Available online at http://icsejournal.com/index.php/JCSE
Journal of Computer Science an Engineering (JCSE)
e-ISSN 2721-0251
Vol. 3, No. 1, February 2022, pp. 01-13

Prediction of Loan Behaviour with Machine Learning Models for Secure Banking

Mayank Anand[1*], Arun Velu[2], Pawan Whig[3]

[1,3]Research Scientist, The Research World, New Delhi, India
[2]Director, Equifax, Atlanta, USA
[1]mayankan@gmail.com*; [2]gctarun@gmail.com ; [3]pawanwhig@gmail.com
*corresponding author

ARTICLE INFO

Article History:
Received August 23, 2021
Revised September 02, 2021
Accepted November 03, 2021

Keywords:
Loan Default
Data Science
Data Mining
Machine Learning

Correspondence:
Telephon: +91 9560291169
E-mail: *mayankan@gmail.com*

ABSTRACT

Given loan default prediction has such a large impact on earnings, it is one of the most influential factor on credit score that banks and other financial organisations face. There have been several traditional methods for mining information about a loan application and some new machine learning methods of which, most of these methods appear to be failing, as the number of defaults in loans has increased. For loan default prediction, a variety of techniques such as Multiple Logistic Regression, Decision Tree, Random Forests, Gaussian Naive Bayes, Support Vector Machines, and other ensemble methods are presented in this research work. The prediction is based on loan data from multiple internet sources such as Kaggle, as well as data sets from the applicant's loan application. Significant evaluation measures including Confusion Matrix, Accuracy, Recall, Precision, F1-Score, ROC analysis area and Feature Importance has been calculated and shown in the results section. It is found that Extra Trees Classifier and Random Forest has highest Accuracy of using predictive modelling, this research concludes effectual results for loan credit disapproval on vulnerable consumers from a large number of loan applications.

1. Introduction

The banking industry's credit lending sector has seen tremendous growth and fierce competition from a slew of new credit start-ups. At the same time, the rise in loan applications and consumption has resulted in an increase in bad credit losses. Credit loans are loans issued by banks or financial institutions to people or customers that are repayable at a certain period with or without interest. Loans Credits are often used for a variety of purposes, i.e. consumer, educational, medical, travel and business purposes. Banks and other financial institutions must use this research to blueprint effectual models that use info about existing facts and produce strong prediction prototypes which assist minimize likelihood of depraved recognition as a result of the increase in loan applications and the rapidly increasing competition. Banks and other financial institutions can gain insights on applicants' habits, money usage patterns, default predictors, and other characteristics using a variety of modern predictive modeling techniques. Numerous studies and research have been performed in order to determine the key variable that influence timely loan repayment; these studies are crucial for both banks and other financial institutions since they facilitate banks optimize profit. According to Manjeet [1], present remain 7 sorts of characteristics that may influence consumer loan default: whether or not the customer has a savings/checking account, occupation, work duration, home ownership, the customer's annual income and debt-income ratio. Chang [2] suggested an candidate's personal attributes, like oldness and boldness, may inspiration debtors' risk behavior. Borrower's age, location, resident/work duration, phone owner, monthly income, loan term, whether or not applicant works in the public sector, house ownership, and loan numbers are among the important characteristics that may impact loan default, according to Steenackers [3]. Bangherpour [4] found that credit age is one of the significant issues in forecasting credit defaulting when using a large dataset from 2001 to 2006,

http://dx.doi.org/10.36596/jcse.v3i1.237

though shop advance to rate was the most effective factor for home loan applications.

Device learning models that may assist capture significant patterns in loan credit data as well as identify indicators that may affect loan default must be developed and tested to ensure their effectiveness. An important factor in determining a prediction system's efficiency and accuracy isthe choice of model. Diverse models have been used to forecast loan defaults, and while there is no single optimal model, certain models clearly outperform others. There are many classification machine learning techniques available today, and in this work we utilize them all to evaluate a bank loan dataset and recommend some important characteristics and variables that may impact loan repayment. There is also information on our model's assessment statistics (Confusion Matrix; Accuracy; Recall; Precision; F1 Score; ROC Area; and Feature Importance).

Credit rating has become an important instrument in today's competitive financial world. To add to the excitement, this issue has garnered greater attention from academics because to recent breakthroughs in data science and some important findings in the field of artificial intelligence (AI). In recent years, research on loan prediction and credit risk assessment has gotten more attention. Due to the unexpectedly high demand for loans, there is a significant increase in the demand for extra credit scores and loan prediction models. Individual credit scores have been assigned using variety of methodologies, and extensive research has been conducted on the subject throughout the years. Unlike in the past, when experts were engaged and models depended on professional judgments to determine a person's creditworthiness, the focus currently is on an automated method of doing so. Scientists and bank officials have been promoting the use of machine learning algorithms and neural networks for credit score calculation and risk assessment for the past several years. Several notable achievements have been made in this subject, which might assist to pave the way for additional study and analysis.

To forecast loan default, Zhu [5] and Ghatasheh [6] employed the Random Forest Classification Algorithm. If you look at other techniques like logistic regression (73%), decision trees (95%), and support vector machines [7], random forest has a much better accuracy (98%). Seventy-five percent. In terms of credit risk prediction, the random forest technique is one of the most effective [8]. There was also a discussion of the algorithms' benefits, which included competitive classification accuracy and simplicity [9]. Explored a number of techniques, including logistic regression, k-nearest neighbors, random forest, neural networks, support vector machines, stochastic gradient boosting, Naive Bayes, and others.

Credit grading for mortgage loans was the subject of a research [11] by Nikhil Madane and Siddharth Nanda. Credit applications that do not satisfy specific requirements are often denied because of the high risk of default. Candidates with low incomes are more likely to get accepted and repay their loans on time. A machine learning approach called Decision Trees was utilized by Pidikiti Supriya et al. [12] in order to construct their model. Cleansing and preprocessing the data was the first step, followed by missing value imputation, exploratory data analysis and finally, the construction and assessment of the model. For example, 78.08 percent precision and 96.4 percent recall were achieved by applying the C4.5 method in decision trees in article [13] when the data partition was 90:10 and 80:20, respectively. So, the 80:20 split was judged to be the most accurate and recallable.

In their work [14], the authors used four models: Multiple Logistic Regression Model M1; Random Forest Model M2; Gradient Boosting Model M3; Multilayer Neural Network Models D1-D4; Multilayer Neural Network Model D1-D4 (deep learning) Data quality check is crucial, i.e. Analyzing and cleaning the data before modeling to eliminate duplicate variables, as shown by these models. Also, the research found that the choice of characteristics and the algorithm are two significant factors when choosing whether or not to provide a loan to a person.

To classify loan risk, Aboobyda Jafar Hamid and Tarig Mohammed Ahmed utilized Data Mining [15]. Three algorithms were used: J48, Bayes Net, and Nave Bayes. The article found that [15]

http://dx.doi.org/10.36596/jcse.v3i1.237

was the best method for the job because of its high accuracy (78.1784%) and low mean absolute error (0.3448). According to Aditi Kacheria et al., the Naive Bayesian method was used to create their model [16]. They also used the k-NN and binning algorithms to increase the data quality and classification accuracy of their data. K-NN was used to cope with missing values, and the binning method was used to remove abnormalities.

Logit method-based models are used by most local banks in the Czech Republic and Slovakia, according to a research by Martin Vojtek and Evzen Kocenda. Others, such as CART or neural networks, help in the selection of variables and evaluation of model quality. Authors have also concluded that either no one or very few people utilize the k-NN method. XGBoost algorithm's performance was compared to logistic regression's performance in Yu Li's work [2]. According to the article, the XGBoost model offers greater model discrimination and model stability than the logistic regression model. As for the rest of the paper, it is divided into three sections: If you want to know how the data was collected and how it was analyzed you may read the methodology in Section 2. Analyses and Results are discussed in Section 3 of the report, while Future Scope ends the report in Section 4.

2. Methodology

The bank's default payment data is represented in the dataset. The collection contains 850 records. A number of preprocessing techniques were used on the aforementioned dataset to produce a well-formed dataset, including cleaning, data integration, data formatting, data normalization, etc. Then, we used a variety of classification methods, including Multiple Logistic Regression, Decision Tree, K-Nearest Neighbors, SVM, Random Forest, and other forms of Ensemble Boosting approaches, to determine the accuracy of our predictive model's predictions. The results were impressive. As a dependent variable, this study used a dichotomous default payment (yes or no). Afterwards, we've compared our categorization findings to the destination's score. This research was implemented using Python on the Jupyter Kernelon a local machine. There are eight major explanatory factors in this study. These variables are described in further depth as follows: (i) X1 : Age (ii) X2 : Educational Background Category (iii) X3 : Employment Status(or Years of Experience) (iv) X4 : Address – Demographic Area converted to Numeric Equivalent (v) X5 : Income, (vi) X6 : Debt Income (vii) X7 : Credit to Debt Rasio (viii) X8 : Other Debt.

3. Results and Discussion

After the pre-processing on data sets, we use the 15 Classification algorithms which included various Classifiers, Boosting, Logistic Regression, SVM and Naive Bayes implemented in python programming language. Out of which, we will show the results of top 5 algorithms which showed optimistic results in the scenario. We trained our classifier on the pre-processed data set using the feature *default* as target. For evaluation, we use seven metrics; Confusion Matrix, Accuracy, F1-Score, Recall, Precision, ROC area and Feature Importance. We have used Confusion matrix and these evaluation metrics as they are the correct metrics used for classification algorithms, i.e. the predicted variable’s value is binary. Evaluation metrics like Accuracy, Precision, Recall, etc. definitely increase the chances of finding and removing the error algorithms are facing. This can further improve the same evaluation metrics performed on the results and lessen the scope of improvement gradually, giving better results.

4.1 CM

As its name suggests, the confusion matrix (CM) is a two-dimensional rectangular array, in which one dimension or column contains predicted values, while the other dimension or row reflects actual values of the classifiers. As a result of the bank loan default study utilized in this article, default goal values are binary, i.e. 0 indicates not having a default, and 1 represents having one.

http://dx.doi.org/10.36596/jcse.v3i1.237

This is a table of 1-5, displays the computed confusion matrix of the top five algorithms using the data in the table below.

Table 1. Confusion Matrix for Extra Trees Classifier

		Predicted Class	
		0 – Not Default	1 - Default
Actual Class	0 – Not Default	81	21
	1 - Default	5	98

Table 1 shows four numeric values which represent True Positive, True Negative, False Negative, False Positive. The values for True Positive and True False Positive are 81 and 98 which totals to 179 true/correct predictions and False Positive and False Negative are 21 and 5 which totals to 26 false/incorrect predictions for the confusion matrix calculated for Extra Trees Classifier.

Table 2. Confusion Matrix for Random Forest Classifier

		Predicted Class	
		0 – Not Default	1 - Default
Actual Class	0 – NotDefault	81	21
	1 - Default	8	95

Table 2 shows four numeric values which represent True Positive, True Negative, False Negative, False Positive. The values for True Positive and True False Positive are 81 and 95 which totals to 176 true/correct predictions and False Positive and False Negative are 21 and 8 which totals to 29 false/incorrect predictions for the confusion matrix calculated for Random Forest Classifier.

Table 3. Confusion Matrix for CatBoost Classifier

		Predicted Class	
		0 – Not Default	1 - Default
Actual Class	0 – NotDefault	77	25
	1 - Default	9	94

Table 3 shows four numeric values which represent True Positive, True Negative, False Negative, False Positive. The values for True Positive and True False Positive are 77 and 25 which totals to 171 true/correct predictions and False Positive and False Negative are 25 and 9 which totals to 34 false/incorrect predictions for the confusion matrix calculated for CatBoost Classifier.

Table 4. Confusion Matrix for Extreme Gradient Boosting

		Predicted Class	
		0 – Not Default	1 - Default
Actual Class	0 – NotDefault	77	25
	1 - Default	9	94

Table 4 shows four numeric values which represent True Positive, True Negative, False Negative, False Positive. The values for True Positive and True False Positive are 77 and 25 which totals to 172 true/correct predictions and False Positive and False Negative are 25 and 9 which totals to 33 false/incorrect predictions for the confusion matrix calculated for Extreme Gradient Boosting.

Table 5. Confusion Matrix for Light Gradient Boosting Machine Classifier

		Predicted Class	
		0 – Not Default	1 - Default
Actual Class	0 – NotDefault	79	23
	1 - Default	10	93

Table 5 shows four numeric values which represent True Positive, True Negative, False Negative, False Positive. The values for True Positive and True False Positive are 79 and 93 which totals to 167 true/correct predictions and False Positive and False Negative are 23 and 10 which totals to 38

http://dx.doi.org/10.36596/jcse.v3i1.237

false/incorrect predictions for the confusion matrix calculated for Gradient Boosting Machine Classifier.

4.2. Accuracy

Accuracy measures the percentage of predicted categorized values that are right. a formula is used to define it

$A = (TP + TN)/(TP + TN + FP + FN)$

Where, TP True Positive TN True Negative, FP False Negative FP False Positive

Table 6. Accuracy of top 5 algorithms in descending order

Algorithm	Accuracy
Extra Trees Classifier	86.17
Random Forest Classifier	85.55
CatBoost Classifier	84.92
Light Gradient Boosting	84.49
Extreme Gradient Boosting	83.87

Table 6 shows five numeric values which represent Accuracy for top five models with the formulae Accuracy = (TP+TN)/(TP+TN+FP+FN) (As mentioned above). These are the Accuracy of top five models from highest to lowest value.

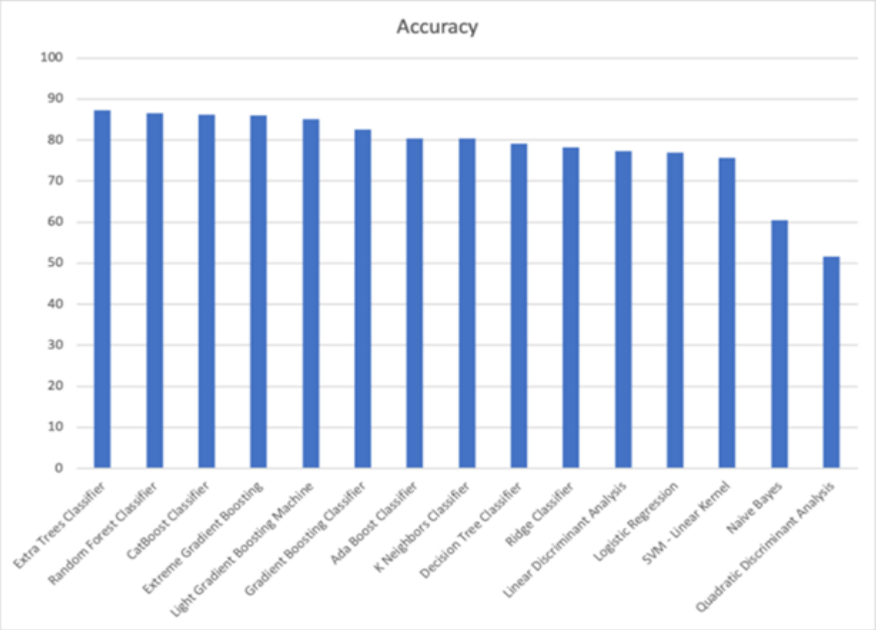

Fig 1. Comparison of Accuracy of all the classification models performed on the data

Fig 1 shows the Accuracy of all the algorithms executed in this research with leftmost being the highest value to rightmost being the lowest value.

4.3 Recall

True Positive Rate, also known as Recall, is the percentage of positive values that were properly predicted out of all positive values.

R = TP/(TP +FN)

http://dx.doi.org/10.36596/jcse.v3i1.237

Table 7. Recall of top 5 algorithms in descending order

Algorithm	Recall
Extra Trees Classifier	88.20
CatBoost Classifier	87.35
Random Forest Classifier	85.59
Light Gradient Boosting	84.68
Extreme Gradient Boosting	82.91

Table 7 shows five numeric values which represent Recall for top five models with the formulae Recall = TP/(TP+FN) (As mentioned above). These are the Recall of top five models from highest to lowest value.

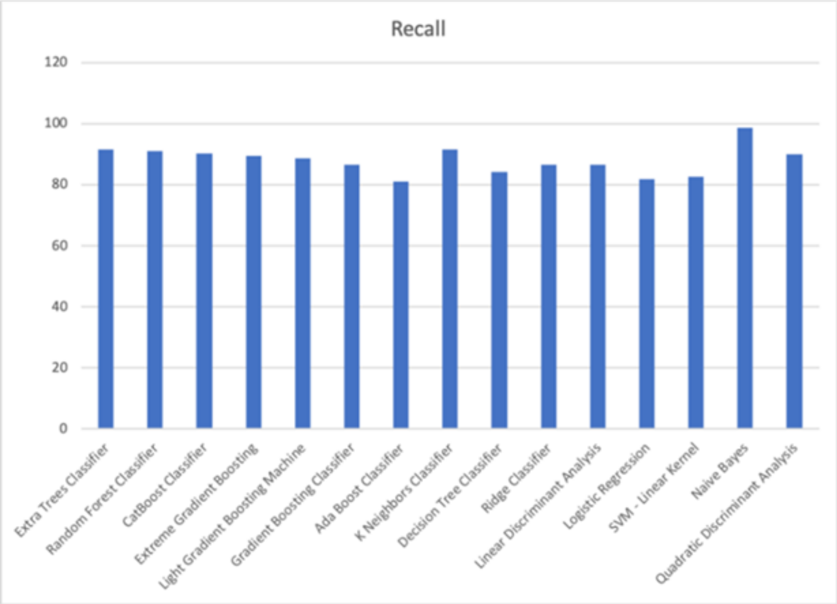

Fig 2. Comparison of Recall of all the classification models performed on the data

Fig 2 shows the Recall of all the algorithms executed in this research with leftmost being the highest value to rightmost being the lowest value.

4.4 Precision

Out of all positive projected classes, the precision represents the proportion of properly predicted positive classes. values.

P = TP/(TP +FP)

Table 8. Precision of top 5 algorithms in descending order

Algorithm	Precision
Random Forest Classifier	85.06
Extra Trees Classifier	84.27
Extreme Gradient Boosting	83.93
Light Gradient Boosting	83.46
CatBoost Classifier	82.57

Table 8 shows five numeric values which represent Precision for top five models with the formulae Precision = TP/(TP+FP) (As mentioned above). These are the Precision of top five models from highest to lowest value.

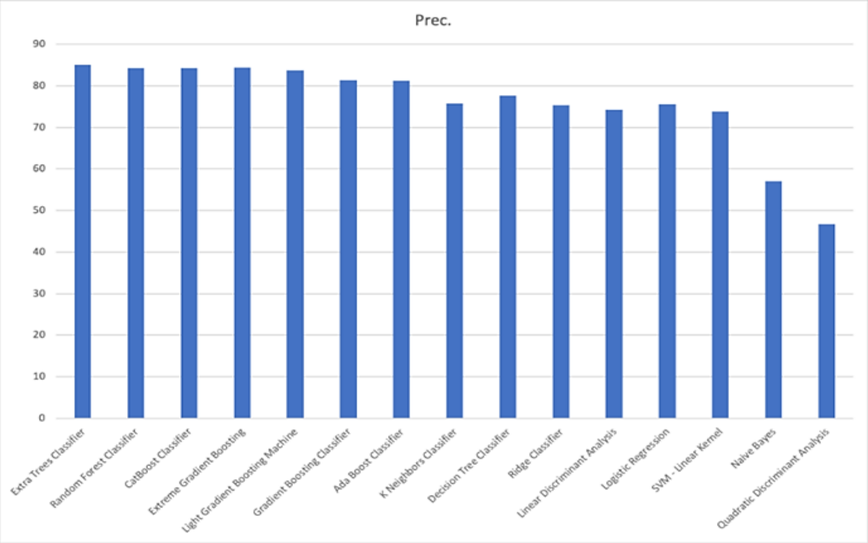

Fig 3. Comparison of Precision of all the classification models performed on the data

Fig 3 shows the Precision of all the algorithms executed in this research with leftmost being the highest value to rightmost being the lowest value.

4.5 F1-Score

Precision and recall are combined to produce an F1-score, which is the harmonic average of precision and recall:

$F1 - Score\ =\ 2 * P * R / \ (P + R)$

Table 9. F1-Score of top 5 algorithms in descending order

Algorithm	F1-Score
Extra Trees Classifier	85.97
Random Forest Classifier	85.03
CatBoost Classifier	84.57
Light Gradient Boosting	83.83
Extreme Gradient Boosting	83.07

Table 9 shows five numeric values which represent F1-Score for top five models with the formulae F1-Score = 2*P*R/(Precision+Recall) (As mentioned above). These are the F1-Score of top five models from highest to lowest value.

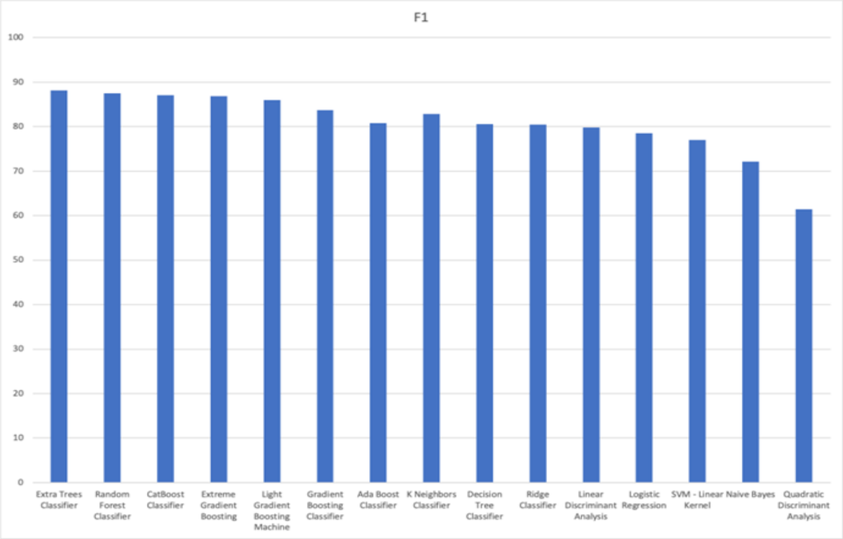

Fig 4. Comparison of F1-Score of all the classification models performed on the data

http://dx.doi.org/10.36596/jcse.v3i1.237

Fig 4 shows the F1-Score of all the algorithms executed in this research with leftmost being the highest value to rightmost being the lowest value.

4.6 ROC

As a visualisation method for classifier performance, the receiver operating characteristic (ROC) curve is used. A classifier's sensitivity and specificity are represented by this. ROC curves are two- dimensional because the FPR (false positive rate) is on the X-axis, and the TPR (true positive rate) is located on the Z-axis. There are two ranges for the range of the ROC curve: 0 and 0. (1, 1). There are just two possible outcomes: (1,1). It's possible to compare models by looking at how much area is beneath the curve (AUC). Models with greater AUCs are more accurate. As shown in Fig 5-9, the ROC-AUC curves for the top five algorithms are shown.

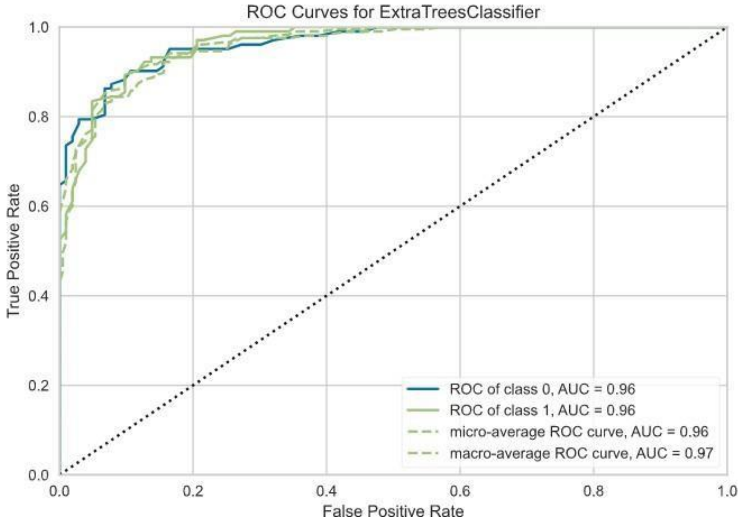

Fig 5. ROC Curve for Extra Trees Classifier

Fig 5 shows the Receiver Operating Curve for the Extra Trees Classifier predictions with a 45 degree line to show the Area under Curve which is the area between ROC and the 45 degree tangent line.

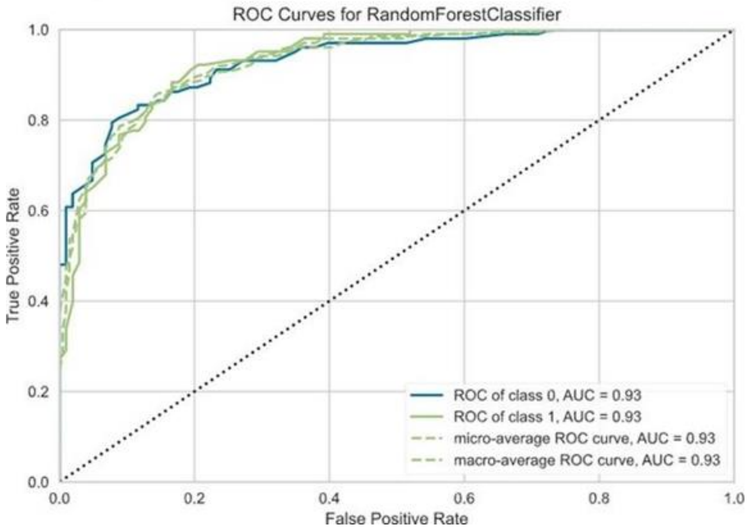

Fig 6. ROC Curve for Random Forest Classifier

Fig 6 shows the Receiver Operating Curve for the Random Forest Classifier predictions with a 45 degree line to show the Area under Curve which is the area between ROC and the 45 degree tangent line.

http://dx.doi.org/10.36596/jcse.v3i1.237

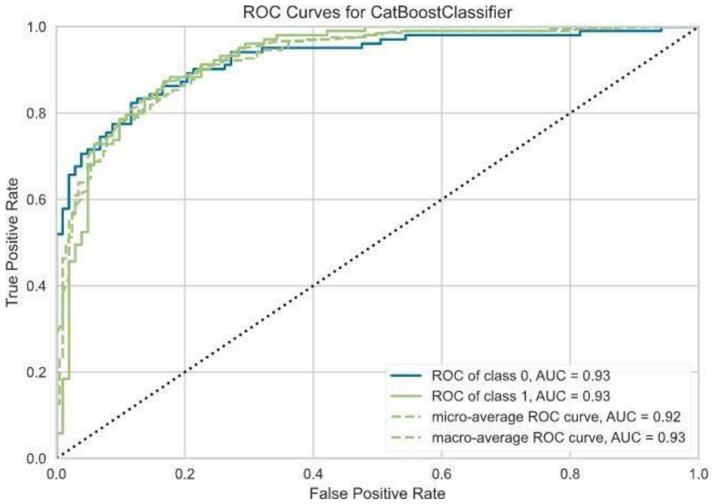

Fig 7. ROC Curve for Cat Boost Classifier

Fig 7 shows the Receiver Operating Curve for the Cat Boost Classifier predictions with a 45 degree line to show the Area under Curve which is the area between ROC and the 45 degree tangent line.

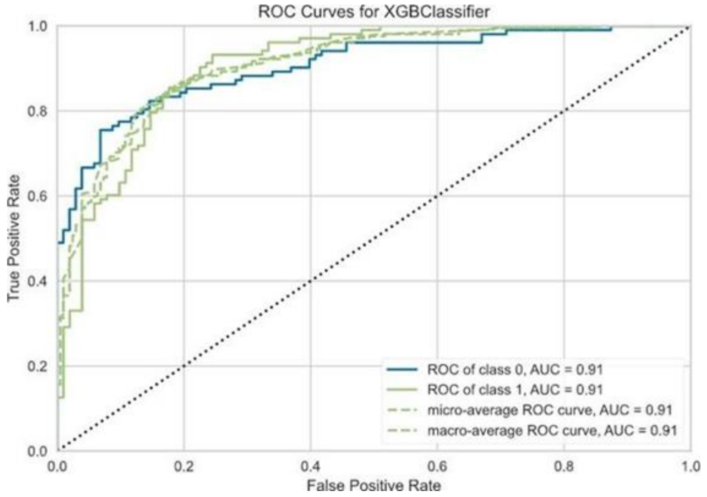

Fig 8. ROC Curve for Extreme Gradient Boosting Classifier

Fig 8 shows the Receiver Operating Curve for the Extreme Gradient Boosting (XGB) Classifier predictions with a 45 degree line to show the Area under Curve which is the area between ROC and the 45 degree tangent line.

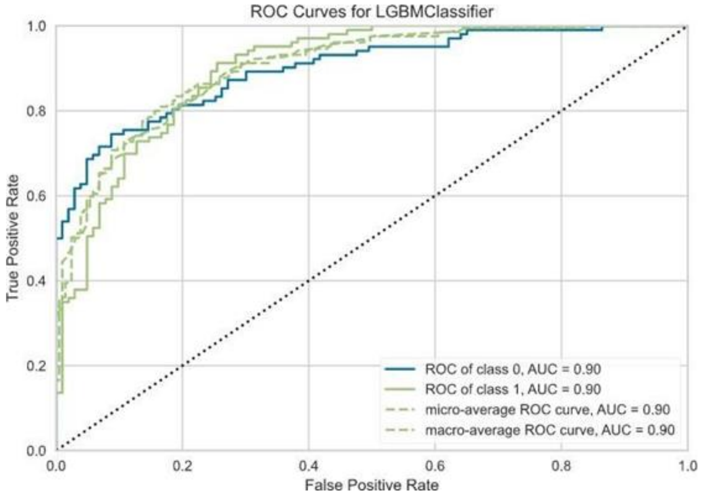

Fig 9. ROC Curve for Light Gradient Boosting Machine Classifier

http://dx.doi.org/10.36596/jcse.v3i1.237

Fig 9 shows the Receiver Operating Curve for the Light Gradient Boosting Machine(LGBM) Classifier predictions with a 45 degree line to show the Area under Curve which is the area between ROC and the 45 degree tangent line.

4.7 Feature Importance

As a result of this research, significant factors that help the classifier properly forecast loan default have been identified. For business intelligence and decision-making, this is a plus. As shown in Fig 10-14, the top five algorithms have the most significant characteristics plotted out. The plot's conclusion is heavily influenced by the customer's job history and loan income.

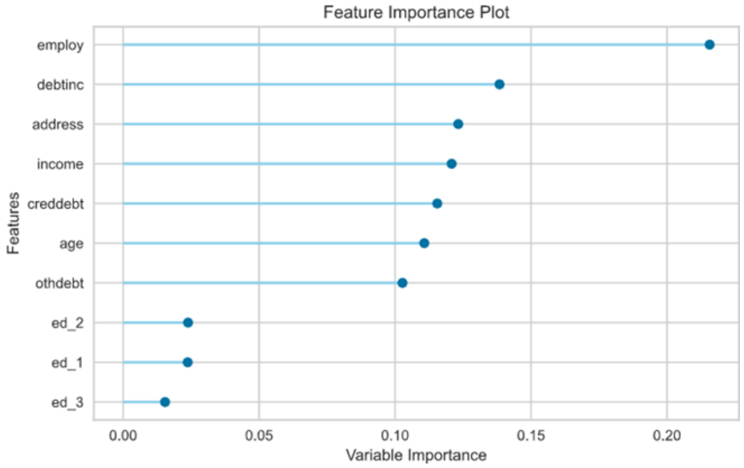

Fig 10. Feature Importance Plot for Extra Trees Classifier

Fig 10 shows the Feature Importance of each individual feature from the dataset used in this research with respect to Extra Trees Classifier predictions.

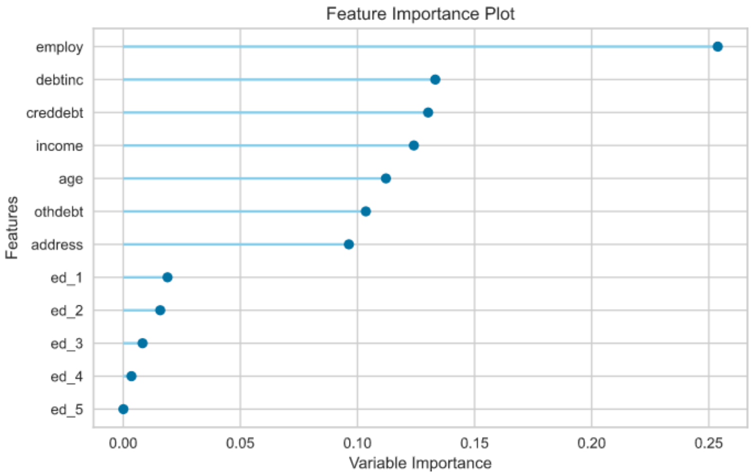

Fig 11. Feature Importance Plot for Random Forest Classifier

Fig 11 shows the Feature Importance of each individual feature from the dataset used in this research with respect to Random Forest Classifier predictions.

http://dx.doi.org/10.36596/jcse.v3i1.237

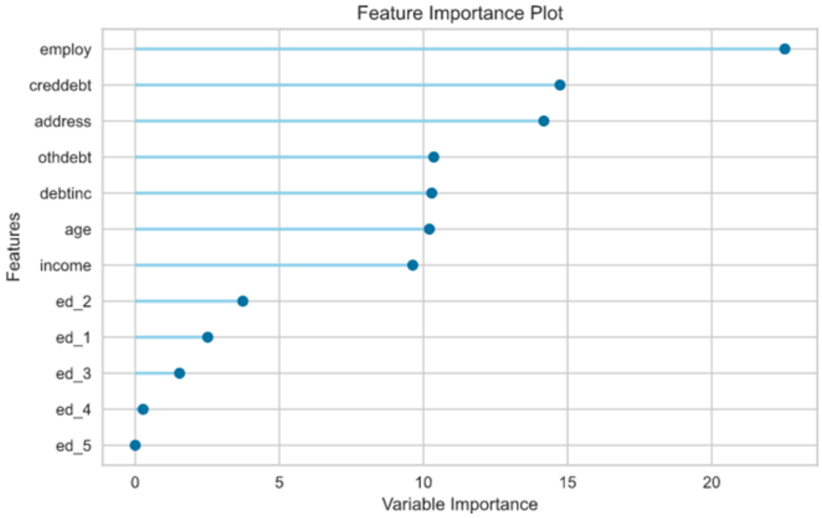

Fig 12. Feature Importance Plot for Cat Boost Classifier

Fig 12 shows the Feature Importance of each individual feature from the dataset used in this research with respect to Cat Boost Classifier predictions.

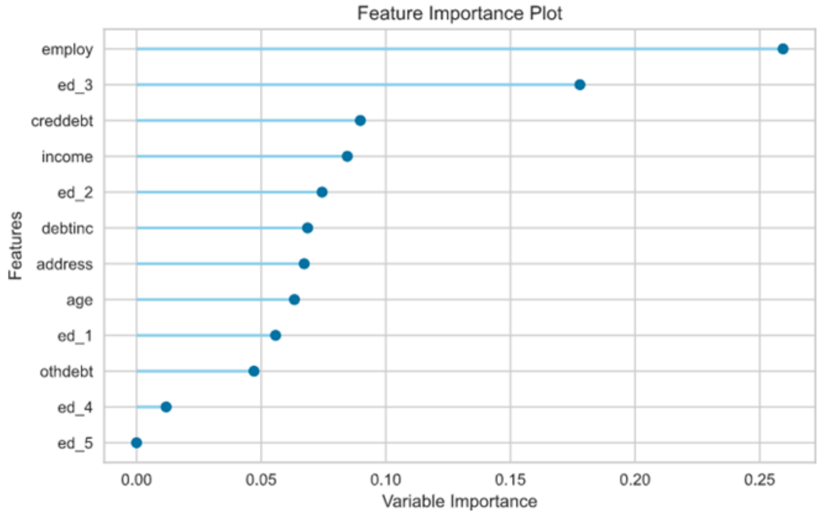

Fig 13. Feature Importance Plot for Extreme Gradient Boosting Classifier

Fig 13 shows the Feature Importance of each individual feature from the dataset used in this research with respect to Extreme Gradient Boosting Classifier predictions.

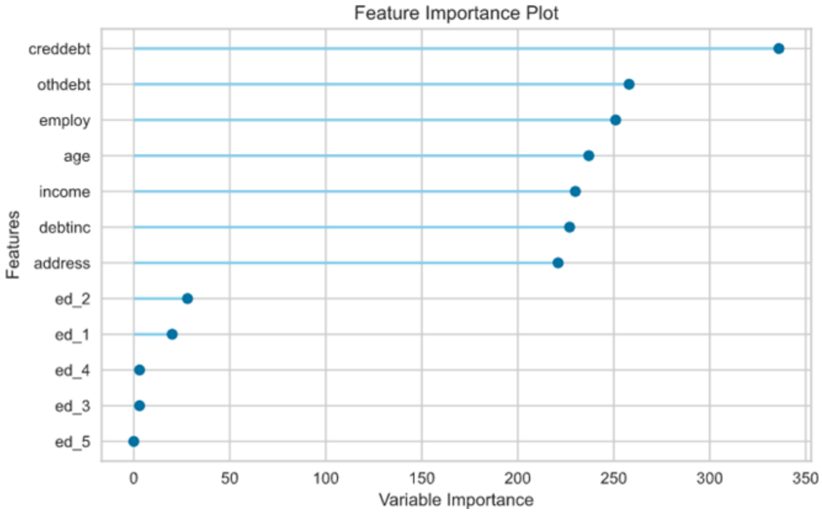

Fig 14. Feature Importance Plot for Light Gradient Boosting Machine Classifier

http://dx.doi.org/10.36596/jcse.v3i1.237

Fig 14 shows the Feature Importance of each individual feature from the dataset used in this research with respect to Light Gradient Boosting Machine Classifier predictions.

4. Conclusion

We successfully employed multiple classification algorithms for bank loan default prediction in thispaper. The goal was to forecast whether or not a loan applicant will default on their payments. The analysis was carried out in Python, and performance indicators like as accuracy, recall, precision, and f1-score were produced. According to the findings, the most crucial parameters employed by ourmodel for forecasting whether or not a customer will default in payment are the customer's employment or job experience in years and debt income. This article uses predictive modelling to detect problematic clients among a large number of loan applicants, resulting in a more effective basis for loan credit approval.

Acknowledgment

I am grateful to The Research World (ThRews) for contributing in my research over this paper.

References

[1] ManjeetKumar, Vishesh Goel, Tarun Jain, Sahil Singhal, DR. Lalit Mohan Goel, “Neural Network Approach To Loan Default Prediction”, International Research Journal of Engineering and Technology (IRJET) , p-ISSN: 2395-0072

[2] Chiang, R.C., Chow, YF. & Liu, M. Residential Mortgage Lending and Borrower Risk: The Relationship between Mortgage Spreads and Individual Characteristics. The Journal of Real Estate Finance and Economics 25, 5–32. 2002. https://doi.org/10.1023/A:1015347516812

[3] Steenackers, A., & Goovaerts, M. J. A credit scoring model for personal loans. Insurance: Mathematics and Economics. Volume 8, Issue 1, March 1989, Pages 31-34. https://doi.org/10.1016/0167-6687(89)90044-9

[4] A. Bagherpour, “Predicting Mortgage Loan Default with Machine Learning Methods” 2017. Computer Science.

[5] L. Zhu, D. Qiu, D. Ergu, C. Ying, and K. Liu, “A study on predicting loan default based on the random forest algorithm,” Procedia Computer Science, vol. 162, pp. 503–513, 2019, doi: 10.1016/j.procs.2019.12.017

[6] N. Ghatasheh, “Business Analytics using Random Forest Trees for Credit Risk Prediction: A Comparison Study”. International Journal of Advanced Science and Technology. Vol.72, pp.19-30. 2014.doi: http://dx.doi.org/10.14257/ijast.2014.72.02

[7] A. Uzair, T. Aziz, H. Ilyas, S. Asim, B. N. Kadhar, "An Empirical Study on Loan Default Prediction Models" Journal of Computational and Theoretical Nanoscience, Volume 16, Number 8, August 2019, pp. 3483-3488(6). DOI: https://doi.org/10.1166/jctn.2019.8312

[8] Li Y (2019), “Credit risk prediction based on machine learning methods”, ICCSE. pp 1011–3

[9] M. S. Irfan Ahmed, P. Ramila Rajalaximi, “An empirical study on credit scoring and credit scorecard for financialinstitutions”, International Journal of Advanced Research in Computer Engineering & Technology (IJARCET). Volume 8, Issue 7, July 2019, ISSN: 2278 – 1323

[10] J. L. Breeden, "Survey of Machine Learning in Credit Risk" (May 30, 2020). Available at http://dx.doi.org/10.2139/ssrn.3616342

[11] M. Madaan et al. "Loan default prediction using decision trees and random forest: A comparative study" IOP Conf. Ser.: Mater. Sci. Eng. 2014. doi: 10.1088/1757-899X/1022/1/012042

[12] Supriya P et al (2019), “Loan prediction by using machine learning models”, IJET. pp144–8

[13] Amin R K et al (2015), “Implementation of decision tree using C4.5 algorithm in decision making of loanapplication by debtor (case study: bank pasar of Yogyakarta special region)”, ICoICT. pp 75–80

http://dx.doi.org/10.36596/jcse.v3i1.237

[14] J. H. Aboobyda and M. A. Tarig, “Developing Prediction Model Of Loan Risk In Banks Using Data Mining Machine Learning and Applications,” An Int. J., vol. 3, no. 1, pp. 1–9, 2016.

[15] M. Madaan et al, “Loan default prediction using decision trees and random forest: A comparative study” IOP Conf. Ser.: Mater. Sci. Eng. 1022 012042. 2021. doi: 10.1088/1757- 899X/1022/1/012042