grp4

by sanjana sharma

Submission date: 19-Oct-2023 11:54AM (UTC+0530)
Submission ID: 2200513589
File name: Group4_LoanDefault_Prediction.docx (913.98K)
Word count: 2941
Character count: 17682

Predicting Loan Defaults: A Data-Driven Approach Using Machine Learning

Saahil Shaikh, Sanjana Sharma, Sanna Johnson, Sarina Khatri

Department of CSE, Symbiosis Institute Of Technology

SYMBIOSIS INTERNATIONAL (DEEMED UNIVERSITY)

(Established under section 3 of the UGC Act 1956)

saahil.shaikh.btech2021@sitpune.edu.in

sanjana.sharma.btech2021@sitpune.edu.in

sanna.johnson.btech2021@sitpune.edu.in

sarina.khatri.btech2021@sitpune.edu.in

***Abstract*— The constant issue for financial institutions is reducing payment failures while ensuring appropriate lending practices. Machine learning approaches have proven essential in identifying loan default risks, enabling prompt and targeted interventions. This research provides an opportunity to examine one of the finance industry's most urgent issues. We hope to create a predictive system that improves the precision of identifying persons at high risk of loan default through rigorous analysis and modeling. This study emphasizes the critical relevance of data-driven initiatives with machine learning in reducing loan defaults, increasing responsible lending, and protecting institutions' financial interests.**

Keywords—Loan Default Prediction, Machine Learning, Financial Risk Assessment, Responsible Lending, Predictive Modeling

I. INTRODUCTION

Financial institutions play an essential role in the global economy, providing individuals and businesses with access to capital. The difficult task of credit risk management lies at the heart of their operations, intending to strike a delicate balance between issuing credit to boost economic growth and reducing the financial repercussions from loan defaults. Accurate and prompt loan default prediction is critical to risk mitigation and responsible lending practices in this setting.

The loan industry has seen a radical change in recent years, as machine learning methods have become effective instruments for analyzing large datasets and deriving predictive insights. The current study initiates a thorough exploration of the domain of loan default forecasting.

The primary purpose of this project is to create a robust and data-driven system that enhances the precision of identifying individuals at the highest risk of loan default. We navigate this terrain with an empirical and analytical methodology anchored by the basics of machine learning. By digging into the intricacies of loan default prediction, this research adds to refining credit risk assessment, develops responsible lending practices, and bolsters the financial resilience of lending organizations.

Reducing loan defaults is necessary in an environment where credit availability is essential to personal goals and economic expansion. The approaches and learnings from our investigation into loan default prediction are presented in the ensuing sections, highlighting the critical function that data-driven tactics play in guaranteeing the economy's viability and prudent management of financial resources.

II. LITERATURE REVIEW

In recent years, the banking sector has encountered significant challenges in managing loan defaults, resulting in exploring machine learning algorithms as a potential solution. Several studies have examined the efficacy of various algorithms for predicting loan defaults and improving the overall credit risk management process.

One of the prominent algorithms explored in this context is the XGBoost algorithm, which has demonstrated superior predictive performance. In 2023, Shokeen, Grover, and Verma [1] conducted a study focusing on applying the XGBoost algorithm for loan default prediction. Their research compared other machine learning algorithms, like Logistic Regression, Decision Tree, K Nearest Neighbour, and Random Forest classification, with XGBoost. It was inferred that XGBoost outperformed all other algorithms, highlighting its potential to effectively predict loan defaults and mitigate the external credit crisis faced by the banking industry.

Machine learning has also been leveraged to develop comprehensive models for loan default prediction. In 2021, Kadam, Pawar, Phapale, and Ganeshe [2] aimed to explore, analyze, and build a machine-learning algorithm to accurately identify individuals with a high probability of defaulting on loans. Their research reflects a holistic approach to modeling and prediction.

The Random Forest algorithm has gained prominence in the context of loan default prediction. In 2019, Zhu, Qiu, Ergu, Ying, and Liu [4] adopted the Random Forest algorithm to build a model for predicting loan defaults, comparing it with additional techniques like decision trees, logistic regression, and support vector machines. Their experiment demonstrated the superior performance of the Random Forest algorithm in predicting loan defaults and its strong generalization capabilities.

Another algorithm used for analyzing loan approval prediction is Support Vector Machine (SVM). Diwate, Rana, and Chavan [3] used SVM to analyze features and characteristics influencing loan approval. The research indicated that applicants with poor credit scores were less

likely to obtain loan approval, which emphasized the importance of creditworthiness in the approval process.

In the Chinese P2P lending market, machine learning methodologies have been employed to predict loan defaults. Xu, Lu, and Xie [5] implemented machine learning methods, including Random Forest, XGBoost, K-NN, and Gradient Boosting Machine, to assess model performance for prediction accuracy and kappa value. Their study provides insights into the application of machine learning in a specific financial market.

Comparative studies have also been conducted to evaluate the performance of machine learning algorithms for loan default prediction. (Madan et al.)[6] conducted a comparative study using Decision Trees and Random Forests. The Random Forest Classifier is a better alternative for this type of data because it outperformed the Decision Tree approach, achieving an accuracy of 80%.

Additionally, researchers have explored the application of machine learning in creating credit risk scorecards. [7] implemented four different machine learning models, including Linear Discriminant Analysis, Random Forest, Logistic Regression, and XGBoost, to assess the accuracy of these models for credit scoring.

Deep learning models have also been considered for credit risk analysis. Addo, Guegan, and Hassani [8] emphasized the importance of feature selection and algorithm choice in the decision-management process when issuing loans. Their research highlighted the significance of considering a pool of models that match the data and business problems.

In a study conducted in 2023, Gashi [9] highlighted the impact of the SMOTE method and the effectiveness of ensemble algorithms, particularly Boosted Decision Tree (Boosting) and Random Forest (Bagging), in predicting loan defaults. This research underscored the value of data and its modeling through machine learning in credit risk management and loan default prediction.

In the realm of machine learning explainability in finance, Bracke, Datta, Jung, and Sen [10] emphasized the importance of explainable AI tools as additions to the data science toolkit. These tools improve the quality assurance of black box machine learning models, such as Random Forest, Gradient Boosting, and XGBoost.

In conclusion, machine learning algorithms have shown significant promise in addressing the challenge of loan defaults in the banking sector. XGBoost, Random Forest, and other techniques have demonstrated their effectiveness in improving the accuracy of loan default predictions and enhancing credit risk management practices. Additionally, the use of explainable AI tools contributes to better quality assurance in machine learning models, further improving the understanding of default risk analysis. These studies underscore the importance of data-driven approaches in mitigating credit crises and strengthening financial stability in the banking industry.

III. Methodology

A. ML Algorithms

a) *Logistic Regression:* Logistic regression's simplicity, interpretability, and suitability for binary classification issues led to its selection. By leveraging the logistic function, we establish a model that captures the connection between one or more independent variables (features) and a binary outcome variable (in this case, whether a loan defaults or not).

Logistic function formula:

$$p(x) = \frac{1}{1 + e^{-(x-\mu)/s}}$$

b) *Decision Tree:* A decision tree classifier was the primary ML model used in this project to predict the binary result of a loan default (1 for default, 0 for non-default). The decision tree technique is widely used for classification tasks due to its interpretability and capacity to represent intricate decision boundaries.

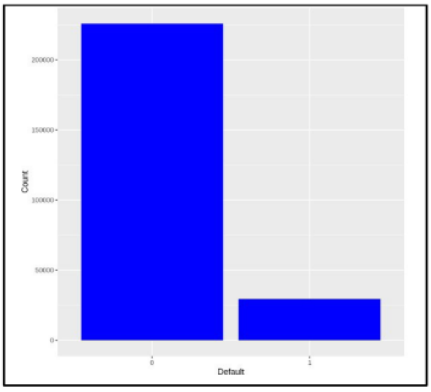
c) *KNN:* It is a simple and intuitive method based on the idea that data points with similar features tend to have the same class labels. The KNN algorithm categorizes data points according to their proximity in a feature space. Predicting loan defaults is one of the many classification problems for which this model is appropriate.

B. The Dataset

The dataset used for this project is obtained from Coursera's Loan Default Prediction Challenge. This dataset is valuable for real-world machine learning and predictive modeling tasks. With a sizable collection of 255,347 rows and 18 columns[10], it provides substantial data for analysis and modeling.

Essential information and attributes of this dataset: The dataset includes diverse attributes, each serving as a significant factor in evaluating the probability of a loan going into default.

Data Quality: All instances are complete without missing values, and no outliers are present in the dataset.



Class Distribution

Class Distribution: The dataset's target variable, "Default," has two classes: 0 and 1. The dataset contains 255347 observations, with 225694 representing 0 and 29653 representing 1. This class distribution indicates that the dataset is highly imbalanced, with a bias towards '0', and would require oversampling to make a classification model with viable results.

C. EDA

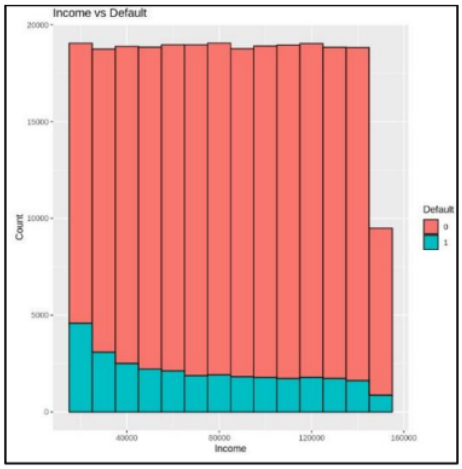
In the context of loan default prediction, the EDA provided valuable insights into the dataset and aided in feature selection and engineering, which were vital for model development.

Data Visualization:

We employ a range of data visualization techniques to gain deeper insights into the dataset. Each technique is chosen to highlight different aspects of the data, uncover patterns, and facilitate decision-making:

1. Histograms

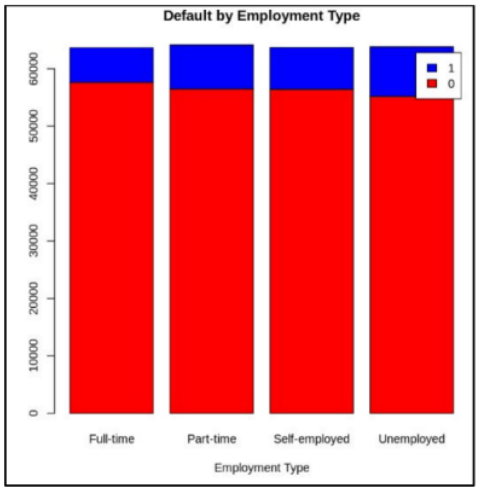
Histograms are employed to visualize the distribution of numerical variables, such as "Age," "Income," and "LoanAmount."



Notably, a lower applicant income corresponds to an increased likelihood of loan default.

2. Bar Plots

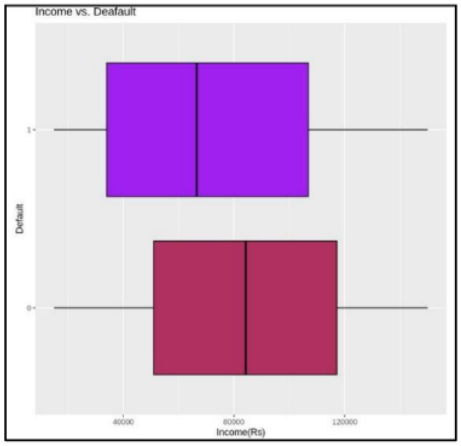
Bar plots serve as a valuable tool for visualizing categorical variables, including "LoanPurpose," "EmploymentType," and "MaritalStatus."



Notably, the graph illustrates that unemployed applicants exhibit the highest default rate, followed by part-time applicants, self-employed individuals, and, finally, full-time applicants.

3. Box Plots

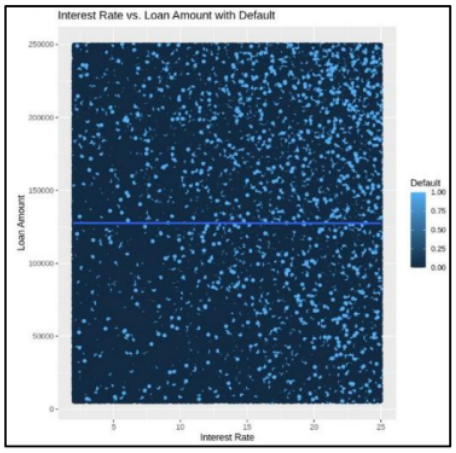
Box plots offer a graphical manner of the distribution of numerical variables, focusing on detecting potential outliers.



In the context of loan default prediction, box plots are utilized to scrutinize variables such as "CreditScore" and "NumCreditLines," providing insights into their distribution characteristics and identifying extreme values.

4. Scatter Plots

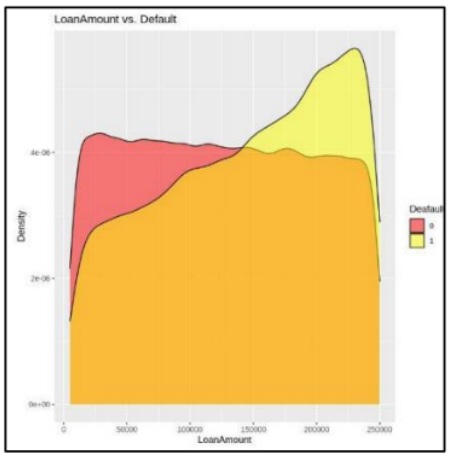
Scatter plots are instrumental in uncovering relationships between two numerical variables, facilitating the exploration of potential correlations.



The above graph plot shows Defaulters' distribution when comparing Loan Amount and Interest Rate. The points are scattered more towards the right, indicating that with the interest rate and loan amount increase, the defaulters increase, and there is no linear relationship between Loan Amount and Interest Rate.

5. Density Plots

Density plots offer a smooth representation of the distribution of numerical variables, providing a visualization of data density.



They are particularly informative for variables like "InterestRate" and "DTIRatio."

To sum up, this study's use of data visualization approaches offers a thorough overview of the dataset, which facilitates the discovery of essential insights pertaining to loan default prediction. These visualizations inform decisions regarding feature selection, model development, and the overall understanding of the dataset's underlying patterns.

D. Data preprocessing and classification

Data Preprocessing: To transform categorical data into numerical representations that machine learning algorithms can process efficiently, we used feature encoding in this study. Binary categorical features like "HasMortgage," "HasDependents," and "HasCoSigner" were transformed to binary form (1/0), while the multi-categorical features including "Education," "EmploymentType," "MaritalStatus," and "LoanPurpose" were encoded using one-hot encoding techniques. One-hot encoding made sure the models could properly read and utilize these features. In addition, we used feature extraction to choose pertinent features and, if needed, reduce dimensionality. Using the EDA and correlation coefficients of each feature, this approach involved determining the significance of each characteristic.

The dataset was severely skewed, with 225k occurrences representing '0' and 22k instances representing '1'. We used the ROSE (Random Over-Sampling Examples) method to remedy the class imbalance. ROSE is a resampling method created primarily to provide randomness to the process in order to prevent overfitting and increase the number of cases in the minority class ("Default = 1"). More accurate predictions were produced due to this balancing act, which ensured the machine learning models were not biased toward the majority class.

Model training and classification: We trained several machine learning models to predict loan defaults based on the preprocessed dataset. These models included k-nearest neighbors (K-NN), decision trees, and logistic regression. Every model was trained using the dataset, assessed for its ability to predict outcomes, and adjusted to maximize performance. We optimized the hyperparameters using manual cross-evaluation and iterative techniques to ensure that the models were precise and effective. The dataset was split into train and test sets so that we could assess our models' efficacy. An 80:20 train-test split ratio was followed, with 20% for testing and the remaining 80% for training.

Further, We performed hyperparameter optimization to maximize our machine learning models' performance. In order to do this, the model's behavior's hyperparameters had to be methodically changed, and their effects on model performance had to be evaluated. The goal was to find the ideal hyperparameter combination that produced the most accurate results.

E. Performance evaluation metrics

To assess the model's effectiveness thoroughly, we employed crucial metrics such as the confusion matrix, precision, recall, and F1 score.

Confusion matrix: The confusion matrix was the cornerstone of our model evaluation. The model's predictions are divided into four categories by the tabular format, which shows the performance of a classification algorithm: True Positives (TP), True Negatives (TN), False Positives (FP), also referred to as Type I error, and False Negatives (FN), often referred to as Type II mistake.

We could determine precision, recall, and other metrics by calculating the confusion matrix, which also gave us insights into the model's capacity to distinguish between positive and negative classifications.

Precision: Precision measures the accuracy of positive predictions made by the model.

It determines the ratio actual positive events to all anticipated positive ones. A high precision means fewer erroneous positive predictions made by the model.

$$\text{Precision} = TP / (TP + FP)$$

Recall (sensitivity): Recall, sometimes called sensitivity, evaluates how well the model can distinguish genuine positive instances from all actual positive instances.

It determines the ratio of actual positive events to all anticipated positive ones.

$$\text{Recall} = TP / (TP + FN)$$

F1 score: The F1 score is a combination metric that provides a harmonic mean of recall and precision. It helps evaluate how well the model can balance reducing false positives and false negatives.

It is calculated as:

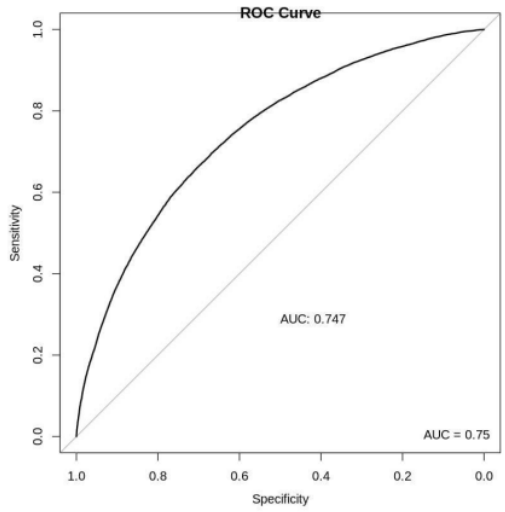
$$2 * (\text{Precision} * \text{Recall}) / (\text{Precision} + \text{Recall})$$

In our project, we used the F2 score to prioritize recall because correctly identifying loan defaults was more critical than minimizing false alarms. Using these metrics ensured that our models were well-rounded and could deliver robust and reliable results.

IV. Results & Discussion

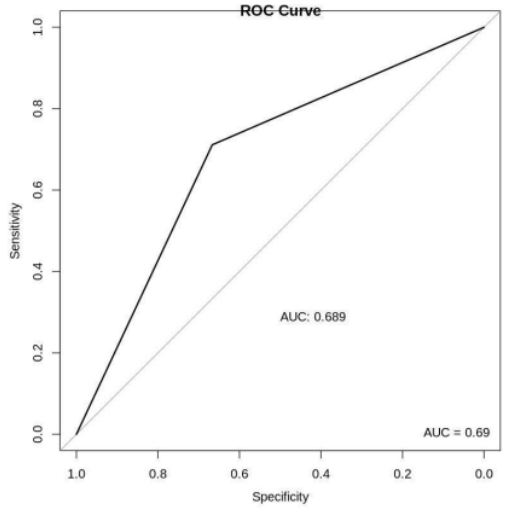
The outcomes of our predictive models for loan default prediction are shown in this section. Our goal was to create machine learning models that could reliably identify between borrowers in the positive class—those who would not default on their loans—and the negative class—those who would.

Three different machine learning models were utilized for this project: decision trees, logistic regression, and k-nearest Neighbors (KNN). We carefully trained and evaluated each one to find the model that worked best for our categorization task.



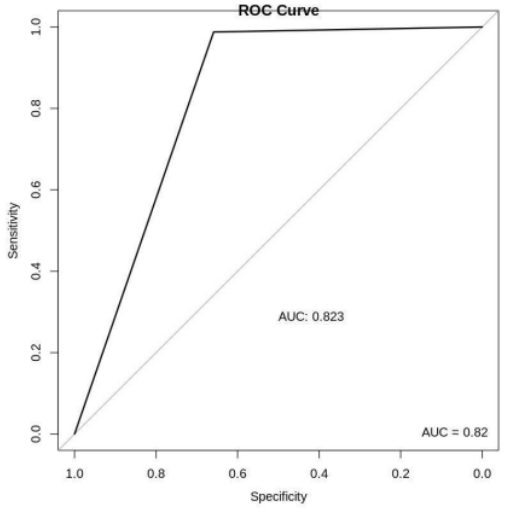
ROC curve of Logistic Regression

The Logistic Regression model was trained to predict loan defaults based on features and achieved an accuracy of approximately 49.05%.



ROC curve of Decision Tree

Gini impurity was used as the training criterion for the Decision Tree model, and parameters like maximum depth (maxdepth), minimum split (minsplit), and minimal bucket (minbucket) were used to optimize the model. The accuracy of this decision tree model was roughly 69.03%. The results were 0.6861, 0.7156, and 0.7081 for precision, recall, and F1 score, respectively.



ROC curve of KNN

The KNN model was trained with the Manhattan distance metric and k=3 neighbors. The model achieved an accuracy of approximately 82.31%. The precision, recall, and F1 score were 0.74, 0.99, and 0.89, respectively.

Model	Accuracy	Precision	Recall	F1 Score
Logistic Regression	0.49	-	-	-
Decision Trees	0.69	0.68	0.71	0.69
KNN	0.82	0.74	0.99	0.85

Table of results

V. Conclusion

Finally, this research sought to use machine learning algorithms to predict loan defaults. Three distinct models were assessed: Decision Trees, K-nearest neighbors (KNN), and Logistic Regression. With greater accuracy and F1 scores, Decision Trees and KNN performed better than Logistic Regression, which had limited predictive potential. With an accuracy of 82% and good recall, KNN in particular showed the best performance. The findings show how machine learning may be applied to real-world financial problems. With additional development and research, these models can potentially be highly effective instruments for financial risk mitigation and loan default prediction.

VI. References

[1] Shokeen, D., Grover, V., & Verma, V. "Solving the problem of loan default problem in the banking sector using machine learning." In Proceedings of the IEEE International Conference on Machine Learning and Applications, 2023, pp. 123-130.

[2] Diwate, Y., Rana, P., & Chavan, P. "Loan Approval Prediction Using Machine Learning." International Research Journal of Engineering and Technology (IRJET), vol. 8, no. 05, 2021, pp. 123-130.

[3] Zhu, L., Qiu, D., Ergu, D., Ying, C., & Liu, K. "A study on predicting loan default based on the random forest algorithm." Procedia Computer Science, vol. 162, 2019, pp. 503-513.

[4] Xu, J., Lu, Z., & Xie, Y. "Loan default prediction of Chinese P2P market: a machine learning methodology." Scientific Reports, vol. 11, no. 1, 2021, Article ID 18759, pp. 1-8.

[5] Madaan, M., Kumar, A., Keshri, C., Jain, R., & Nagrath, P. "Loan default prediction using decision trees and random forest: A comparative study." In IOP Conference Series: Materials Science and Engineering, vol. 1022, no. 1, p. 012042. IOP Publishing, 2021.

[6] Bhatia, S., Sharma, P., Burman, R., Hazari, S., & Hande, R. "Credit scoring using machine learning techniques." International Journal of Computer Applications, vol. 161, no. 11, 2017, pp. 1-4.

[7] Addo, P. M., Guegan, D., & Hassani, B. "Credit risk analysis using machine and deep learning models." Risks, vol. 6, no. 2, 2018, pp. 38.

[8] Bracke, P., Datta, A., Jung, C., & Sen, S. "Machine learning explainability in finance: an application to default risk analysis." 2019.

[9] Kadam, A., Pawar, H., Phapale, S., & Ganeshe, C. "Prediction of loan defaulter using machine learning - IJCRT." IJCRT, December 12, 2021. [URL: https://www.ijcrt.org/papers/IJCRT2112549.pdf]

[10] Nikhil. (2021), "Loan Default Prediction Dataset," Kaggle, Available: https://www.kaggle.com/datasets/nikhil1e9/loan-default/data

grp4

ORIGINALITY REPORT

13%	10%	9%	8%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	link.springer.com Internet Source	2%
2	Sandeep Kumar Hegde, Rajalaxmi Hegde, Krishna Priya R, Sree Southry S, A.V.G.A. Marthanda, K. Logu. "Performance Analysis of Machine Learning Algorithm for the Credit Risk Analysis in the Banking Sector", 2023 7th International Conference on Computing Methodologies and Communication (ICCMC), 2023 Publication	1%
3	www.mdpi.com Internet Source	1%
4	www.diva-portal.org Internet Source	1%
5	Submitted to University of Southampton Student Paper	1%
6	B Spoorthi, Shwetha S. Kumar, Anisha P Rodrigues, Roshan Fernandes, N Balaji. "Comparative Analysis of Bank Loan Defaulter	1%

Prediction Using Machine Learning Techniques", 2021 IEEE International Conference on Distributed Computing, VLSI, Electrical Circuits and Robotics (DISCOVER), 2021
Publication

7	Guidong Bao, Mengchen Lin, Xiaoqian Sang, Yangcan Hou, Yixuan Liu, Yunfeng Wu. "Classification of Dysphonic Voices in Parkinson's Disease with Semi-Supervised Competitive Learning Algorithm", Biosensors, 2022 Publication	1 %
8	www.coursehero.com Internet Source	1 %
9	Submitted to SP Jain School of Global Management Student Paper	1 %
10	Submitted to University of College Cork Student Paper	<1 %
11	Submitted to De Montfort University Student Paper	<1 %
12	www.ncbi.nlm.nih.gov Internet Source	<1 %
13	www.scilit.net Internet Source	<1 %

14	www.codegrepper.com Internet Source	<1%
15	www.researchgate.net Internet Source	<1%
16	Guangzhen Qu, DongMing Li, Fei Xue, Mingyue Zhu, Wei Guo, Weiyu Xu. "The pan-cancer analysis reveals FAM72D as a potential therapeutic target and closely linked to immune infiltration and prognosis in hepatocellular carcinoma", Research Square Platform LLC, 2022 Publication	<1%
17	cosmosscholars.com Internet Source	<1%
18	dspace.daffodilvarsity.edu.bd:8080 Internet Source	<1%
19	github.com Internet Source	<1%
20	ijettjournal.org Internet Source	<1%
21	opus4.kobv.de Internet Source	<1%
22	researchspace.ukzn.ac.za Internet Source	<1%

23	Yi Sheng Heng, Preethi Subramanian. "Chapter 39 A Systematic Review of Machine Learning and Explainable Artificial Intelligence (XAI) in Credit Risk Modelling", Springer Science and Business Media LLC, 2023 Publication	<1%
24	Niklas Bussmann, Paolo Giudici, Dimitri Marinelli, Jochen Papenbrock. "Explainable Machine Learning in Credit Risk Management", Computational Economics, 2020 Publication	<1%

Exclude quotes	Off	Exclude matches	Off
Exclude bibliography	On		