JOURNAL OF CRITICAL REVIEWS

ISSN- 2394-5125 VOL 08, ISSUE 05, 2021

Machine Learning for Loan Prediction Dataset with Data Analysis

Patlannagari Hasitha Reddy, Nusrath Mohammad

Department of Computer Science and Engineering

Sree Dattha Group of Institutions, Hyderabad, Telangana, India.

Abstract

With the enhancement in the banking sector lots of people are applying for bank loans but the bank has its limited assets which it must 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. So, in this project we try to reduce this risk factor behind selecting the safe person to save lots of bank efforts and assets. This is done by mining the Big Data of the previous records of the people to whom the loan was granted before and based on these records/experiences the machine was trained using the machine learning model which give the most accurate result. The main objective of this paper is to predict whether assigning the loan to person will be safe or not. This work is divided into four sections such as data collection, comparison of machine learning models on collected data, training of system on most promising model, and testing.

Loan Prediction is very helpful for employee of banks as well as for the applicant also. The aim of this Paper is to provide quick, immediate, and easy way to choose the deserving applicants. It can provide special advantages to the bank. The loan prediction system can automatically calculate the weight of each features taking part in loan processing and on new test data same features are processed with respect to their associated weight. A time limit can be set for the applicant to check whether his/her loan can be sanctioned or not. Loan prediction system allows jumping to specific application so that it can be check on priority basis. This project is exclusively for the managing authority of Bank/finance Company, whole process of prediction is done privately no stakeholders would be able to alter the processing. Result against particular Loan Id can be send to various departments of banks so that they can take appropriate action on application. This helps all others department to carried out other formalities.

Keywords: Loan prediction, machine learning, data analysis.

1. 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 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.

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 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, 2] 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 applications that deserve to be approved on a priority basis.

2. Literature survey

Sheikh et al. studied a very important approach in predictive analytics is used to study the problem of predicting loan defaulters: The Logistic regression model. The data is collected from the Kaggle for studying and prediction. Logistic Regression models have been performed and the different measures of performances are computed. The models are compared since the performance measures such as sensitivity and specificity.

Tumuluru et al. used the Machine Learning (ML) algorithms to extract patterns from a common loan-approved dataset and retrieve patterns in forecasting future loan defaulters. Customers' past data, such as their age, income, loan amount, and tenure of work, will be used to conduct the analysis. To determine the maximum relevant features, i.e., the factors that have the most impact on the prediction outcome, various ML algorithms such as Random Forest, Support Vector Machine, K-Nearest Neighbor and Logistic Regression, were used. These mentioned algorithms are evaluated with the standard metrics and compared with each other. The random forest algorithm achieves better accuracy.

Lohani et al. aimed to minimize the credit risks of defaulting. This study has applied logistic regression as a tool to predict whether an applicant is eligible for the loan or not. The data is collected from the Kaggle for studying and prediction.

Shaheen et al. applied on different machine learning techniques on customer's loans dataset obtained from a public bank's database that contains customer's loans and personal data. The data is processed and analyzed using Apache Spark, a machine learning tool for big data processing. The result of the proposed system is evaluated by seven performance metrics to compare the performance of each classifier and find out the best performing one among them. It is found out that the ensemble machine learning techniques has better performance than single base classifiers in predicting the loan default.

Sharma et al. studied the learning techniques as well as the raw datasets utilized for training and test sets. The system model's precision is also discussed. This work also provides a quick overview of a few datasets that can be used to anticipate loan/mortgage analysis. Recent and future trends are also spotlighted.

Gupta et al. used 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.

Maheswari et al. used statistical measures to preprocess the data and build an effective model that will predicts the loan defaulter accurately.

Lai et al. demonstrated that the AdaBoost model can achieve a 100% accuracy for predicting loan default, outperforming other models including XGBoost, random forest, k nearest neighbors, and multilayer perceptrons. This result showed the promising application of machine learning techniques in the financial industry.

Ereiz et al. demonstrated the prediction using machine learning models is very high but depends on the quality of the data. Several algorithms (to be more specific - BigML's OptiML) were used to identify the best suited for the lending business.

Kumar et al. reduced the risking factor of banks behind finding the appropriate person for loan approval by the bank. This work even reduced the time of loan approval analysis. This work first used data mining techniques to analyze previous records to which the bank has already sanctioned loan based on the analysis made out of these records this work train the deep learning model. The new data is treated as testing data, and the output of the customer is calculated accordingly.

3. Proposed system

Fig. 1 shows the proposed block diagram of loan description.

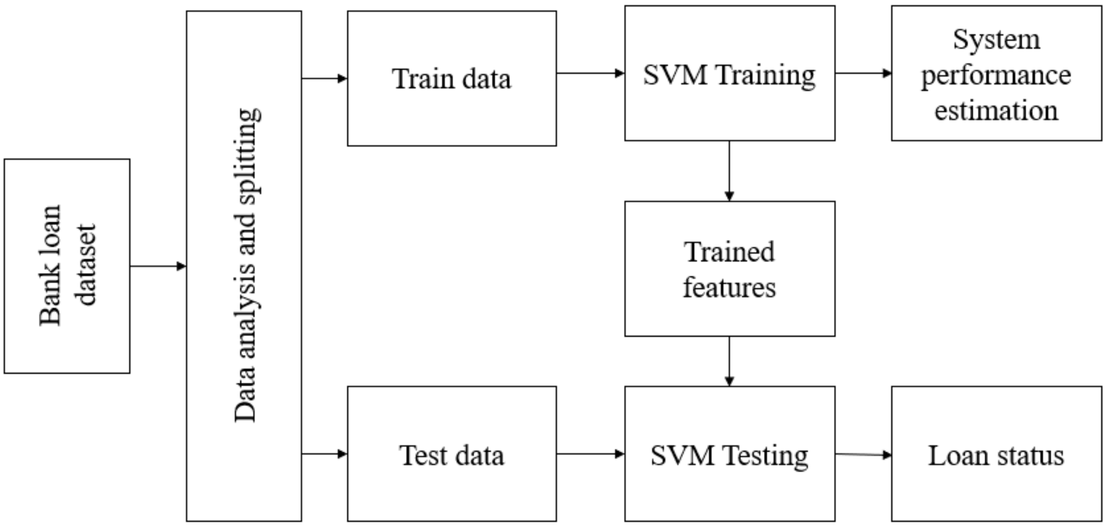


Fig. 1: Block diagram of proposed system.

3.1 Dataset Description

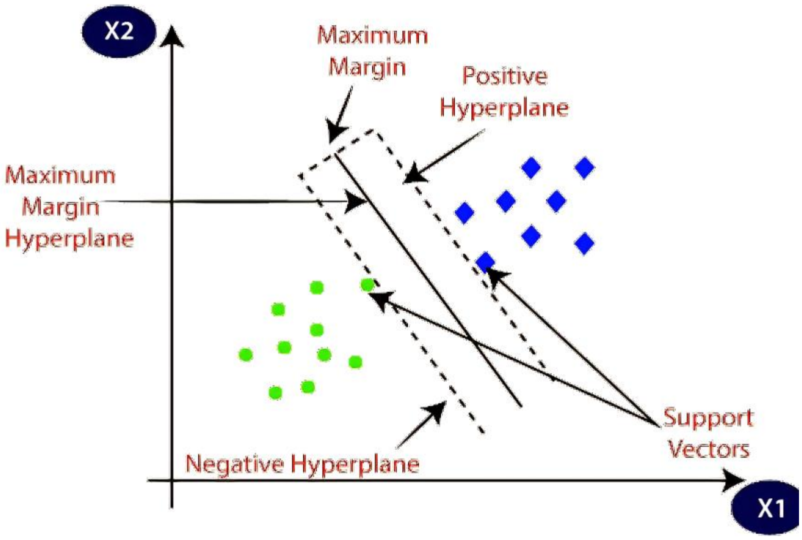
13-Columns: Loan_ID, Gender, Married, Dependents, Education, Self_Employed, ApplicantIncome, CoapplicantIncome, LoanAmount, Loan_Amount_Term, Credit_History, Property_Area, Loan_Status,

615-Rows

3.2 Support Vector Machine Algorithm (SVM)

Support Vector Machine or SVM is one of the most popular Supervised Learning algorithms, which is used for Classification as well as Regression problems. However, primarily, it is used for Classification problems in Machine Learning. The goal of the SVM algorithm is to create the best line or decision boundary that can segregate n-dimensional space into classes so that we can easily put the new data point in the correct category in the future. This best decision boundary is called a hyperplane.

SVM chooses the extreme points/vectors that help in creating the hyperplane. These extreme cases are called as support vectors, and hence algorithm is termed as Support Vector Machine. Consider the below diagram in which there are two different categories that are classified using a decision boundary or hyperplane:



Applications

- Face recognition
- Weather prediction
- Medical diagnosis
- Spam detection
- Age/gender identification
- Language identification
- Sentimental analysis
- Authorship identification
- News classification

3.3 Advantages of proposed system

- SVM works relatively well when there is a clear margin of separation between classes.
- SVM is more effective in high dimensional spaces.
- SVM is effective in cases where the number of dimensions is greater than the number of samples.
- SVM is relatively memory efficient.

4. Results

Module Description

- Bank Dataset
- Data analysis and Splitting
- Train data
- Test data
- SVM Training
- Trained features

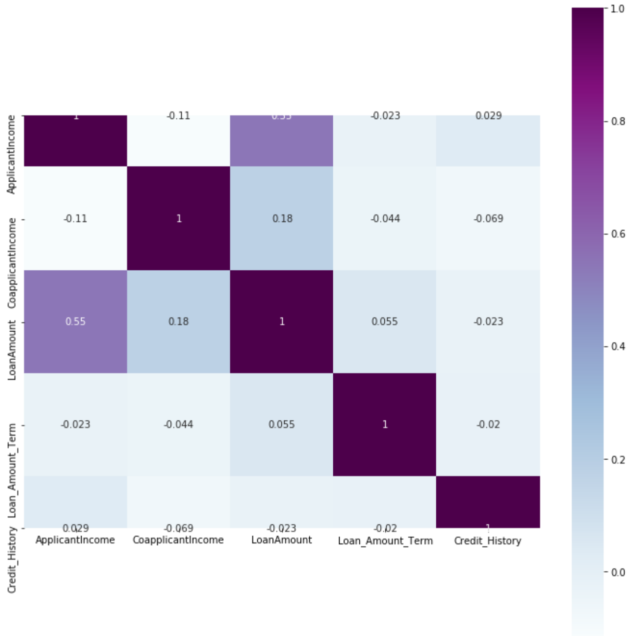
ISSN- 2394-5125 VOL 08, ISSUE 05, 2021

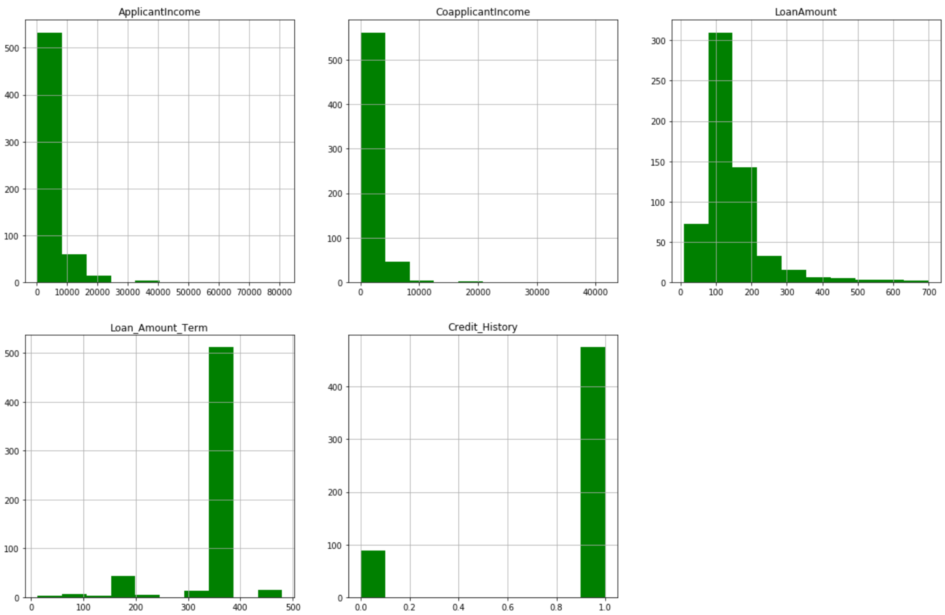
- SVM Testing
- System performance estimation
- Loan status

Sample dataset

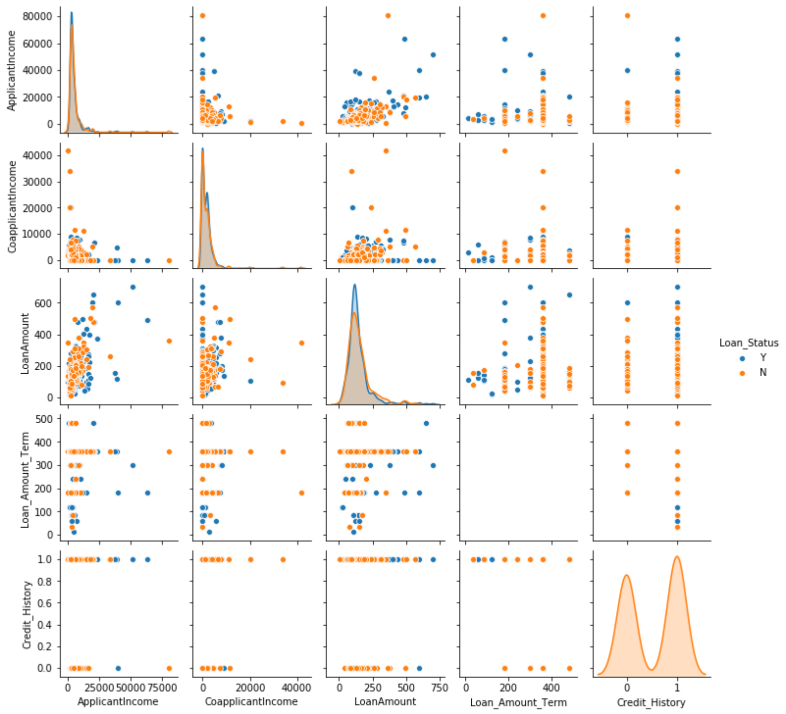
```
    Loan_ID Gender Married Dependents     Education Self_Employed  \
0  LP001002   Male      No          0      Graduate            No
1  LP001003   Male     Yes          1      Graduate            No
2  LP001005   Male     Yes          0      Graduate           Yes
3  LP001006   Male     Yes          0  Not Graduate            No
4  LP001008   Male      No          0      Graduate            No

   ApplicantIncome  CoapplicantIncome  LoanAmount  Loan_Amount_Term  \
0             5849                0.0         NaN             360.0
1             4583             1508.0       128.0             360.0
2             3000                0.0        66.0             360.0
3             2583             2358.0       120.0             360.0
4             6000                0.0       141.0             360.0

   Credit_History Property_Area Loan_Status
0             1.0         Urban           Y
1             1.0         Rural           N
2             1.0         Urban           Y
3             1.0         Urban           Y
4             1.0         Urban           Y
(981, 13)
Index(['Loan_ID', 'Gender', 'Married', 'Dependents', 'Education',
       'Self_Employed', 'ApplicantIncome', 'CoapplicantIncome', 'LoanAmount',
       'Loan_Amount_Term', 'Credit_History', 'Property_Area', 'Loan_Status'],
      dtype='object')

            TRAINING DATA DETAILS
            Total  number of records present in the dataset - 614
            Total  number of columns present in the dataset - 13

             TESTING DATA DETAILS
            Total  number of records present in the dataset - 367
            Total  number of columns present in the dataset - 12

TOTAL NUMBER OF RECORDS IN THE COMBINED DATASET  -  981


Number of value counts for - Gender
Male      775
Female    182
Name: Gender, dtype: int64
Number of Missing values: 24


Number of value counts for - Married
Yes    631
No     347
Name: Married, dtype: int64
Number of Missing values: 3


Number of value counts for - Dependents
0     545
2     160
1     160
3+     91
Name: Dependents, dtype: int64
Number of Missing values: 25


Number of value counts for - Education
Graduate        763
Not Graduate    218
Name: Education, dtype: int64
Number of Missing values: 0


Number of value counts for - Self_Employed
No     807
Yes    119
Name: Self_Employed, dtype: int64
Number of Missing values: 55
```

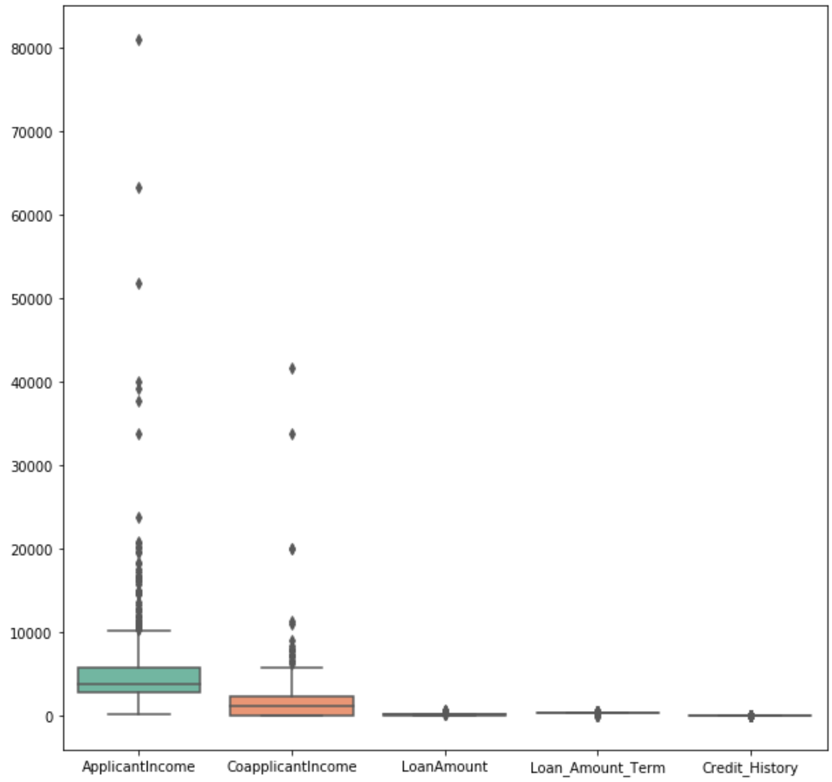
```
Number of value counts for - Property_Area
Semiurban    349
Urban        342
Rural        290
Name: Property_Area, dtype: int64
Number of Missing values: 0


Number of value counts for - Loan_Status
Y    422
N    192
Name: Loan_Status, dtype: int64
Number of Missing values: 367


Number of missing values in  ApplicantIncome : 0
Number of missing values in  CoapplicantIncome : 0
Number of missing values in  LoanAmount : 27
Number of missing values in  Loan_Amount_Term : 20
Number of missing values in  Credit_History : 79
```





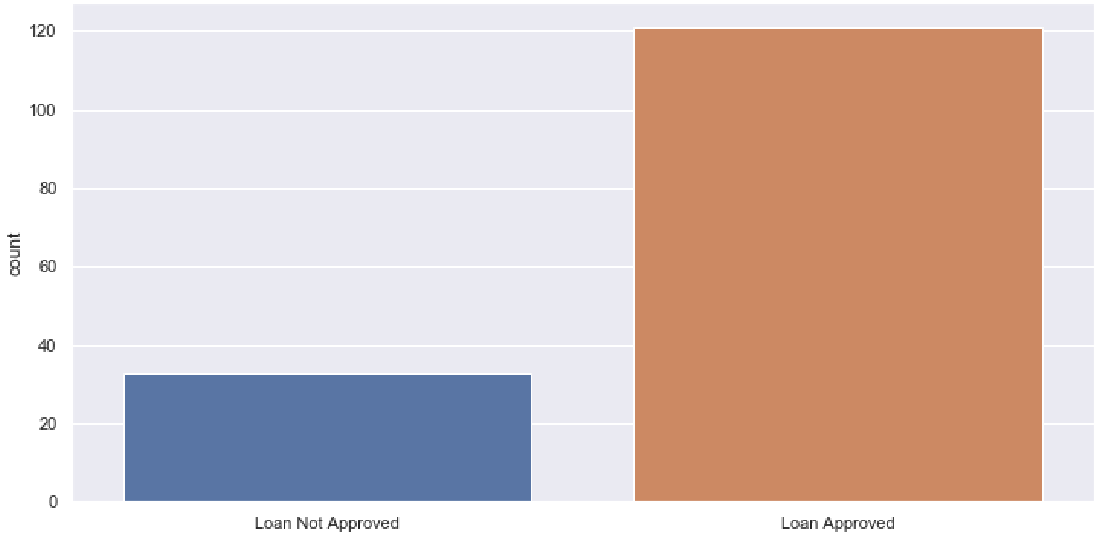

Loan_Status
N
Y
400
300
200
100
0
0.0
1.0
Credit_History


ApplicantIncome
CoapplicantIncome
LoanAmount
Loan_Amount_Term
Credit_History
Loan_Status
Y
N

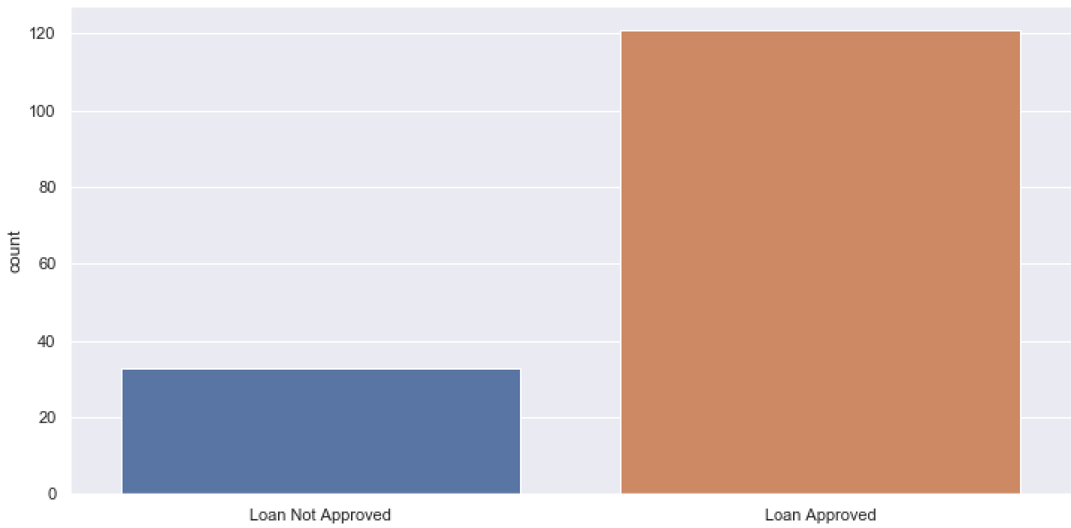



```
Accuracy : 73.171%

 Classification report for classifier LogisticRegression(C=1.0, class_weight=None, dual=False, fit_intercept=True,
                   intercept_scaling=1, l1_ratio=None, max_iter=100,
                   multi_class='auto', n_jobs=None, penalty='l2',
                   random_state=None, solver='sag', tol=0.0001, verbose=0,
                   warm_start=False):
              precision    recall  f1-score   support

           0       0.00      0.00      0.00        33
           1       0.73      1.00      0.85        90

    accuracy                           0.73       123
   macro avg       0.37      0.50      0.42       123
weighted avg       0.54      0.73      0.62       123


Confusion matrix:
[[ 0 33]
 [ 0 90]]
TOTAL NUMBER OF TESTING RECORD -  123
NUMBER OF CORRECTLY PREDICTED OUTPUTS -  90
```

```
Accuracy : 79.221%

 Classification report for classifier LogisticRegression(C=1.0, class_weight=None, dual=False, fit_intercept=True,
                   intercept_scaling=1, l1_ratio=None, max_iter=100,
                   multi_class='auto', n_jobs=None, penalty='l2',
                   random_state=None, solver='sag', tol=0.0001, verbose=0,
                   warm_start=False):
              precision    recall  f1-score   support

           0       0.67      0.51      0.58        43
           1       0.83      0.90      0.86       111

    accuracy                           0.79       154
   macro avg       0.75      0.71      0.72       154
weighted avg       0.78      0.79      0.78       154


Confusion matrix:
[[ 22  21]
 [ 11 100]]
TOTAL NUMBER OF TESTING RECORD -  154
NUMBER OF CORRECTLY PREDICTED OUTPUTS -  122
```




```
Accuracy : 79.221%

 Classification report for classifier LogisticRegression(C=1.0, class_weight=None, dual=False, fit_intercept=True,
                  intercept_scaling=1, l1_ratio=None, max_iter=100,
                  multi_class='auto', n_jobs=None, penalty='l2',
                  random_state=None, solver='sag', tol=0.0001, verbose=0,
                  warm_start=False):
             precision    recall  f1-score   support

           0       0.67      0.51      0.58        43
           1       0.83      0.90      0.86       111

    accuracy                           0.79       154
   macro avg       0.75      0.71      0.72       154
weighted avg       0.78      0.79      0.78       154


Confusion matrix:
[[ 22  21]
 [ 11 100]]
TOTAL NUMBER OF TESTING RECORD -  154
NUMBER OF CORRECTLY PREDICTED OUTPUTS -  122
```



5. Conclusion

This application can help banks in anticipating the fate of credit and its status and relies upon that they can make a move in introductory long periods of advance. Utilizing this application banks can diminish the quantity of awful advances from bringing about cut off misfortunes. A few AI calculations and bundles were utilized to set up the information and to fabricate the arrangement model. AI bundle libraries help in fruitful information examination and highlight determination. Utilizing this technique bank can without much of a stretch distinguish the necessary data from immense measure of informational collections and aides in fruitful advance forecast to diminish the quantity of awful credit issues. Information mining strategies are helpful to the financial part for better focusing on and procuring new clients, most significant client maintenance, programmed credit endorsement, which is utilized for extortion avoidance, misrepresentation identification progressively, giving section-based item, investigation of the client, exchange designs after some time for better maintenance and relationship, hazard the executives and showcasing.

References

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

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

[3] M. A. Sheikh, A. K. Goel and T. Kumar, "An Approach for Prediction of Loan Approval using Machine Learning Algorithm," 2020 International Conference on Electronics and Sustainable Communication Systems (ICESC), 2020, pp. 490-494, doi: 10.1109/ICESC48915.2020.9155614.

[4] P. Tumuluru, L. R. Burra, M. Loukya, S. Bhavana, H. M. H. CSaiBaba and N. Sunanda, "Comparative Analysis of Customer Loan Approval Prediction using Machine Learning Algorithms," 2022 Second International Conference on Artificial Intelligence and Smart Energy (ICAIS), 2022, pp. 349-353, doi: 10.1109/ICAIS53314.2022.9742800.

[5] B. P. Lohani, M. Trivedi, R. J. Singh, V. Bibhu, S. Ranjan and P. K. Kushwaha, "Machine Learning Based Model for Prediction of Loan Approval," 2022 3rd International Conference on Intelligent Engineering and Management (ICIEM), 2022, pp. 465-470, doi: 10.1109/ICIEM54221.2022.9853160.

[6] S. K. Shaheen and E. ElFakharany, "Predictive analytics for loan default in banking sector using machine learning techniques," 2018 28th International Conference on Computer Theory and Applications (ICCTA), 2018, pp. 66-71, doi: 10.1109/ICCTA45985.2018.9499147.

[7] Sharma, A., Kumar, V. (2023). An Exploratory Study-Based Analysis on Loan Prediction. In: Ranganathan, G., Fernando, X., Rocha, Á. (eds) Inventive Communication and Computational Technologies. Lecture Notes in Networks and Systems, vol 383. Springer, Singapore. https://doi.org/10.1007/978-981-19-4960-9_33.

[8] A. Gupta, V. Pant, S. Kumar and P. K. Bansal, "Bank Loan Prediction System using Machine Learning," 2020 9th International Conference System Modeling and Advancement in Research Trends (SMART), 2020, pp. 423-426, doi: 10.1109/SMART50582.2020.9336801.

[9] P. Maheswari and C. V. Narayana, "Predictions of Loan Defaulter - A Data Science Perspective," 2020 5th International Conference on Computing, Communication and Security (ICCCS), 2020, pp. 1-4, doi: 10.1109/ICCCS49678.2020.9277458.

[10] L. Lai, "Loan Default Prediction with Machine Learning Techniques," 2020 International Conference on Computer Communication and Network Security (CCNS), 2020, pp. 5-9, doi: 10.1109/CCNS50731.2020.00009.

ISSN- 2394-5125 VOL 08, ISSUE 05, 2021

[11] Z. Ereiz, "Predicting Default Loans Using Machine Learning (OptiML)," 2019 27th Telecommunications Forum (TELFOR), 2019, pp. 1-4, doi: 10.1109/TELFOR48224.2019.8971110.

[12] Kumar, A., Dugyala, R., Bhattacharya, P. (2021). Prediction of Loan Scoring Strategies Using Deep Learning Algorithm for Banking System. In: Singh, P.K., Polkowski, Z., Tanwar, S., Pandey, S.K., Matei, G., Pirvu, D. (eds) Innovations in Information and Communication Technologies (IICT-2020). Advances in Science, Technology & Innovation. Springer, Cham. https://doi.org/10.1007/978-3-030-66218-9_13