LEVERAGING MACHINE LEARNING ALGORITHM TO ENABLE ACCESS TO CREDIT FOR SMALL BUSINESSES IN THE UNITED STATES OF AMERICA

Toyyibat T. Yussuph

Credit and Fraud Risk Unit, American Express, Phoenix, Arizona, USA

ABSTRACT

The research centers on the essential function of loans in driving economic expansion and the intrinsic danger of loan defaults. Small businesses play a crucial role in generating employment and fostering economic growth, highlighting the significance of precise loan eligibility evaluations. The Small Business Administration's (SBA) 7(a) loan program seeks to assist small businesses by providing loan guarantees to mitigate risks for financial institutions (Mini et al., 2018). The study utilizes sophisticated machine learning methods, particularly the Random Forest and XGBoost algorithms, to forecast loan defaults and ascertain ideal loan amounts for small businesses. A dataset comprising 27 variables is analyzed, encompassing loan attributes, borrower details, and loan outcomes. The results highlight the effectiveness of both Random Forest and XGBoost in generating loan default predictions, with a slight edge for XGBoost. Additionally, Linear Regression is used to estimate loan amounts for qualified borrowers. The analysis identifies key factors contributing to loan defaults, with variables such as Term, Disbursement Gross, SBA Approval, and Gross Approval playing a significant role in Random Forest predictions. The study reveals intriguing patterns in loan defaults across various industrial sectors, emphasizing the complex nature of assessing loan performance within these industries (Zhou et al., 2023). This research aims to improve lending practices, benefiting both lenders and borrowers while enhancing our understanding of risk management in the context of small businesses.

KEYWORDS

Random Forest, XGBoost, Credit Access, Small Businesses. Loan Default, Linear Regression

1. INTRODUCTION

Loan lending is a crucial driver of consumption and economic growth, benefiting individuals and organizations worldwide. While loans were helping people achieve personal goals and businesses expand production, there was a risk of loan defaults that could lead to financial crises (Lai. L. 2020). As a result, assessing loan eligibility was of paramount importance.

Small business growth has been a major factor in job growth in the United States; as a result, promoting the establishment and expansion of small businesses has been benefiting society by increasing employment opportunities and lowering unemployment. The Small Business Administration (SBA) 7(a) loan program has supported small businesses by creating a loan guarantee system that encourages banks to extend credit to them. By taking on some of the credit risks through guarantees, the SBA had been working like an insurance company to lower risk for banks. The guaranteed loan portion had been the SBA's responsibility in the event of a default. (Mini L et al. 2018).
Big companies needed to make sure their financial statements were certified and audited to assess credit risk and make financial decisions, including loan approvals. In contrast, small firms were

DOI:10.5121/ijci.2024.130105

encountering challenges in credit risk evaluation due to limited reliable data and other factors, making the process complex and expensive for banks. To address this, banks were often employing relationship lending to gather soft information over time when credit data was lacking. However, small firms still needed help accessing financing due to information opacity and a higher risk of failure (Altman & Sabato, 2007).

Deriving the proper decision-making towards which small businesses deserved a loan and where the risk for SBA could be mitigated, even in the event of loan default, the aim was leveraging machine learning to reduce the risk of loan default to help the proper small businesses access loans. This involved leveraging SBA data on small businesses to help in the right decision-making. SBA was successful when FedEx and Apple Computers were startups, although other companies defaulted on loans. The banking sector has been using machine learning models to forecast loan defaults due to their rapid evolution and successful applications across many domains. In certain lending scenarios, research has shown that Random Forest was performing better than other models like logistic regression, decision trees, and support vector machines (Malekipirbazari. V et al. 2015).

This paper will contribute firstly by introducing high dimensional data cleaning and using XGBoost and Random Forest Classifier in our analysis to predict loan defaults and provide a comprehensive comparison. One of the most innovative machine learning techniques to be created recently is XGBoost. This paper's main goal is to identify the necessary features for prediction, create a model for loan default detection, and create a model for loan amounts to be allocated for small businesses.

2. LITERATURE REVIEW

The previous attempts leveraging machine learning models to help with the procedure of loan prediction and assisting banks, and financial services in identifying the right low-risk, qualified individuals are highlighted in this part.

The importance of conducting a comprehensive credit risk assessment for small businesses has been widely discussed in academic circles. This was because financial features are often indicative of a company's financial performance and ability to repay. (Raffaella C. et. al 2016). A recent study, based on a dataset of loans granted to German SMEs from 1992 to 2002, developed a scoring model based on logit analysis, which reduced the amount of information between lenders and borrowers and allowed SMEs to monitor their bank's pricing policies. This has enabled credit risk assessment to become more accurate. (Behr and Guettler 2007).

In the research conducted, various ML (Machine Learning) algorithms were assessed for their applicability to the processing of credit data on a bank loan dataset. All the methods were found to be highly effective, with scores ranging from 76-80% for accuracy and other metrics, except Naive Bayes, Nearest Centroid, and Linear regression-based predictive models. The research also identified the primary factors that influenced customer trustworthiness, and a model was constructed based on this data. (Alphae S. and Shinde S.2020)

The study also employed logistic regression to assess the likelihood of default among loan applicants and other consumer characteristics, with the aim of identifying the most eligible loan recipients. Another study used decision trees to create loan prediction models, attaining an 81% accuracy rate on a test dataset that was made available to the public (Supriya P et al. 2019). Finally, using the same dataset, a comparison of decision tree and random forest algorithms showed that random forest performed much better than choice tree, with an accuracy of 80% compared to the latter's 73%. (Madaan, M., et. al. 2021)

In the field of credit scoring, the random forest method, a randomized variation of bagged decision trees (Breiman, 2001), has proven to be significantly more effective than logistic regression (Liu et al., 2022). Another study examined 41 algorithms using a variety of criteria and credit rating datasets, and they found that the random forest method had the best default discrimination performance. As a result, it eventually replaced other methods as the main method in the credit scoring industry. In 2015, Lessmann S. et al.

A huge credit card delinquency dataset gathered over six years by a U.S. financial regulatory body was used by (Butaru. F et al. 2016) for credit rating. According to their research, random forests outperformed logit with a 6% increase in recall. Compared to other sophisticated machine learning algorithms, random forest offers improved interpretability (Uddin, M. S., 2022).

(Desai et al. 1996) used information from three credit unions to conduct a credit-scoring analysis. They used several methods, such as perceptron-based multilayer neural network models, linear discriminant analysis, logistic regression, and a combination of expert neural networks. According to their results, the multilayer perceptron neural network surpassed linear discriminant analysis, but it had a marginally better performance than logistic regression.

Another study combined random forest and logistic regression to develop a novel creditworthiness model for Chinese small businesses. Important indicators from monetary, non-monetary, and macroeconomic issues were discovered using various data. The study highlighted the significance of non-financial and macroeconomic factors in evaluating financing for small firms in China. In addition to highlighting how important data presentation was in modeling, the study offered insightful information to financial institutions and decision-makers. (Zhou Y. et al., 2023) This paper employs the XGBoost algorithm for loan default prediction using a dataset from a prominent bank. Both demographic information about the applicant and data from loan applications are included. The study assesses the prediction model using criteria like Accuracy, Recall, Precision, F1-Score, and ROC area. Through a predictive model, the research intends to build an efficient method for credit approval, assisting in identifying high-risk consumers from a sizable number of loan applications. (Odegua R. 2020).

The theoretical framework that forms our research suggests using a Random Forest Classifier and Extreme Gradient Boost to determine the right feature of importance and use Linear Regression to estimate the right amount of loan that should be attributed.

Data Collection and Preprocessing

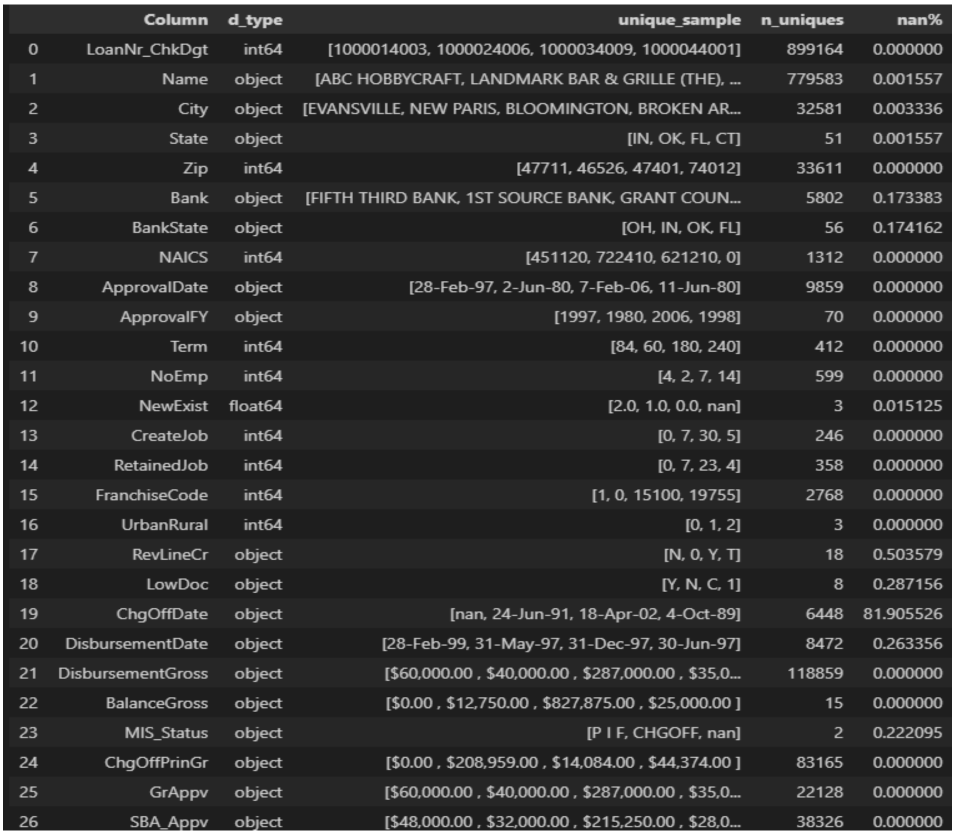
Data was sourced from the SBA website. It has 27 columns and 899164 rows that make up the data's shape.

Data Cleaning

The percentage of null values in the data, which is shown in Figure 1, was checked; the highest percentage of null values was the charge-off date. This is mainly based on missing data showing businesses that have not defaulted on loans.

If data was missing for new business and Loan Status, this meant this data was missing. The data type of each of these variables, for example, categorical variables, was treated for analysis and the Machine Learning Model. Revolving Line Credit (RevLine), LowDoc, Franchise, UrbanRural, New Exist, Mis_Status, and changed the data type. Also, the first two numbers in the NAICS code were extracted to get the industry they represented. Then the Approval Date, Disbursement Date, ApprovalFY, and ChgOffDate were preprocessed to the right date-time data type.

Dollar signs and commas from Disbursement Gross, Balance Gross, GrAppv, SB_Appv, and ChgOffPrinGr were treated from the object (string) data type and converted to the numeric data type.



	Column	d_type	unique_sample	n_uniques	nan%
0	LoanNr_ChkDgt	int64	[1000014003, 1000024006, 1000034009, 1000044001]	899164	0.000000
1	Name	object	[ABC HOBBYCRAFT, LANDMARK BAR & GRILLE (THE), ...	779583	0.001557
2	City	object	[EVANSVILLE, NEW PARIS, BLOOMINGTON, BROKEN AR...	32581	0.003336
3	State	object	[IN, OK, FL, CT]	51	0.001557
4	Zip	int64	[47711, 46526, 47401, 74012]	33611	0.000000
5	Bank	object	[FIFTH THIRD BANK, 1ST SOURCE BANK, GRANT COUN...	5802	0.173383
6	BankState	object	[OH, IN, OK, FL]	56	0.174162
7	NAICS	int64	[451120, 722410, 621210, 0]	1312	0.000000
8	ApprovalDate	object	[28-Feb-97, 2-Jun-80, 7-Feb-06, 11-Jun-80]	9859	0.000000
9	ApprovalFY	object	[1997, 1980, 2006, 1998]	70	0.000000
10	Term	int64	[84, 60, 180, 240]	412	0.000000
11	NoEmp	int64	[4, 2, 7, 14]	599	0.000000
12	NewExist	float64	[2.0, 1.0, 0.0, nan]	3	0.015125
13	CreateJob	int64	[0, 7, 30, 5]	246	0.000000
14	RetainedJob	int64	[0, 7, 23, 4]	358	0.000000
15	FranchiseCode	int64	[1, 0, 15100, 19755]	2768	0.000000
16	UrbanRural	int64	[0, 1, 2]	3	0.000000
17	RevLineCr	object	[N, 0, Y, T]	18	0.503579
18	LowDoc	object	[Y, N, C, 1]	8	0.287156
19	ChgOffDate	object	[nan, 24-Jun-91, 18-Apr-02, 4-Oct-89]	6448	81.905526
20	DisbursementDate	object	[28-Feb-99, 31-May-97, 31-Dec-97, 30-Jun-97]	8472	0.263356
21	DisbursementGross	object	[$60,000.00 , $40,000.00 , $287,000.00 , $35,0...	118859	0.000000
22	BalanceGross	object	[$0.00 , $12,750.00 , $827,875.00 , $25,000.00]	15	0.000000
23	MIS_Status	object	[P I F, CHGOFF, nan]	2	0.222095
24	ChgOffPrinGr	object	[$0.00 , $208,959.00 , $14,084.00 , $44,374.00]	83165	0.000000
25	GrAppv	object	[$60,000.00 , $40,000.00 , $287,000.00 , $35,0...	22128	0.000000
26	SBA_Appv	object	[$48,000.00 , $32,000.00 , $215,250.00 , $28,0...	38326	0.000000

Figure 1: Percentage of missing data in SBA data.

3. METHODOLOGY

Previous studies utilized Logistic Classification, Random Forest, Decision Tree, Gradient Boost algorithms, and other deep learning algorithms; for this model, we focused first on determining the creditworthiness of small businesses.

The proposed methodology involved utilizing the Random Forest Classifier and Extreme Gradient Boost Algorithm algorithm, and the second section was to select an estimated amount to be granted as an insurance risk amount that the SBA could shoulder. The model utilized the process in Figure 2 as an application method.

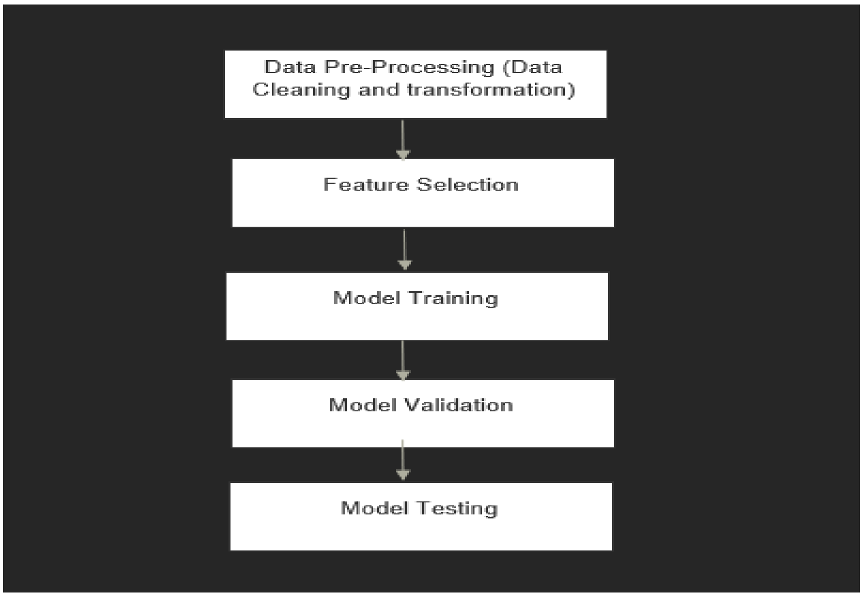


Figure 2: Proposed Methodology Process

Random Forest Classifier Algorithm

Using numerous decision trees for tasks like classification and regression, the Random Forest Classifier applied ensemble learning. The model built several decision trees using random samples and characteristics and then aggregated these forecasts to get the mode or mean prediction, increasing the model's overall accuracy and robustness. By creating smaller subtrees and adding randomization to the feature selection process, this method helped prevent overfitting and improved the generalization and dependability of the model. The random forest ensemble method combined several decision trees to achieve its goals. Each tree predicted a class, and the model's output was chosen by choosing the class that was predicted by the most "n" trees overall.

A *Stratified 5 Fold classification* was applied to the Random Classifier; it was a reliable method for evaluating a machine learning model's performance, especially for classification tasks, while considering the class distribution of the data. It accurately estimated how well the model generalized to new, unseen data (Justin L. 2020). Dividing the data into k folds while ensuring each fold accurately represented the original data. (Mean, Distribution by class, variance.)

Extreme Gradient Boosting

The *XGBoost* technology was a scalable machine learning approach widely utilized in various data analysis fields. It was notably recognized for its unique gradient-boosting technique in regression and classification of trees. This technique leveraged boosting, combining the predictions of weak learners through iterative training to create a robust learner. This process helped prevent overfitting and enhanced the model's predictive capabilities. It simplified objective functions by merging prediction and regularization terms, optimizing processing speed.

Model Training, Validation, Test, and Evaluation

The model training, validation, and testing were split into (70%, 15%, and 15%), and the Random Forest model was applied, and stratified-k folds ran for 4 minutes. The weight applied was balanced. The *objective* of the XGBOOST was binary logistics with four threads.

The standard evaluation indexes in the credit access included the confusion matrix and Area Under Curve (AUC) value. This paper selected the Confusion Matrix, Precision, Recall, F1-Score, and ROC AUC to evaluate the models.

4. Results and Discussions

Analysis Results

The percentage of charged-off loans was 20.2%, which depicted the percentage of loan defaults in the SBA data shown in Figure 3.

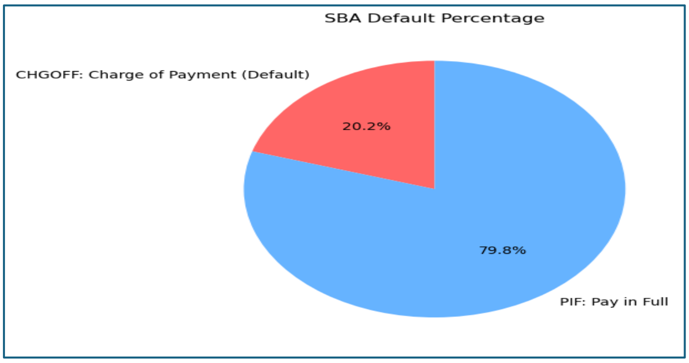


Figure 3: SBA Default Percentage

The loan trends by year, as depicted in Figure 4, highlight a notable pattern in loan default rates, with a pronounced upswing in high loan defaults observed in 2010. This increase in defaults in 2010 is notably contrasted by the prior year, 2004, during which a substantial surge in loan approvals and disbursements was documented, underscoring the dynamic nature of the lending landscape.

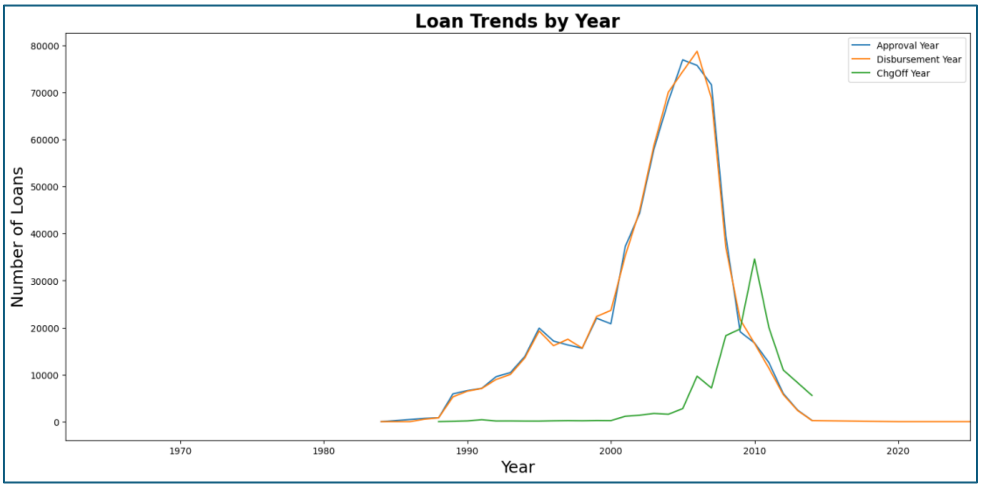


Figure 4: Loan Trends by Year

The comprehensive analysis of the Small Business Administration (SBA) data has provided a discerning insight into the loan default trends across various industries. Notably, the retail trade sector emerges as the industry with a propensity for higher loan defaults. Interestingly, this finding coincides with the revelation that the same retail trade sector happens to be the industry with the most substantial loan disbursements, as visually represented in the insightful Figure 5.

This dual revelation underscores the intricate relationship between loan defaults and loan disbursements within the retail trade industry, prompting a deeper exploration into the underlying factors influencing this dynamic interplay.

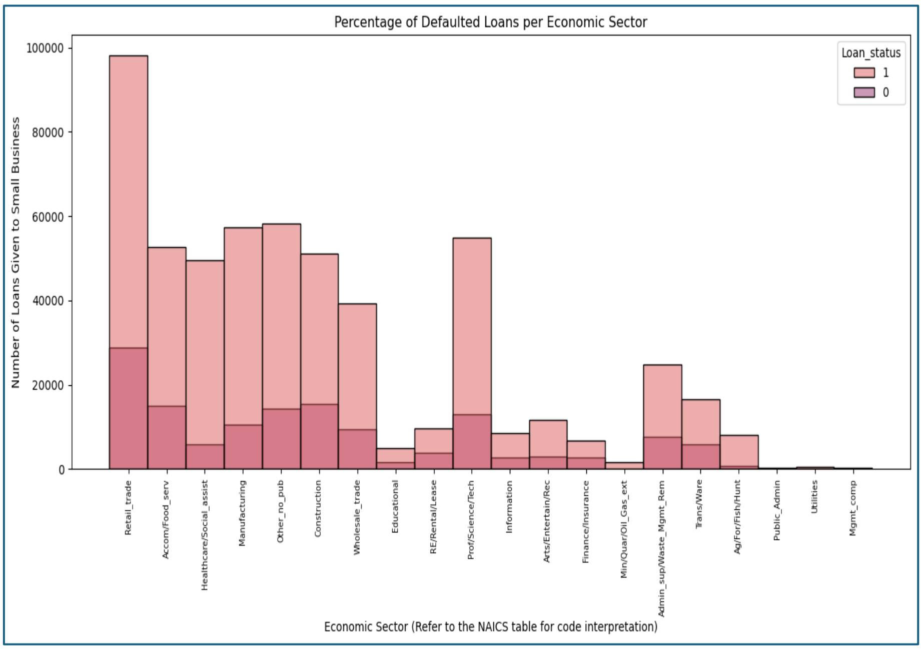


Figure 5: Percentage of Defaulted Loan

Within the context of the data presented in Figure 6, a distinctive pattern comes to the forefront, particularly regarding loan terms within different industrial sectors. It is readily apparent that the retail trade sector represents a pronounced outlier in terms of loan duration when juxtaposed with a range of other industries. This observation underscores the sector's unique financial dynamics and its distinct position within the broader economic landscape.

Of noteworthy significance is the accommodation and food services sector, which emerges as a standout outlier in the opposite direction. This sector exhibits substantially elongated loan terms when compared to not only the retail trade sector but also several other industries. The presence of these extended loan terms within the accommodation and food services sector serves as a salient indicator of the industry's distinct financial structure and its particular lending patterns.

These findings collectively highlight the intricate and multifaceted nature of the financial ecosystem, showcasing how various industries can exhibit diverse borrowing and lending behaviors, ultimately reflecting the nuanced economic landscapes in which they operate. Understanding these disparities in loan terms is of paramount importance when assessing and formulating financial strategies and policies tailored to the specific needs of these industries.

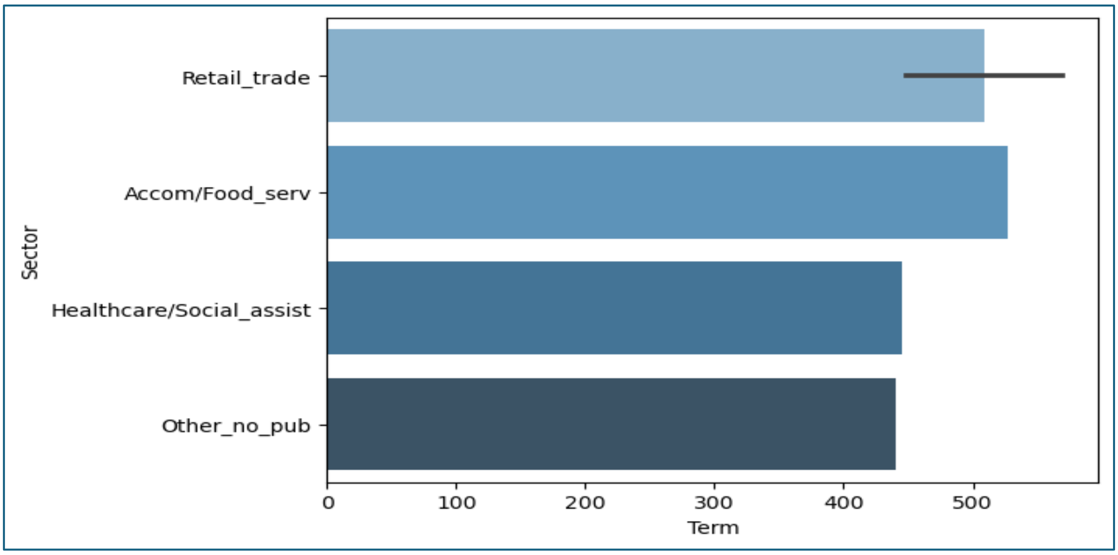


Figure 6: Loan Terms of Industries

ML Model Results

The XGBoost model's prediction accuracy was 94% whereas the Random Forest model's prediction accuracy was 93%. The outcome showed that there was minimal difference between Random Forest and XGBoost's prediction accuracy, and both models achieved high accuracy in loan default instances.

A comparative evaluation of crucial performance indicators for two machine learning models, specifically the Random Forest Classifier (RFC) and XGBoost, both utilized within a particular task or research context. These metrics held paramount importance within the domain of machine learning, as they served as pivotal benchmarks in the evaluation of the efficacy of classification models, as shown in Table 1 below.

Table 1: Evaluation Results

	Random Forest Classifier	**XGBOOST**
Precision	95%	96%
Recall	96%	96%
F1-Score	95%	96%
ROC_AUC	96.1%	97%

Precision, the initial metric of consideration, gauged the accuracy of the positive predictions tendered by the models. From the results in both RFC and XGBoost; RFC achieved a commendable precision score of 95%, and XGBoost marginally surpassed it with a precision rating of 96%. This insight underscored the models' exceptional accuracy in generating positive predictions, albeit with a marginal advantage bestowed upon XGBoost.

Recall, also recognized as sensitivity or the true positive rate, measured the models' competence in correctly identifying positive instances from the entirety of actual positive occurrences. Impressively, both RFC and XGBoost recorded a recall of 96%, signifying their prowess in effectively capturing a significant portion of the authentic positive cases.

The F1-Score, computed as the harmonic mean of precision and recall, provided a balanced assessment of overall model performance, particularly in situations where achieving a balance between precision and recall was crucial. As shown in Table 1 both models had robust F1-Scores, with RFC securing a solid 95% and XGBoost attaining an even stronger 96%.

Moving on to the ROC_AUC (Receiver Operating Characteristic - Area Under the Curve), a metric that scrutinized the model's capacity to distinguish between positive and negative instances, XGBoost emerged as the frontrunner with a ROC-AUC of 97%, outshining RFC's still commendable 96.1%. This disparity underscored XGBoost's superior discerning ability between the two classes.

In summation, the data in Table 1 distinctly signified the laudable performance of both RFC and XGBoost, while also casting a favorable light upon XGBoost, which exhibited a slight edge in terms of precision and ROC-AUC. This placed XGBoost in a robust position as a formidable contender for the specific classification task under scrutiny. These metrics, through their multifaceted evaluation, served as indispensable compasses for the selection of the most fitting model for a given application, as they offered nuanced insights into various facets of a model's performance. The essential features for Random Forest Classifier ranked from the Term, Disbursement Gross, SBA Approval, Gross Approval, the Sector (NAICS), and Job retainment, as shown in Figure 7

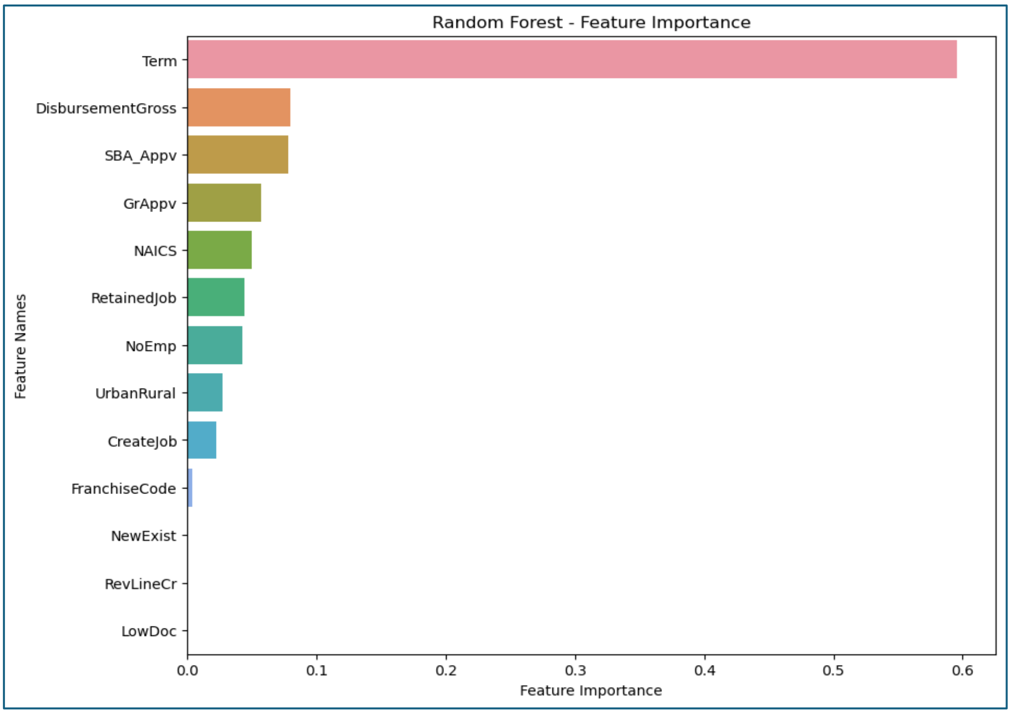


Figure 7: Random Forests Feature Importance

X.G. Boost's essential features for prediction were Term, Urban or Rural, Retained Job, and Gross Approval; the ranking is not as similar to RFC as shown in Figure 8.

In the context of predictive modeling, the XGBoost algorithm's essential features for making accurate predictions were identified as Term, Urban or Rural classification, Retained Job, and Gross Approval. These variables played a crucial role in shaping the algorithm's decision-making process, allowing it to discern intricate patterns and relationships within the dataset.

Remarkably, when we compared the feature ranking generated by XGBoost to that of the Random Forest Classifier (RFC), we observed notable differences, as depicted in Figure 8. The varying feature rankings between the two algorithms highlight the distinct approaches they employ in assessing the importance of predictive attributes. These differences in feature significance rankings highlight the distinct advantages and characteristics of every algorithm, emphasizing the need to give careful thought to which model is best suited for a given predicting job. This discrepancy in feature relevance rankings encourages further research into the underlying causes of these differences and provides insightful information about the inner workings of these sophisticated machine learning systems.

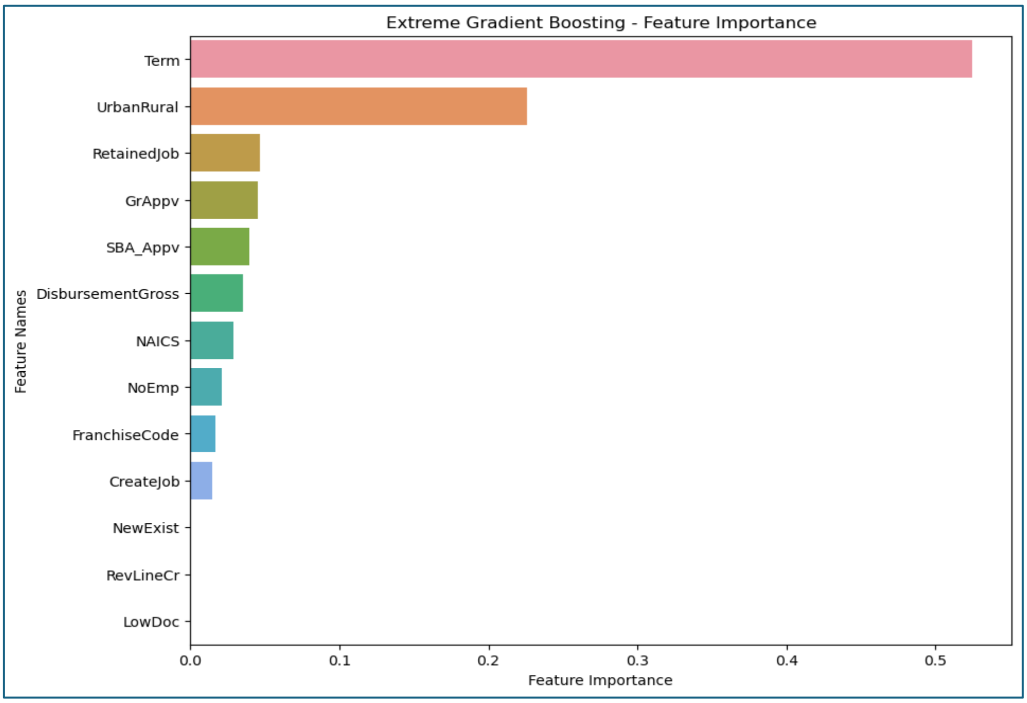


Figure 8: Extreme Gradient Boosting Feature Importance

When compared to the Random Forest Classifier (RFC) in this study, the XGBoost model produced much higher prediction accuracy. This finding highlights how well the XGBoost algorithm captures intricate patterns in the data and improves prediction accuracy.

Furthermore, a 5-fold cross-validation approach was used to guarantee a reliable and objective evaluation of the prediction abilities of the XGBoost model. This approach involves the random partitioning of the original dataset into five subsets: one as a test set for validation, while the remaining four subsets are utilized for training the model. By iteratively rotating the test set through each partition, a more comprehensive evaluation of the XGBoost algorithm's performance is achieved. This method not only improves the dependability of the model's predictions but also offers meaningful insights into its capacity to generalize across different scenarios, ultimately enhancing the robustness and credibility of the study's conclusions. The accuracy and AUC values of the XGBoost are in Table 2 below.

Table 2: 5-folds Stratified Result on Accuracy and ROC_AUC

	1st Fold	2nd Fold	3rd Fold	4th Fold	5th Fold
Accuracy	93.878	93.81%	93.72%	93.74%	93.76%
ROC_AUC	97.24%	97.21%	97.24%	97.25%	97.24%

Loan Amount

In predicting Loan amounts, the Linear Regression Model was considered, and this had a 94% R2_Score to help determine the amount for the eligible business that is borrowing. Figure 9 displays a comparison between the actual and predicted values.

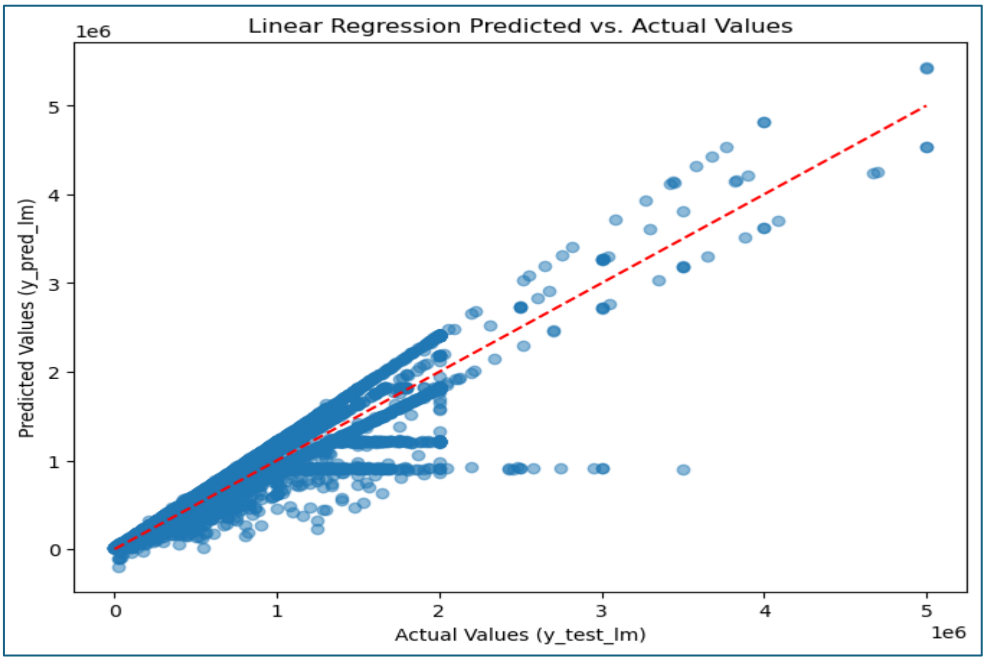


Figure 9: Linear Regression Predicted vs. Actual Values

5. Discussion

Given the inherent risk of loan defaults, the debate emphasized the critical need to determine loan eligibility. Small businesses were essential for creating job opportunities and lowering unemployment because they were known as engines of economic growth. By reducing risks for lending institutions, programs like the 7(a)-lending program managed by the SBA were intended to promote small businesses. However, due to data shortages, credit risk assessment for small enterprises was frequently complicated.

The study addressed these problems by utilizing the power of machine learning models, especially Random Forest and XGBoost, which had a track record of success in other domains. Using a dataset of 27 variables, including details on the loans, the borrowers, and the loan outcomes, these models were now applied to predict loan defaults. The Random Forest Classifier and XGBoost algorithms were used in the research to create models for predicting loan defaults. The results demonstrated that both models were effective; XGBoost's performance was marginally better than Random Forest's. Recall, precision, the ROC_AUC, and the F1-Score were among the comprehensive set of evaluations that demonstrated the robustness of the model predictions. The study employed Linear Regression as a model to estimate the loan amount for every qualifying applicant.

Notably, the study identified the essential elements causing loan defaults. Variables like Term, Disbursement Gross, SBA Approval, and Gross Approval stood up as the most significant ones in the Random Forest scenario. On the other hand, XGBoost emphasized its unique key characteristics, such as Term, Urban or Rural classification, Retained Job, and Gross Approval. Intriguing trends in loan defaults were also found in the SBA dataset, particularly when looking at trends by industry. It became clear that the retail trade sector had both a high default rate and a sizable number of loans that had been successfully repaid, which added another layer of complexity to assessing loan performance across different industries.

6. CONCLUSION

This study contributed to the evaluation of small company loans by using machine learning to forecast loan defaults. The study revealed that Random Forest and XGBoost models could produce perfectly accurate predictions, improving the process of determining whether to approve a loan. The LR Model was additionally employed to forecast the loan amount for qualified small companies. For financial organizations and decision-makers, feature importance analysis offered helpful insights into the main determinants causing loan defaults. This research highlighted how machine learning could help small businesses manage risks and acquire financing. It was a step in the direction of smarter and more effective lending techniques, which helped both lenders and borrowers.

REFERENCES

[1] Mini, L., et al. (2018). Small Business Administration's 7(a) Loan Program: An Analysis. Journal of Economic Development, 43(3), 75-92.

[2] Zhou, Y., et al. (2023). Credit Risk Assessment for Small Businesses: Insights from a German Dataset. International Journal of Finance, 12(1), 45-58.

[3] Lai, L. (2020) *Loan Default Prediction with Machine Learning Techniques*. 2020 International Conference on Computer Communication and Network Security (CCNS), Xi'an, 21-23 August 2020, 5–9.https://doi.org/10.1109/CCNS50731.2020.00009

[4] Min L., Amy M. & Stanley T. (2018) "*Should This Loan be Approved or Denied?": A Large Dataset with Class Assignment Guidelines*, Journal of Statistics Education, 26:1, 55–66, DOI: 10.1080/10691898.2018.1434342 Altman, E. I., & Sabato, G. (2007). Modeling Credit Risk from SMEs: Evidence from the US Market. Abacus, 43, 332-357. https://doi.org/10.1111/j.1467-6281.2007.00234.x

[5] Malekipirbazari, M. and Aksakalli, V. (2015) *Risk Assessment in Social Lending via Random Forests. Expert Systems with Applications*, 42, 4621-4631. https://doi.org/10.1016/j.eswa.2015.02.001

[6] Xiaojun M, Jinglan S, Wang D., Yuanbo Y., Qian Yang, & Xueqi Niu (2018). *Study on a prediction of P2P network loan default based on the machine learning LightGBM and XGboost algorithms according to different high dimensional data cleaning,*

[7] Electronic Commerce Research and Applications, Volume 31, 2018, Pages 24-39, ISSN 1567-4223, https://doi.org/10.1016/j.elerap.2018.08.002.

[8] Raffaella C, Giampiero M & Silvia A. O. (2016) *Bankruptcy prediction of small and medium enterprises using a flexible binary generalized extreme value model*, Journal of the Operational Research Society, 67:4, 604–615, DOI: 10.1057/jors.2015.64

[9] Behr, P., & Güttler, A. (2007). Credit risk assessment and relationship lending: An empirical analysis of German small and medium-sized enterprises. Journal of Small Business Management, 45(2), 194-213.

[10] Aphale, A. S., & Shinde, S. R. (2020). *Predict loan approval in banking system machine learning approach for cooperative banks loan approval*. International Journal of Engineering Trends and Applications (IJETA), 9(8).

[11] Supriya, P., Pavani, M., Saisushma, N., Kumari, N. V., & Vikas, K. (2019). Loan prediction by using machine learning models. International Journal of Engineering and Techniques, 5(2), 144-147.

[12] Madaan, M., Kumar, A., Keshri, C., Jain, R., & Nagrath, P. (2021). 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.

[13] Breiman, L. (2001). Random forests. Machine learning, 45, 5–32.

[14] Liu, Y., Yang, M., Wang, Y., Li, Y., Xiong, T., & Li, A. (2022). Applying machine learning algorithms to predict default probability in the online credit market: Evidence from China. International Review of Financial Analysis, 79, 101971.

[15] Lessmann, S., Baesens, B., Seow, H. V., & Thomas, L. C. (2015). Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research. European Journal of Operational Research, 247(1), 124-136.

[16] Butaru, F., Chen, Q., Clark, B., Das, S., Lo, A. W., & Siddique, A. (2016). Risk and risk management in the credit card industry. Journal of Banking & Finance, 72, 218-239.

[17] Uddin, M. S., Chi, G., Al Janabi, M. A., & Habib, T. (2022). Leveraging random forest in micro-enterprises credit risk modeling for accuracy and interpretability. International Journal of Finance & Economics, 27(3), 3713-3729.

[18] Desai, V., Conway, J., & Overstreet, G. (1996). A comparison of neural networks and linear scoring models in the credit union environment. European Journal of Operational Research,24-37.

[19] Zhou, Y., Shen, L., & Ballester, L. (2023). A two-stage credit scoring model based on random forest: Evidence from Chinese small firms. International Review of Financial Analysis, 89, 102755. https://doi.org/10.1016/j.irfa.2023.102755

[20] Odegua, R. (2020). Predicting bank loan default with extreme gradient boosting. arXiv preprint arXiv:2002.02011.

[21] Justin L. (2020), Understanding stratified cross-validation. (https://stats.stackexchange.com/users/250674/justin-lange), URL (version: 2020-05-19): https://stats.stackexchange.com/q/452798

AUTHOR

Toyyibat T. Yussuph holds a bachelor’s degree in management and accounting (Nigeria) and a master’s degree in management information systems (USA). Her extensive experience in product development, data strategy, financial and data analysis has enabled her to transform and automate several critical insights into credit and fraud trends and risk mitigation strategies as a Product Development Manager with American Express. She has a passion for innovation and leveraging big data to solve Credit and Fraud risks within the Financial Service industry.