlations tend to increase. The Moody's data has undergone significant periods of stress including the Asian market collapses, GFC. Our historical data has gone through most of the aftereffects of the GFC in the later years. This possibly explains our results on the principle of devaluation of collaterals with time.

7. CONCLUSIONS

Logistic regression was a better Machine learning method than random forests to model probability of default.

With increasing thresholds, the false negative rate rises while false positive rates and acceptance rates decline. Thus, along with lower default rates banks need to strike a balance between the type I and type II errors. As long as the banks adhere to the limits set by Basel committee and satisfy the capital requirements banks can either take a conservative business approach or an aggressive business approach by changing acceptance threshold within the limits. As correlations between defaults increase during periods of financial distress a bank can follow a conservative approach in this period while a period of economic growth the bank can take slightly higher risk.

Loss given default is mean reverting through time due to depreciation changes in collateral valuations. Thus, banks must monitor recovery rates periodically and take them into consideration while calculating capital adequacy and operational losses.

Historical data has a lower mean exposure than the 2015 mean exposure indicating 2015 loans were issued recently than the historical set which is congruent to the inverse time dependency relationship of exposure at default. Banks should study its implications to adjust capital requirements for a portfolio of loans.

REFERENCES

[1] Global Association of Risk Professionals. www.garp.org, Book 1 (Pearson) John C. Hull - Valuation and risk models (10th edition - 2020), Book 2 (Pearson) - Credit risk measurement and management (2021).

[2] 'Creditor recovery: The macroeconomic dependence of industry equilibrium' Nada Mora Journal of Financial Stability Volume 18, (2015), Pages 172-186.

[3] 'Machine Learning Algorithms: A Review', Ayon Dey / (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 7 (3), 2016, 1174-1179

[4] 'Regression Analysis - 2nd Edition (elsevier.com)', Rudolf Freund, William Wilson, Ping Sa (2006)

[5] 'Introduction to Algorithms for Data Mining and Machine Learning - 1st Edition (elsevier.com)', Xin-She Yang (2019)

[6] Moody's Ultimate Recovery Database (moodys.com)

[7] An Explanatory Note on the Basel II IRB Risk Weight Functions - July 2005 (bis.org)