IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.301
Volume 3, Issue 10, May 2023

Prediction of Loan Approval using Machine Learning Algorithm

Mr. Patil Mukesh D.[1], Dhiraj Patil[2], Neha Kale[3]
Department of Electronics and Telecommunication
Sinhgad Institute of Technology and Science, Narhe, Pune, India

Abstract: *In today's world, obtaining loans from financial institutions has become a very common phenomenon. Giving credit is one of the core businesses in banking and the importance of credit risk management was highlighted in the 2008 financial crisis. Loan defaults has become important as banks try to follow laws and regulations, grant credits to qualified customers, mitigate credits to unqualified customers and to make their application processes efficient. Every day many people apply for loans, for a variety of purposes for different tasks. But not all the applicants are reliable, and not everyone can be application get approved. Every year, there are many cases where people do not repay the loan amount to the bank which results in huge financial loss for bank. The risk associated with making a decision on a loan approval is immense. Hence, the idea of this project is to gather loan data from the Lending Club website &from bank websites by using machine learning techniques on this data to extract important information and predict if a customer would be able to applicable for loan or not. In other words, the goal is to predict if the customer would be a defaulter or not*

Keywords: Supervised learning, Predictive analytics, Logistic regression, Classification tree, Random Forest, Extreme gradient boosting, XGBoost

I. INTRODUCTION

Loan prediction is a common problem for such lending companies. This is the type of problem banks and credit card companies face whenever customers ask for a loan. This thesis focusses on using the Lending Club dataset which is freely available on their website. The objective is to make loan predictions and whether investors should lend to a customer or not. Data from 2019-2021 will be used because most of the loans from that period have already been repaid or defaulted on. Lending Club is the platform, or rather the marketplace, where investors and borrowers meet virtually.
The Lending Club processes the application with their own data science methods. However, on the side of the investor, there is nothing to ensure the creditworthiness of the borrower and the level of risk involved in any given case. Customer first apply for loan after that company or bank validates the customer eligibility for loan. Company or bank wants to automate the loan eligibility process (real time) based on customer details provided while filling application form. These details are Gender, Marital Status, Education, Number of Dependents, Income, Loan Amount, Credit History and other. This project has taken the data of previous customers of various banks to whom on a set of parameters loan were approved. So, the machine learning model is trained on that record to get accurate results. Our main objective of this project is to predict the safety of loan. To predict loan safety, the SVM and Naive bayes algorithm are used. First the data is cleaned so as to avoid the missing values in the data set.

II. RELEVANCE

Loan approval is a very important process for banking organizations. The system approved or reject the loan applications. Recovery of loans is a major contributing parameter in the financial statements of a bank. It is very difficult to predict the possibility of payment of loan by the customer. Using Machine learning we predict the loan approval. The main objective of this project is to predict whether assigning the loan to particular person will be safe or not.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.301
Volume 3, Issue 10, May 2023

III. MOTIVATION

Applying machine learning to loan predictions, showcases a useful application of this branch of artificial intelligence to solve real-world and business problems. The goal is to predict if the customer would be a defaulter or not. Loan approval is a very important process for banking organizations. The system approved or reject the loan applications. Recovery of loans is a major contributing parameter in the financial statements of a bank. It is very difficult to predict the possibility of payment of loan by the customer. Using Machine learning we predict the loan approval.

IV. PROBLEM DEFINITION

If a model can identify credit-worthy customers that were not recognized by traditional credit scores, while minimizing their risk of default on the loans, this can be a lucrative niche market or micro-market, pushing higher the profit margin of the financial institution or investor. Although the prospect of more customers seems positive, it is important to be careful as to not lend to people that will default on the loan. Thus, a conservative approach and strict evaluation metrics were kept in mind throughout the project. The loan default prediction is a problem of binary classification (should the investor lend or not). Logistic Regression is a good model for this problem.

V. LITERATURE SURVEY

Mohammad Ahmad Sheikh, Amit Kumar Goel, Tapas Kumar, "An Approach for Prediction of Loan Approval using Machine Learning Algorithm", 2020 [1]. The enhancement in the banking sector lots of people are applying for bank loans but the bank has its limited assets which it has to grant to limited people only, so finding out to whom the loan can be granted which will be a safer option for the bank is a typical process. The main objective of this paper is to predict whether assigning the loan to particular person will be safe or not. This paper is divided into four sections:(i) Data Collection (ii) Comparison of machine learning models on collected data (iii) Training of system on most promising model(iv) Testing. To reduce this risk factor behind selecting the safe person so as to save lots of bank efforts and asset using machine learning.

Ya-qi Chen, Jianjun Zhang, Wing W. Y. Ng., "Loan Default Prediction Using Diversified Sensitivity Under sampling", 2018 [2]. The loan default prediction is to predict rather the borrower will delay the repayment or not. This is an important problem for banking and finance companies. In this study, we focus on dealing with the data imbalance problem to enhance the performance of the loan default prediction. The approach in this study is a hybrid under sampling method that combines the clustering, the stochastic sensitivity measure and the radial basis function neural networks. A real loan default data from a P2P company in China is used to valid the performance of our method. It mainly focused on dealing with the data imbalance problem to enhance the performance of the loan default prediction.

Lila Lai, "Loan Default Prediction using Machine Learning", 2018 [3]. Loan business is one of the major income sources for bank. However, loan default problem is a major issue for loan business. With the rise of big data era and the development of machine learning techniques, nowadays we have more options for classifying and predicting loan default, other than manual processing. With a real-world dataset from a prestigious international bank, we demonstrate that the AdaBoost model can achieve a 100% accuracy for predicting loan default, outperforming other models including XGBoost, random forest, nearest neighbors, and multilayer perceptions.

Bhoomi Patel, Harshal Patil, Jovita Hembram, Shree Jaswal, "Loan Default forecasting using Data Mining", 2020[4].

Estimation or assessment of default on a debt is a crucial process that should be carried out by banks to help them to assess if a loan applicant can be a defaulter at a later phase so that they process the application and decide whether to approve the loan or not. The conclusion derived from such assessments helps banks and other financial institutions to lessen their losses and eventually increase the number of credits. Hence, it becomes vital to construct a model that will take into account the different aspects of an applicant and derive a result regarding the concerned applicant. All available means to loan the money from their illicit activities are used for criminal activities in today's technology-based realm. The increasing number of bad debts resulting from commercial banks' loans reflects the growing problem of distraught banks within the economic system. We have used data mining algorithms to predict the likely defaulters from a dataset that contains information about home loan applications, thereby helping the banks for making better decisions in the future.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.301
Volume 3, Issue 10, May 2023

VI. MACHINE LEARNING&MODEL BUILDING

Machine learning is divided into three categories:

- Supervised Learning
- Unsupervised Learning
- Reinforcement Learning

6.1 Model Building

The major objective of this project is to derive patterns from the datasets which are used for the loan sanctioning process and create a model based on the patterns derived in the previous step. Classification data mining algorithms are used to filter out the probable loan defaulters from the list. For analysis purposes, essential inputs like gender, married, dependents, education, self-employed, applicant income, coapplicant income, loan amount, loan amount term, etc., are collected and used to find the appropriate attributes.

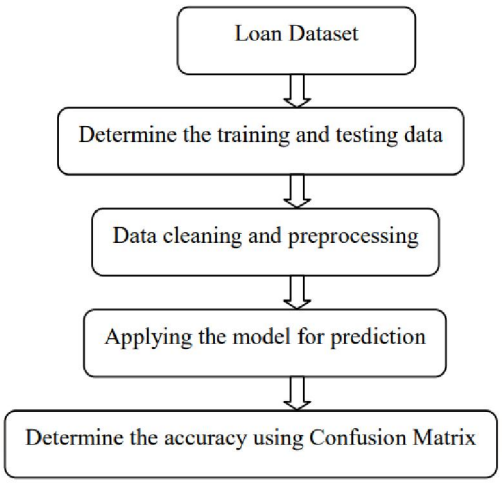

Fig. 1. Architecture of the Proposed Loan Prediction Model

6.2 Algorithms used:

Logistic Regression:

After splitting the data to training set and test set, the first step in creating the logistic regression model was to review collinearity. Collinearity can affect a logistic regression model and therefore, variables with high correlation should be removed. In the data set, variables “AMT GOOD PRICE” and “AMT CREDIT” had almost a perfect correlation of 0.985. Therefore, it was decided that “AMT CREDIT” would not be used in the logistic regression model. After correlation was revised, standardizing the training and test sets was next. As the numeric variables had many values, scaling was done to prevent one significant number having too much power in the prediction

K-Neighbors Classifier

An approach of classifying the data which would be used in estimating the likelihood of a data point in being a member of one group or the other based upon the nearest available group of data points can be described as a k nearest neighbor algorithm, which is often called as KNN algorithm. KNN algorithm usually will not construct a model until a query is imposed on the dataset, which makes K-nearest neighbor a predominant example of a "lazy learning" algorithm. In knn algorithm, if we need to determine whether a point will come under either group A or B, the algorithm will look at the nearest data points and the group does they belong to. If we consider a sample of data, the range is randomly determined. in case, if the majority of the points belong to group B, in such instance the data point will be having the most likelihood of being a member of group B and vice versa.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.301
Volume 3, Issue 10, May 2023

SVC:

In this approach, each data item is plotted in a n-dimensional space, where n represents the number of features with each feature represented in a corresponding coordinate. A hyper plane is determined to distinguish the classes (possibly two) based on their features

Decision Tree Classifier

The outcome variable "TARGET" was a categorical value, therefore the decision tree built was a classification tree. In creating the classification tree, numeric variables were not standardized. Standardised values do not change the outcome of the prediction in decision trees. Another difference to building a logistic regression model, oversampling and under sampling of the data was needed in classification tree model. As mentioned, the data was split 70% to training set and 30% to test set.

VII. SYSTEM DESIGN AND ARCHITECTURE

7.1 System Architecture

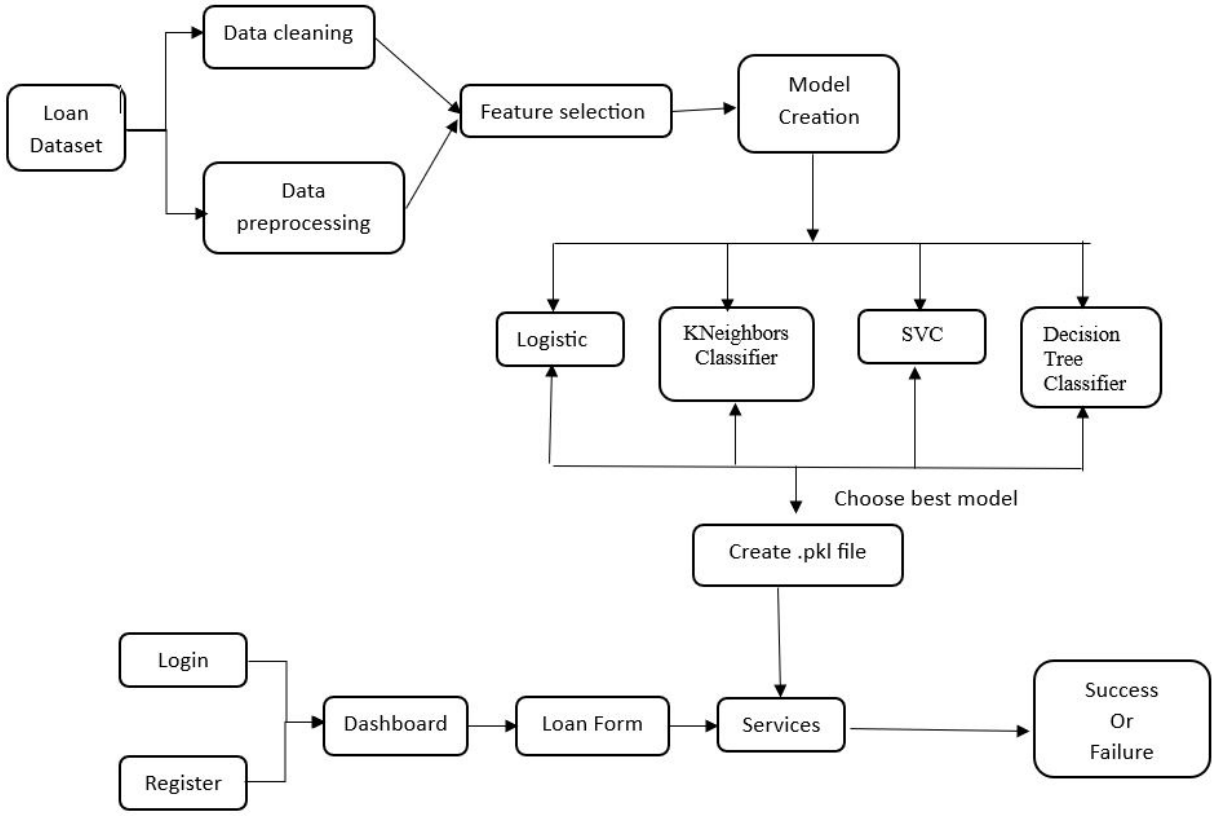

Fig. 2. Loan Prediction Architecture

Loan Dataset:

Loan Dataset is very useful in our system for prediction of more accurate result. Using the loan Dataset, the system will automatically predict which costumer's loan it should approve and which to reject. System will accept loan application form as an input. Justified format of application form should be given as an input to get processed.

Data cleaning and processing:

InData cleaning the system detect and correct corrupt or inaccurate records from database and refers to identifying incomplete, incorrect, inaccurate or irrelevant parts of the data and then replacing, modifying or detecting the dirty or coarse data. In Data processing the system convert data from a given form to a much more usable and desired form i.e., make it more meaningful and informative.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.301
Volume 3, Issue 10, May 2023

Feature Selection:

A good feature subset is one that contains features highly correlated with (predictive of) the class, yet uncorrelated with (not predictive of) each other. A feature is said to be redundant if one or more of the other features are highly correlated with it. Correlation was implemented and the feature set was selected which was highly correlated to the 'TARGET' field.

Choose best model

To measure the success rate of the model, the best metric was the precise prediction of a loan default. The profitability of the investor or the financial institution depends on the decision of the model. These are the error types which were used for determining a conservative evaluation of the loan default rate.

- True Positive (TP)
- True Negative (TN)
- False Positive (FP)
- False Negative (FN)

Models evaluated through:

- Precision
- Recall
- F1 score
- Loss
- Accuracy

Create .pkl file

Pickle is a useful Python tool that allows you to save your ML models, to minimise lengthy re-training and allow you to share, commit, and re-load pre-trained machine learning models

VIII. RESULT & FUTURE SCOPE

This section shows a comparative study of all the models that were built. These models are evaluated through accuracy, precision, and f1-score. Through experiments, the model was found which best suits the dataset and serves the purpose of giving an investor a model which would increase their chances of a profit**.**

Before Feature selection:

we can see that best model is Logistic Regression at least for now, but the models are just memorizing the data so it is overfitting problem.

Sr. No.	Algorithm Name	Precision	Recall	F1 Score	Loss	Accuracy
1	Logistic Regression	0.930	0.429	0.587	6.827	0.811
2	KNeighbors Classifier	0.667	0.364	0.471	9.249	0.743
3	SVC	1.000	0.013	0.026	11.158	0.690
4	Decision Tree Classifier	0.929	0.422	0.580	6.900	0.809

Table 8.1

After Feature selection:

To prevent Overfitting problem of model, we introduce feature selection which increases accuracy of model as well as reduce time for calculation due less no. of feature.

Sr. No.	Algorithm Name	Precision	Recall	F1 Score	Loss	Accuracy
1	Logistic Regression	0.850	0.447	0.586	7.033	0.805
2	KNeighbors Classifier	0.615	0.421	0.500	9.377	0.740

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.301
Volume 3, Issue 10, May 2023

3	SVC	0.867	0.342	0.491	7.912	0.780
4	Decision Tree Classifier	0.895	0.447	0.596	6.740	0.813

Table 8.2

IX. FUTURE SCOPE

Here in this paper, we have only considered home loan prediction, a system could be made for predicting defaulters of other loans as well. Different machine learning techniques (Random Forests, Neural Networks etc.) can be implemented and compared to get better results. Also, whether the non-defaulter would turn out to be a fraudster or not could be predicted

X. CONCLUSION

Loan prediction is an important and challenging problem. In this paper, we compare the performance of four Machine learning models which are Logistic Regression, KNeighbors Classifier, Support Vector Classifier and Decision Tree Classifier using ROC and AUC as evaluation metrics. The outcomes show that Decision Tree Classifier and Logistic Regression had greater accuracy than other algorithm based on our results, we believe that machine learning methods have a huge potentiality to be applied to process the loan prediction problem.

REFERENCES

[1]. Mohammad Ahmad Sheikh, Amit Kumar Goel, Tapas Kumar, "An Approach for Prediction of Loan Approval using Machine Learning Algorithm", 2020.
[2]. Ya-qi Chen, Jianjun Zhang, Wing W. Y. Ng., "Loan Default Prediction Using Diversified Sensitivity Under sampling", 2018.
[3]. Lila Lai, "Loan Default Prediction using Machine Learning", 2018.
[4]. Bhoomi Patel,Harshal Patil,Jovita Hembram, Shree Jaswal, "Loan Default forecasting using Data Mining", 2020.
[5]. Kacheria, A., Shivakumar, N., Sawkar, S. and Gupta, A., "Loan Sanctioning Prediction System".
[6]. A. Gahlaut, Tushar, and P. K. Singh, "Prediction analysis of risky credit using Data mining classification models", 2017 8th International Conference on Computing, Communication and Networking Technologies (ICCCNT)", 2017.
[7]. Nikhil Madane, Siddharth Nanda, "Loan Prediction using Decision tree", Journal of the Gujrat Research History, Volume 21 Issue 14s", 2019.
[8]. Pidikiti Supriya, Myneedi Pavani, Nagarapu Saisushma, Namburi Vimala Kumari, k Vikas, "Loan Prediction by using Machine Learning Models", International Journal of Engineering and Techniques, Volume 5 Issue 2", 2019