Customer Loan Eligibility Prediction Using Machine Learning Boosting Algorithms

Divya Evney[1] Prof. Ajit Shrivastava[2] Prof. Rohit Bansal[3]
[1]M.Tech. Scholar [2]Professor [3]Assistant Professor
[1,2,3]Department of Computer Science Engineering
[1,2,3]SISTec-R, Bhopal (M.P), India

Abstract — The In recent days, data mining has become very important for gaining vital information in Loan Granting Industries Like Home Loan, Personal Loan, Business Loan any many more Loan Services Available Now Days. For This kind of Industries Machine Learning and Deep Learning industries. Any Housing Finance company in all kind of Different loans. Presence across all urban, semi urban and rural area. Customer first applies for a home loan. Company validates the customers eligibility for the loan. Company wants to automate the loan eligibility process. Gender, Marital Status Number of Dependents, Income Loan Amount Credit History and Many More. To enhance their business in better way these types of facilities, may enhance business as well as customer satisfaction. This Research deals with predicting the eligibilities for applicants who applied for any kind of loans from any financial institution. Here various data mining or machine learning can support for this kind of work like many classifications Algorithm Regression, Naïve Bayes, Support Vector Machine Decision tree & Random Forest and many more. A comparison has been done between the actual and predicted expenses of the prediction premium and eventually, a graph has been plotted on this basis which will enlighten us to choose the best-suited Algorithm. The Selected Algorithm will be applied for our proposed work i.e., Loan Prediction. for prediction, correctness has been measured by the Coefficient of determination. Gradient Boosting Classifier gives the best result in terms of Accuracy i.e. 0.9125 which can be used in its best possible way for the correct prediction of the Loan Prediction Guarantee for companies as well as Customers.

Keywords: Machine Learning, Classification, Logistic Regression, Gradient Boosting Machine & Random Forest

I. INTRODUCTION

A recent development of machine learning techniques and data mining has led to an interest of implementing these techniques in various fields. The banking sector is no exclusion and the increasing requirements towards financial institutions to have robust risk management has led to an interest of developing current methods of risk estimation. Potentially, the implementation of machine learning techniques could lead to better quantification of the financial risks that banks are exposed to.

A. Credit Risk in Banking

Here The various financial risks banks confront can be broadly classified as credit risk, market risk, liquidity risk and interest rate risk [4]. Author's explains credit risk as the risk of a debtor defaulting his or her loan, which leads to losses for the lender. Authors elaborates that credit risk includes that a group of borrowers or a counterparty fails to meet its obligations, or an investment deteriorates and defaults and explains that loans are the most common source of credit risk for banks. However, financial instruments such as bonds, swaps, options and interbank transactions all include credit risk.

B. Domains Distribution of Loan

Fig. 1: Domains Distribution of Loan

Personal Loan: Most banks offer personal loans to their customers and the money can be used for any expense like paying a bill or purchasing a new television. Generally, these loans are unsecured loans.

Home Loan: When you wish to purchase a house, applying for a home loan can help you to a great extent. It provides you the financial support and helps you buy the house for yourself and your loved ones.

Credit Card Loan: Loan against credit card is like a personal loan that is taken against your credit card. These are usually pre-approved loans that do not require any additional documentation.

C. Scope

The scope of this Dissertation is to implement and investigate how different Machine Learning Algorithm impact default prediction. The model evaluation techniques used in this project are limited to precision, sensitivity, F-score and AUC score. The reasons for choosing these metrics will be explained in more detail When we Explain Result Analysis.

All rights reserved by *www.ijsrd.com*

D. KDD Process

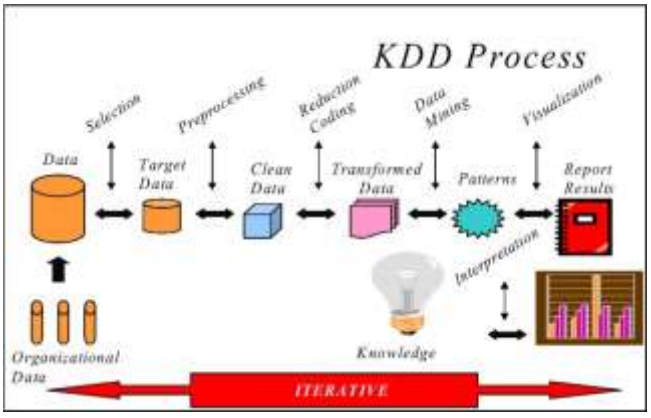

Fig. 2: KDD Process

II. Related Work

Here Here Author's found out that the decision tree model outperformed other models in average classification rate, but neural networks had lower type II errors and therefore had better overall results. Their research used a Turkish bank data set with 1260 loans. The sensitivities and specificities of their study were: Discriminant analysis had a sensitivity of 57.81%, a specificity of 67.09%, logistic regression's sensitivity was 65.06% and specificity 60.10%, neural networks had a sensitivity of 74.10% and a specificity of 51.23%, decision tree's sensitivity was 62.05% and specificity 68.47%.

Health Here Author's compared 41 different learning algorithms in their study and utilized eight retail credit scoring data sets. Their study compared individual classifiers (e.g., logistic regression) to more advanced classifiers (e.g., random forest). They mention that logistic regression is an industry standard for credit scoring predictions. In addition, they state that several classifiers, such as random forest performs better than logistic regression. Wang et al. (2018) conducted a default probability study on a peer-to-peer lending data set. They compared ensemble mixture random forest model (EMRF) with a standard mixture cure model, the Cox proportional hazards model and logistic regression model. They stated that their EMRF model outperformed all the other models based on the mean area under the ROC curve.

III. Problem Identification

Many As we all know that, banks face various financial risks, including different Loan risk. This research focuses on studying default risk, which is one of the credit risk components obliged in the Basel II regulation as explained. Since managing Loan risk is crucial for banks and calculating default risk is obliged, the objective for this research is to understand how loan granting is regulated, and how machine learning is utilized in loan granting.

IV. Algorithms

Step 01: Store Data from Kaggle Repository
Step 02: Import Prior Libraries
Step 03: Now Import our Required Dataset
Step 04: Apply Descriptive Statistics for the Continuous Data
Step 05: Dynamics Statistics
Step 06: Applying Data Cleaning & Missing Values
Step 07: Applying Machine Learning Algorithms
Step 08: Apply Different Model
  a) Linear Regression
  b) Gradient Boosting
Step 09: Repeat Step07 for many times with different Algorithms
Step10: Finally Compare Results with performance parameters like RMSE Score & R2 Score.
Step11: Stop

V. Flow Diagram

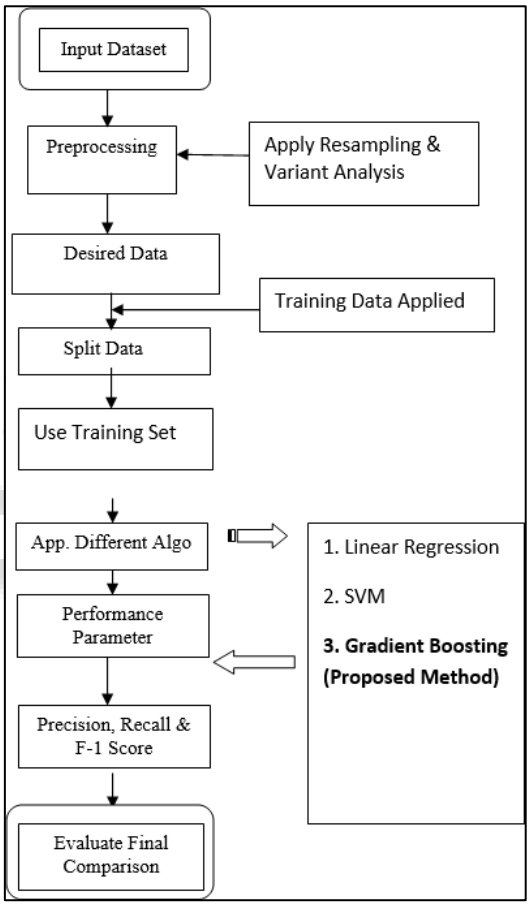

Fig. 3: Flow Diagram

In In figure 3 At first step, we need to fetched Data from any external source or we can collect Data from Local Market but for better Analysis we are Fetching our Data from Kaggle. That is very reliable Data Source through Word Wide. In Next Step we need to Fetched Different Libraries for processing our Data. At very next Step that is Third Step we need to process our Data for next step Here we have many processing Mechanism. Fourth Step we need to repeat it for different Data split. Then after we will reach at step 5 where we will apply Different Machine Learning Algorithms and Finally, we will apply our own Proposed Methods i.e., Gradient Boosting with Weighted Average values. Here we will adding average values of previous implemented mechanism. At Final Step i.e., Sith step we will have to find Performance Measures i.e., F-1 Score, Recall & Precision. At

All rights reserved by *www.ijsrd.com*

Final Step we will compare these given Results. We can say that Our Proposed Methods gives better Result.

VI. PROCESS OF IMPLEMENTATION

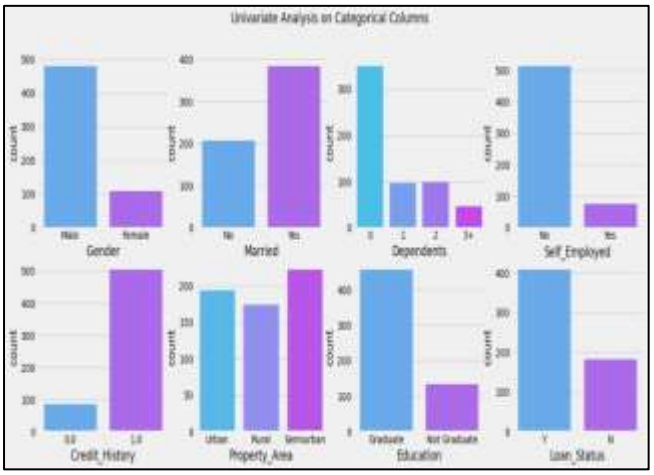

Fig. 4: Uni Variate Analysis Over Categorical

In Figure 4 researchers explained How thesecolumns play vital roles here:

- Case-1: Number of Male Applicant is more than Female Applicants.
- Case-2: Married Applicants are more in numberthan unmarried Applicants.
- Case-3: The number of Depends increasing less the applicants count.
- Case-4: Self-employed Attempted less for Loan Case-5: Credit History play vital roles here we cansee that if your credit limit is less your possibility isless.
- Case-6: Property Area: According to observationurban property values is good values than semi and rural area.

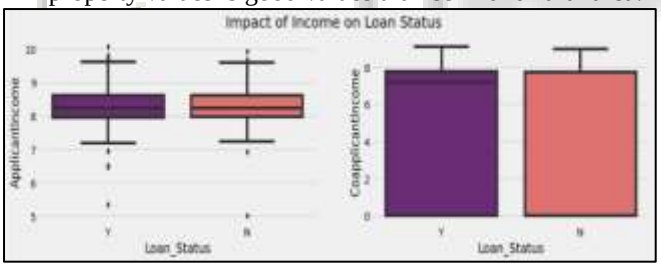

Fig. 5: Polar Analysis

In figure 5 Here Unmarried People rejected loan 78 Times and selected for 131 Times. In second Case a Married People Rejected Loan 108 Time andSelected for 278 Times. Here we are getting clearcut picture that married people has more credibility than unmarried People.

A. Output Results

In the table 1 we are trying to show the Accuracy Results of Different Algorithms which we implemented.

Algorithms	Training Accuracy
Linear Regression	0.77
Support Vector Machine	0.87
Gradient Boosting (Proposed Methods)	0.91

Table 1:

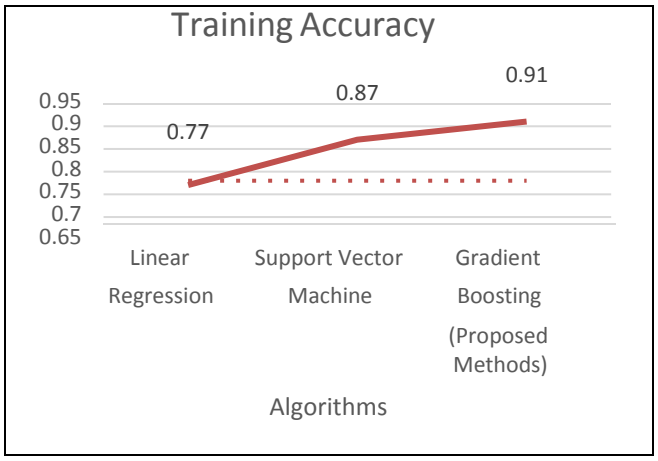

Fig. 6: Comparison Line Graph (Training)

In figure 6 by the analysis of above graph we can say that we find that Liner Regression, Support Vector Machine, Gradient Boosting (Proposed Method) gives Different Results respectively. Here

We are going to propose some tuning mechanism so that Results get improved.

VII. CONCLUSION

Here researches conclude We conclude that when we Implemented Number of Machine Learning Algorithms for finding best results in terms of performance. Modelling our data with Logistic Regression and Gradient Boosting. But Gradient Boosting did a little better as comparison to the Logistic Model. As Training data is less so Gradient Boosting is a good Model, as the cross-Validation Scores also came very good. We used Accuracy, Precision and Recall to evaluate Model Performance. We Finds the major Features of given Data set are Education, Marital Status, Income Group, Living Area. Most Important Factors to Predict the Guarantee of Loan is given above. Finally, we Implemented Different Machine Learning Algorithms Like Linear Regression, Support Vector Machine & Finally our Proposed Method gives results respectively i.e., 83%, 87 % and finally 91%. When we looked into our proposed methods then we can claim that with existing system our proposed methods give better results in terms of Accuracy. Apart from Accuracy we explained various Graphs from which we come for better understanding about our Implemented Models.

VIII. FUTURE SCOPE

In The future works focus on applying some other techniques to improving the performances of these methods for up to maximum extent. Another concept that can be implemented Deep learning in place of machine learning technology. The reason behind this is best and efficient techniques using nowadays. Deep learning is also introduced nowadays which is becoming more popular for classification purpose. If we have imbalanced data, we need to apply Resampling Techniques to make it balanced. If we have skewed data then we need to apply some Data Transformation techniques. We must perform Univariates and Bivariate Analysis to understand the Better. Finally, try more and more Predictive Models, compare them using various Evaluations Metrics is

All rights reserved by *www.ijsrd.com*

a good way of finding the best models. So, we can also implement deep learning in future work also.

REFERENCES

[1] Bullivant, G. (Ed.). (2016). Credit management (6th ed.). Taylor & Francis Group.
[2] Bandyopadhyay, A. (2016). Managing portfolio credit risk in banks. Cambridge University Press. https://doi.org/10.1017/CBO9781316550915.
[3] Messan Komi, J un Li ,"Application of Data Mining Methods in Diabetes Prediction",2017 2nd International Conference on Image, Vision and Computing.
[4] Bessis, J. (2015). Risk management in banking (4th ed.). John Wiley & Sons, Incorporated
[5] Silva, E.C., Lopes, I.C., Correia, A. & Faria, S. (2020). A logistic regression model for consumer default risk. Journal of Applied Statistics, 47(13- 15), 2879-2894. https://doi.org/10.1080/02664763.2020.1759030.
[6] Camilla Cal`ı and Maria Longobardi. "Some mathematical properties of the ROC curve and their applications". In: Ricerche di Matematica 64 (Oct. 2015). doi: 10.1007/s11587-015-0246-8.
[7] Tianqi Chen and Carlos Guestrin. "Xgboost: A scalable tree boosting system". In: Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining. ACM. 2016, pp. 785–794.
[8] Andreas Christmann. Support Vector Machines. eng. Information science and statistics. 2008. isbn: 1-281-92704-X.[9] Feature Analysis Visualizer. Recursive Feature Elimination. https://www.scikit-yb.org/en/latest/api/features/rfecv.html. Accessed: 2019-05-24.
[9] James Finance. "Machine Learning in Credit Risk Modeling: Efficiency shouldnot come at the expense of Explainability". In: (July 2017).
[10] Jean D. Gibbons. Nonparametric Measures of Association. Thousand Oaks, California: SAGE Publications, Inc.,

All rights reserved by *www.ijsrd.com*