Customer Loan Default Prediction Using Machine Learning

Udochukwu Clement Ekeocha
Dept. of Computer Science
York St. John's University
London campus, United Kingdom.
udochukwu.ekeocha@yorksj.ac.uk

***Abstract*— Given the substantial impact of customer loan defaults on the stability and financial viability of institutions within the financial sector, it becomes imperative to ensure an accurate assessment of default risk during the loan approval process. In order to do this, the study aims to create a machine learning model that can accurately predict customers who are likely to miss loan repayments using a dataset of loan applications. One significant process involves the advanced feature selection techniques with the primary objective of identifying the most influential variables that significantly affect loan default rates. These variables include important factors like the frequency of delinquencies, employment types, location, interest rate, loan size, and tenure. The study also conducts a comparative analysis of different machine learning algorithms, such as Extreme Gradient Boosting, Random Forest, and Logistic Regression. Multiple metrics, including precision, recall, F1 score, and the area under the receiver operating characteristic (ROC) curve, are used to assess these algorithms. It is important to note that these analyses take the dataset's imbalance into account. With respect to the aforementioned metrics for predicting customer loan defaults, the results demonstrate that the random forest classifier model outperforms others, exhibiting an AUC of 0.82, an F1 score of 0.89, and a recall of 0.69. According to the study's findings, financial institutions can use machine learning techniques to identify high-risk borrowers and reduce loan default rates. Furthermore, it suggests that future studies could investigate the use of data sources with better-aligned characteristics like credit scores to improve the precision of loan default prediction models.**

Keywords*— *machine learning, extreme gradient boosting, accuracy, loan default, classification, customer default, random forest, logistic regression, supervised learning, confusion matrix.

I. INTRODUCTION

The financial sector is crucial to the economy because it provides vital financial services, like loans to people and businesses. One of the biggest problems lending financial institutions face is managing the risk of loan default. When borrowers are unable to repay their loans within the specified time frame, loan default occurs, which can have serious consequences for both borrowers and lenders [7]. Borrowers who default on loans may suffer monetary losses, legal repercussions, and credit score damage, making it more difficult for them to obtain loans in the future. Loan defaults for lenders may result in sizable monetary losses, business interruptions, potential regulatory noncompliance, and legal actions against the borrower. To address the problem of loan defaults, lenders have been forced to rely on loan default predictors developed using machine learning algorithms and techniques to identify borrowers during the application stage that are most susceptible to fall short on their loans.

In recent years, machine learning algorithms have grown in popularity as a reliable resource for loan default prediction. Machine learning models are particularly useful in this context because they can analyze large and complex datasets and identify patterns and relationships that may not be apparent using traditional statistical techniques. During the development of an effective loan default prediction model using machine learning, lenders would need to have access to high-quality data. This data should include information on the borrower's credit history, employment status, income, age, occupation, and other relevant factors. By feeding this data into the prediction model built with machine learning algorithms, lenders can identify high-risk borrowers and take proactive steps to mitigate the risk of loan default early on.

A. BACKGROUND

Credit granting has ancient origins in civilizations like Egypt, Babylon, and Assyria, stretching back millennia. European merchants during the Middle Ages popularized credit for trade across borders. The Industrial Revolution further revolutionized credit as global product demand necessitated the "buy now - pay later" model. This facilitated business growth by accommodating increased demand beyond profit margins. Njuguna and Sowon [11] emphasized credit's rise, prompting lenders to assess creditworthiness and loan specifics. This positioned credit management as a pivotal business aspect in the mid-20th century. Credit, as per Njuguna and Sowon [11], underpins commerce and the economy, enabling business expansion and personal consumption. However, credit entails risks necessitating understanding and control. Credit risk management is crucial to identify and mitigate such calculated risks [11].

In Bandyopadhyay [3]'s work, credit risk management is vital for financial institutions' stability and growth. Collaboratively, they and regulators curb risks, ensuring market efficiency and minimizing losses. Amid possible high defaults, banks might relax standards during low interest to boost revenue. Bandyopadhyay [3] emphasizes three factors elevating credit risk management's importance. Realities like non-performing assets, competition, and volatile collateral underscore this. After 2008 crisis, rising non-performing assets expose lenders to credit risks. Collateral shifts introduce uncertainty in recovery, while competition narrows margins. Regulatory changes and risk tolerance accentuate its relevance. Data and machine learning build credit-scoring models, precisely forecasting risk for balance. Credit scores guide creditworthiness with better scores yielding favorable terms and vice versa [18]. Scores guide decisions from pre-screening to account management, considering payment, utilization, debt, employment, and demographics. After risk ratings, applicants are categorized, higher risk facing scrutiny or denial.

B. OUTLINE OF THE STUDY

This study explores the use of machine learning in predicting customer loan defaults in the financial sector. It begins with an introductory chapter, analyzing prior methodologies and insights, and then presents the methodology, data preprocessing, feature selection, and algorithms used. The results chapter evaluates the model's effectiveness, revealing its capabilities and limitations. The conclusion discusses real-world implications for the lending industry and challenges in managing credit risk with machine learning. The study aims to determine if machine learning can effectively identify high-risk borrowers and reduce financial losses from defaults.

II. LITERATURE REVIEW

Machine learning techniques have been increasingly utilized to develop more accurate models for predicting loan default risk and assessing creditworthiness.

Sheikh et al. [16] evaluated classification algorithms like random forests and support vector machines on loan application data, finding superior performance over logistic regression and decision trees for approval prediction. Khandani et al. [9] similarly found machine learning outperformed conventional techniques for credit risk modeling, with random forest excelling in accuracy and predictive power.

Multiple studies have benchmarked algorithms using loan and credit data. Machado and Karray [10] developed a hybrid model combining machine learning with financial ratios, outperforming lone models. Aniceto et al. [2] found random forest and gradient boosting machine effective for predicting bank loan and peer-to-peer lending defaults respectively. Tiwari [19] demonstrated random forest's superiority for predicting defaults using loan data.

Some research also focused on using machine learning with government and peer-to-peer lending data alongside macroeconomic factors to improve prediction using optimized ensembles and profit-driven models [6]. Further research by Shivanna and Agrawal [17] explored cloud-based machine learning solutions by finding benefits in terms of scalability and faster modeling. Their result provided evidence of Azure's data science virtual machine outperforming other models in terms of accuracy and AUC.

Multiple works emphasized data balancing, cleaning and preprocessing as crucial steps prior to modeling. Performance showed to be superior on balanced datasets and eliminating discriminatory variables showed not to reduce predictive power [1], [6].

In summary, research showed that meticulous data preparation and algorithm selection allows financial institutions to leverage these advances for improved lending decisions [1], [19].

A. RESEARCH GAPS AND CONTRIBUTION

Prior studies in loan default prediction using machine learning have concentrated on classification model development and accuracy assessment. However, these often omit exploring customer attributes influencing default propensity and overlook suitable evaluation metrics for imbalanced data.

In contrast, this research aims to identify early warning indicators for loan default, enabling proactive risk mitigation. Emphasizing metrics like recall, precision, F1 score, and ROC curve, the study offers comprehensive performance insights on imbalanced datasets.

Addressing gaps in existing research, this work deepens the understanding of default prediction features and advocates for appropriate evaluation metrics. Enhancing model accuracy, it empowers informed credit risk management decisions.

In conclusion, this section showcases studies using machine learning to predict loan defaults, highlighting algorithm superiority (e.g., Random Forest, Gradient Boosting, Support Vector Machine) and guiding future research to target early detection and proper metrics for unbalanced datasets.

III. METHODOLOGY

The study focuses on predicting customer loan default using machine learning techniques. This chapter outlines a thorough framework for predicting loan default, covering tools used, research hypothesis formulation, problem-solving approach, experiment design, data collection methods, and chosen machine learning algorithms with rationale for their suitability in credit scoring prediction

A. EXPERIMENT DESIGN

The design workflow, depicted in Figure 1, outlines the planned research approach. The primary goal is to uncover borrower attributes impacting loan default probability during application. This involves constructing a predictive model using Logistic Regression, Random Forest, or Extreme Gradient Boosting.

The workflow includes essential steps ensuring model accuracy and reliability, aligning with research aims. These steps will unveil key factors linked to loan default.

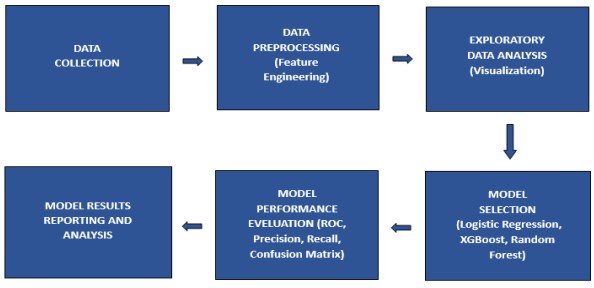

Figure 1: Proposed workflow of research model development

1) DATA COLLECTION

This research uses Kaggle's "TVS_Loan_Default" dataset, published by Sayantan Jana in 2020, to develop a risk model for evaluating customer default likelihood after loan products are sold. The dataset has a usability score of

8.24 and is anonymized to protect applicant privacy. It contains 119,528 rows and 32 columns, with the target variable being the V32 outcome. Variables include loan amount, tenure, bounce rates, gender, interest rates, and geographical location.

2) DATA PREPROCESSING

Zelaya [20] emphasizes the importance of data preprocessing in machine learning model development, accounting for 80% of the process. This step reduces the inclusion of problematic or irrelevant data, ensuring the model's output is free of false or misleading information.

a: MISSING VALUES

Missing values in data preprocessing are crucial for removing inconsistencies and ensuring model development's accuracy. Researchers use various methods to handle missing values, including replacing them with statistical measures, replacing them with 0 depending on the task, or completely deleting them. However, this approach may lead to the loss of valuable information, as more information yields better results. It is crucial to consider alternative strategies to effectively handle missing values while preserving data integrity and completeness.

b: OUTLIERS

Sedefozcan [15] emphasizes the significance of data preprocessing, particularly outliers, in machine learning model development. Outliers are data points that differ significantly from other observations, affecting model quality and performance. Outlier detection and handling are crucial during the data preprocessing stage. Methods include the Z-score approach, which normalizes variables and sets a threshold range for outliers. Additionally, visualizations like histograms, box plots, and scatter plots are used during EDA to identify outliers in datasets.

c: FEATURE ENGINEERING

Feature engineering is crucial in machine learning and statistical modeling, transforming raw data into relevant features for predictive models. It aims to improve model performance and accuracy by optimizing input data for the chosen algorithm [14]. Researchers can uncover meaningful patterns, relationships, and insights by carefully choosing and developing features. Techniques include one-hot encoding, bag of words, binging, and N-grams, which help in sequence prediction tasks like sentiment analysis. These techniques help researchers uncover meaningful patterns and relationships in their data [14].

d: NORMALIZATION

Normalization is a widely used data preparation technique in machine learning, adjusting column scales to ensure comparability and minimize outliers. It enhances the performance and stability of machine learning models by standardizing features [4]. Common normalization methods include Min-Max Normalization, Decimal Scaling, Log Scaling, and Z-Score. Min-Max Normalization adjusts feature values to a predefined range, while Decimal Scaling scales numerical attributes within a specific range. Log Scaling compresses a wide range of values into a narrower range, particularly useful for power law data with small values and few occurrences. Z-Score scaling measures standard deviations away from the mean, allowing for better comparison of values across different features and removing outliers' influence on the overall distribution. By normalizing data to a standard scale, it is easier to identify patterns, analyze relationships, and make more accurate predictions in statistical and machine learning models [8].

e: IMBALANCED DATASET

Data imbalance occurs when the label column has an imbalanced distribution of categories, often affecting machine learning algorithms [5]. Severe imbalances can hinder learning processes, as the majority class has a more significant impact. Two methods to reduce dataset imbalance are oversampling and undersampling [13]. Oversampling involves random additions to the minority class to balance the imbalance but may introduce insignificant objects or noise. Undersampling deletes members of the majority class to maintain the imbalance, but may remove significant samples, potentially affecting model prediction efficiency.

3) EXPLORATORY DATA ANALYSIS

Exploratory Data Analysis (EDA) is an analytical approach that summarizes dataset characteristics using statistical graphs and visualization techniques. It goes beyond modeling and hypothesis testing, uncovering insights and patterns before formal statistical modeling [12]. Different charts used in EDA include bar charts, scatter plots, Histograms, or Box plots.

4) MODEL SELECTION

This research focuses on predicting customer loan default using machine learning algorithms. The three algorithms used are Random Forest, Logistic Regression, and Extreme Gradient Boosting. Random Forest is a user-friendly ensemble method with uncorrelated decision trees, providing enhanced predictive capabilities [8]. Logistic Regression is a widely used statistical model for classification, estimating dependent variable probabilities and modeling relationships between variables. Its simplicity, efficiency, and adaptability make it valuable for various applications [5]. Extreme Gradient Boosting, created by Tianqi Chen in 2016, is a popular tree-based algorithm with exceptional performance across various machine learning tasks. It has been widely adopted and demonstrated in Kaggle competitions, with XGBoost-powered models consistently outperforming other approaches [5]. XGBoost employs tree boosting, transforming feeble predictors into powerful ones, resulting in 90%+ model accuracy performance.

5) MODEL PERFORMANCE EVALUATION

The loan dataset in this research work demonstrates a significant class imbalance, making relying solely on

accuracy as an evaluation metric inadequate]. In such imbalanced scenarios, relying solely on accuracy can lead to misleading conclusions. To address this challenge, alternative model evaluation metrics are explored, including Precision, Recall, F1-score, the Receiver Operating Characteristic (ROC) curve, and the confusion matrix. These metrics provide a more comprehensive and accurate assessment of a classification model's performance on the imbalanced data, allowing for informed interpretation and decision-making.

The confusion matrix provides detailed information about classification errors, while Precision evaluates the effectiveness of a classifier's outcomes [8]. Recall measures the completeness of a classifier's results, while F1-score measures the model's ability to capture positive cases while maintaining accuracy with the cases it identifies [8]. The ROC curve and AUC are valuable tools for evaluating and comparing the performance of various classification models, particularly in scenarios involving imbalanced data [14].

IV. ANALYSIS AND MODEL DEVELOPMENT

This chapter discusses the development and analysis of classification models for prediction using machine learning algorithms Logistic Regression, Random Forest, and XGBoost.

A. DATA OVERVIEW

This study uses a 119,528-row CSV file of TVS customer loan applicant data to analyze and model their financial traits. Key indicators include equated monthly installments (EMI), business relationship duration, and previous bounces, which reveal financial stability, discipline, and potential impact on credit score. These insights help predict loan defaults and improve borrowing prospects. Figure 2 below shows the target variable V32 with 116,914 (97.81%) non-defaults and 2,614 (2.19%) defaults.

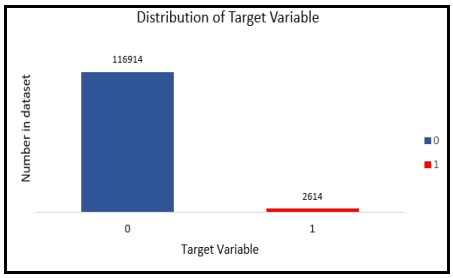

Figure 2. Distribution of Target Variable.

B. DATA VISUALIZATION

Data Visualization plays a crucial role in aiding machine learning analysts comprehend and examine datasets by presenting them in a user-friendly format. By visually representing the TVS dataset in this research, data visualization aids the identification of outliers, and patterns that may not be seen through other analytical methods.

Figure 3: Bar plots showing the bivariate analysis between customer property and target variable

C. DATA PREPROCESSING

1) FEATURE SELECTION

All features in the dataset with more than 30% missing values and features that lack comprehensive understanding are removed to improve model bias and robustness.

2) MISSING VALUES

First, all categorical features with object data types are converted to numerical data types. After which all categorical missing values are inputted using mode while the numerical features are inputted using mean.

3) OUTLIERS

Outliers are identified in ‘V28’, ‘V25’, ‘V29’, ‘V30’, ‘No_Of_Unsecured_Loans’, ‘No_Of_Loans’, ‘EMI’, and ‘No_Of_Bounce’. An example is displayed below

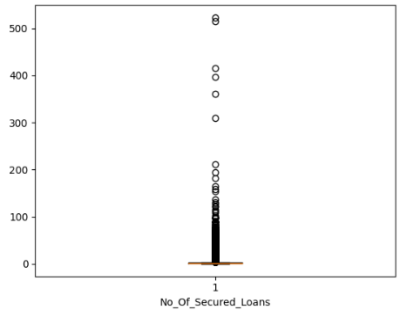

Figure 4: Outlier identification in No_of_Secured_Loans using boxplots

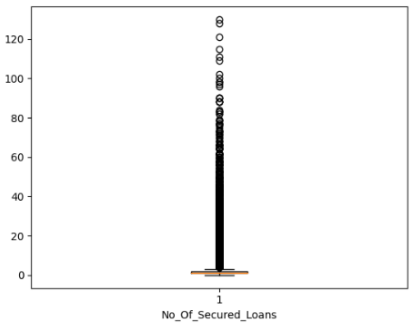

Figure 5: No_of_Secured_Loans Representation Post-Outlier Removal

4) DATA NORMALIZATION

The last step in the preprocessing stage is the normalization of the entire dataset to a common scale due to the presence of different ranges in the features of the dataset. The method of normalization used in this research work is the min-max method which scales the data to a fixed range of 0 to 1.

D. MODEL CREATION

This study focuses on developing prediction models using three algorithms: logistic regression, random forest, and XGBoost. Six models were considered, including the basic and optimized models using RandomizedSearchCV. The dataset is divided into two sets, with 80% of observations assigned to the training set and 20% to the test set. The test set evaluates model performance, while the training set undergoes oversampling using SMOTE, a technique used to address class imbalance in datasets. SMOTE generates synthetic samples of the minority class, enhancing predictive performance while preserving the original dataset's characteristics and patterns. The base models are tuned further using RandomizedSearchCV and some of the parameters considered are listed below.

1) LOGISTIC REGRESSION

In the course of tuning the Logistic Regression model, the following parameters below are considered:

i. Regularization ‘Penalty’: ['l1', 'l2']

ii. Optimization 'solver': ['liblinear', 'saga']

iii. Maximum iterations: [100, 200, 300]

iv. Cross-Validation Fold ‘cv’: 5

2) RANDOM FOREST

Some parameters considered in the hyper-tuning of the base RandomForest model are as follows:

i. No. of Decision Trees 'n_estimators': [100, 200, 300]

ii. Max Depth for each Tree 'max_depth': [None, 5, 10]

iii. Min Samples per leaf node 'min_samples_leaf': [1, 4]

iv. Bootstrap samples 'bootstrap': [True, False]

3) EXTREME GRADIENT BOOSTING

The parameters considered in the hyper-tuning of the XGBoost model include:

i. Maximum Tree Depth - 'max_depth': [3, 4, 5],

ii. Step-Size - 'learning_rate': [0.1, 0.01, 0.001],

iii. Number of Trees - 'n_estimators': [100, 200, 300],

iv. Lasso Regularization - 'reg_alpha': [0, 0.1, 0.5],

v. Ridge Regularization - 'reg_lambda': [0, 0.1, 0.5]

E. MODEL PERFORMANCE EVALUATION

This section evaluates logistic regression, random forest, and XGBoost models' performance using metrics like ROC curve, F1-score, precision, and confusion matrix on unseen data.

Models	True Positive (TP)	False Positive (FP)	True Negative (TN)	False Negative (FN)
Logistic Regression Base Model	20000	3800	320	160
Logistic Regression Tuned Model	20000	3700	320	160
Random Forest Base Model	20000	3800	310	170
Random Forest Base Model	23000	3100	39	440
XGBoost Base Model	22000	1200	110	370
XGBoost Tuned Model	20000	3700	320	160

Table 1: Confusion matrix of the base and tuned versions of the classification models

Models	AUC	F1-Score	Recall	Precision
Logistic Regression Base Model	0.80	0.89	0.65	0.075
Logistic Regression Tuned Model	0.81	0.89	0.67	0.08
Random Forest Base Model	0.82	0.89	0.69	0.073
Random Forest Base Model	0.81	0.97	0.08	0.11
XGBoost Base Model	0.77	0.97	0.17	0.18
XGBoost Tuned Model	0.78	0.89	0.67	0.08

Table 2: Performance evaluation report of the base and tuned versions of the classification models

F. RESULTS

The study re-classifies six developed models into three distinct models based on evaluation performance metrics. The best-performing version is chosen as the representative model. Logistic Regression outperforms the base model in AUC, recall, and precision. Random Forest outperforms the base model, while XGBoost outperforms the tuned version.

Table 3 represents the results of the best versions of each model for effective comparison.

Models	AUC	F1-Score	Recall	Precision
Logistic Regression	0.81	0.89	0.67	0.08
Random Forest	0.82	0.89	0.69	0.07
XGBoost	0.78	0.89	0.67	0.08

Table 3: Model Results

Table 3 shows the AUC, F1-score, Recall, and Precision metrics for various models. Logistic Regression achieved an AUC score of 0.81, indicating moderate ability to distinguish between positive and negative classes. The F1-Score of 0.89 indicates a good balance between Precision and Recall. The model correctly identified 67% of positive instances but scored a low precision of 0.08. Random Forest achieved an AUC score of 0.82, demonstrating better accuracy in separating positive and negative classes. The model also achieved a high ability to capture instances of interest but showed a lower tendency to incorrectly classify negative instances as positive. Finally, XGBoost achieved an AUC score of 0.78, indicating a moderate ability to distinguish between positive and negative classes. The F1-Score of 0.89

is similar to Logistic Regression and Random Forest, indicating a good balance between Precision and Recall.

From the comparison, the Random Forest model is the best choice for identifying customers likely to default on their loans during the loan application process. It achieves the highest AUC score of 0.82, a high F1-Score of 0.89, and a high recall of 0.69. However, it has a high rate of false positives, with a precision of 0.07. Further investigation is needed to identify the most influential features for prediction.

Furthermore, one of the research questions proposed in the introduction involved providing insights into the features or characteristics of customers during loan application that might influence the chances that the customer would default on the loan repayment. The answer to this research question is given using the feature importance plot of the Random Forest model as seen in figure 6.

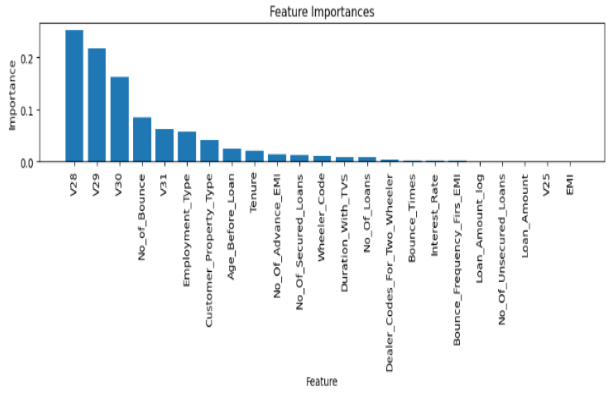

Figure 6: Most Significant Features in Developing the Random Forest Classification Model

The Random Forest model's feature importance plot reveals three key customer features: "V28," "V29," and "V30." These features capture delinquencies at 30 days, 60 days, and 90 days for EMIs. This information is crucial for lenders to identify and manage default risks, improve loan portfolio management, and reduce financial losses.

V. CONCLUSION

This study combines literature review and empirical study to compare machine learning algorithms for loan default prediction. The Random Forest model outperformed Logistic Regression, and XGBoost in identifying customer characteristics likely to influence default decisions. The Random Forest model achieved an AUC of 0.82, F1-score of 0.89, recall of 0.69, and precision of 0.07. These performance metrics provide advantages for lending organizations, such as capturing 69% of customers likely to default on loan repayments and distinguishing between defaulters and non-defaulters. The low precision score of 0.07 indicates a high number of false positives, but this trade-off allows organizations to focus on addressing false positives while still benefiting from the model's overall ability to identify defaulters effectively. The study also highlights the importance of identifying relevant features for accurately predicting customers likely to default on loan repayment, with "V28" having the most influence on customers defaulting on their loan repayment plans. This knowledge can aid financial institutions in improving risk assessment processes and allocating resources effectively.

BIBLIOGRAPHY

[1] Alam, T.M. et al. (2020) 'An Investigation of Credit Card Default Prediction in the Imbalanced Datasets', IEEE Access, 8, pp. 201173–201198. Available at: https://doi.org/10.1109/ACCESS.2020.3033784.

[2] Aniceto, M.C., Barboza, F. and Kimura, H. (2020) 'Machine learning predictivity applied to consumer creditworthiness', Future Business Journal, 6(1), pp. 1–14.

[3] Bandyopadhyay, A. (2016) Managing Portfolio Credit Risk in Banks. Cambridge University Press. Available at: https://doi.org/10.1017/CBO9781316550915.

[4] Bhanja, S. and Das, A. (2018) 'Impact of data normalization on deep neural network for time series forecasting', arXiv preprint arXiv:1812.05519 [Preprint].

[5] Brownlee, J. (2019) A Gentle Introduction to Imbalanced Classification - MachineLearningMastery.com. Available at: https://machinelearningmastery.com/what-is-imbalanced-classification/ (Accessed: 31 July 2023).

[6] De Castro Vieira, J.R. et al. (2019) 'Machine learning models for credit analysis improvements: Predicting low-income families' default', Applied Soft Computing, 83, p. 105640. Available at: https://doi.org/https://doi.org/10.1016/j.asoc.2019.105640.

[7] Chang, D. (2023) What Does It Mean to Default on a Loan? | The Motley Fool. Available at: https://www.fool.com/the-ascent/personal-loans/what-does-it-mean-to-default-on-loan/ (Accessed: 3 August 2023).

[8] Himberg, T. (2021) 'Loan Default Prediction with Machine Learning'.

[9] Khandani, A.E., Kim, A.J. and Lo, A.W. (2019) 'Consumer credit-risk models via machine-learning algorithms', Journal of Banking & Finance, 34(11), pp. 2767–2787. Available at: https://doi.org/https://doi.org/10.1016/j.jbankfin.2010.06.001.

[10] Machado, M.R. and Karray, S. (2022) 'Assessing credit risk of commercial customers using hybrid machine learning algorithms', Expert Systems with Applications, 200, p. 116889. Available at: https://doi.org/https://doi.org/10.1016/j.eswa.2022.116889.

[11] Njuguna, R. and Sowon, K. (2021) 'Poster: A Scoping Review of Alternative Credit Scoring Literature', in ACM SIGCAS Conference on Computing and Sustainable Societies. New York, NY, USA: Association for Computing Machinery (COMPASS '21), pp. 437–444. Available at: https://doi.org/10.1145/3460112.3471972.

[12] Patil, P. (2018) What is Exploratory Data Analysis? | by Prasad Patil | Towards Data Science. Available at: https://towardsdatascience.com/exploratory-data-analysis-8fc1cb20fd15 (Accessed: 31 July 2023).

[13] Pykes, K. (2020) Oversampling and Undersampling. A technique for Imbalanced… | by Kurtis Pykes | Towards Data Science. Available at: https://towardsdatascience.com/oversampling-and-undersampling-5e2bbaf56dcf (Accessed: 31 July 2023).

[14] Rosencrance, L. (2021) What is Feature Engineering for Machine Learning? Available at: https://www.techtarget.com/searchdatamanagement/definition/feature-engineering (Accessed: 31 July 2023).

[15] Sedefozcan (2022) Feature Engineering & Data Pre-Processing: Outliers | by Sedefozcan | Medium. Available at: https://medium.com/@sedefftaskin92/feature-engineering-data-pre-processing-outliers-e072f7bdcc63 (Accessed: 31 July 2023).

[16] Sheikh, M.A., Goel, A.K. and Kumar, T. (2020) 'An approach for prediction of loan approval using machine learning algorithm', in 2020 International Conference on Electronics and Sustainable Communication Systems (ICESC), pp. 490–494.

[17] Shivanna, A. and Agrawal, D.P. (2020) 'Prediction of defaulters using machine learning on Azure ML', in 2020 11th IEEE annual information technology, electronics and mobile communication conference (IEMCON), pp. 320–325.

[18] Thomas, L., Crook, J. and Edelman, D. (2017) Credit scoring and its applications. SIAM.

[19] Tiwari, A.K. (2018) 'Machine learning application in loan default prediction', JournalNX, 4(05), pp. 1–5.

[20] Zelaya, C.V.G. (2019) 'Towards explaining the effects of data preprocessing on machine learning', in 2019 IEEE 35th international conference on data engineering (ICDE), pp. 2086–2090.