Credit Score Evaluation of Customer Using Machine Learning Algorithms

Saurabh Agrawal
KJSCE,Vidyavihar
saurabh.agrawal@somaiya.edu

Purnima Ahirao
KJSCE,Vidyavihar
purnima.ahirao@somaiya.edu

Saurabh Kumar
KJSCE,Vidyavihar
saurabh.kumar@somaiya.edu

Pinak Dere
KJSCE,Vidyavihar
pinak.dere@somaiya.edu

***Abstract*— Credit Score is one of the most important and critical features or attributes of an individual which states their financial behavior and Market Credibility. Credit Score measures of an individual help banking and lending institutions to identify their customers as a good or a bad customer and also help them to identify the defaulters in terms of granting loans and credit cards to their customers. Credit Score may also help an individual to calculate their financial capability and eventually help them to explore various investment options so as to increase their market credibility. Machine Learning Algorithms can play important role in predicting the status of a customer as good or bad and at the same time can help them to plan a progressive financial eligibility for future prospects. The paper proposes a loan eligibility prediction system based on Credit Score calculator using machine learning algorithms. The accuracies of each algorithm can be compared and a system can be developed with the one having highest accuracy. The proposed system can help to identify an Individual's Credit Worthiness and extend support to financial institutions in deciding whether to grant or deny loan to the customer. The system can also help an individual to understand their credit worthiness and provide various investment recommendation so as to increase their overall credit score.**

Keywords—: Credit Worthiness, Credit Score, Logistic Regression, Machine Learning Algorithms

I. INTRODUCTION

Credit Score is one of the most important and critical attribute of an Individual to understand their Credit Worthiness and Market Credibility. There is a high rise in the financial sectors such as loan, investments, stocks which involves a lot of money, so a strategy build on the basis of just human experiences is not advisable. In money lending business there are always defaulters who may cause large losses which is not at all favorable on the part of the financial institutions. The institutions have in fact experienced many such losses historically. The money lenders have to decide some chosen set of criteria before providing loan to certain person, so as to minimize the losses and maximize the profit. Building an efficient model that can envisage the Client's Credit Worthiness and Credit Score can aid financial institutions in order to identify the defaulters and separate out them from good borrowers. This can help to maximize their profit and to come out among the top rated financial institutions. Recently with high rise in new technologies like data science, machine learning it's been easier to predict things and work accordingly. An ensemble strategy is combining or adding multiple classifiers that individually tests or perform to give results and get a single predictive output based on all the data. Ensemble model decrease the variance, bias and greatly improve the capability to predict with accuracy. Ensemble models are in high demands due to their impressive features and have a lot of potential to become one of the highly used methods in the financial field.

II. LITERATURE SURVEY

In paper [1] Customer financial Behavior is classified as a Good or a bad using algorithms of Machine Learning like Neural Network, Decision Tree, Logistic Regression, SVM is calculated. Attributes such as Age, Gender, Income, Dependents, Current Address, Married, Self Employed, City, Previous Loan, Loan Period, Job Title, House(own/Rented). Accuracies obtained were ANN-0.80 LR-0.82 Decision Tree-0.78 SVM-0.79. In paper[9] Customer Behavior as a Good(1) or bad(0) is classified using Logistic Discriminant Analysis, Artificial Neural Network, XGBoost, SVM was calculated using attributes such as Gender, Married, Income, Education, Dependents, Previous Loan ,House(own/rented),credit card, Number of Previous loans, Job type. Accuracies obtained were LDA-0.73 ANN-0.82 SVM-0.84. In paper[6] Credit score of a customer is classified using its Social Networking information considering peer to peer lending, online micro lending market etc. using Random Forest, Adaboost, Light GBM. Attributes considered were Social Stability, Social Exposure, Social Quality, Gender, Age, Income, Education, Job, Years at Residence, Credit History, City, Dependents, type of loan. Accuracies obtained were Random Forest-0.74 Adaboost-0.78. In paper[8] Credit Score is calculated to predict Good or a bad Customer using Oracle Data Miner, ROC Curves and Graphical analysis. Attributes considered were Office Details, other loans, Credit history, Nationality, Current address, income, dependents, Education, Gender, Marriage, Self Employed, DOB. Accuracies obtained were LR-0.82. SVM-0.79 ANN-0.76.In paper [7] Customer behavior as a Good or a bad is calculated using machine learning tools like Neural Network, Decision tree, Adaptive boosting, SVM, Logistic Regression. Gender, Marriage, Income, Dependents, Property Details, House(own/rented), type of loan, City, Monthly expense, previous loans, dependent income, savings, credit history of an individual were taken into consideration. Accuracy obtained were ANN-0.82 Decision tree-0.78 Adaptive boosting-0.76 SVM-0.84 LR-0.88.

Table 1. Accuracy Comparison

Research Paper Name	Attributes	Accuracy
A comparison of data mining techniques for credit scoring in banking: A managerial perspec- tive	gender, age, marital status, educational level, occu- pation, job position, income, customer type and credit cards from the other banks	LR-62.33 % CART-65. 58% Neural Networks-61.52%
EXPLORATORY ANALYSIS FOR CREDIT SCORE WITH APPLIED MACHINE LEARNING	age, gender, job,housing,credit amount,duration	CART and LR combined - 84%
An Empirical Study on Credit Scoring Model for Credit Card by using Data Mining Technology	sex,age, occupation, education,household register, marriage, personal monthly income and expenditure, family monthly income, family size, family economic status, card using frequency, monthly amount of card swiping, loan balance,	Logistic Model-86.3% CART-86.6%
Prediction analysis of risky credit using Data mining classification models	Credit, Balance_credit_acc, Duration, Rate, Age, Occupation,etc.	Random Forrest-79.68% Linear Regression-76.86% Support Vector Machine-76.72%
CREDIT SCORING USING LOGISTIC REGRESSION	Serious Delinquency in 2 years, Revolving Utilization of unse- cured Lines, Age, Debt Ratio, Monthly Income, Number of De-pendents,	LR-78%
An improved Bank Credit Scoring Model A Naïve Bayesian Approach	gender, age, tribe/sector, mari- tal status, consistency, years as resident, educational level.	Classification Tree-82% KNN-72.8% Naïve Bayes-83.3%
Using Data Mining Predictive Models to Classify Credit Card Applicants	age, annual income, gender, marital status, number of children, number of other credit cards, loan amount, loan duration, monthly salary.	LR-72.35% CART-72.72% Neural Network-76.58%

Table 1. describes the machine learning model's accuracies and attributes comparison.

III. PROPOSED WORK

The system proposes to provide credit score calculator for an individual and the financial institution as well. The financial institution can use the system to separate out good customers from defaulters and accordingly help them to decide whether the customer can be granted loan or not. The user on other hand can see their credit score and the probable investment options available to them to increase their credit worthiness and become eligible for availing loan facility.

A. Exploratory Data Analysis

The implementation phase of Project starts with Exploratory data analysis of training and testing dataset. We have used python language and Jupyter notebook for exploratory data analysis and various python inbuilt languages such as Numpy Panda, sklearn. Exploratory data analysis steps includes data cleaning and data preparation steps which includes removing outliers , treating null values, feature extraction and converting the categorical values into numerical values. Then we have splitted the dataset into training and testing dataset in the ratio of 80:20. Then we have selected the target variables. After that we have trained our model using training datasets and then validated our trained data with the test data by using numerous algorithms of Machine Learning like Logistic Regression, Random Forest, Classification and Regression Trees (CART), Naïve Bayes, KNN, Decision Trees and selected the Algorithm with highest Accuracy to predict our system.

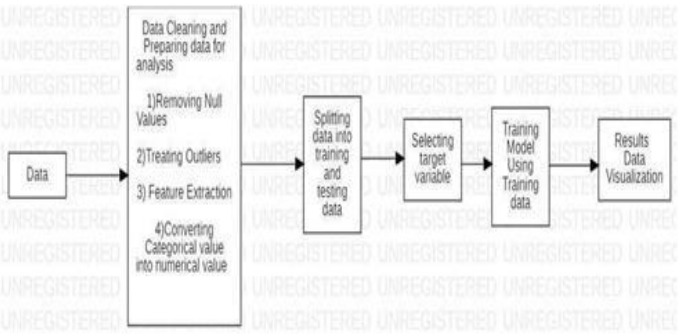

Fig1. Data Modelling

Fig 1 depicts how the exploratory data analysis of dataset would be performed in order to get model trained.

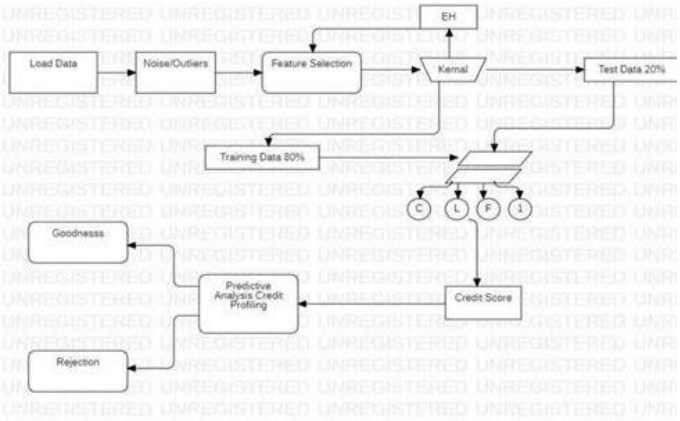

Fig2. Process Flow

Fig 2 represents the process flow of Exploratory data analysis and represents the entire model.

B. Dataset

The dataset we obtained is from a banking Institution. We have received two dataset first dataset contains the target variable and second dataset is use to predict the target variable. We have splitted our first dataset into training and testing dataset in the ratio of 80:20 respectively. The attribute list of the testing dataset is as follow:

Table 2. Dataset Attributes

Attributes	Values
Gender	Male/Female
Married	Yes/No
Dependents	0-5
Qualification	Graduate/Not Graduate
Self Employed	Yes/No
Applicant- Income	Integer Value
Co-Applicant Income	Integer Value
Previous Loans	Yes/No
Previous Loan Amount	Integer
Previous Loan Period	Number of Months
Property	Urban/Semi Urban/Rural
Monthly Debt Ratio (Monthly Living Cost/Monthly Gross Income)	Integer Value

Table 2. represent the list of attributes taken into consideration from the dataset.

C. Machine Learning Techniques

1) Logistic Regression

The model that will be used for designing the credit score system is Logistic Regression Model. It is one of the most commonly used Machine Learning Algorithm utilized in Credit Scoring Model. While comparing Logistic regression and linear regression it is observed that both alter in their results and outcome, as the outcome of latter is continuous and the outcome of former one is discrete. Logistic regression establishes the relationship between the independent variables where the number of variables can be single or multiple. The probability calculated will have value which will either be 0 or 1. Build upon independent variable's values we calculate the value of probability of the dependent variable which could be either 0 or 1. Logistic Regression is generally practiced for foretelling and prediction. Marketing, illness, profit, probability of failure of a given process, and estimated with the help of Logistic regression approach.

LR FORMULA

$$P(x) = \frac{e^{\beta_0 + x \cdot \beta}}{1 + e^{\beta_0 + x \cdot \beta}}$$

We can also re-express the above equation as

$$log(\frac{P(x)}{1 - P(x)}) = \beta_0 + x \cdot \beta$$

Fig.3. Logistic regression formula.

Fig.3. describes logistic regression formula which calculates the probabilistic score.
where,
x-target variable
P- probabilistic score of target variable x
βo & β -parameters of the model
The above formula is taken in reference with [9]

Table 3. Performance Matrix of Logistic Regression

		PREDICTED CLASS		
		Non-default	***Default***	SUMMARY
Actual Class	***Non-default***	288	2	**290**
	Default	9	68	**77**
	SUMMARY	**297**	**70**	**367**

Table 3. represents the performance matrix of Logistic Regression in credit scoring model.

2) Linear Discriminant Analysis

Linear Discriminant Analysis is a Machine Learning Algorithm which, a classification method used to establish a linear relationship between attributes which classifies between 2 or more classes. The resultant combination is taken such that attributes are very important are considered which

is further is applied for reduction of dimensions before the eventual classification.

Any individual who wants to determine the discriminant dimension in the response pattern space while Dimensions reduction, Linear discriminant analysis is the best choice applicable that, where we maximize the between-class to within-class variance of the applicable data ratio.

Thus, the said algorithm is simple and has good accuracy when compared to other algorithms. In the proposed model, our aim is to predict the credit worthiness of the customer to receive loan and credit. We do this by calculating the probabilistic value whether the customer will be a defaulter or not. So for this purpose we can use Linear Discriminant Analysis to determine whether the customer is good or bad using the customer details.

Table 4. Performance Matrix of Linear Discriminant Analysis

		PREDICTED CLASS		
		Non-default	***Default***	***SUMMARY***
Actual Class	***Non-default***	275	15	**290**
	Default	27	50	**77**
	SUMMARY	**302**	**65**	**367**

Table 4 represents the Performance Matrix of Linear Discriminant Analysis in the Credit Scoring Model

3) Decision Tree

Decision Trees is a tree based learning algorithms and has high efficiency and used mostly in the field of machine learning techniques. Tree based techniques have higher accuracy, good stability and ease of research. Decision Tree algorithms establishes non-linear relationships between attributes such that the accuracy obtained is good. Any conversion in the input do not directly change the outputs as Decision trees are non-linear. Here, we divide the dataset into small subsets and decision tree is designed in an incremental manner. These steps lead to creation of decision nodes and tree nodes.

Table 5. Performance Matrix of Decision Tree

		PREDICTED CLASS		
		Non-default	***Default***	***SUMMARY***
Actual Class	***Non-default***	225	65	**290**
	Default	16	61	**77**
	SUMMARY	**241**	**126**	**367**

Table 5. represents the performance matrix of Decision Trees in credit scoring model.

4) Random Forest Algorithm

Random Forest is an all-round Algorithm which has features regression as well as classification tasks. It also deal with missing data, removing outliers, commence dimensional reduction technique and other learning methods. It designs a strong model by combining weak models. Each tree assigns a classification to classify a new object based on its attributes, i.e. voting is carried out for that specific selected class. This Algorithm then selects the tree having the most votes (from the list of all trees in forest) and then takes average of different outputs to compute the results in the case of regression.

The algorithm used is as follows:

1. Let N be the number of cases in the training set. Then, sample of these N cases is taken in irregular sequence but with replacement. The sample obtained is used as the training set for further developing the tree.
2. If number of attributes is M, m a number (where m is less than M) is stated such that at each node, m variables are taken at random. While we grow the forest further the m value is set constant.
3. Every tree is grown to its maximum size possible and no pruning is performed.
4. Taking the aggregate of the predictions of the n tree trees (i.e., majority votes for classification, average for regression technique) we can estimate new data.

Table 6. Performance Matrix of Random Forest

		PREDICTED CLASS		
		Non-default	***Default***	***SUMMARY***
ActualClass	***Non-default***	262	28	**290**
	Default	9	68	**77**
	SUMMARY	**271**	**96**	**367**

Table 6. represents the performance matrix of Random Forest in credit scoring model.

5)Naïve Bayes

Thomas Bayes formulated Bayes' Theorem, The Naive Bayesian algorithm is based on the Bayesian theorem with assumed independence among the predictors, to estimate the probability terms required for classification a training dataset is used. This performance is quantified by the accuracies of the predicted required probability terms, which is always a challenging part as the training dataset is not easily available. Notwithstanding this limitation and despite the availability of numerous classifiers, Naive Bayesian Algorithm stands amongst the most admired classifiers and graded among the top 10 performing Machine Learning algorithms for its clarity, empirically and efficacious. Naive Bayesian has been applied and witnessed efficacious in various domains and area of interest including Finance, Agriculture, medicine, and biometrics.

Table 7. Performance Matrix of Naive Bayesian

		PREDICTED CLASS		
		Non-default	***Default***	***SUMMARY***
Actual Class	***Non-default***	282	8	**290**
	Default	19	58	**77**
	SUMMARY	**301**	**66**	**367**

Table 7. represents the performance matrix of Naive Bayesian in credit scoring model.

D. Comparison of Credit Scoring Model

For comparing the performance of all the algorithms, we compare Performance Matrix obtained for all the algorithms. The testing dataset used for this purpose has 367 rows of data where there are 290 non-default values (Customers whose loan has been approved) and 77 default values (Customers whose loan has been rejected).

In the performance matrix we have 2 classes, Actual Class and Predicted Class. Actual Class represents the actual values of target attribute in testing dataset whereas the Predicted Class represents the values obtained after prediction. The summary at extreme right represents the non-default and default values for actual class and summary at bottom represents the non-default and default values for predicted class.

To calculate the performance of credit scoring potential of developed credit scoring models, we have to calculate the error which is incurred. It is obvious that the misclassification costs linked with two types of errors i.e.-

1.Type-I - A good client having good credit is confused or classified wrongly as a bad client

2.Type-II- A bad client having bad credit is confused or classified wrongly as a good client.

Then after comparison, the performance of Logistic Regression was found out to be best where 288 non-default values were correctly predicted out of 290 and 68 default values were correctly predicted out of 77. Here, type I error was obtained to be as 0.68% and type II error was obtained to be as 10.38% and overall error thus incurred was 5.53% which was lowest when compared to all other algorithms.

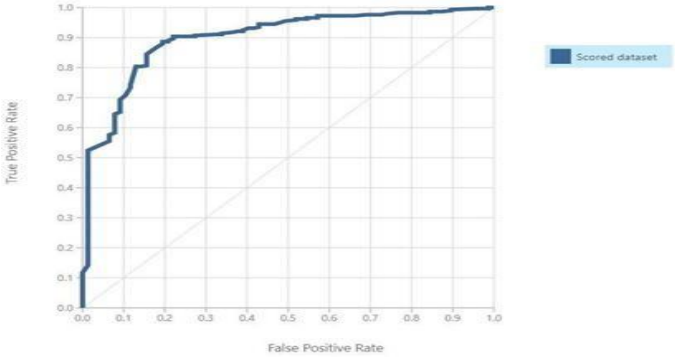

Fig 4. ROC Curve

Fig 4. represents ROC- curve of a credit scoring model using Logistic Regression. Area under the ROC curve represents the model accuracy.

IV. Results and Discussion

Applying different algorithm on data sets the following results can be discussed.

Accuracies of Algorithm used in the model can be summarized as follow:

Table 8. Algorithm Accuracy Comparison

Algorithm	**Accuracy**
Logistic Regression	0.94
KNN	0.73
SVM	0.81
Naïve Bayes	0.92
Decision Tree	0.78
Random Forest	0.90
LDA	0.84

The above accuracies have been obtained using Scikit-Learn Machine Learning Library. We have divided our dataset into training and testing dataset, attributes in both the dataset were stored in four variables, two variables for each dataset. One variable of each dataset would contain all the attributes other than the target variable and the second variable contains the target variable. Scikit-Learn(sklearn) machine learning library features various inbuilt functions, classification, regression and etc. Using sklearn library we have imported its various inbuilt functions and algorithms functions. All the above accuracies results are obtained using sklearn library inbuilt algorithm functions. Initially we imported each algorithm function from sklearn library, this algorithm is then trained with the model preprocessed data with the help of (.fit) method. Once model is proper-ly trained we use the (.predict) method on the testing dataset(on attributes ex-cluding the target variable) and obtain the results to get the target attribute value for our testing data, thereafter we compare our predicted results with the actual results and using accuracy score method we calculate our algorithm accuracies.

A. Data Visualization

Here we have observed each of our dependent attribute vs the target variable comparison and results obtained are as follow:

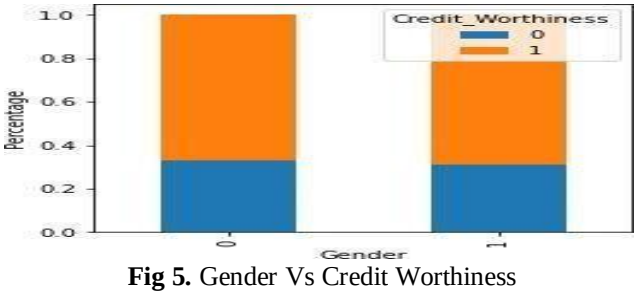

Fig 5. Gender Vs Credit Worthiness

Fig 5. represents comparison of Gender and Credit worthiness

0-Female 1-Male

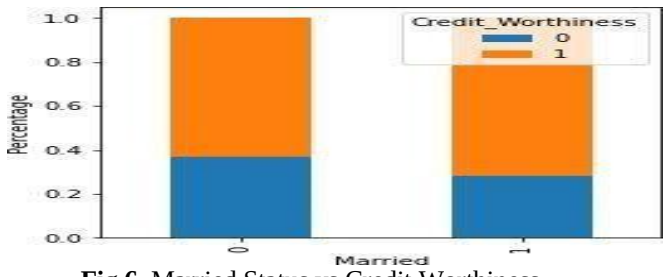

Fig 6. Married Status vs Credit Worthiness

Fig 6 represents comparison of Married Status and Credit worthiness

0-Not Married 1-Married

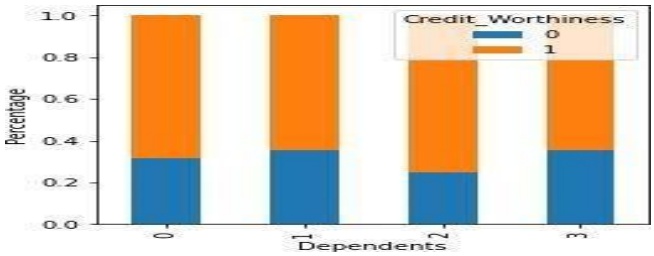

Fig. 7. Dependents vs Credit Worthiness

Fig 7. represents comparison of Dependents and Credit worthiness

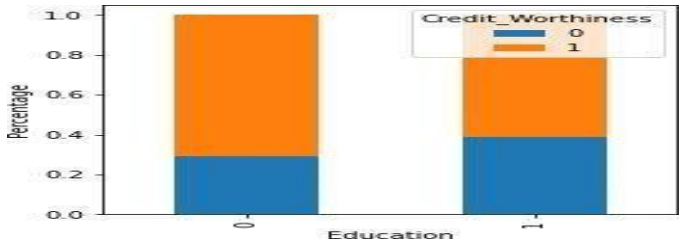

Fig. 8. Education vs Credit Worthiness

Fig 8 represents comparison of Education and Credit worthiness

0-Graduate 1-Not Graduate

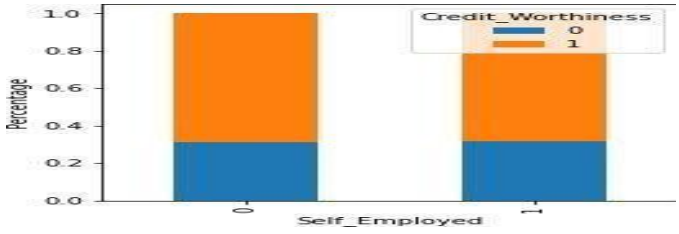

Fig. 9. Self Employed vs Credit Worthiness

Fig 9 represents comparison of Self Employed and Creditworthiness

0-No 1-Yes

Fig. 10. Property Area vs Credit Worthiness

Fig 10 represents comparison of Property Area and Credit worthiness

0-Rural 1-Semi Urban 2-Urban

V. Conclusion

The System was able to predict Credit Score by using numerous algorithms of Machine Learning like Logistic Regression, Classification and Regression Trees (CART), Random Forest, Naïve Bayes, KNN, Decision Trees to predict the Credit Score and Credit Worthiness of an individual and compared the accuracies of the Algorithm. Our Study found that Logistic Regression among all other Machine Learning Algorithms has the highest accuracy in Credit Score Predicting Model and thus we selected Logistic Regression Algorithm as a tool for Credit Score Prediction System.

VI. References

[1] Huseyin Ince and Bora Aktan, **"A comparison of data mining techniques for credit scoring in banking: A managerial perspective",** 14 Oct 2010.

[2] Jasmina Nalić and Amar Švraka , **"Using Data Mining Approaches to Build Credit Scoring Model"** ,17th International Symposium INFOTEH-JAHORINA, 21- 23 March 2018.

[3] Durgesh Kumar Singh and Noopur Goel, **"Analysing Data Mining Techniques on Bank Customers for Credit Score"**, 2020 8th International Conference on Reliability, Infocom Technologies and Optimization, June 2020.

[4] Archana Gahlaut, Tushar , Prince Kumar Singh , **"Prediction analysis of risky credit using Data mining classification models",** 8th ICCCNT 2017 IIT Delhi , July 3-5, 2017.

[5] Wei Li and Jibiao Liao , **" An Empirical Study on Credit Scoring Model for Credit Card by using Data Mining Technology"** , 2011 Seventh International Conference on Computational Intelligence and Security.

[6] Beibei Niu, Jinzheng Ren , Xiaotao Li , **"Credit Scoring Using Machine Learn- ing by Combing Social Network Information: Evidence from Peer-to-Peer Lending"** , 17 December 2019.

[7] Ali Al-Aradi , " Credit Scoring via Logistic Regression" ,February 28, 2014.

[8] Ansen Mathew , " Credit Scoring Using Logistic Regression" ,24 th May 2017.

[9] Pratik Sharma, Sunil Bhatia, Rohit Burman, Santosh Hazari, Rupali Hande, "Credit Scoring using Machine Learning Techniques" , March 2017.

[10] ROHIT KUMAR, MAYUKH BISWAS, SANGRAM MONDAL, "Exploratory Analysis For Credit Score With Applied Machine Learning", May 2020.