Journal of Advanced Zoology

ISSN: 0253-7214
*Volume **44** Issue **S-6 Year 2023** Page **863:868***

An Approach for Loan Approval Prediction Using Machine Learning

Anusha G[1*], K Thanusha Reddy[2], G Tanmayee[3], G Roopa[4], Vani Krishnaswamy[5]

[1,2,3,4,5]*School of Computing and Information Technology*
Email: anushagopinath165@gmail.com[1], thanushareddykonikanti@gmail.com[2], tanudurgag@gmail.com[3], gownirooparedy22@gmail.com[4], vanikrishnaswamy@gmail.com[5]

**Corresponding author's E-mail:anushagopinath165@gmail.com*

Article History

Received: 06 June 2023
Revised: 05 Sept 2023
Accepted: 30 Nov 2023

Abstract

Banking sector is one such field where the company needs more accurate results after analysis. There are many people applying for bank loans from banks or other finance companies each day. But the banks cannot provide loan to every individual who is applying for loan. There is a very complex task that the bank employees do to study an analyze if the applicant is genuine or not. To find this out, there are a lot of factors to be considered. Going through this huge amount of data can be a really difficult task and yet one cannot be sure if the applicant will be able to pay back the loan within the given time or not. Objective of the paper is to make thorough analysis of the test data and make predictions if the applicant is genuine or not. For this process, we are using Machine Learning where the trained data is used to make predictions.

Keywords: *Banking Sector, Finance Companies, Loan, Trained Data, Test Data, Machine Learning, Prediction*

1. Introduction

There are many applicants who apply for loan each day. But there is no surety that the applicants are genuine and the money is in safe hands. Even though there are a lot of people applying for loan, loan cannot be granted to every applicant who has applied for loan. There has to be a filtration process to check the details of applicants and see if the loan can be granted or not. Banks have a fixed percentage of money to be issued as loan. Hence filtration among applicants is necessary. For banks, the interest that comes from the applicants of loan serves as a huge business part. Loan is kind of an asset for the bank as the income from loan interest brings good profit to the bank. Hence the loan has to distributed after a good thorough analysis.

Selection of approval of loan is a major business part of all the banks and finance companies. The employees working at the banking sectors face a lot of problem to process large data. There is a necessity to filter and give accurate result whether to approve the loan or not for a particular person. Processing huge data requires more time and effort. By the implementation of this project, we can reduce the time, effort and manpower required at the banking sectors. This problem is done by mining the data collected.

A. Scope of the project

The main contributions of this project therefore are:

- Data Analysis
- Dataset Pre-processing
- Training the Model
- Testing of Dataset

B. Domain overview

To predict the future or to do classification on information, machine learning is popular in doing such problems. The algorithms which we use in machine learning are trained over examples or instances from past experience and historical data is also analysed. As these algorithms are trained repeatedly, they are able to find patterns, which help to make future predictions. In machine learning algorithms data is the pillar. By using historical data, we are creating much more data by training algorithms.

Literature Review

There are numerous researches performed on loan prediction, few of them are listed in our related works. The authors in [1] created a model which could predict for high-dimensional data samples, identify hidden patterns in the input data and was a cost-effective method. But the disadvantages in this paper was that there was requirement of re-training for novel types of cases that are fraud and classification performance is poor when compared with other datamining methods.

In this paper [2] the authors have performed the creation of a model which was good for binary classification problems and had robustness toward a small number of data points. The drawback in this paper is that it is tough to process the results due to the transformation of the input data and it does not represent the information that is underlying in the domain.

The main theme of the paper [3] is to design a model that is able to give a solution for non-linear classification problems and acceptable for real-time operations. But the drawback of this paper was that the model was showing abnormal behaviors for various fraud cases and it suffered a minor reduction in the fraud detection.

This paper [4] provides a solution where the model gives estimates the absent data and maintains accuracy when a large piece of data is missing. But the drawback in this paper is that it is computationally expensive and if the boundaries are not informative, Support Vector Machine will not do well.

Proposed System

In the proposed system, normalization is used, which is a process of attribute value decomposition, it is done so that they will be within a particular range of minimum size.

It is used to change large database to a smaller one, also take out irregularities. Min-Max normalization is a popular method to handle such problems. It is the reason behind accomplishment of linear transformation of real data to a simple one.

The first step of random forest algorithm is selecting randomly ‘K’ features out of all ‘m’ features. Next, by implementing best split criteria we find out the root node of those randomly selected ’k’ features. In the third step, by implementing the best split criteria, child nodes are calculated. To do this we can use functions like ginindex () or infogain () which gives the best result. We repeat the first three steps till we form a tree having a root node and leaf node as a target. This enter step 1 to step 4 is done to generate ‘n’ randomly generated trees. These randomly generated trees form random forest.

In the proposed system, the main parameter that influence the classification accuracy is the amount of variables in the dataset and it states the fact that in datamining process, feature selection has a crucial role. There are two parameters that are crucial in this algorithm, in each tree there are how many random variables used and the number of trees used in each forest. We split the data into child nodes in a specific tree by using a function called information gain.

The drawbacks of the existing system are that the robustness of algorithms are not assessed. It also cannot learn if the environment is nonstationary. Advantages of the proposed system:

- Reduced excessive number of rules.
- Doubled the performance compare to the existing system.
- Can consider the noise data.
- Can achieve higher precision.
- Reduce the total cost and processing time.

2. Materials And Methods

There are three major modules implemented in this project:

MODULE 1: PRE-PROCESSING

In real world data, we frequently have lot of missing values. The reason might be failure to record data or corruption of data. Pre-processing should be done to handle such data as lot of machine learning algorithms cannot work if there is any missing data.

Features which have numeric value such as amount, age can be replaced by using Mean/Median/Mode strategy. We can replace the missing values by calculating mean, mode or median. It just like an approximation which can join variance to the dataset. If the data is linear we can use another method which can approximate it with deviation of neighbour values. When the absent values are of a sting column then it can be restored with a new category. This feature will have limited number of

possibilities, such as sex. As they have limited number of classes, we can give a new class for the missing's, for example 'u' for unknown. This method will append more data into the set resulting a change in variance. As they are categorical, we need to find a method to convert it into numerical format so that the algorithm can understand.

MODULE 2: MODEL SELECTION

In many supervised algorithms, random forest is one among them. It is flexible and simple to use. It is used for both classification and regression. Generally, a forest has trees. It is said that if a forest has more trees, then that forest is considered as a strong forest.

In random forest algorithm, data samples are randomly collected and decision trees are generated and by means of voting we choose the best solution from the predictions we get from each tree.

On a randomly split data, decision trees are generated as it is a divide and conquer approach. A group of decision tree classifier is known as forest. Using an attribute selection indicator such as gain ratio, information gains these individual decision trees are.

Every tree depends on a single random sample. In classification problems, every tree votes and the majority class are selected as the last result but in regression, the average of all the three outputs is taken. It is powerful and simple than all the non-linear classification algorithm.

The K-NN classification algorithm decides on the presumption being true for the functionality of algorithm. K-NN takes closeness into account like proximity, distance with some basic mathematics methods.

With given k-value we can set limits of each class and all the training readings remain constant. These limits will separate RC from GS. Similarly, let us see the effect ok 'k' value on the class limits.

MODULE 3: MAKING PREDICTIONS

There are many distance metrics, let us discuss a few popularly used metrics. Initially let us know the concept behind these metrics and then learn which algorithm use these distance metrics.

In normal vector space, distance is a metric. A normal vector space is where a norm is defined on a vector space. Consider X is a real valued function where X is a vector space and real valued function ||x|| satisfies these following conditions:

- Triangle inequality- calculated distance between two points is always a straight line if distance is a norm.
- Scalar factor- when you multiply with a positive number, its length varies if the direction of vector doesn't change.
- Zero vector- it will have length as zero.

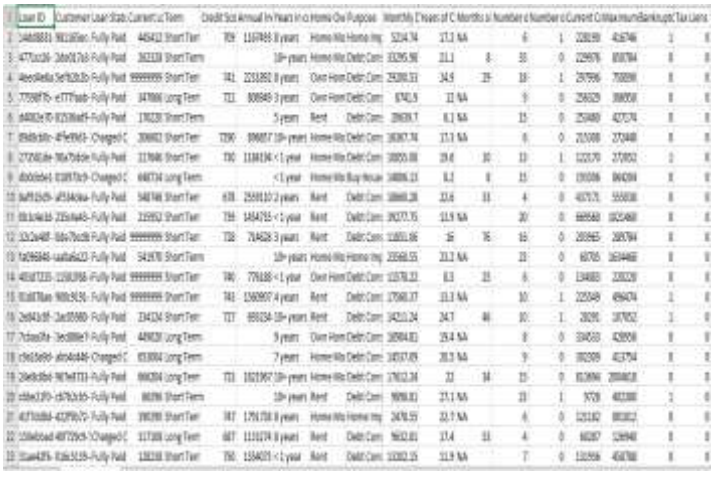


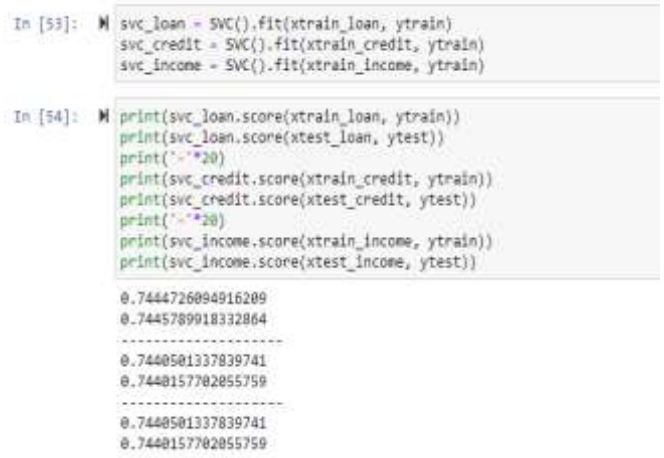
Fig. 1 Example of the database maintained for training.

This figure here is the example of the data frame taken from the publicly available resource.

3. Results and Discussion

To grant the loan we set a value. If the value is greater than the threshold value, then we grant the loan. If the value is less than threshold value, the application will be denied. To do this we used different algorithms on the data frame. At the end we will get the accuracy of the algorithms used for trained and

test data set. Following are the pictures of accuracy value of the different scores. Each algorithm has different impact on the given data frame. The following figures are the outputs that we got using the algorithms in Jupyter Notebook.



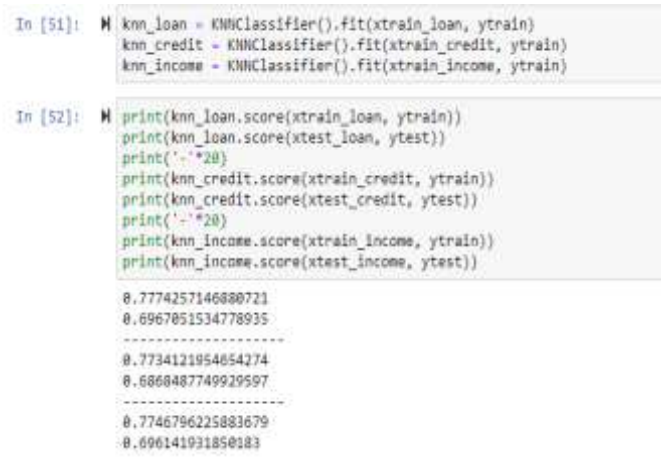
```
In [53]: svc_loan = SVC().fit(xtrain_loan, ytrain)
         svc_credit = SVC().fit(xtrain_credit, ytrain)
         svc_income = SVC().fit(xtrain_income, ytrain)

In [54]: print(svc_loan.score(xtrain_loan, ytrain))
         print(svc_loan.score(xtest_loan, ytest))
         print('-'*20)
         print(svc_credit.score(xtrain_credit, ytrain))
         print(svc_credit.score(xtest_credit, ytest))
         print('-'*20)
         print(svc_income.score(xtrain_income, ytrain))
         print(svc_income.score(xtest_income, ytest))

0.7444726094916209
0.7445789918332864
--------------------
0.7440501337839741
0.7440157702055759
--------------------
0.7440501337839741
0.7440157702055759
```

Fig. 2 Support Vector Classification Output

In figure 2 we have the precision for trained and test values of loan score, credit score and income score using Support Vector Classification.



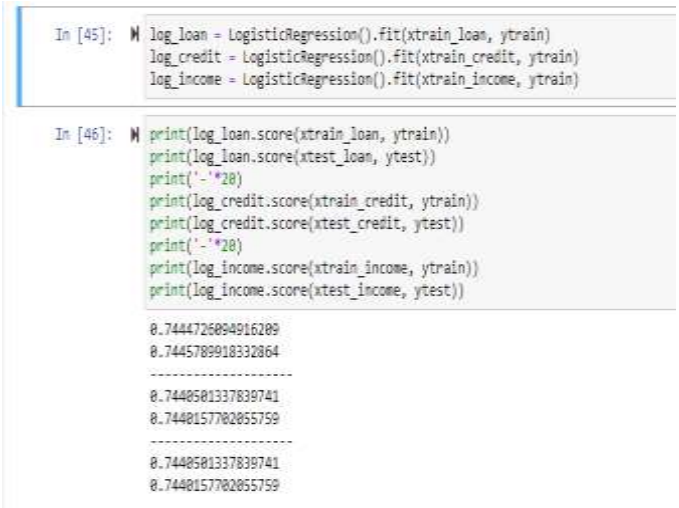
```
In [51]: knn_loan = KNNClassifier().fit(xtrain_loan, ytrain)
         knn_credit = KNNClassifier().fit(xtrain_credit, ytrain)
         knn_income = KNNClassifier().fit(xtrain_income, ytrain)

In [52]: print(knn_loan.score(xtrain_loan, ytrain))
         print(knn_loan.score(xtest_loan, ytest))
         print('-'*20)
         print(knn_credit.score(xtrain_credit, ytrain))
         print(knn_credit.score(xtest_credit, ytest))
         print('-'*20)
         print(knn_income.score(xtrain_income, ytrain))
         print(knn_income.score(xtest_income, ytest))

0.7774257146880721
0.6967051534778935
--------------------
0.7734121954654274
0.6868487749929597
--------------------
0.7746796225883679
0.696141931850183
```

Fig. 3 KNN Output

In figure 3 we have the precision for trained and test values of loan score, credit score and income score using KNN.



```
In [45]: log_loan = LogisticRegression().fit(xtrain_loan, ytrain)
         log_credit = LogisticRegression().fit(xtrain_credit, ytrain)
         log_income = LogisticRegression().fit(xtrain_income, ytrain)

In [46]: print(log_loan.score(xtrain_loan, ytrain))
         print(log_loan.score(xtest_loan, ytest))
         print('-'*20)
         print(log_credit.score(xtrain_credit, ytrain))
         print(log_credit.score(xtest_credit, ytest))
         print('-'*20)
         print(log_income.score(xtrain_income, ytrain))
         print(log_income.score(xtest_income, ytest))

0.7444726094916209
0.7445789918332864
--------------------
0.7440501337839741
0.7440157702055759
--------------------
0.7440501337839741
0.7440157702055759
```

Fig. 4 Logistic regression Output

In figure 4 we have the precision for trained and test values of loan score, credit score and income score using logistic regression classification.

Fig. 5 Random Forest Output

In figure 5 we have the precision for the trained and test values of loan score, credit score, and income score using random forest classification.

4. Conclusion

By considering the changing trend in the banking sector, we come into a conclusion that loan approval prediction using machine learning is very necessary as it reduces the work of the employees of banks or any other finance companies. A task that consumes a lot of time and effort can be done easily with the help of this project. It is a very time-consuming task for the employees go through huge data sets to make analysis for an applicant. A detailed study on loan approval prediction has been presented in this paper which proves that the bank employees can save a lot of time by using this system than the traditional methodology where the employees to go through huge data and make predictions.

Acknowledgment

Eventually we are grateful to God Almighty and a number of individuals whose professional guidance, assistance and encouragement have made it a pleasure endeavour to undertake this Course. We would like to express our sincere gratitude to the esteemed REVA University, Bangalore for the wonderful opportunity given to us in carrying out this project. We wish to express heartfelt thanks to my research supervisor Prof. Vani K, for her valuable guidance, motivation during my project. She has been a source of inspiration and her tireless guidance and incredible effort has enabled to complete this course successfully. Her Simplicity, honesty, patience, generosity and perfectness are qualities that have inspired us a lot. We express our deepest gratitude to Dr. P. Shyama Raju, Chancellor, REVA University, Bangalore, for the environment and infrastructure provided to carry out and complete my project under one roof in REVA University campus, Bangalore. We owe our deepest gratitude to Dr. Surendra Rao Shankapal, I/C Vice-Chancellor REVA University, Bangalore, for his continuous support, encouragement. It is pleasure to express our gratitude whole heartily thanks to Dr. Sunil Kumar S. Manvi, Director, School of Computing and Information Technology, REVA University, Bangalore, to timely process the research related tasks.

References:

[1] Kuldeep Randhawa, Manjeevan Seera, Chu Kiong Loo, Chee Peng Lim, Asoke K. Nandi-"Credit card fraud detection using AdaBoost and majority voting" published in 2018.

[2] Andrea Dal Pozzolo, Olivier Caelen, Giacomo Boracchi, Cesare Alippi, Gianluca Bontempi-"Credit Card Fraud Detection: A Realistic Modeling and a Novel Learning Strategy" published in 2018.

[3] Changjun Jiang, Guanjun Liu, Jiahui Song, Lutao Zheng, and Wenjing Luan-"Credit Card Fraud Detection: A Novel Approach Using Aggregation Strategy and Feedback Mechanism" published in 2018.

[4] G. Arutjothi, Dr. C. Senthamarai-"Prediction of Loan Status in Commercial Bank using Machine Learning Classifier." Published in 2017.

[5] Shiyang Liao, Ruiyun Yu, Junbo Wang, Koichi Sato, and Zixue Cheng, " CNN for situat ions understanding based on sent iment analysis of twit ter data," Procedia computer science, 111:376–381, 2017.CrossRef.

[6] K I Rahmani, Amit Kumar Goel, M.A. Ansari, "An Efficient Indexing Algorithm for CBIR,"IEEE- Internat ional Conference on Computat ional Intelligence & Communication Technology ,13-14 Feb 2015.

[7] Gurlove Singh, Amit Kumar Goel ,"Face Detect ion and Recognit ion System using Digital Image Processing" , 2nd Internat ional conference on Innovat ive Mechanism for Indust ry Applicat ion ICMIA 2020, 5-7 March 2020, IEEE Publisher.

[8] O. Adewumi and A. A. Akinyelu, "A survey of machine-learning and nature-inspired based credit card fraud detection techniques," International Journal of System Assurance Engineering and Management, vol. 8, pp. 937–953, 2017.

[9] N. S. Halvaiee and M. K. Akbari, "A novel model for credit card fraud detection using Artificial Immune Systems," Applied Soft Computing, vol. 24, pp. 40–49, 2014.
[10]Alippi, G. Boracchi, and M. Roveri, "Hierarchical change-detection tests," IEEE Trans. Neural Netw. Learn. Syst., vol. 28, no. 2, pp. 246–258, Feb. 2016.
[11] A. Abuhamad, P. Falkensammer, F. Reichartseder, and Y. Zhao