Efficient Analysis of Financial Risks Using Multinomial Logistic Regression

DR. G. ARUTJOTHI[1], DR. C. SENTHAMARAI[2]
[1]*Department of Computer Science, Vivekanandha College of Arts and Sciences for Women(A), Tiruchengcode, TamilNadu, India*
[2]*Department of Computer Applications, Govt. Arts College (Autonomous), Salem-7, TamilNadu, India*

Abstract- ***Banks are a basic part of financial development. The banking industry has credit risk like different businesses in the finance division. Predicting credit risk is the biggest problem for the financial area in most countries around the world. Credit risk prediction and loan lending process are difficult for credit managers. This research work is focused on making a prediction model using machine learning techniques. The credit risk prediction model will change the high impact of the financial industry. The primary motivation behind this paper is to analyze the relative execution between tuned Multinomial Logistic Regression and Multinomial logistic regression fashions for default classification and credit score chance assessment. The financial information from a dataset of 30000 records in the UCI repository for prediction as default markers. The research goal is to find credit risk probabilities are assessed by some classifier metrics. It is shown that Multinomial Logistic Regression (MLR) significantly outperforms than other Classifier models, especially under the state of credit risk prediction model will change the high impact of the financial industry.***

Indexed Terms- ***Machine Learning, Logistic Regression, Credit Risk, Classifier Matrics***

I. INTRODUCTION

Banks are a basic part of the financial development of every country's economy[1]. Predicting credit risk is the biggest problem for the financial area in most countries around the world. Credit risk arises when the borrower fails to repay the loan amount[2]. Credit managers used credit history for making customer loan proposals, in this case, customers will fall in the credit default group. It depends on the customer's credit history record. So, this work mainly focuses on finding the best parameter for making a credit risk prediction model for credit managers. In this paper, we will focus on the machine learning algorithms that are used to make these decisions. Algorithms are used to make different models for different problems.

In the same way, we have to make a model for the loan lending process to the credit manager. This paper uses a Multinomial Logistic Regression and data mining classifier algorithm for making a credit risk prediction model. In the past several years many works have focused on developing credit risk models to provide loans to an enterprise. We observe several existing models and their working strategies to obtain our objective of finding the best parameter for the credit risk model. This paper attempts to find the best prediction algorithm based on the evaluation metrics score.

This paper is organized as follows: In Section 2, describes the algorithms that we use in this paper. Section 3 a brief survey on credit risk prediction models. Section 4 discusses the proposed model. Section 5 gives the results and discussion. Section 6 provides the conclusion.

II. MACHINE LEARNING

Machine Learning a field very important in computer science and statistics is the process that automatically analyzes model building without the intervention of humans. It is one of the parts of the data mining process. The overall goal of machine learning is to extract patterns from a large amount of data and to convert these patterns into understandable ones for

further use[3]. Many problems are used in machine learning techniques today. My research work uses the logistic regression technique.

2.1 REGRESSION

Regression is a statistical method[4]; it's far used for plenty of problems. The process of regression work is correlation and strength between dependent variables into independent variables. Regression is the best modeling method of machine learning in the current situation. There are many types of regression models in Machine learning that is, Linear regression, Logistic regression, Ridge Regression, Lasso regression, polynomial regression, and Bayesian Linear regression.

2.2 LOGISTIC REGRESSION

Logistic regression is used when the dependent variables are discrete. For example, 0,1 or true, false., this means the target variables have only two values. Logit function to measure the relationship between the target variable and independent variables. Statistical methods such as regression can be used to model the prediction of continuous values. Regression analysis aims to find the best model for explaining the relationship between the output and input data. In general, regression analysis establishes the relationship between the dependent (response) variable Y and one or more independent variables (inputs, regressors, or descriptive variables) X1, X2,..., Xn. [6]. Logistic regression is a type of regression model that is used to categorize dependent variables into two classes. [6]. There are two reasons to use regression analysis:

- For prediction purposes, computing the output measurement from input data is inexpensive.
- Before predicting new unknown input data, input training data is used to classify input data. There are several different types of logistic regression.

i) Binary logistic regression
ii) Multinomial logistic regression

2.3 MULTINOMIAL LOGISTIC REGRESSION

A target variable can have three or more possible types which are not ordered is called multinomial logistic regression (i.e. types that have no quantitative significance) like " A" vs "B" vs "C". My research work was carried fully out on this method. Binary logistic regression can predict only binary output while multinomial logistic regression can be dealt with one out of K-possible outcomes, where K can be target classes.

Features $(X_1 + X_2 + \ldots X_n)$ = Target $Y(Y_0 + Y_1 + Y_2 + \ldots Y_k)$

2.4 STANDARDIZATION

Standardization is a preprocessing method that is used on are before model to the data set. Preprocessing is one of the important steps in machine learning because scaled input data can give better results than normal datasets. This research work is used in the standardization method because this is suitable for differing scaled input data. Standardizing works rescaling the distribution of values and mean of observed values is 0 and the standard deviation is 1. Subtracting the mean value from the data is called centering, while dividing by the standard deviation is called scaling. So, the method is sometimes called "center scaling".

Standardization can be calculated by the following formula

$$y = (x - mean) / standard\_deviation$$

Where the *mean* and *standard_deviation* are calculated by the following formula

$$mean = sum(x) / count(x)$$

$$standard\_deviation = v\,(\,sum(\,(x - mean)\,\hat{}\,2\,)\,/\,count(x))$$

To estimate a more robust dataset using mean and standard deviation.

III. RELATED WORK

We studied various articles regarding performance evaluation of data mining and Machine Learning algorithms on different tools, some of them are described here. There is a large number of quantitative methods to estimate the credit score of loan applicants and to evaluate credit Risk. Developing SVM classifier models are powerful learning systems that are suitable for default classification and the estimation of probabilities of default (PD)[3]. As well as developed statistical models that give good

performance for a credit risk evaluation[4], quantitative methods are common in banks' credit risk estimation. The paper[5] mainly focuses on a machine-learning model for a small bank with a large financial dataset. And also, they focus only on credit default, not on credit risk. They used the Classification and Regression Tree(CART) algorithm only.

[6] This paper used three approaches that are, C4.5 decision tree, logistic regression, and random forest which all find the consumer's delinquency with data from six different banks. In this research[7] work with the CART decision tree for making loan decisions. They compare their results with k-nearest Neighbor Classifier (K-NN) and ANN but CART-based default prediction is outperform other techniques. [2] This work proposed a new credit scoring model, which is based on the hybrid feature selection method and C4.5 classifier. This relief-based hybrid system not only has a strong mathematical basis but also has higher accuracy and effectiveness.

In this work [8] developed a combination of ANFIS(Adaptive Network-based fuzzy inference system), Fuzzy clustering, and Fuzzy system algorithm-based dynamic model. This dynamic model works well in Iran's banking sector. This model replaces the static model. They compare their results with different bank datasets. [9]This paper developed a binary classifier for the prediction of default probability based on machine learning techniques. They find that tree-based models are more stable than multilayer neural networks.

IV. METHODOLOGY AND DATA

This paper mainly focuses on finding credit defaulters using the credit scoring model. This proposed work uses a Multinomial Logistic Regression classifier to evaluate credit risk. The K-Nearest Neighbour classifier(K-NN), Logistic Regression(LR), Decision Tree(DT) clasifier also used to evaluate the credit risk. Managing and analyzing financial data is more difficult because the data volume is huge [11]. To make a good decision on loan proposals an efficient model is designed using machine learning techniques. Regression techniques are also compared to find the best model.

4.1 PROPOSED ARCHITECTURE

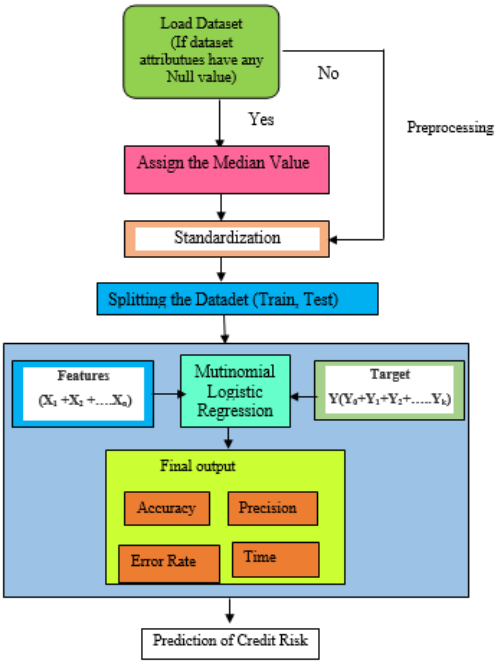

Figure 1: Proposed Architecture

Figure.1. Shows at proposed architecture, the whole dataset is taken from Banking Industry; it is analyzed to find useful information. This is tough or critical work in the banking industry. The proposed work finds the credit defaulters and makes the decision whether the loan can be approved or rejected for the new potential customer.

A standardization technique is used in this proposed work for transforming the dataset to scaled one. Multinomial logistic regression strategies classify the prevailing credit score patron statistics into defaulter and legitimate customers. This proposed model is used to make decisions on the loan lending process.

4.2 MODEL ALGORITHM:

Input Credit Dataset D, D={x1, x2,x3,…xn)
Output: Target Output j
Start
Step1: Upload the dataset D
Step 2: Processing the dataset
Step 3: Normalize the Dataset using Standard Scaler
Step 4: Split the Dataset (train and test)
Step 6: Create MLR model
i) Set the values of the parameters

{ ‘Penalty: elasticnet’, ‘Max_iter: 120’,
‘solver:saga’, ‘Random_State: 42’}
Step 7: Obtain the final Value
Step 8: Find the Customer Probability Defaults
Stop the process

The MLR model have number of parameters, from that this research work found four parameters for tuning the model. That parameters are penalty, random_state, solver and max_iter. The parameter tuned MLR model gives high accuracy than other classifiers.

4.3 DATASET USED

In this paper, credit data from Australian bank clients (available at the UCI repository) is used. The data set contains records on 30000 customers, with each record containing 25 features. The classification problem is to find default and non-default customers. Python software is used for analysis and evaluation. The metrics used for analysis are accuracy, Precision, run time, and RMSE(Root Mean Squared Error). The name of the dependent attribute is the default. This attribute defines more than three class labels. The class labels of the defaults are shown in Figure2.

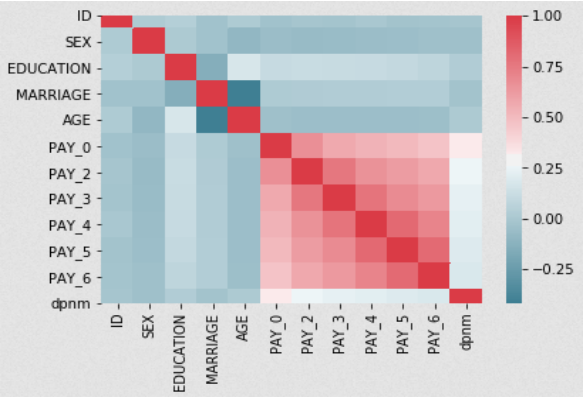

Figure 2. A simplified multi-class label

Some important attributes are analyzed using Exploratory Data Analysis techniques. The below figures are represented in the original dataset-based analysis shown in figure 3, figure 4.

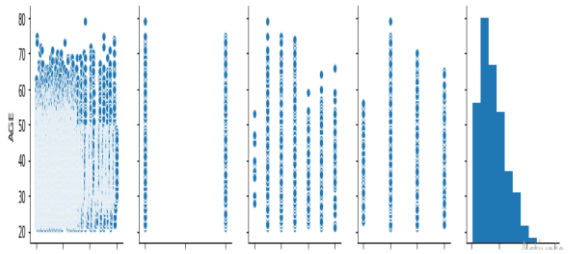

Figure 3. A simplified customer age group

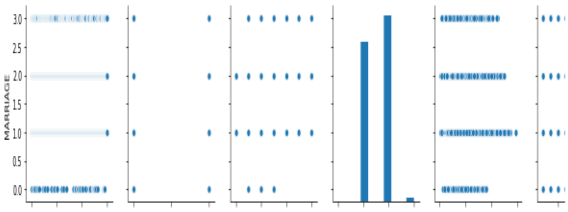

Figure 4. Marriage type

V. RESULTS AND DISCUSSION

The MLR credit scoring model was successful in classifying default and non-default loans. Hence, the lender can reduce the risk of investment failure by selecting profitable borrowers after processing the loan applications through this model. This model can correctly classify the default and no-default as 82% in the test dataset. Table I presents the classification results of this model.

Table I shows the result of the classifier models and regression model using the standardization method. We set the default parameters MLR which is used to find the minimum error rate of the training set. The dataset is partitioned in different ways, we get the lowest time is 70%: 30% of the dataset. Table 1 shows the result for classifier algorithms.

Table I : CLASSIFIER RESULTS

Classifier Model	Data Splitting 70%: 30%			
	Accuracy	Precision	Error Rate(RMSE)	Time in Seconds
Tuned MLR	8282	82	0.43	57
K-NN	77.7	72.8	0.22	62.74
LR	69.3	71.2	0.46	61
MLR	80	80.1	0.78	2.11

DT	70.3	80.4	0.29	6.11

Figure 5 shows the comparision of classifier results. The graph is drawn with the classifiers on the x-axis and the percentage of classifier metrics on the y-axis.

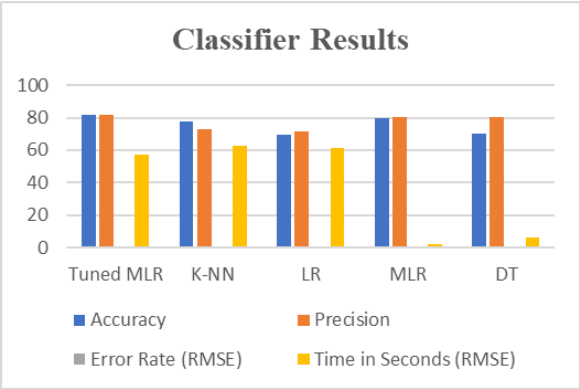

Figure 5: Comparision of Classifier Results

The maximum performance is obtained by the KM-MLR model with the precision of 82% and accuracy of 82%. After creating a proposed model we get highest result than other models and the improvement is 2%. Multinomial Logistic Regression(MLR) models has achieved slightly better performance with an precision of 82%, After creating a tuned model we get highest result than other models and improvement is 2% which is calculated to subtract the maximum accuracy and proposed model accuracy

The Tuned multinomial logistic regression credit scoring model successfully demonstrates the applicability of MLR in credit scoring for the classification and prediction of loan defaulters. Machine learning techniques are used to develop credit scoring systems. The MLR-based credit scoring system provides high accuracy than other classifiers. The data is used to develop MLR credit scoring model with splitting criteria. Fig 5 shows the highest accuracy of training data. This proposed model presented in this study can be effectively used by loan lenders to predict the loan applicant. Lenders can use this model to predict the default customer of the loan applicant.

CONCLUSION

In this paper, we have proposed a model to identify the default customer of a bank loan applicant effectively. The proposed model shows an 82% accuracy result in classifying credit applicants using Python. The credit manager can use this model to make a loan decision on loan proposals. Further, the comparison study has been made with different classifiers. 70%: 30% percentage-based Tuned MLR system gives the highest accuracy and also the precision. This model can be used to avoid the huge loss of financial institutions.

FUTURE WORK

The current paper only focuses on accuracy. The result suggests MLR is best suitable for large datasets. However, this study considers only MLR based credit scoring model. Hence, future research is to optimized MLR for all types of datasets and compare it to other machine learning classifiers with optimization techniques.

REFERENCES

[1] Arutjothi, G. and C. Senthamarai. “Assessment of Probability Defaults Using K-Means Based Multinomial Logistic Regression.” *International Journal of Computer Theory and Engineering* (2022): n. pag.

[2] Arutjothi,G.,Dr.C.Senthamarai. "Credit Risk Evaluation using Hybrid Feature Selection Method." Software Engineering and Technology 9.2 (2017): 23-26.

[3] Jan-Henning Trustorff • Paul Markus Konrad • Jens Leker, “Credit risk prediction using support vector machines “, Rev Quant Finan Acc (2011) 36:565–581. DOI 10.1007/s11156-010-0190-3

[4] Sun L, “A re-evaluation of auditors opinions versus statistical models in bankruptcy prediction”, Rev Quantitative Finance Account (2007), t 28:55–78

[5] Khandani, Amir E., Adlar J. Kim, and Andrew W. Lo. 2010. Consumer credit-risk models via

machine-learning algorithms. Journal of Banking and Finance 34: 2767–87

[6] Butaru, Florentin, Qingqing Chen, Brian Clark, Sanmay Das, Andrew W. Lo, and Akhtar Siddique. 2016. Risk and risk management in the credit card industry. Journal of Banking and Finance 72: ytgfc18–39.

[7] Galindo, Jorge, and Pablo Tamayo. 2000. Credit risk assessment using statistical and machine learning: Basic methodology and risk modeling applications. Computational Economics 15: 107–43

[8] Moradi, S., Mokhatab Rafiei, F. A dynamic credit risk assessment model with data mining techniques: evidence from Iranian banks. *Financ Innov* 5, 15 (2019).

[9] Addo, Peter & Guegan, Dominique & Hassani, Bertrand. (2018). Credit Risk Analysis Using Machine and Deep Learning Models. Risks. 6. 38. 10.3390/risks6020038.

[10] M. Sustersic, D. Mramor, and J. Zupan, "Consumer credit scoring models with limited data," Expert Systems with Applications, Vol. 36, No. 3, 2009, pp. 4736-4744

[11] http://mlr.cs.umass.edu/ml/datasets.html