



# Predictions of Loan Defaulter-A Data Science Perspective

Miss Sanjiwani Subhashrao Gawande<sup>1</sup>, Prof. Vaishali B. Bhagat<sup>2</sup>

<sup>1,2</sup>Department of Computer Science & Engineering, P. R. Pote (Patil) College of Engineering & Management, Amravati-444605 2020-2021

**Abstract:** In our financial framework, banks have numerous items to sell yet fundamental kind of revenue of any banks is on its credit line. So they can procure from revenue of those advances which they credits. Past research in this period has shown that there are such countless techniques to examine the issue of controlling advance default. A vital methodology in prescient investigation is utilized to examine the issue of anticipating defaulters: The information is gathered from the Kaggle for examining and expectation. The advancement of innovation and execution of Data Science in banking, changes the substance of banking industry. The vast majority of the banking, monetary areas and social loaning stages are effectively contributing on loaning. Be that as it may, monetary foundations may confront enormous capital misfortune on the off chance that they affirmed the credit without having any earlier appraisal of default hazard. Monetary organizations consistently need a more exact prescient framework for different purposes. Foreseeing credit defaulters is a urgent assignment for the financial business. Banks have massively enormous measure of information like client's information, exchange conduct, and so on Information Science is a promising zone to handle the information and concentrate the secret examples utilizing AI strategies. Considering the magnitude of risk and financial loss involved, it is essential for banks to give loans to credible applicants who are highly likely to pay back the loan amount.

**Keywords:** Classification, Pre-processing, Prediction, Features selection, Generic algorithm, PSO algorithm, Naïve Bayes, decision tree, SVM, Random Forest.

## 1. INTRODUCTION

Finance sector is one of the earliest applications of Data Science. Financial institutions meet bad debts and losses every year. However, they initially perform paper work while sanctioning a loan which yields to get lot of data. Since from last few years banking improves their analysis of identifying the probability of risk through customer profiling, past expenditures, and customer transaction behavior, etc. Data Science is a combination of various statistical tools, algorithms and machine learning techniques to extract the hidden patterns from the data which helps to turn into in- sights. Now-a-days, financial organizations takes the advantage of data science applications, to study the individual customers banking profile, and providing the appropriate services by segmenting the customers based on their credit history.

Banks will receive number of loan applications every day. Loan is the main asset of banks to improve their profitability. However, assessing the risk is one of the major concerns for banks. This paper classifies that, the customer will be defaulter or not, by performing data science process, i.e., data pre- processing, Exploratory data analysis, building the models using machine learning algorithms and finally evaluate the models using various validation metrics.

## 2. RELATED WORK

Fraud detection and credit risk applications are the most popular application of Data Science, particularly well-suited to classification technique. Prediction of loan defaults mostly employs classification algorithms. In classification, data is processed into train and test. Training data is used to build model for prediction and test set is used to evaluate the model. In the first step is gathering information, data from previously approved loan datasets are gathered together. In paper [1] uses Exploratory Data Analysis (EDA), is to provide basic insights of any dataset. The main objective of EDA is to extract the essential patterns and visualize them in graphs, plots. In [2] proposed Decision Tree Induction Algorithm to predict the attributes relevant for loan credibility? In this paper a prototype model is built which can be used by the organization in making the right decision to approve or reject the loan.

In [3] authors were proposed clustering mechanism to improve the accuracy of defaulters in banks based on probability. The experimental results were obtained using KNN algorithm and it is implemented in R.

In [4] states authors stated problem statement with the class imbalance problem. Various approaches are discussed to handle the class imbalance problem. This classification follows binary method which results the output in one of the two variables either default or not. In [5] proposed Naive Bayesian classifier to classify the loan defaulters which is quickly produce the results. It assumes that all the input variables are independent and calculate the prior and posterior

probabilities. The naïve Bayes classifier is particularly appropriate when the dimensionality of the inputs is high. Along with its simplicity Naïve Bayes is one of the most sophisticated classification techniques. It well suited to credit-risk manager domain.

To the best of our knowledge, this paper addresses the very popular and novel research works studied in detail is shown in Table 1. The growth of the time series data is increasing dramatically. Furthermore, there are several tools to predict or forecast the time series accurately. Although this is not a clear research objective, but it is interesting to be able to develop more real-time forecasting algorithms and tools.

### 3. PROPOSED SYSTEM

This The Data Science process revolves around using machine learning and other analytical methods to produce insights and predictions from data in order to achieve a business objective. The entire process involves several steps like data cleaning, preparation, modeling and model evaluation shown in Fig. 1.

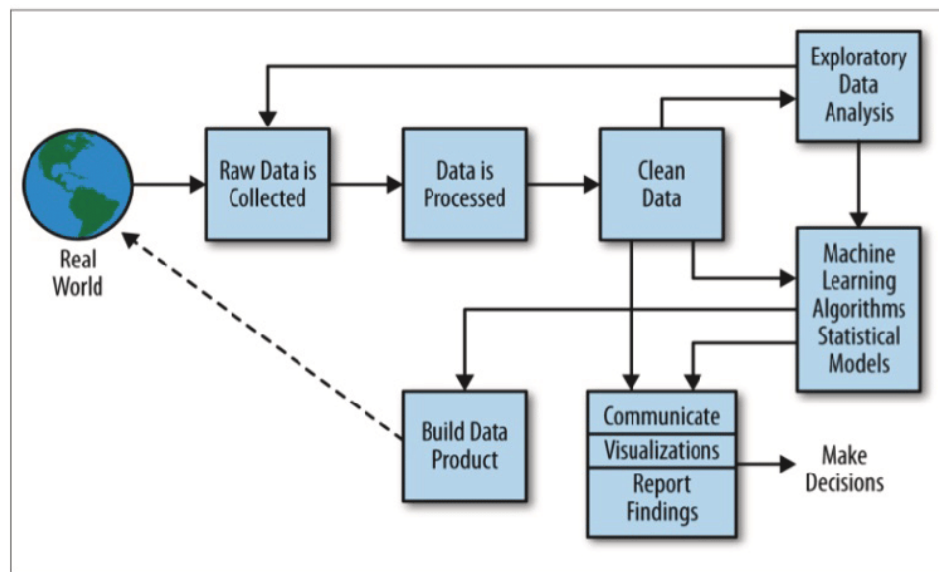


Fig 1. Data Science Process

#### A. Business Understanding

The data collection or the analysis completely depends on the objective need to satisfy. For that initially need to gain the domain knowledge of the particular aspect.

#### B. Data Understanding

Once the business problem is identified and understood, the next step is to gather the data from the various resources. In these days most of the data can be available through multiple resources, the forward step is to understand the data. Somehow in few cases we cannot gather data directly, need to gather data from the ground level. Both the cases can be achieved when we have domain knowledge on particular objective.

#### C. Data Preparation

Data preparation is the most time- consuming process in the process of data science. In data preparation, there are several steps to pre-process the data like, selecting the relevant features, identifying the noisy data, imputing the missing fraction of data using imputation methods, finding outliers and handling them. Creating new set of data from the existing features. This is the major step in data science process because how good will the data was pre-processed results the good models for accurate outcome.

#### D. Exploratory Data Analysis (EDA)

In EDA, the data will be present more effectively. Using statistical measures or metrics we can able to find the meaningful patterns. Further, we can visualize the results in plots, graphs, and histogram, scatter plot, etc. A lot more work can be done in EDA, to tune the data like dimensionality reduction, sampling data and class imbalanced problems to model the data.



#### E. Data Modeling

Data modeling stage is closely related with problem understanding and data understanding, because we need to understand whether the problem statement can be solved through a classification or regression technique. Without understanding the problem we cannot perform the data modeling. After the model built, need to tune the parameters more precisely to get best accurate results.

#### F. Model Evaluation

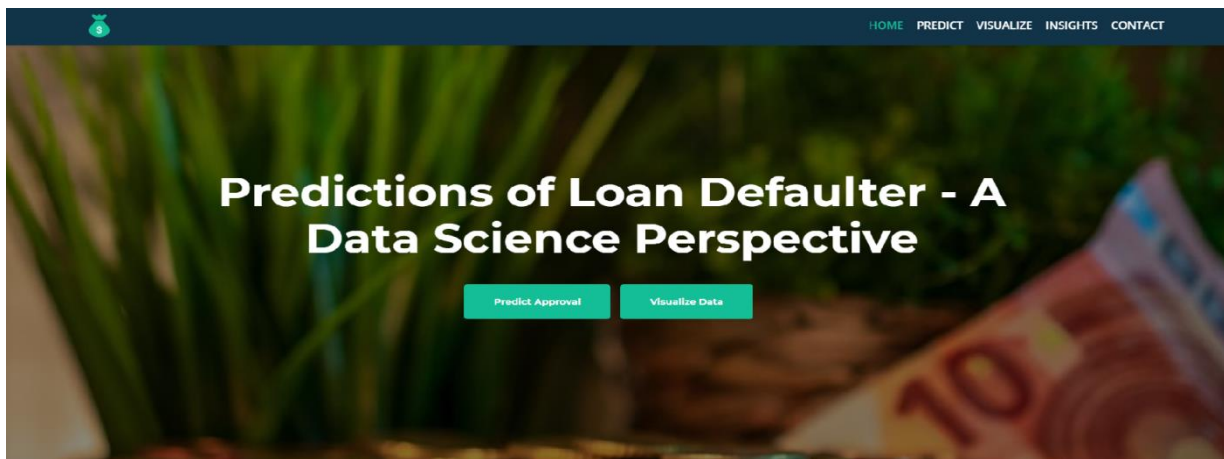
It is very important to evaluate a model before applying in real-time. The cost and time will depend on this stage. The model is evaluated by using existing data or with new data, how well the model is evaluated, results stable consistency when applied on different platforms.

#### G. Model Deployment

Finally, the model will be deployed into real time application which should produce results according to real- time basis. If any of the above steps goes improperly, all the steps are iteratively repeated frequently.

## 4. IMPLEMENTATION

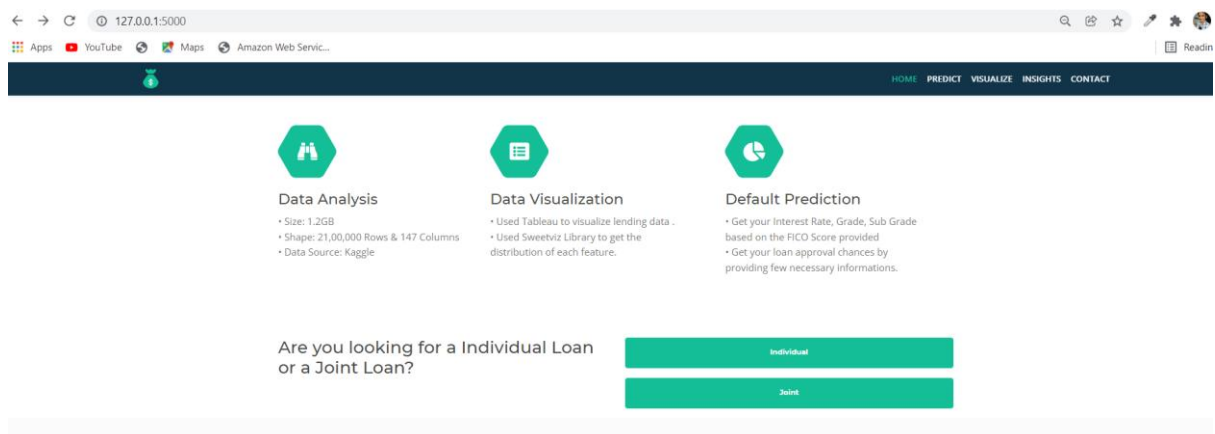
### 1) Home Page of application :



First page of application i.e Home Page. It contains two options Predict Approval and Visualize Data .In that Predict Approval, there are two options available which are Individual Loan and Joint Loan.

### 2) Second page for prediction menu:

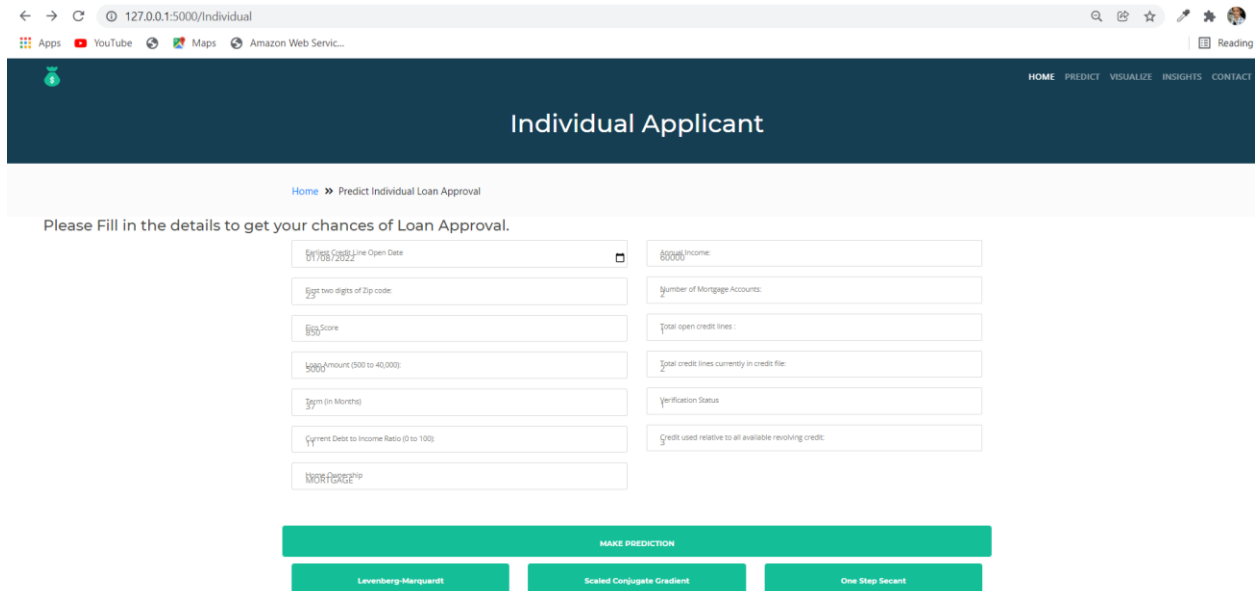
It contains two options Individual Loan or a Joint Loan .If customer wants to apply as individual they can apply.



### 3) Input data for Loan Approval:

The customer have to fill up this form for Loan Approval like Fico Score, Loan Amount, Term in months, Annual

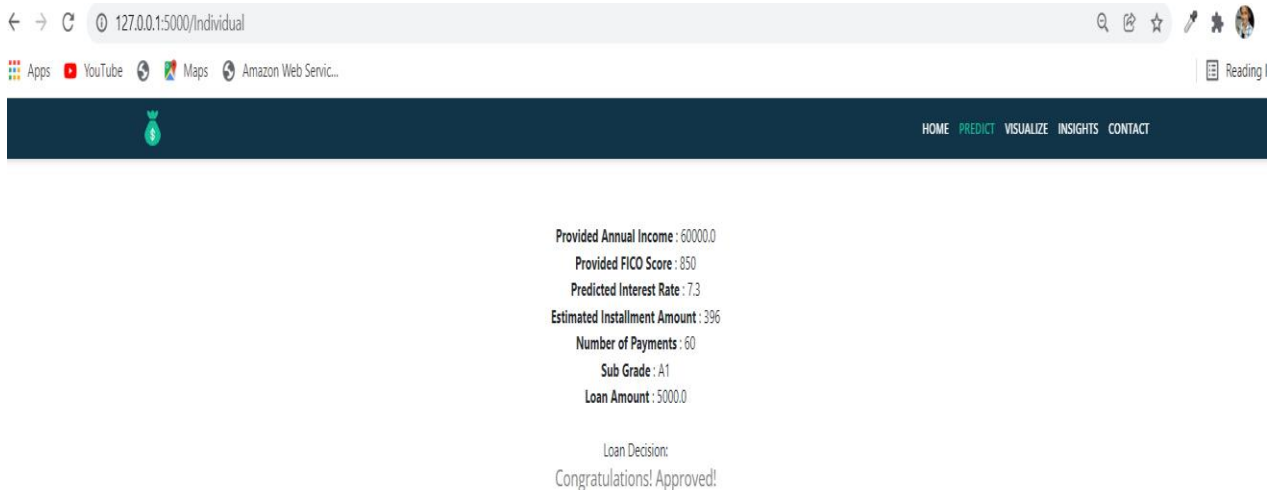
income etc.. After filling this form it will make a prediction. If Fico Score, Annual income and Loan Amount is not in under criteria then Loan will be denied.



#### 4) Output from Input data for Loan Approval:

The customer have to fill up this form for Loan Approval like Fico Score, Loan Amount, Term in months, Annual income etc.. After filling this form it will make a prediction. If Fico Score, Annual income and Loan Amount is in criteria then Loan will be Approved.

If Fico Score, Annual income and Loan Amount is not in under criteria then Loan will be denied.



#### 5) Data of LM, SCG and OSS:

This data presented an investigation of the use of supervised neural network models for customer loan default prediction under different training algorithms scaled conjugate gradient back propagation, The Levenberg-Marquardt algorithm and One-step secant back propagation (SCG, LM and OSS). The table Shows the data of default customer with their bank.



| Bank Name             | Wilful Defaulter | Amount | Name            |
|-----------------------|------------------|--------|-----------------|
| State Bank of India   | 685              | 43,887 | John Doe        |
| Punjab National Bank  | 325              | 22370  | James Smith     |
| Bank of Baroda        | 355              | 14661  | Maria Garcia    |
| Bank of India         | 184              | 11250  | Maria Rodriguez |
| Central Bank of India | 69               | 96663  | Mary Smith      |
| United Bank of India  | 128              | 7028   | Maria Hernandez |
| UCO Bank              | 87               | 6813   | William         |
| Canara Bank           | 96               | 6549   | Charles bed     |
| Andhra Bank           | 84               | 5276   | Joseph doe      |
| Indian Bank           | 57               | 5165   | Samuel          |
| Corporation Bank      | 49               | 4339   | Henry desuza    |

## 5. CONCLUSION

This paper presented an investigation of the use of supervised neural network models for customer loan default prediction under different training algorithms scaled conjugate gradient back propagation, Levenberg-Marquardt algorithm and One-step secant back propagation (SCG, LM and OSS). This paper also compared between two filtering functions and evaluation of the ensemble models. Several parameters were used in the experiment to do this comparison; training time, iteration, MSE and R. The slowest algorithm was OSS. The best algorithm was LM because it had the largest R. The accuracy percentages of all models were calculated. First the filtering function was applied on the original dataset producing another two datasets. Then for each dataset three supervised neural network models, each one using different training algorithms. The results in Table 4 shows that LM algorithm and (PLsFilter) filtering function gave the best model. The ensemble models accuracy percentage were calculated and recorded., showing that the ensemble model of the three algorithms (LM, OSS and SCG) of dataset (DS2) was the best model. This paper discussed how data science can impact the banking sector to improve their analysis of identifying risk by preprocessing the historical data of customers and building the model using machine learning techniques. Due to huge volume data processing, built the models in cross validation approach using GridSearchCV. The classification techniques logistic regression, random forest and KNN model are built, so far three algorithms results similarly. Among them Logistic regression with SGD training results better predictions than the others.

## 6. REFERENCES

1. M. S. Sivasree, "Loan Credibility Prediction System Based on Decision Tree Algorithm," Int. J. Eng. Res. Technol., 2015.
2. Aida Krichene, "Using a naive Bayesian classifier methodology for loan risk assessment," Journal of Economics, Finance and Administrative Science, 2017
3. Bagherpour, "Predicting mortgage loan default with machine learning methods," Univ. California / Riverside, 2017.
4. Namvar, M. Siami, F. Rabhi, and M. Naderpour, "Credit risk prediction in an imbalanced social lending environment," arXiv Prepr. arXiv1805.00801, 2018
5. Goyal and R. Kaur, "Loan Prediction Using Ensemble Technique.," Int. J. Adv. Res. Comput. Commun. Eng., vol. 5, no. 3, pp. 523–526, 2016.
6. X.Francis Jency, V.P.Sumathi, Janani Shiva Sri, "An Exploratory Data Analysis for Loan Prediction Based on Nature of the Clients"
7. Dr. K. Chitra1, B. Subashini, "Data Mining Techniques and its Applications in Banking Sector", International Journal of Emerging Technology and Advanced Engineering Website: www.ijetae.com (ISSN 2250-2459, ISO 9001:2008 Certified Journal, Volume 3)
8. Semiu A, Akanmu Abdul Rehman Gilal August 2013), "A Boosted Decision Tree Model for Predicting Loan Default in P2P Lending Communities"
9. S. Raschka and V. Mirjalili, Python machine learning. Packt Publishing Ltd, 2017.



10. Pandit, "DATA MINING ON LOAN APPROVED DATSET FOR PREDICTING DEFAULTERS," Rochester Institute of Technology, 2016.
11. G. Sudhamathy, "Credit risk analysis and prediction modelling of bank loans using R," Int. J. Eng. Technol, vol. 8, pp. 1954–1966, 2016.
12. M. Li, A. Mickel, and S. Taylor, "Should This Loan be Approved or Denied?: A Large Dataset with Class Assignment Guidelines," J. Stat. Educ., vol. 26, no. 1, pp. 55–66, 2018.
13. Mahesh Marodkar, "Loan Defaulter's Application in R programming".
14. Shaath, "Credit Risk Analysis and Prediction Modelling of Bank Loans Using R".
15. S. Chebrolu, A. Abraham, and J. P. Thomas, "Feature deduction and ensemble design of intrusion detection.
16. Gaurav Akrani., Kaylan City Life (20-Apr-2011), Available: <http://kalyancity.blogspot.com/2011/04/functions-of-banks-important-banking.html>.