Proceedings of the SMART–2020, IEEE Conference ID: 50582
9th International Conference on System Modeling & Advancement in Research Trends, 4th–5th, December, 2020
Faculty of Engineering & Computing Sciences, Teerthanker Mahaveer University, Moradabad, India

2020 9th International Conference System Modeling and Advancement in Research Trends (SMART) | 978-1-7281-8908-6/20/$31.00 ©2020 IEEE | DOI: 10.1109/SMART50582.2020.9336801

Bank Loan Prediction System using Machine Learning

Anshika Gupta[1], Vinay Pant[2], Sudhanshu Kumar[3] and Pravesh Kumar Bansal[4]
[1,2,3]*Faculty of Engineering and Computing Sciences, iNurture, TMU, Moradabad*
[4]*Department of Computer Science, Government Engineering College, Bharatpur*
E-mail: [1]anshikagupta0308@gmail.com, [2]vinay.p@inurture.co.in,
[3]sudhanshu.k@inurture.co.in, [4]bansal086@gmail.com

***Abstract*—With the advancement in technology, there are so many enhancements in the banking sector also. The number of applications is increasing every day for loan approval. There are some bank policies that they have to consider while selecting an applicant for loan approval. Based on some parameters, the bank has to decide which one is best for approval. It is tough and risky to check out manually every person and then recommended for loan approval. In this work, we use a machine learning technique that will predict the person who is reliable for a loan, based on the previous record of the person whom the loan amount is accredited before. This work's primary objective is to predict whether the loan approval to a specific individual is safe or not.**

Keyword: **Loan Dataset, Logistic Regression, Random Forest, Django.**

I. Introduction

As the data are increasing daily due to digitization in the banking sector, people want to apply for loans through the internet. Artificial intelligence (AI), as a typical method for information investigation, has gotten more consideration increasingly. Individuals of various businesses are utilizing AI calculations to take care of the issues dependent on their industry information. Banks are facing a significant problem in the approval of the loan. Daily there are so many applications that are challenging to manage by the bank employees, and also the chances of some mistakes are high. Most banks earn profit from the loan, but it is risky to choose deserving customers from the number of applications. One mistake can make a massive loss to a bank.

Loan distribution is the primary business of almost every bank. This project aims to provide a loan [1, 8] to a deserving applicant out of all applicants. An efficient and non-biased system that reduces the bank's time employs checking every applicant on a priority basis. The bank authorities complete all other customer's other formalities on time, which positively impacts the customers. The best part is that it is efficient for both banks and applicants. This system allows jumping on particular applications that deserve to be approved on a priority basis.

There are some features for the prediction like- 'Gender', 'Married', 'Dependents', 'Education', 'Self_ Employed', 'ApplicantIncome', 'CoapplicantIncome', 'LoanAmount', 'Loan_Amount_Term', 'Credit_History', 'Property_Area', 'Loan_Status'.

II. Literature Survey

A prediction is a statement about what someone thinks will happen in the future. People make predictions all the time. Some are very serious and are based on scientific calculations, but many are just guesses. Prediction helps us in many things to guess what will happen after some time or after a year or after ten years.

Predictive analytics is a branch of advanced analytics that uses many techniques from data mining, statistics, modeling, machine learning, and artificial intelligence to analyze current data to make predictions. "Adyan Nur Alfiyatin, Hilman Taufiq [2] and their friends work on the house price prediction. They use regression analysis and Particle Swarm Optimization (PSO) to predict house price". One other similar work on the Mohamed El Mohadab, Belaid Bouikhalene [3] and Said Safi to predict the rank for scientific research paper using supervised learning. Kumar Arun, Garg Ishan and Kaur Sanmeet [1] work on bank loan prediction on how to bank approve a loan. They proposed a model with the help of SVM and Neural networks like machine learning algorithms.

This literature review helps us carry out our work and propose a reliable bank loan prediction model.

III. Proposed Methodology

The process to predict the bank loan of the applicants is as shown in figure 1. There is a different phase in each step, which is described here.

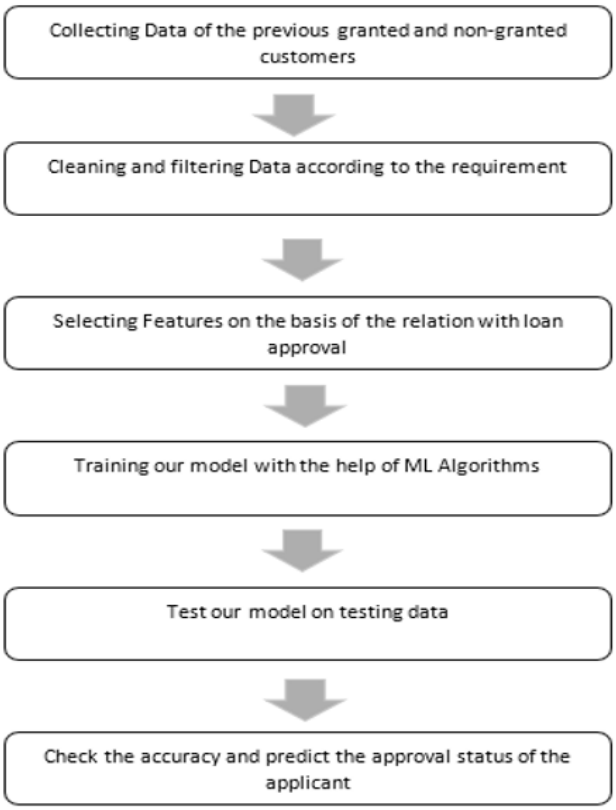

Fig. 1: Proposed Methodology

IV. Dataset Descriptions and Pre-processing

The bank loan prediction system dataset is taken from kaggle competition which belong to different age group and gender of the applicants. There are thirteen attributes in the data set, such as education, married status, income, assets, etc. as shown in figure 2.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 614 entries, 0 to 613
Data columns (total 13 columns):
 #   Column            Non-Null Count  Dtype
---  ------            --------------  -----
 0   Loan_ID           614 non-null    object
 1   Gender            601 non-null    object
 2   Married           611 non-null    object
 3   Dependents        599 non-null    object
 4   Education         614 non-null    object
 5   Self_Employed     582 non-null    object
 6   ApplicantIncome   614 non-null    int64
 7   CoapplicantIncome 614 non-null    float64
 8   LoanAmount        592 non-null    float64
 9   Loan_Amount_Term  600 non-null    float64
 10  Credit_History    564 non-null    float64
 11  Property_Area     614 non-null    object
 12  Loan_Status       614 non-null    object
dtypes: float64(4), int64(1), object(8)
memory usage: 62.5+ KB
```

Fig. 2: Data Describtion

There are total 981 records of the applicants with the values of their concerning attributes in categorical and numerical data. The total count value of the attribute is also shown in figure 3. In the pre-processing and feature engineering of the data, we handle the missing value and also normalize the data so we can further process it into ML algorithm. The dataset is further divided into training and testing. The model is trained on machine learning algorithms and predicts the system on test data which is discussed in the Next section in details.

Value Count some Important parameters

```
Y    422
N    192
Name: Loan_Status, dtype: int64
Male      489
Female    112
Name: Gender, dtype: int64
Yes    398
No     213
Name: Married, dtype: int64
0     345
1     102
2     101
3+     51
Name: Dependents, dtype: int64
No     500
Yes     82
Name: Self_Employed, dtype: int64
1.0    475
0.0     89
Name: Credit_History, dtype: int64
Semiurban    233
Urban        202
Rural        179
Name: Property_Area, dtype: int64
Graduate        480
Not Graduate    134
Name: Education, dtype: int64
```

Fig. 3: Value Count of the Parameters in the Dataset

A. Machine Learning

Predictive analytics [10] is used to predict the data about future events. It includes many techniques such as data mining, machine learning [4, 9] and modeling. Machine learning is a type of artificial intelligence that allows a software application to learn from the data & become more accurate in predicting outcomes without human intervention. Machine learning and deep learning help to design and develop such a machine that automatically learns and predicts your data and situation. Machine learning is often divided into different subcategories according to the type of problems being comes. Some ML type is as follows:

1) Supervised Learning

Supervised learning is the point at which the model is getting prepared on a labelled dataset. In this kind of learning both training and testing, datasets are labelled. The output of prediction is always coming either 1 (yes) or 0 (No).

2) Unsupervised Learning

In unsupervised learning, the input data are not labelled and also do not have any prior information about the data.

Here the task of the machine is to find the hidden pattern from the data by using cluster analysis. The dataset is labelled so that here we used supervised learning approach.

In our work, we used a supervised learning approach.

B. Algorithms used for Prediction

1) Logistic Regression

It is a classification set of rules used to assign observations to a discrete set of instructions. Logistic regression is also a predictive analysis, like other regression analyses methods. Logistic regression is basically used for define the relationship between dependent binary variable and nominal or other independent variable. Now a day's logistic regression is used in many research areas like medical science, machine learning and social science. It also used by many e-commerce applications to predict the mind set of customer to buy the product.

2) Random Forest

Random Forest is a robust system learning algorithm that is used for a ramification of responsibilities along with classification and regression. Random forests method overcome the over fitting issue of decision trees during training. It is an ensemble method made up of a large number of small decision trees [5,7] called estimators where each tree produces the prediction. The random forest model combines the predictions of the estimators to produce a more accurate prediction.

C. Correlation between Parameters

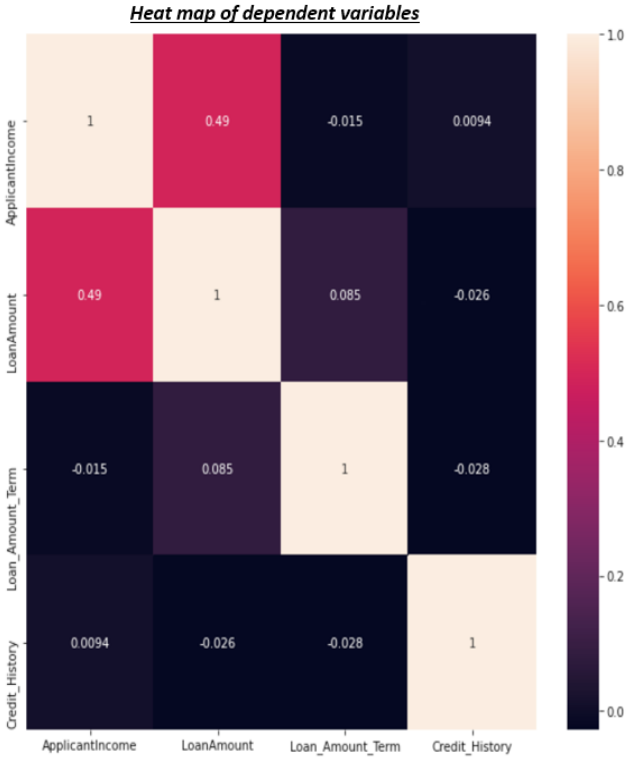

Fig. 4: Heat Map

Heat map is a data visualization technique that shows the magnitude of a phenomenon as color in two dimensions as shown in figure 4. Color intensity shows the relationship between each other. The color variation may be by hue or intensity, giving an obvious visual to the reader about how the phenomenon is clustered or varies over space. From this heat map, it is negative relation in Loan amount term with application come attribute.

V. Experiment and Result Analysis

In this section, we use a machine learning algorithm on a loan prediction dataset and deploy the result using HTML, CSS, Django at the local server. Figure 5 shows the loan prediction system of the applicants based on the value enter by the bank employee.

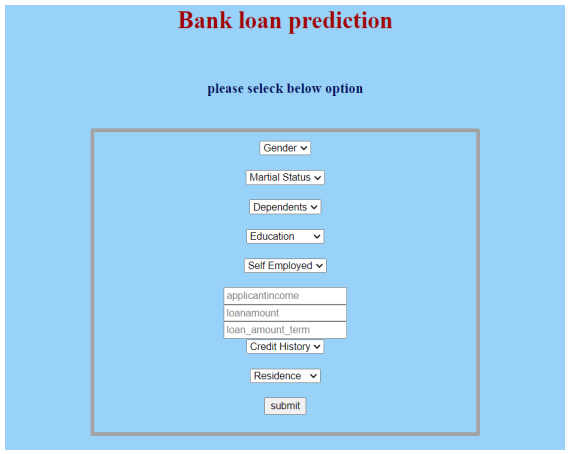

Fig. 5: Final Layout

The first attribute is to select the gender of the applicant either male or female. The second is the marital status then dependents attributes mean that the applicant is dependent financially on someone or not. Other attributes are the education of the customer, employment status, applicant income, loan amount term, credit history, residential area, etc. of the applicant. Finally, it shows the status of the loan i.e. it is safe or risky as shown in fig 6.

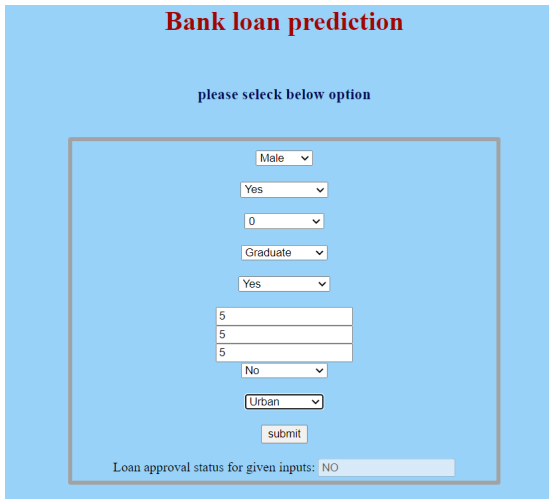

Fig. 6: Final Result

VI. Conclusion and Future Scope

Today's fast-growing IT industry needs to discover new technology and update the old technology that helps us to reduce human intervention and increase the efficiency of the work. This model is used for the banking system or anyone who wants to apply for a loan. It will be very helpful in bank management. From the analysis of the data, it is very clear that it reduces all the frauds done at the time of loan approval. Time is also very precious for everyone through this not only the bank but also the waiting time of the applicant will also reduce. As it seems, it will not deal with some special cases when only one parameter is enough for the decision, but it is quite efficient and reliable in some instant.

In the future, this prediction module can be more improved and integrated. The system is prepared on the previous training data but in the future, it is possible to make changes to software, which can accept new testing data and should also take part in training data and predict accordingly.

References

[1] Kumar Arun, Garg Ishan, Kaur Sanmeer, Loan Approval Prediction based on Machine Learning Approach.

[2] Adyan Nur Alfiyatin, Hilman Taufiq, Ruth Ema Febrita, Wayan Firdaus Mahmudy, 'Modeling House Price Prediction using Regression Analysis and Particle Swarm Optimization': International Journal of Advanced Computer Science and Applications (Vol. 8, No. 10, 2017).

[3] Mohamed El Mohadab, Belaid Bouikhalene, Said Safi, 'Predicting rank for scientific research papers using supervised learning'Applied Computing and Informatics 15 (2019) 182–190.

[4] K. Hanumantha Rao, G. Srinivas, A. Damodhar, M. Vikas Krishna: Implementation of Anomaly Detection Technique Using Machine Learning Algorithms: Internatinal Journal of Computer Science and Telecommunications (Volume2, Issue3, June 2011).

[5] J. R. Quinlan. Induction of Decision Tree. Machine Learning, Vol. 1, No. 1. pp. 81-106., 1086.

[6] G. Arutjothi, C. Senthamarai: Prediction of loan status in commercial bank using machine learning classifier, International Conference on Intelligent Sustainable Systems (ICISS), 2017.

[7] J.R. Quinlan. Induction of decision trees. Machine learning Springer, 1(1):81–106, 1086.

[8] Vishnu Vardhan case study of bank loan prediction,https://medium.com/@vishnumbaprof/case-study-loan-prediction-ac035f3ec9e4.

[9] S.S. Keerthi and E.G. Gilbert. Convergence of a generalize SMO algorithm for SVM classifier design. Machine Learning, Springer, 46(1):351–360, 2002.

[10] J.M. Chambers. Computational methods for data analysis. Applied Statistics, Wiley, 1(2):1–10, 1077.