iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 11 **Issue:** III **Month of publication:** March 2023

DOI: https://doi.org/10.22214/ijraset.2023.49748

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 11 Issue III Mar 2023- Available at www.ijraset.com

Feature Selection for Loan Repayment Prediction System Using Machine Learning

Jishnu Goyal[1], Abhay Sota[2], Varun Arora[3]
[1, 2]*Department of Computer Science and Engineering, Maharaja Agrasen Institute of Technology, Delhi, India*
[3]*These authors contributed equally to this work and share first authorship*

Abstract: ***It is essential for banks to evaluate and predict the repayment ability of the loaners in order to minimise the risk of loan payment default. Due to this, there are systems created by the banks to process the loan request based on the loaners' status, such as employment status, credit history, etc. This paper attempts to determine the most significant factors/features which help in predicting whether a loan applicant would be able to repay their loan. Feature selection provides an effective way to solve this problem by removing irrelevant and redundant data, which can reduce computation time, improve learning accuracy, and facilitate a better understanding of the learning model or data. In order to properly assess the repayment ability of all groups of people, several frequently-used evaluation measures for feature selection are applied, and different sets of features using different feature selection methods are generated. Afterwards, those sets are tested against different machine learning models, to figure out the most effective feature set that should be analysed in order to figure out the repayment ability of an applicant. The data used in this study was gathered from a Kaggle Dataset which contained the details of over 300,000+ loaners and whether they were able to repay their loans or not. After data cleaning and feature engineering, the dataset still appeared quite imbalanced , so, along with accuracy, other measures such as precision, recall, and F1 Score were also considered. Results of the study indicate that, days employed, number of family members , number of children , income of the person were some of the most significant factors for determining a borrower's performance.***

I. INTRODUCTION

Many people struggle to get loans from trustworthy sources such as banks, due to insufficient credit histories.In 2015, the federal Consumer Financial Protection Bureau (CFPB) reported that one of every 10 American adults is "credit invisible," meaning they don't have a credit history with one of the three major credit bureaus. Usually students and unemployed adults, who don't have enough credit history fall under this category, as supported by the following data "20 percent of people ages 18 to 22 have no credit report, according to data by credit rating company VantageScore". Apart from evaluating the borrower based on their credit score, there are other ways to measure or predict their ability to repay. For example, employment is generally a big factor which affects the person's repayment ability since an employed adult has more stable incomes and cash flow. Factors such as marital status, property owned, number of dependents, might also affect the study of the repayment ability.

In this project, feature selection methods are used to choose the set of factors that play a crucial role in determining the repayment ability of an applicant. Feature selection refers to the process of obtaining a subset from an original feature set according to a certain criterion. This crucially helps in compressing the data , where the redundant and irrelevant features are removed [2][3]. This pre-processing reduces the data to train and test, which in turn reduces the time taken by the model to learn, thus it simplifies the results [1][3]. The dataset, 'Loan Application Prediction Analysis' from Kaggle.com, was used in this project .This open dataset contains 300,000+ anonymous clients' with 122 unique features [1][4]. Due to such a large dataset, feature selection is highly recommended in order to reduce the training time and simplify the results. The study of correlation between these features and repayment ability of the clients, would help lenders evaluate borrowers from the most significant dimensions and would also help borrowers, especially those who do not have sufficient credit histories, to find a credible loaner.

II. LITERATURE REVIEW

Numerous pieces of literature about loan prediction have been published already and are available for public usage. A recent paper published in 2019 , ' Loanliness : Predicting Loan Repayment Ability by Using Machine Learning Methods '[1] aims at predicting the loan repayment ability of a loaner by training the dataset with different models and then choosing the best model with highest accuracy [1][6][8]. It stresses on different data balancing techniques such as up-sampling and down-sampling to balance the dataset [1].

Out of all the models, K-clustering achieved the highest accuracy with 71.57%[1]. 'Credit Risk Analysis and Prediction Modelling of Bank Loans Using R' by Sudhamathy G. focused on preprocessing and used clustering and classification techniques in R to prepare the data for further use [8].The decision tree classifier was then built using the preprocessed dataset which achieved 0.833 precision [8]. The 'Loan Prediction by using Machine Learning Models' paper also emphasizes on pre-processing where it uses Outlier detection and removal, as well as imputation removal processing in the pre-processing stage [1][6]. To predict the chances of current status regarding the loan approval process, SVM, DT, KNN, and gradient boosting models were used [6]. According to the results, experimentation concluded that the Decision Tree has significantly higher loan prediction accuracy than the other models.It yielded an accuracy of 81.1% [6].

Another paper takes a different approach and uses Exploratory Data Analysis (EDA) as a method for predicting loan amounts based on the nature of the client and their needs [7]. Annual income versus loan purpose, customer trust, loan tenure versus delinquent months, loan tenure versus credit category, loan tenure versus credit category, loan tenure versus the number of years in current job, and chances for loan repayment versus homeownership were the major factors concentrated during the data analysis [7]. The purpose of this study was to infer the constraints that the customer faces when applying for a loan, as well as to make a prediction about repayment[1][7][8]. It also revealed that borrowers are more interested in short term loans than long term loans [7].

III. RESEARCH METHODOLOGY

Dataset was collected from Kaggle containing information regarding all the loans approved and whether the borrower was able to repay the loan or not. It contained data of 300,000+ respondents, out of which more than 90% of borrowers were able to repay their loans. At the initial stages, the dataset was cleaned. All the null values were either removed or converted into a not-null value depending upon the field and the datatype. Due to such imbalance in the dataset, the result was likely to be skewed. As a result, SMOTE analysis was applied for Up-Sampling the dataset and to increase its balance. After this , the feature selection technique of 'Correlation' was applied to select the features which are highly correlated to the class. It was made sure that certain features with negative correlation were also selected since Up-Sampling was applied earlier which increases the impact of the minority values. Upon the feature selection, different machine learning models like Logistic Regression, Naïve Bayes and Random Forest were applied. Apart from Correlation, 'Mutual Information' Feature Selection was also applied, and certain features were selected, followed by the application of Logistic Regression. It yielded 91.3% accuracy. The result generated showed that Employment Days, Income, Number of Members, Number of Children were some of the most significant features. In addition, the Random Forest model generated the highest accuracy, followed by Naïve Bayes and Logistic Regression.

IV. DATASET

A. Introduction

The dataset was obtained from Kaggle. Initially it had information about 300,000+ people to whom loans were granted and whether they were able to repay their loans or not. It contained 122 columns/dimensions including the TARGET variable which indicated the loan repayment status. The number of people who were able to repay their loans is extremely high when compared to the loan repayment defaulters. Fig 1 shows the frequency of loan defaulters and loan payers.

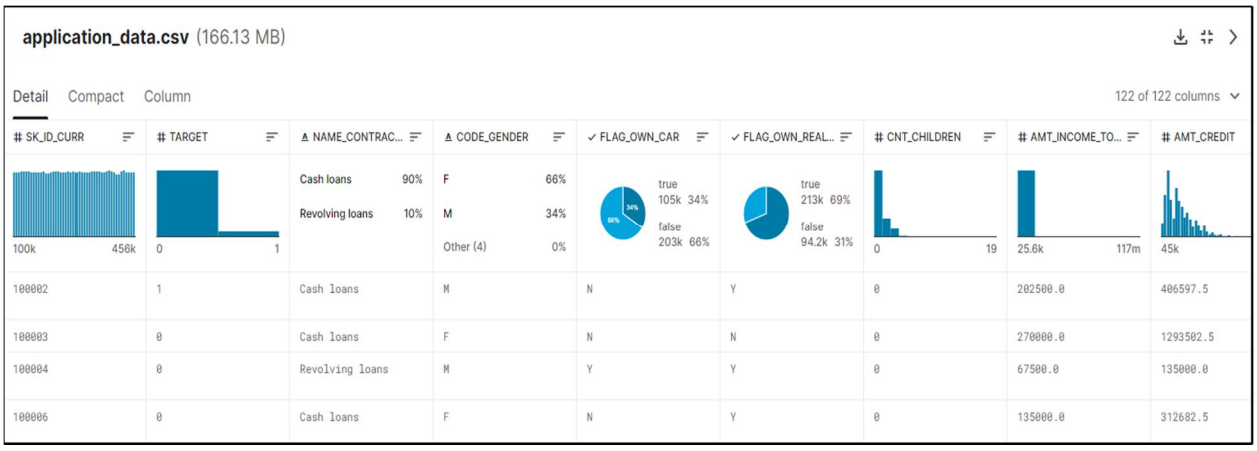


Fig 1. Frequency of 0 and 1 in 'TARGET' field

B. Challenges with the dataset

1) *Null Values and Unknown Features:* The dataset was inspected and it was found that many entries contain invalid values such as NaN(not a number). There are also three features 'EXT SOURCE 1', 'EXT SOURCE 2', and 'EXT SOURCE 3' and their representation was unknown. Existing models are evaluated using the three features and by removing the invalid values.
2) *Imbalanced Dataset:* According to the dataset, more than 90% people whose loans were sanctioned managed to repay their loans. This resulted in an imbalanced dataset which would not yield appropriate accuracy results , since, accuracy is equivalent to the proportion of majority class data in the test set. Due to this reason, SMOTE analysis was used.
3) *Large Dataset:* The dataset obtained is relatively large in terms of number of features as well as the amount of data, so the training process is quite slow, especially when more sophisticated machine learning models were built and applied.

V. METHODS

A. Feature Selection Techniques

1) *Manual Feature Selection Through Correlation:* 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 .
2) *Mutual Information Feature Selection:* Mutual information from the field of information theory is the application of information gain to feature selection. It is calculated between two variables and measures the reduction in uncertainty for one variable given a known value of the other variable. This feature selection was implemented between our dependent variables and the class.

B. Machine Learning Models :

1) *Logistic Regression:* Logistic regression is often a great baseline model to try on machine learning problems. It is a supervised learning algorithm that is appropriate to conduct when the dependent variable is binary. The procedure is quite similar to linear regression, but its response variable is binomial. The result is the impact of each variable on the odds ratio of the observed event of interest. Since it does not enforce strong assumptions on the distribution of the underlying data, it is a great model to start with.
2) *Random Forest:* Random Forest is a supervised learning algorithm. It is like an ensemble of decision trees with a bagging method. The general idea of the bagging method is that a combination of learning models improves the overall result. The Random Forest algorithm randomly selects observations and features to build several decision trees and then averages the results. It also usually performs well on an imbalanced dataset. Another benefit of running this model is that it prevents overfitting problems for most of the time, as it creates several random subsets of the features and only constructs smaller subtrees.
3) *Naive Bayes:* Naive Bayes uses the "naive" assumption on the features, which means that the features are conditionally independent with each other given the class variable. It is a generative method, which is different from the previous two algorithms. The performance between the basic discriminative methods and this generative method can be compared.

VI. EXPERIMENTS AND RESULTS

A. Frequency of the 2 possible 'TARGET' Values

The frequency count of each of the possible outcomes in the data set, helps understand the balancing nature of the data. However, in this case , the results showed that the data set was highly imbalanced.

```
df['TARGET'].value_counts()
```

Value	Frequency
0 (repayers)	282686
1 (defaulters)	24825

Fig 1. Frequency of 0 and 1 in 'TARGET' field

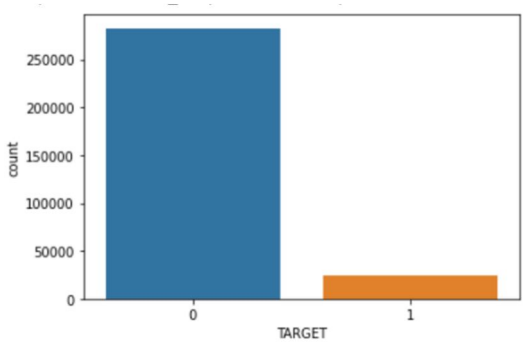


Fig 2 . Graph demonstrating frequency of 0 and 1

B. Logistic Regression (After Correlation)

The classification results generated by the model are as follows -

```
[ ]  print(classification_report(y_test,y_pred1))
```

	precision	recall	f1-score	support
0	0.96	0.69	0.80	56554
1	0.15	0.65	0.25	4949
accuracy			0.68	61503
macro avg	0.56	0.67	0.52	61503
weighted avg	0.89	0.68	0.76	61503

Fig 3 . Logistic Regression Model Results

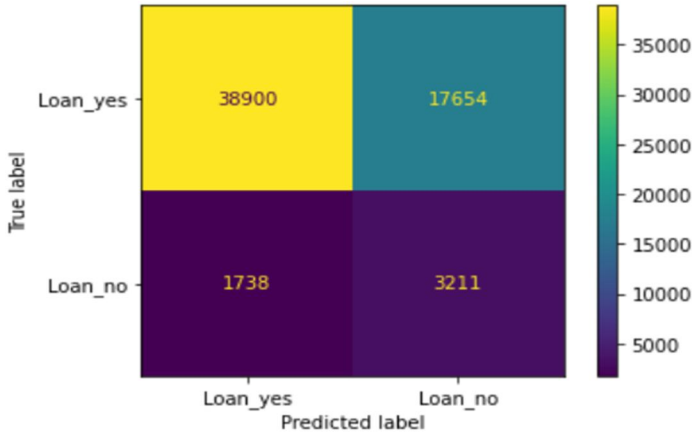


Fig 4 . Logistic Regression Confusion Matrix

C. Random Forest (After Correlation)

The classification results generated by the model are as follows –

```
print(classification_report(y_test, y_pred2))
```

	precision	recall	f1-score	support
0	0.93	0.95	0.94	56554
1	0.20	0.14	0.17	4949
accuracy			0.89	61503
macro avg	0.56	0.55	0.55	61503
weighted avg	0.87	0.89	0.88	61503

Fig 5 . Random Forest Model Results

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 11 Issue III Mar 2023- Available at www.ijraset.com

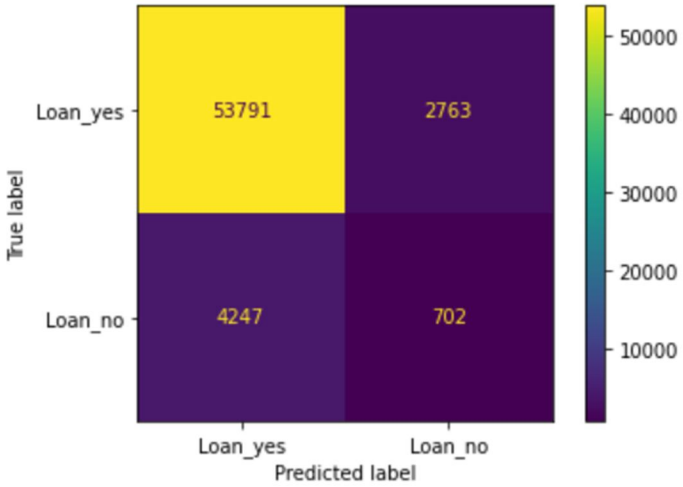


Fig 6 . Random Forest Confusion Matrix

D. Naive Bayes (After Correlation)

The classification results generated by the model are as follows -

```
print(classification_report(y_test,y_pred))
```

	precision	recall	f1-score	support
0	0.94	0.86	0.90	56554
1	0.17	0.32	0.22	4949
accuracy			0.82	61503
macro avg	0.55	0.59	0.56	61503
weighted avg	0.87	0.82	0.84	61503

Fig 7 . Naive Bayes Model Results

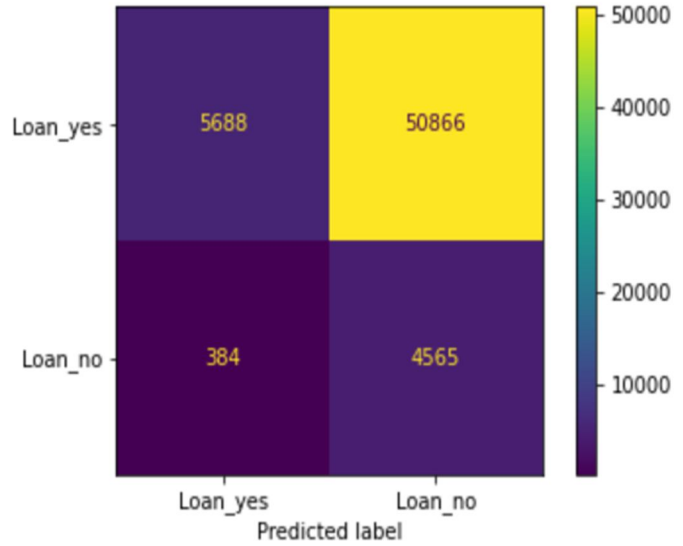


Fig 8 . Naive Bayes Confusion Matrix

E. Results Generated (After Correlation)

After applying different machine learning models on the dataset selected by Correlation, the results obtained clearly depict that Random Forest obtains the best accuracy out of the implemented models, with a 88.6% accuracy. It is closely followed by Naive Bayes with a 82.06% accuracy. However, Logistic Regression didn't seem as accurate as the others.

Machine Learning Model	Accuracy	Precision	Recall	F1 score
Logistic Regression	68.469%	0.96/0.15	0.69/0.65	0.80/0.25
Random Forest	88.602%	0.93/0.20	0.95/0.14	0.94/0.17
Naive Bayes	82.06%	0.94/0.17	0.86/0.32	0.90/0.22

Table 1 . Prediction Results Of Machine Learning Models

F. Logistic Regression (After Mutual Information)

The classification results generated by the model are as follows -

	precision	recall	f1-score	support
0	0.94	0.57	0.71	56330
1	0.11	0.62	0.19	4988
accuracy			0.57	61318
macro avg	0.53	0.59	0.45	61318
weighted avg	0.88	0.57	0.67	61318

Fig 9 . Logistic Regression Model Results

Fig 10 . Logistic Regression Confusion Matrix

G. Random Forest (After Mutual Information)

The classification results generated by the model are as follows -

	precision	recall	f1-score	support
0	0.92	0.98	0.95	56330
1	0.12	0.03	0.05	4988
accuracy			0.90	61318
macro avg	0.52	0.51	0.50	61318
weighted avg	0.85	0.90	0.88	61318

Fig 11 . Random Forest Model Results

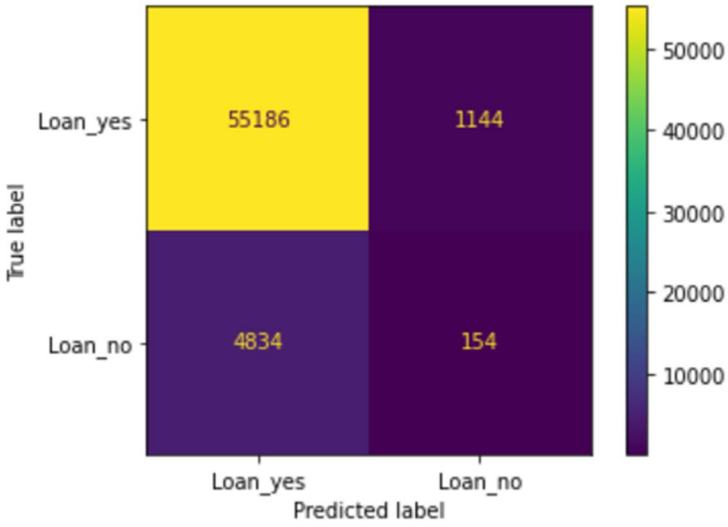


Fig 12 . Random Forest Confusion Matrix

H. Naive Bayes (After Mutual Information)

The classification results generated by the model are as follows -

```
print(classification_report(y_test, y_pred3))
```

	precision	recall	f1-score	support
0	0.94	0.37	0.53	56330
1	0.09	0.74	0.17	4988
accuracy			0.40	61318
macro avg	0.52	0.56	0.35	61318
weighted avg	0.87	0.40	0.50	61318

Fig 13 . Naive Bayes Model Results

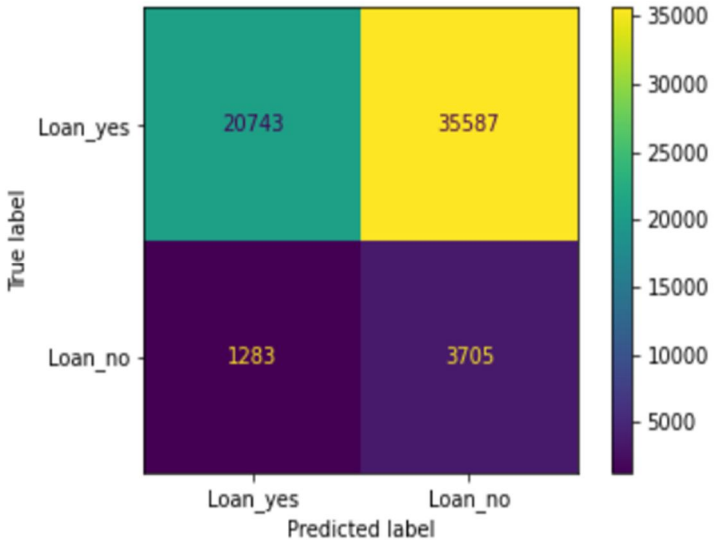


Fig 14 . Naive Bayes Confusion Matrix

I. Results Generated (After Mutual Information)

After applying different machine learning models on the dataset selected by Mutual Information, the results obtained clearly depict that Random Forest obtains the best accuracy out of the implemented models, with a 90.25% accuracy. Naive Bayes and Logistic Regression didn't generate the expected results and had an accuracy of 39% and 57% respectively.

Machine Learning Model	Accuracy	Precision	Recall	F1 score
Logistic Regression	57.14%	0.94/0.11	0.57/0.62	0.71/0.19
Random Forest	90.25%	0.92/0.12	0.98/0.03	0.95/0.05
Naive Bayes	39.68%	0.94/0.09	0.37/0.74	0.53/0.17

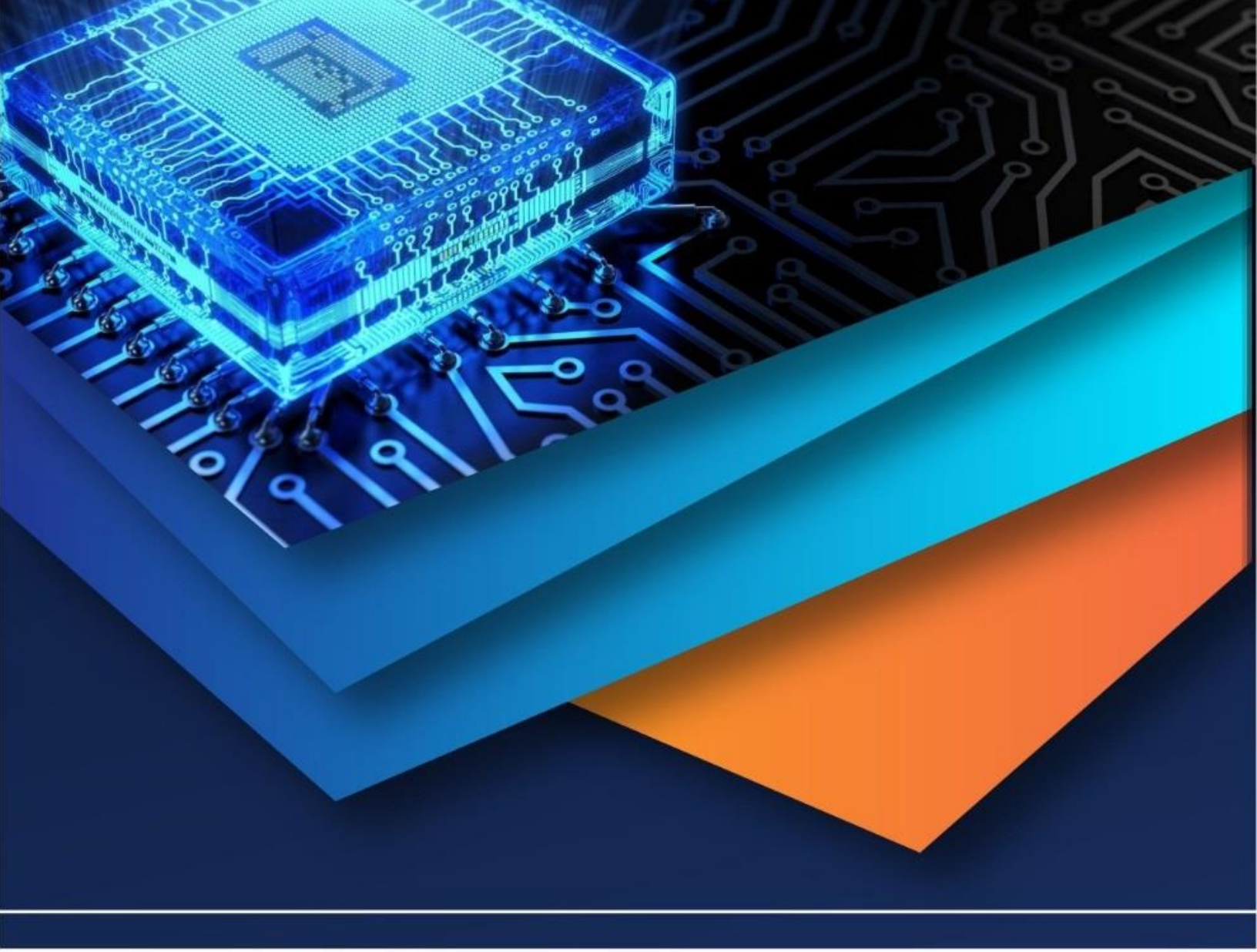
Table 2 . Prediction Results Of Machine Learning Models

VII. CONCLUSIONS AND FUTURE WORK

This study, has revealed a curated set of features such as Employment Days, Income, Number of Members, Number of Children are some of the most significant features to be considered while predicting someone's ability to repay the loan. Our research reveals that upon using Correlation, and different models, Random Forest has yielded the best results , followed by Naïve Bayes and Logistic Regression. Also, while using Mutual Information Classification as feature selection, Random Forest has yielded the best results with 90.25% accuracy. On the other hand , Logistic Regression and Naïve Bayes were not as effective with 57% and 39.7% respectively. However, since the dataset was imbalanced, such high accuracies were expected. Hence, in our future work, our intention would be collect a balanced dataset and apply different feature selection techniques on it.

REFERENCES

[1] Yiyun Liang, Xiaomeng Jin, Zihan Wang : "Loanliness: Predicting Loan Repayment Ability by Using Machine Learning Methods" (2019)
[2] Ritika Purswani, Sakshi Verma, Yash Jaiswal, Prof. Surekha M : "Loan Approval Prediction using Machine Learning" (June 2021)
[3] Jie Cai, Jiawei Luo, Shulin Wang, Sheng Yang : "Feature selection in machine learning : a new perspective" (2018)
[4] Kaggle Dataset : Link
[5] Aboobyda Jafar Hamid and Tarig Mohammed Ahmed : "Developing Prediction Model of Loan Risk in Banks using Data Mining"
[6] Pidikiti Supriya, Myneedi Pavani, Nagarapu Saisushma- "Loan Prediction by using Machine Learning Models" (April 2019)
[7] X. Francis Jency, V.P.Sumathi, Janani Shiva Sri - "An Exploratory Data Analysis for Loan Prediction Based on Nature of the Clients"
[8] Sudhamathy G.-"Credit Risk Analysis and Prediction Modelling of Bank Loans Using R", (IJET), Oct-Nov 2016
[9] Aphale & Shide . "Predict Loan Approval in Banking System Machine Learning Approach For Cooperative Banks Loan Approval" . International Research Journal of Engineering and technology, (2020)
[10] Chandra & Rekha: "Exploring the Machine Algorithm for Prediction the Loan Sanctioning" (2019)
[11] Khan et al.: "Loan Approval Prediction Model: A Comparative Analysis. Advances and Applications" (2021)
[12] Nikhil Madane, Siddharth Nanda: "Loan Prediction using Decision tree" ,Journal of the Gujrat Research History - December 2019
[13] Shrishti Srivastava, Ayush Garg, Arpit Sehgal, Ashok kumar – "Analysis and comparison of Loan Sanction Prediction Model using Python" International journal of computer science engineering and information technology research(IJCSEITR), (June 2018)
[14] Anchal Goyal, Ranpreet Kaur- "A survey on ensemble model of Loan Prediction" , International journal of engineering trends and application(IJETA), (Feb 2016)
[15] Li Y (2019) - Credit risk prediction based on machine learning methods The 14th Int. Conf. on Computer Science & Education (ICCSE) pp 1011–3
[16] Ahmed M S I and Rajaleximi P R (2019) - An empirical study on credit scoring and credit scorecard for financial institutions Int. Journal of Advanced Research in Computer Engineering & Technol. (IJARCET)
[17] Shoumo S Z H, Dhruba M I M, Hossain S, Ghani N H, Arif H and Islam S (2019) "Application of machine learning in credit risk assessment: a prelude to smart banking" TENCON 2019 – 2019 IEEE Region 10 Conf.
[18] Alshouiliy K, Alghamdi A and Agrawal D P 2020 AzureML based analysis and prediction loan borrowers creditworthy The 3rd Int. Conf. on Information and Computer Technologies (ICICT)
[19] Li M, Mickel A and Taylor S 2018, "Should this loan be approved or denied?": a large dataset with class assignment guidelines Journal of Statistics Education
[20] Vaidya A 2017 Predictive and probabilistic approach using logistic regression: application to prediction of loan approval The 8th Int. Conf. on Computing, Communication and Networking Technologies (ICCCNT)
[21] S. Vimala, K.C. Sharmili, —Prediction of Loan Risk using NB and Support Vector Machinel, International Conference on Advancements in Computing Technologies (ICACT 2018),
[22] A. Goyal and R. Kaur, "Accuracy Prediction for Loan Risk Using Machine Learning Models"



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY