International Journal of Research in Engineering, Science and Management
Volume-3, Issue-3, March-2020
www.ijresm.com | ISSN (Online): 2581-5792

Statistical Modelling on Loan Default Prediction Using Different Models

Apurva Datkhile[1], Komal Chandak[2], Sakshi Bhandari[3], Himali Gajare[4], Mandar Karyakarte[5]
[1,2,3,4]*B.E. Student, Dept. of Computer Engineering, Vishwakarma Inst. of Information Technology, Pune, India*
[5]*Professor, Dept. of Computer Engineering, Vishwakarma Institute of Information Technology, Pune, India*

***Abstract*: To allow markets and society to work smoothly, it is necessary to extend credit to individuals. It is useful for banks to assess whether or not to sanction a loan for an individual to estimate the probability of failure of their loan. We implement an efficient forecasting methodology that lets the banker predict credit risk for loan applicants. A prototype is defined in the paper which can be used by organizations to make a correct or correct decision to approve or reject a consumer's request for a loan. The paper uses four different models- Naive Bayes, Random Forest, Logistic Regression and Decision Tree Algorithm, which analyses the credit risk for optimum results.**

***Keywords*: Defaulter, Naive Bayes, Random Forest, Prediction.**

1. Introduction

Now a day's bank plays a crucial role in the market economy. Organizational success or failure primarily depends on the ability of industry to determine credit risk. Before giving the loan to borrowers, the bank decides whether the borrower is bad (defaulter) or good (non-default).

For any organization or bank, predicting the borrower status i.e. in future borrower will be defaulter or non-defaulter is a challenging task. The loan default prediction is a binary classification problem. Loan amount, customer's history determines his credibility for obtaining the loan. The problem is classifying customers as defaulter or non-defaulter. Moreover, designing such a model is a very challenging task since loan demands are rising. A prototype of the model is described in the paper that banks or any other organization can use to make the correct decision to approve or reject the customer's loan request. This work involves creating four separate machine learning models and analyzing them for optimum results.

2. Related works

Based on past literatures, different data mining techniques such as artificial neural network, linear regression, Naïve Bayes and random forest regression have been used to evaluate the risk of customers and their likelihood to default (Bu-yun ZHANG, Shi-wei LI & Chuantao YIN, 2017; Ali AghaeiRad, Ning Chen & Bernardete Ribeiro, 2016; Ajay, Venkatesh & Jacob, 2016).

Among the literature, the most used methodology was the artificial neural network (ANN) to analyze the risks associated with credit clients. It is a technique that uses interconnected neurons to solve a problem just like the human brain works [1].

The artificial neural neural network was used by Bu-yun Zhang, Shi-wei LI and Chuan-tao Yin (2017) in their A Classification Approach to Neural Networks for Credit Card Default Detection study and the results showed that their neural network has the highest processing capabilities when it involves massive complex financial data [2].

Some researchers used the Bayesian network, which is a graphic representation model that indicates how likely variables are to be interconnected. Xia et al. (2017) conducted a study using the Bayesian method of assessing credit scoring, and developed a model that could be used as a decision support system for banks to adhere when authorizing credit facilities. Furthermore, the random forest regression model has proven to provide useful insights as it is a framework that uses a set of decision trees for the purpose of prediction. It is a robust technique that researchers use as they study the banking domain. Research done by Ajay, Venkatesh & Jacob (2016) have shown that the random forest method is on top of the level of accuracy in predicting credit card default. Few techniques which have a good record based on past studies will be applied to the data to evaluate the credit risk factors and to come up with a predictive model.

3. Proposed methodology

In this paper different models like Naive Bayes, Random Forest, Logistic Regression and Decision Tree are built on the dataset. All these models have parameters which significantly affect their accuracy.

In random forest, the Gini Impurity of a node is the probability that a randomly chosen sample in a node would be incorrectly labeled if it was labeled by the distribution of samples in the node. The decision tree traverses through the features and searches for the value to split on. It repeats this splitting process in a recursive procedure till it has all nodes belonging to the same class.

$$I_G(n) = 1 - \sum_{i=1}^{J} (p_i)^2$$

The Gini Impurity decreases with each level. {Sqrt(n features)} is used for splitting each node with a subset of all the

features. The accuracy of random forest on this dataset is 0.93. Although random forest overfits, it is able to generalize much better to testing data as compared to decision tree. The random forest has lower variance which is good and possesses low bias which is a feature of decision tree. In decision tree, Information gain is used to decide which feature to split on at each step in building the tree. For each node, this parameter measures the amount of information it provides.

$$H(S) = \sum_{x \in X} p(x) \log_2 \frac{1}{p(x)}$$

The parameter with the highest information gain is used for splitting. The splitting will continue until each parameter has information gain of zero value. This algorithm is resistant to outliers and requires less preprocessing of data. The accuracy of decision tree is 0.89.

4. Data cleaning and preparation

A. *Data Description*

Give me some credit dataset from kaggle.com is used for statistical modelling. The dataset consists of 12 attributes (11 numerical,1 categorical). The number of instances in the dataset is 150001. The dependent variable is Serious Delinquencies in 2 years and is a binary variable. All independent variables will be evaluated and some or all of them will be used to build the model. The table shows the attributes used in the dataset.

Table 1
Data description

Variable Name	Data Description
Serious Delinquencies in 2yrs	Y/N
Credit Utilization	Percentage
Age	Integer
Number of Time 30-59 Days Due	Integer
Debt Ratio	Percentage
Monthly Income	Real
Number of Credit cards and Loans	Integer
Number of Times 90 Days Late	Integer
Number of Real Estate Loans	Integer
Number of Times 60-89 Days Due	Integer
Number of Dependents	Integer

B. *Data Cleaning and Preparation*

Interval variable statistical summary is done to obtain the initial findings of the dataset. Attributes like Monthly Income and Number of Dependents contain missing values which are imputed with median values. The values of all the attributes are plotted to understand the distribution of values and detect outliers. Outlier treatment is done with top coding to make the distribution more normal. Binning is used to convert the attributes age, monthly income, revolving utilization of unsecured lines, debt ratio, number of dependents, number of time 30-59 days past due not worse, number of open credit lines and loans, number of times 90 days late, number real estate loans or lines, number of time 60-89 days past due not worse to categorical. Weight of Evidence (WOE) and Information Value (IV) are calculated. WOE is calculated to assign a unique value for each group of categorical variables. IV is useful to know the predictive power of the attribute which is used for feature selection.

$$\text{WoE:} \quad \left[ln\left(\frac{\text{Distr Good}}{\text{Distr Bad}}\right) \right] \times 100.$$

$$\text{IV:} \quad \sum_{i=1}^{n} (\text{Distr Good}_i - \text{Distr Bad}_i) * ln\left(\frac{\text{Distr Good}_i}{\text{Distr Bad}_i}\right)$$

The correlation between each feature is shown in figure 1. This is required for interpretation. When variables are serving the same function one of them is eliminated.

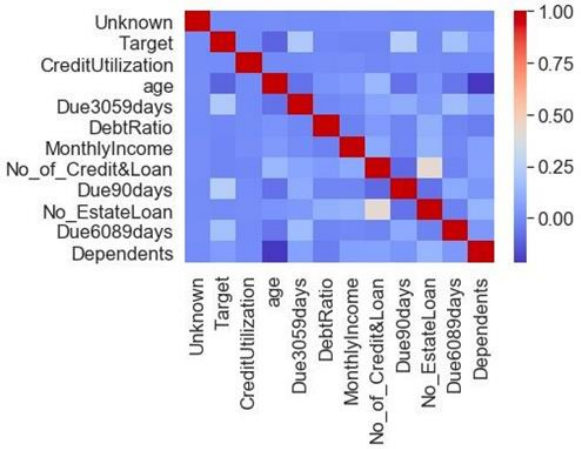

Fig. 1. Correlation plot for this dataset

C. *Experimental Setup*

Attribute selection is done and all attributes except number real estate loans or lines are considered. Attribute selection reduces memory requirements and increases the accuracy of the model. The target variable is Serious delinquencies in 2 years, containing values true or false. Logistic Regression, Naive Bayes, Random Forest, Decision Tree algorithms are run on the same dataset to compare the results of all four models. Criterion used for Random Forest is gini.

1) *Logistic regression*

A logistic regression is run using each variable against the binary target variable for the result. ROC curve for each variable is plotted. The variable containing the largest area under the curve has the largest relevancy and contributes the most for the result. The feature containing the largest Information gain ratio has the lowest importance. The subset of optimal features is arranged in descending order to obtain the highest relevancy features of the dataset.

2) *Naive Bayes*

The Naive Bayes(NB) algorithm uses the Bayes theorem and assumes independence among the variables.

X: Data tuple
H: Hypothesis
P(H/X): Posterior probability
P(X/H): Prior Probability

$$P(X/H) = P(H/X)P(H)/P(X)$$

3) Random Forest

The most important aspect of random forest is variable importance ranking. It creates recursive partitioning trees using a majority vote. A number m is specified which is much smaller than the total number of attributes. At each node, m variables are selected at random out of the total number of attributes, and then split is performed.

4) Decision tree

Recursive binary splitting technique can be used to perform split at a node. In this method, all the attributes are taken into consideration and various split points are tried and tested. They are tested using a cost function and the split with the best cost is selected.

5. Result and analysis

In this, we calculate the results of prediction on all the models which are trained on the training dataset. The data is split in 80-20 proportion i.e. the training data is 80% of the whole data and testing is 20% for all models. The performance is calculated on the basis of Accuracy, ROC, Gini. As shown in the figure 2, the accuracy for Logistic regression, random forest, naive bayes, decision tree is 93.77%, 93.44%, 89.86%, 89.51% respectively.

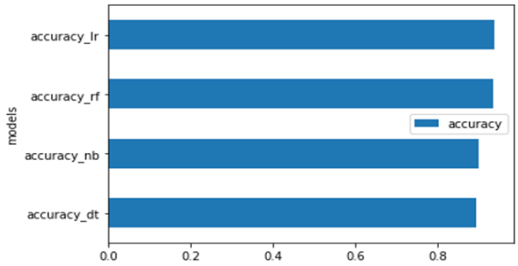

Fig. 2. Accuracy plot

6. Conclusion

In the proposed work, four learning models are constructed using nine attributes to predict the credit risk of the consumer. Accuracy, ROC, Gini are the various criteria used as measures to check the correctness of the algorithms.

References

[1] Y. Sayjadah, I. A. T. Hashem, F. Alotaibi and K. A. Kasmiran, "Credit Card Default Prediction using Machine Learning Techniques," 2018Fourth International Conference on Advances in Computing, Communication Automation (ICACCA), Subang Jaya, Malaysia, 2018, pp. 1-4.

[2] ZHANG, Bu-yunLI, Shi-weiYIN, Chuan-tao, "A Classification Approach of Neural Networks for Credit Card Default Detection," DES tech Transactions on Computer Science and Engineering, 2017.

[3] A. K. I. Hassan and A. Abraham, "Modeling consumer loan default prediction using ensemble neural networks," 2013 International Conference on Computing, Electrical and Electronic Engineering (ICCEEE), Khartoum, 2013, pp. 719-724.

[4] Y. Yan, T. Liu and H. Jiang, "A Credit Scoring Model of the Self Employed People," 2009 International Conference on Management and Service Science, Wuhan, 2009, pp. 1-4.

[5] X. Zhang, Y. Yang and Z. Zhou, "A novel credit scoring model based on optimized random forest," 2018 IEEE 8th Annual Computing and Communication Workshop and Conference (CCWC), Las Vegas, 2018, pp. 60-65.