Prediction of Loan Defaulter Using Different Approaches

Dev Jethva
Pandit Deendayal Energy University
Gandhinagar, Gujarat, India
Email: 23mds003@sot.pdpu.ac.in

Vipul Vasava
Pandit Deendayal Energy University
Gandhinagar, Gujarat, India
Email: 23mds014@sot.pdpu.ac.in

ABSTRACT - The problem of loan defaulters has emerged as a crucial issue in the modern financial environment, placing significant strain on the stability of financial institutions and the nation's economy. This study makes use of a thorough analysis of many economic, social, and psychological aspects in an effort to clarify the complex interactions of factors that contribute to the phenomena of loan defaulting. This study explores the root causes of defaulting behavior using a multi-dimensional approach, illuminating the complexity of borrower-lender relationships as well as the larger socioeconomic environment. This approach also explores how macroeconomic changes, regulatory structures, and cultural factors affect a person's propensity to default, emphasizing the necessity of an all-encompassing regulatory framework that includes both preventative and corrective actions. If the banking system not adapt with the growing world in the system of analysis which can provide the benefits to the banking system. Loan defaulters is the problem that banking system can face. Study of this problem can give the benefits of grooving loan customers and through that bank can make a profit.

Keywords- *Machine Learning, Data Science, Random Forest, KNN, SVM, Loan prediction, data mining,* Defaulter, Naive Bayes, Gradient Boosting, Prediction.

I. INTRODUCTION

This project has taken the data from the Kaggle. So, the machine learning model is trained on that record to get accurate results. we will also develop a basic understanding of risk analytics in banking and financial services and understand how data is used to minimise the risk of losing money while lending to customers. The main aim of this project work is for loan safety. One of the first industries to use data science was finance. Financial organizations experience losses and bad debits each year. However, they begin by doing paperwork. when approving a loan, a lot of data is obtained. Since the banking industry has improved their study of by using customer profile to determine the likelihood of risk, past purchases, client transaction patterns, etc. The field of data science combines a variety of statistical tools, with algorithms and machine learning methods, you may latently pattern in the data that can be used to generate new insights. Currently, financial institutions adopt the use data science techniques to

banking advantage to research the giving the banking information for each individual consumer, and providing the customers with the most relevant services based on their financial history. We have data of loan defaulter with the record of 307511 and 122 attributes.

II. Literature survey

Logistic regression is a popular and very useful algorithm from machine learning to classification problems. Profit The thing about logistic regression is that it is predictive analysis. It is used to describe data and explain relationships between one binary variable and one or more nominal, follow-up, and dose-level variables that are independent in nature. Even income verified loans sanctioned loans higher probability of loan default. There is a high chance of subsoil degradation of loan default, so we can consider the subcategory function instead of a generic class function. It is identified by the home ownership feature that we cannot consider our home ownership loan approval priority. Many papers and the project related to the problem statement the accuracy of the model is between 75 to 80 percentage.

III. Methodology

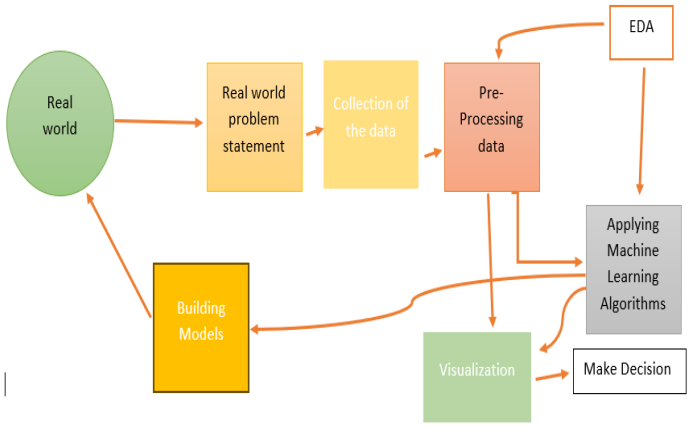

Fig. 1. Methodology

A. Business Understanding

First, we have to decides that which is the problem in the real world so that we can go with the problem statement which is very important factor because the business understanding is depends on the next two steps which is collection of the data and the analysis.

B. Understand the data

We know that the business problem so after the understand the business problem we have to understand the data that which type of the data we need for further analysis and the model building so that we can give the best model and the accurate result from the data.

C. Pre-processing Data

After understanding the data, we have selected the feature which is more important for the analysis. Understand the noise from the data that what kind of data noise are there with data and also handling with the missing values and the what kind of value should be there in the features and dealing with null values from we get know about the noise and null values in the data. Preprocessing is the significant part of the data science if we not do the step we are

not able to give an accurate answer or the result.

D. EDA (Exploratory data analysis)

Tabular data can be represented Through the graph which give the more analysis, more information and meaningful insights from the graph. EDA can be used for the imbalance data so, that we cannot compromised with the result and get accurate result.

E. Modelling data

Modelling the data is the most important stage that we decide how to solve the problem statement with the classification model or with the regression model so we need to understand the problem statement.

F. Evolution of Model

The model can be evaluated by the data. Data which is collected and pre-processed which can be split in parts training and testing. Then applying to the model train the model so that model can give accurate result on the real time data.

G. Deployment of the Model

Model which is evaluated is deployed so that it can give the Realtime result. Given above all the steps are important if we miss the any than we can be suffered in the real time problem result and need to start from first.

IV. Implementation

A. Collected the data

Collection of data is most time consuming from the real world but now a days many techniques and technology are used for the collection of the data. This paper uses the application data which is available on the Kaggle. Dataset consist of loan defaulter with the record of 307511 and 122 attributes.

B. Pre-processing of the data

from we get know about the noise and null values in the data. Then we are removing the columns with missing values more than 40%. We count 372235 missing values after removing columns so we fill with Mean for the appropriate columns. remove the noise and null data from the dataset. After that we applying EDA for the analysing and summarising the main characters from that

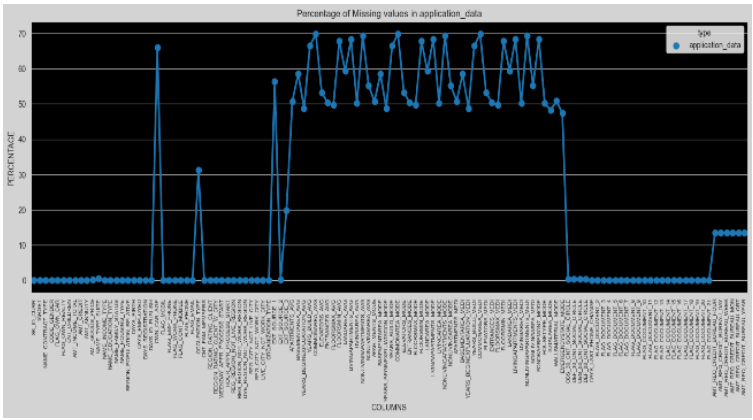

Fig. 2. Missing Values before Preprocessing

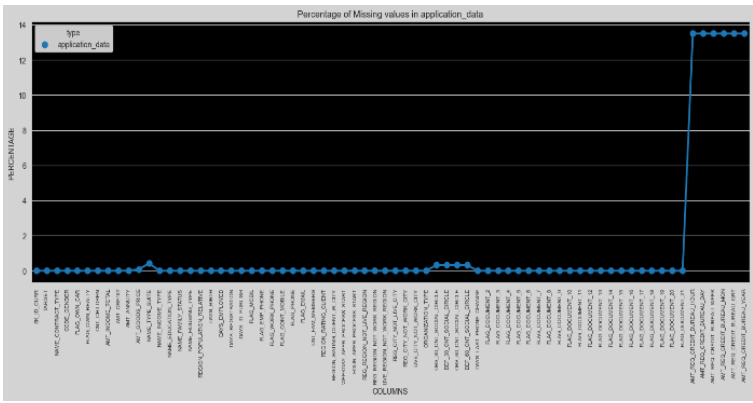

Fig. 3. Missing Values After Preprocessing

V. Model Building

Based on the loan issue date taking into account the earliest date of loan issuance dataset separation is carried out. In this work, a model was constructed using the Grid Search CV technique on a high volume of

records, as opposed to using the default algorithm process.

A. SVM (Support Vector Machine)

In the field of neural networks and machine learning, support vector machines have become a potent contender for classification. Support vector machines, as per the research, are supervised learning models that employ the principles of most learning algorithms to analyse binary classes. A hyper-plane is a component of an SVM model that is used to divide up the various observations for classification purposes. SVM is called for its fast execution time and accurate results in classification.

B. Naive Bayes

Naive Bayesian is considered quite robust and is popular for its simplicity in problem of classification. Using Naive Bayes, the researchers conducted many experiments.

C. KNN (K-Nearest Neighbour)

Regression and classification issues can be resolved with the KNN technique. Feature-scaling is required for this technique. Although the approach produces identical results to logistic regression, its primary limitation is that larger datasets require more processing time.

D. Random forest

One method of ensemble classification is Random Forest. According to earlier research, random forests are more effective than logistic regression in handling substantial datasets. This work model is asked on training dataset that is not standardized, yielding minimal poor accuracy yet comparable categorization metrics outcomes Degradation using logistics.

E. Gradient Boosting Algorithm

Gradient Boosting is a machine learning technique for both regression and classification issues. It creates a prediction model by building an ensemble of poor prediction models. The weak classification algorithm is sequentially applied to modified versions of the data multiple times in the boosting algorithm, producing a sequence of weak classifiers. Gradient Boosting is an ensemble learning technique that builds a strong classifier by combining multiple weak decision trees. Together, these decision trees create a potent gradient boosting model.

VI.Results of Experiment

Providing all the information and the performance of the models.

Table 1
Results of Experiment

Model	Performance of the Models			
	Test Accuracy	Train Accuracy	Precision	Recall
Support Vector Machine	91.76	90.86	1.0	0.015
Naïve Bayes	15.33	17.9	0.101	1.0
K-Nearest Neighbour	91.43	91.06	0.66	0.076
Random forest	91.76	100.0	1.0	1.0
Gradient Boosting Algorithm	91.57	98.71	1.0	0.86

We observe that the SVM and Random Forest models are given the more effective performance for the data of the loan defaulter.

A. ROC Curve

a. SVM

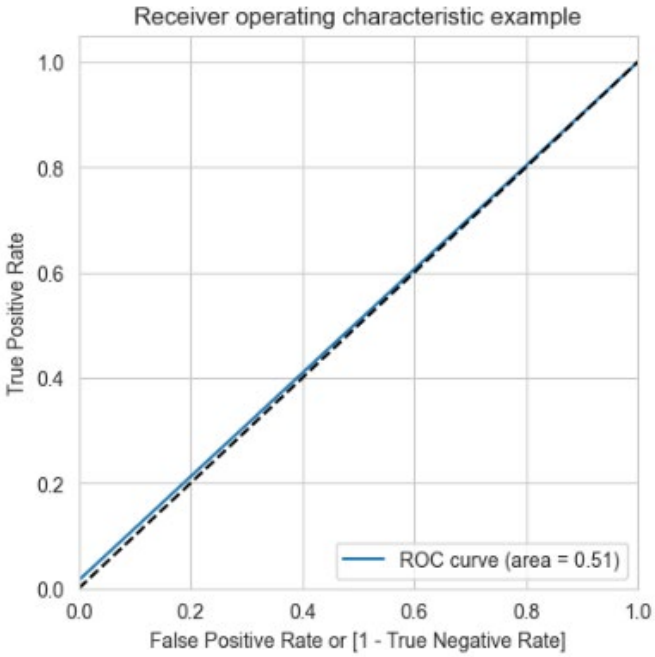

Fig. 4. SVM ROC Plot

b. Naïve Bayes

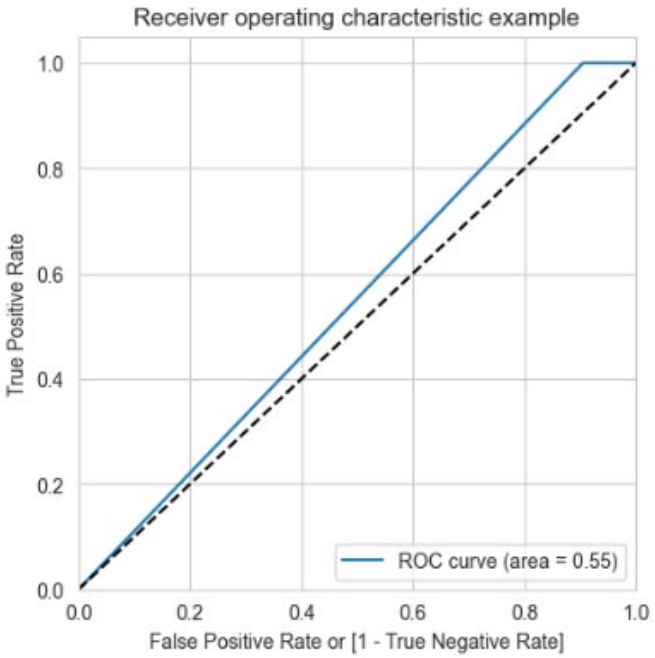

Fig. 5. Naïve bayes ROC Plot

c. K-Nearest Neighbor (KNN)

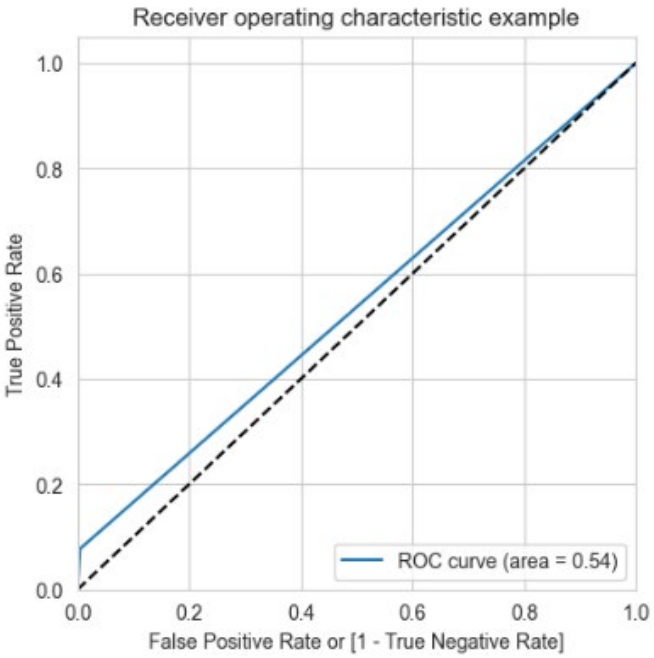

Fig. 6. KNN ROC Plot

d. Random Forest Classifier

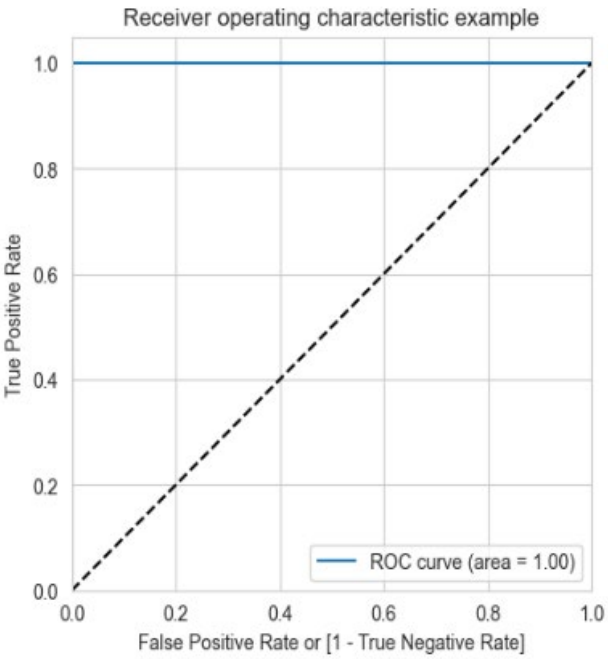

Fig. 7. Random Forest Classifier ROC Plot

e. Gradient Boosting Algorithm

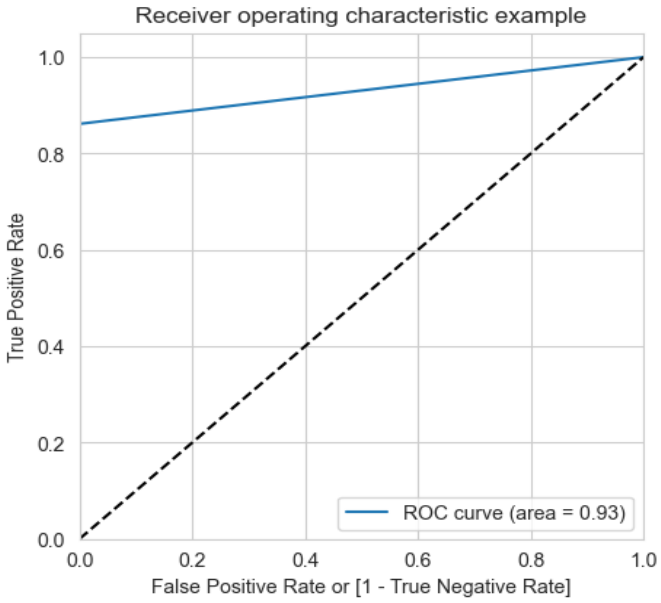

Fig. 8. Gradient boosting ROC Plot

VII. CONCLUSION

In this paper, it was discussed how applying data science to preprocess customer historical data and construct a model using machine learning techniques can help the banking industry better analyse risk. Owing to the massive amount of data processing, GridSearchCV was used to build the models using a cross-validation approach. The KNN model and random forest classification techniques are constructed; thus far, the two algorithms produce results that are nearly identical.

Observation:

A. If the income can be verified and approve loan have higher probability of the loan default.
B. 8% of the population can be a loan defaulter
C. Loan defaulter is more in the less rated city area compared to the rated city area.

VIII. REFERENCES

[1] Aslam, Uzair & Aziz, Hafiz Ilyas Tariq & Sohail, Asim & Batcha, Nowshath. (2019). An Empirical Study on Loan Default Prediction Models. Journal of Computational and Theoretical Nanoscience.16.3483-3488. 10.1166/jctn.2019.8312.

[2] Aslam, Uzair; Tariq Aziz, Hafiz Ilyas; Sohail, Asim; Batcha, Nowshath Kadhar, Journal of Computational and Theoretical Nanoscience, American Scientific Publishers, Volume 16, Number 8, August 2019, pp. 3483-3488(6)

[3] Lifang Zhang, Jianzhou Wang, Zhenkun Liu, what should lenders be more concerned about? Developing a profit-driven loan default prediction model, Expert Systems with Applications, Volume 213, Part B,2023, 118938, ISSN 0957-4174

[4] Gross, Jacob P.K.; Cekic, Osman; Hossler, Don; and Hillman, Nick (2010) "What Matters in Student Loan Default: A Review of the Research Literature," Journal of Student Financial Aid: Vol. 39: Iss. 1, Article 2

[5] Narayana Darapaneni, Pramod Srinivas, Keerthi Reddy, Anwesh Reddy Paduri, Lakshmikanth Kanugovi, Pavithra J, Sudha B G, Bharath S, "Tree Based Models: A Comparative and Explainable Study for Credit Default Classification", 2022 IEEE 9th Uttar Pradesh Section

International Conference on Electrical, Electronics and Computer Engineering (UPCON), pp.1-8, 2022.

[6] P. Maheswari and C. V. Narayana, "Predictions of Loan Defaulter - A Data Science Perspective," 2020 5th International Conference on Computing, Communication and Security (ICCCS), Patna, India, 2020

[7] Everett, Craig R., Group Membership, Relationship Banking and Loan Default Risk: The Case of Online Social Lending (September 1, 2015). Banking and Finance Review, 7(2)

[8] Y. -Q. Chen, J. Zhang and W. W. Y. Ng, "Loan Default Prediction Using Diversified Sensitivity Under sampling," 2018 International Conference on Machine Learning and Cybernetics (ICMLC), Chengdu, China, 2018, pp. 240-245, Doi: 10.1109/ICMLC.2018.8526936.

[9] S. Barua, D. Gavandi, P. Sagle, L. Shinde and J. Ramteke, "Swindle: Predicting the Probability of Loan Defaults using Cat Boost Algorithm," 2021 5th International Conference on Computing Methodologies and Communication (ICCMC), Erode, India, 2021, pp. 1710-1715, Doi: 10.1109/ICCMC51019.2021.9418277.

[10] A. Shivanna and D. P. Agrawal, "Prediction of Defaulters using Machine Learning on Azure ML," 2020 11th IEEE Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON), Vancouver, BC, Canada, 2020, pp. 0320-0325, Doi: 10.1109/IEMCON51383.2020.9284884.