Explainable Artificial Intelligence in the Credit Verification Process

Senzekile Mofokeng
Alma Mater Europeae
Europe, Slovenia
senzekile.mofokeng@almamater.si

Abstract

The objective of this study is to evaluate the transparency of the credit verification process when machine learning algorithms are used to predict customer credit facility defaults. XGBoost, was utilised for enhancing credit score evaluation on secondary credit verification data obtained from Kaggle. Meanwhile, the Local Interpretable Model-Agnostic Explanation (LIME) provides valuable insights into model operations, enabling the identification of critical areas within images or highlighting important features. The results indicate that the most important feature is the duration, also known as the term of the loan. The second important feature is the paydays, which is the number of days in which repayments are made, and the third most important feature is whether the customer owns a house.

Keywords: Credit verification, XGBoost, LIME

1 Introduction

1.1 Introduction

The current drama between the OpenAI board is interesting; the substance of the debate is whether OpenAI is committed to creating safe AI [1]. OpenAI is an organisation that has heightened research on generative AI, thus introducing ChatGPT in early 2021. The scholarly debate on safe AI has been topical and has been discussed in multiple disciplines. In law, the question of who is accountable when AI errs [2]. In social science, concerns are raised about fairness and whether AI can be trusted [3]. Computer scientists have been at the forefront of designing trustworthy AI by reducing bias and making it more transparent [4]. An industry that could substantially benefit from transparent AI is the financial services industry; as this industry is highly regulated, customer trust is key to sustaining profitability, and financial risk needs to be always managed. A process that poses the greatest financial risk to the point where a financial services organisation can be closed is the credit verification process [5]. The credit verification process assesses a customer's historic credit profile to predict whether a customer will not default on future credit facilities or loan [5].

The objective of this study is to evaluate the transparency of the credit verification process when machine learning algorithms are used to predict customer credit facility defaults.

Trust plays a vital role in the recommendations made by AI systems in critical sectors such as healthcare, banking, and criminal justice. A key challenge lies in comprehending the intricate nature of machine learning models. While these models can decipher complex relationships between input variables and outcomes, understanding their underlying processes can be complex [6]. Commonly known as "black box models," algorithms may struggle to meet legal standards [7]. It is widely acknowledged that explainable AI (XAI) is essential for establishing trust in classifier algorithms. Nonetheless, there exist varied theoretical frameworks and approaches in different research studies, XAI effectively elucidates biased and unbalanced datasets [6].

1.2 Research Problem

Trust plays a vital role in the recommendations made by AI systems in critical sectors such as healthcare, banking, and criminal justice. A key challenge lies in comprehending the intricate nature of machine learning models. While these models can decipher complex relationships between input variables and outcomes, understanding their underlying processes can be complex [6]. Commonly known as "black box models," algorithms may struggle to meet legal standards [7]. It is widely acknowledged that explainable AI (XAI) is essential for establishing trust in classifier algorithms. Nonetheless, there exist varied theoretical frameworks and approaches in different research studies, XAI effectively elucidates biased and unbalanced datasets [6].

1.3 Research Objectives

The study's main goal is to evaluate the efficacy of the eXplainable Artificial Intelligence (XAI) model known as Local Interpretable Model-agnostic Explanations (LIME) in explaining the results of our experimental trials. It aims to demonstrate the viability of incorporating LIME into the assessment of credit scores.

1.4 Significance of the research

This study is practically significant as it can assist managers in organisations in managing credit verification risk using XAI.Furthermore, the study will assist managers in managing financial risk specifically caused by offering customer's loans which they cannot afford to pay back [7]. The study is theoretically relevant as it furthers knowledge in XAI specifically using LIME in evaluating credit verification models. Furthermore, it seeks to pinpoint any obstacles and constraints that may arise when utilising LIME in credit scoring analysis [6].

2 Literature Review

2.1 Introduction

The literature review's structure is as follows: It commences with an exploration of explainability, interpretability, and understandability. Subsequently, it addresses the classification models utilized in credit verification. Following this, it provides an elucidation of LIME, and it concludes with a comprehensive summary of the chapter.

2.2 Explainability, Interpretability and

Understandability

This section explores the interconnected relationships among explainability, interpretability, and fidelity within the field of machine learning. Despite often being used interchangeably, these terms have nuanced differences. Explainable AI focuses on explaining the reasoning behind decisions rather than delving into the decision-making process itself [5]. Explainability involves the ability to express a machine learning model and its results in a way that is easily understandable to individuals. It requires a comprehensive examination of the logical constructs that underlie the system's decision-making processes. By ensuring that insights from a machine learning model can be effectively communicated using precise and accessible language, explainability plays a crucial role [8].

Interpretable AI provides insight into how decisions are made but may not necessarily provide explanations for the specific criteria selected [9]. Interpretability allows for understanding the results of learning models by revealing the rationale behind their decisions [10]. Interpretable systems are considered explainable when humans can comprehend their processes, highlighting the close relationship between explainability and interpretability [8]. Interpretability and fidelity are fundamental aspects of explainability [11].

A comprehensive explanation should be human-comprehensible (interpretability) and accurately represent the model's behaviour across the entire feature space (fidelity). Interpretability handles the social aspect of explainability, while fidelity aids in confirming other model requirements or uncovering new explanations [5]. Simply put, the fidelity of an explanation refers to how accurately and reliably the model's behaviour is explained. An explanation can be deemed explainable if it is easily understood by humans and effectively explains the model [5]. The primary objective of eXplainable AI (XAI) is to enhance the interpretability of deep learning and machine learning models [12].

2.3 Extreme Gradient Boosting (XGBoost)

XGBoost, a sophisticated machine learning algorithm utilized for enhancing credit score evaluation, has become well-known for its ability to provide enhanced predictive accuracy while managing extensive and intricate datasets. This gradient boosting-based ensemble learning algorithm has achieved recognition, particularly for its characteristics such as regularization, parallel processing, and decision tree optimization, making it particularly suitable for credit scoring tasks [6].

Operating as a tree ensemble model, XGBoost tackles the limitations of individual trees by consolidating their predictions through a linear combination. This results in a progressive enhancement of predictive capabilities through an iterative error learning process and innovative data assimilation techniques. Noteworthy for its regularization method, XGBoost allows for the adjustment of variable weights, addressing overfitting concerns and refining variable selection, especially in scenarios with numerous dimensions. The ultimate model integrates all trees within the ensemble, offering a comprehensive model outlook [13]

Furthermore, XGBoost provides explainability through three key facets: global explanations, local feature-based explanations, and instance-based explanations. Global explanations present a broad overview of essential model elements, while local feature-based explanations illustrate how a specific attribute influences the model's prediction for a particular scenario. Local instance-based explanations depict potential variations in the model's predictions when a specific instance is altered [14]. In essence, XGBoost is a robust and flexible classification model that is relatively straightforward to grasp. This attribute enables a comprehensive understanding of how the algorithm generates predictions, proving crucial in high-stakes realms like credit scoring [6].

2.4 XAI Models

Complex machine learning models often lead to black-box models, necessitating explanations through either post-hoc, ante-hoc, or instance-based approaches [12]. Post-hoc explanations involve utilizing additional models such as Shapley Additive explanations (SHAP) and Local Interpretable Model-Agnostic Explanations (LIME). These methods, commonly known as eXplainable Artificial Intelligence (XAI) models, are frequently applied to elucidate underlying machine learning credit scoring models. On the other hand, ante-hoc explanation involves inherently interpretable models like Decision Trees. Instance-based explanations rely on specific instances to explicate the behaviour of a black-box model [12]. Explainable Artificial Intelligence (XAI) is a research area within artificial intelligence that aims to enhance the interpretability of machine learning models. Understanding these models is crucial as it enables us to grasp their inner workings, identify the most critical attributes influencing them, and comprehend the rationales behind their predictions [6].

2.5 Local Interpretable Model -Agnostic Explanation (LIME)

The Local Interpretable Model-Agnostic Explanation (LIME) framework is a publicly available resource designed to enhance trust in machine learning models by elucidating their decision-making mechanisms [6]. LIME is structured to concentrate on specific data points, aiming to render models interpretable while remaining model-agnostic. This framework provides valuable insights into model operations, enabling the identification of critical areas within images or highlighting important features. Its key functionalities span image interpretation, text analysis, and evaluation of tabular data [15]. Lime was discovered to have explained credit verification models using XGBoost effectively [6].

2.6 Summary of the Literature Review

The study will focus on using XGBoost as it have been found to provide the best results in credit verification [6]. LIME is also quoted to be able to explain XGBoost models [6].

3 Methodology

3.1 Datasets

The dataset used for the experiment in this study is for credit scoring for borrowers in banks. The dataset can be accessed publicly from Kaggle (Kapturov, 2024). The data was selected as it was the most recent data set at the time of the experiment, which was conducted on May 20th, 2024. It had a usability score of 10 and 4522 rows and 17 columns. There were no missing values in the data. This study's experiment used Python 3, running Microsoft Windows 11.

3.2 Performance Measures

The traditional performance measures in the Area of credit scoring embraced the evaluation measurements of this study. These measurements are average accuracy, Type I and Type

II errors and F1-score. A combination of these measurements, rather than a single measure, is used to measure the predictive performance of the proposed credit scoring model. From the confusion matrix table, the following calculations are defined:

Average Accuracy (ACC) =TP+TN/TP+FN+TN+FP

Type I error = FP/ TN+FP

Type II error = FN/ TP+FN

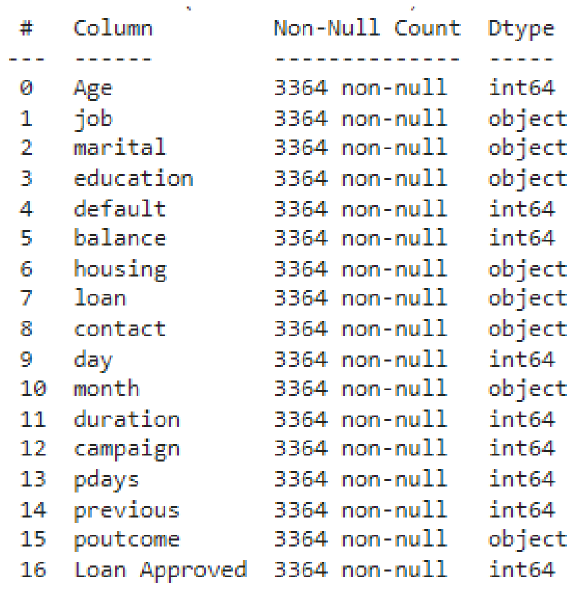
True Positive (TP) stands for a customer who has been approved for a loan and has been correctly classified by the model as a customer with an approved loan. True Negative (TN) stands for a customer who has been disapproved from receiving a loan and has been correctly classified by the model as a customer with a disapproved loan. False Positives (FP) (Type I) stands for a customer who has been incorrectly classified by the model as an approved customer, whereas in reality, they should have been disapproved. False Negatives (FN) (Type II) stands for a customer who has been incorrectly classified by the model as being disapproved for a loan, yet, in fact, should have been approved. Accuracy is calculated as (TP+TN)/ (TP+TN+FP+FN). Recall in the confusion matrix is therefore calculated as TP/(TP+FN). Precision is calculated as TP/(TP+FN) and F1 score is calculated as 2* [(Recall*Precision)/ (Precision+ Recall)].

3.3 Workflow

Data preparation included data exploratory data analysis. There were outliers in the data set. However, all the outliers were assumed to be valid data points; therefore, no outlier imputation was performed. The resilience of XGBoost to missing data and

its ability to handle both categorical and continuous variables make it a powerful tool for this stage. The data was separated in training and testing data sets. The training data set comprised 80% of the data with 3364 rows and 17 columns. The testing dataset comprised of 20% of the original data with 1157 rows and 17 columns. The target variable is whether the loan was approved or disapproved. This column was converted into numerical data 1 representing approved and 0 representing disapproved.

Table 1: Column description for training data



```
 #   Column         Non-Null Count  Dtype
---  ------         --------------  -----
 0   Age            3364 non-null   int64
 1   job            3364 non-null   object
 2   marital        3364 non-null   object
 3   education      3364 non-null   object
 4   default        3364 non-null   int64
 5   balance        3364 non-null   int64
 6   housing        3364 non-null   object
 7   loan           3364 non-null   object
 8   contact        3364 non-null   object
 9   day            3364 non-null   int64
 10  month          3364 non-null   object
 11  duration       3364 non-null   int64
 12  campaign       3364 non-null   int64
 13  pdays          3364 non-null   int64
 14  previous       3364 non-null   int64
 15  poutcome       3364 non-null   object
 16  Loan Approved  3364 non-null   int64
```

During model training, various models will be trained to identify the best one to support the selection of the XGBoost model. The gradient-boosting architecture used by XGBoost allows the algorithm to learn complex patterns and correlations in credit data. Its ability to combine multiple weak models into strong ones leads to the creation of a highly accurate credit-scoring model.

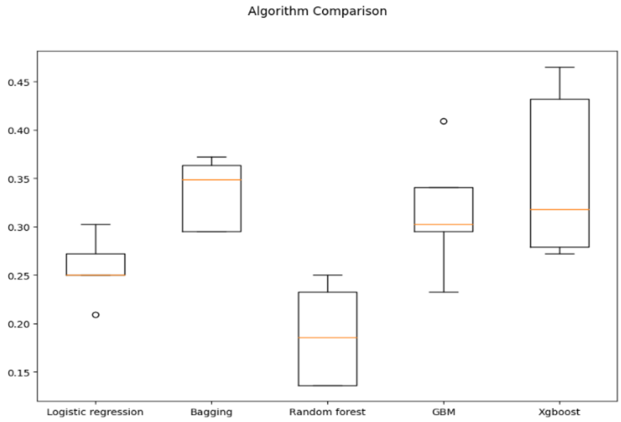


Figure 1: Algorithm comparison

The XGBoost credit scoring model will be evaluated using various performance indicators such as recall in the confusion matrix and accuracy. The model with higher recall is considered better, and in this case, XGBoost has the second-highest recall score. Despite its complexity, the XAI model LIME will simplify the outcomes for easier understanding. Therefore, XGBoost has been chosen for further analysis due to its longer range and the availability of more alternatives.

4 Results and Discussions

4.1 Introduction

The credit verification process initially employed Logistic Regression, Bagging, Random Forest, Gradient Boosting Machine, and Extreme Gradient Boosting models. These models yielded preliminary results indicating the need for data tuning due to imbalance. Subsequently, three models underwent tuning: Gradient Boosting with oversampled data, Adaboost classifier with oversampling, and XGBoost with oversampled data. Oversampling was necessary as the dataset consisted of 90% unapproved loans and 10% approved loans, warranting the need to rectify the data imbalance. The ensuing discussion will focus on the outcomes of the three models, with particular emphasis on XGBoost.

4.2 Interpretation of Findings

The trained models indicate a high accuracy rate, as illustrated in Table B below. The model with the highest accuracy, recall precision, and F1 score is the Adaboost Classifer tuned with oversampled data, followed by the Gradient Boosting tuned oversampled data with XGBoost using oversampled data, which indicates a lower accuracy, precision, and F1 score in the training environment.

Table 2: Training Performance comparison

Performance Matrix	Gradient Boosting tuned with	Adaboost classifier tuned with	XGBoost using oversampled data

	oversampled data	oversampled data	
Accuracy	0.977	0.999	0.954
Recall	0.972	1.000	1.000
Precision	0.981	0.999	0.916
F1	0.977	0.999	0.956

The validation performance comparison indicates a lower accuracy rate for the two models as illustrated in Table C below. Gradient Boosting tuned with oversampled data and Adaboost classifier tuned with oversampled data. Further these models have decreased in all scores recall, precision and F1 score have all decreased to below 50% indicating that these models are not doing well with the validation data. The XGBoost using oversampled data has remained unchanged with an accuracy rate of 95%, a recall of 100%, a precision of 92% and an F1 score of 96%. This indicates that the model holds a high accuracy rate in the validation environment. A contra argument is that the model may be overfitting and would require further investigation to rule out this assumption.

Table 3: Validation performance comparison

Performance Matrix	Gradient Boosting tuned with oversampled data	Adaboost classifier tuned with oversampled data	XGBoost using oversampled data
Accuracy	0.872	0.883	0.954
Recall	0.438	0.479	1.000
Precision	0.416	0.461	0.916
F1	0.427	0.470	0.956

The feature importance indicated in Figure B below indicates that the most important feature is the duration, also known as the term of the loan. The second important feature is the paydays, which is the number of days in which repayments are made and the third most important feature is the whether the customer owns a house.

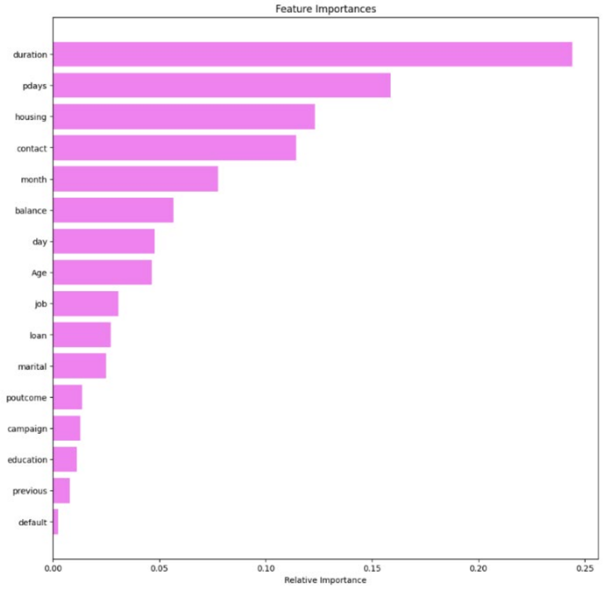


Figure 2: Important feature comparison

4.3 Conclusions and Implications

The objective of this study was to indicate the explainability of credit verification processes using XGBoost with LIME as an explainable AI. The results from XGBoost indicate a higher accuracy in both training and validation models. The first three important features in the model are the duration of the loan, the number of days in which the loan payments were made, and finally, whether the customer is the owner of a house. The LIME model is likely to indicate similar results.

5. Limitations and Future Research

The research used publicly available data, which provided a simulation of the real environment. The research findings would be different if real data had been used, which would provide better insight into how explainable AI models can explain complex models. The research project would further provide different insights if deep learning models were utilised for credit verification purposes. As explainability is defined as humans understanding complex models in human language, an interesting future study would be the interpretation of the results of a credit verification model using natural language processing models.

Acknowledgements

To Prof Matjaz Gams for his unwavering patience and ability to facilitate knowledge transfer.

References

[1] Roush. (2024). *More Than 700 OpenAI Employees Threaten To Quit—And Join Microsoft—Unless Board Resigns*. Forbes. https://www.forbes.com/sites/tylerroush/2023/11/20/more-than-500-openai-employees-threaten-to-quit-over-sam-altmans-removal/?sh=785fa8794ebc

[2] Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J., & Mané, D. (2016). Concrete Problems in AI Safety. 277(2003), 1–21. http://arxiv.org/abs/1606.06565

[3] Kellogg, K. C., Valentine, M. A., & Christin, A. (2020). Algorithms at work: The new contested terrain of control. Academy of Management Annals, 14(1), 366–410. https://doi.org/10.5465/annals.2018.0174

[4] Shin, D. (2021). The effects of explainability and causability on perception, trust, and acceptance: Implications for explainable AI. International Journal of Human Computer Studies, 146(October 2020). https://doi.org/10.1016/j.ijhcs.2020.102551

[5] Bharati et al. (2023). A Review on Explainable Artificial Intelligence Methods, Applications, and Challenges. IEEE Transactions on Engineering Management, 11(4), 1007–1024. https://doi.org/10.52549/ijeei.v11i4.5151

[6] Alblooshi, M., Alhajeri, H., Almatrooshi, M., & Alaraj, M. (2024). Unlocking Transparency in Credit Scoring: Leveraging XGBoost with XAI for Informed Business Decision-Making. International Conference on Artificial Intelligence, Computer, Data Sciences, and Applications, ACDSA 2024, February, 1–6. https://doi.org/10.1109/ACDSA59508.2024.10467573

[7] Salter, R. (2023). Explainable Artificial Intelligence and its Applications in Behavioural Credit Scoring Subtitle if any.

[8] Adadi, A., & Berrada, M. (2018). Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI). IEEE Access, 6, 52138–52160. https://doi.org/10.1109/ACCESS.2018.2870052

[9] Vishwarupe, V., Joshi, P. M., Mathias, N., Maheshwari, S., Mhaisalkar, S., & Pawar, V. (2022). Explainable AI and Interpretable Machine Learning: A Case Study in Perspective. Procedia Computer Science, 204(2021), 869–876

[10] Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. Artificial Intelligence, 267, 1–38. https://doi.org/10.1016/j.artint.2018.07.007

[11] Gilpin, L. H., Bau, D., Yuan, B. Z., Bajwa, A., Specter, M., & Kagal, L. (2018). Explaining explanations: An overview of interpretability of machine learning. Proceedings - 2018 IEEE 5th International Conference on Data Science and Advanced Analytics, DSAA 2018, 80–89. https://doi.org/10.1109/DSAA.2018.00018

[12] Dastile, X., Celik, T., & Vandierendonck, H. (2022). Model-Agnostic Counterfactual Explanations in Credit Scoring. IEEE Access, 10(April), 69543–69554. https://doi.org/10.1109/ACCESS.2022.3177783

[13] Carmona, P., Dwekat, A., & Mardawi, Z. (2022). No more black boxes! Explaining the predictions of a machine learning XGBoost classifier algorithm in business failure. Research in International Business and Finance, 61, 101649.

[14] Demajo, L. M., Vella, V., & Dingli, A. (2021). An explanation framework for interpretable credit scoring. International Journal of Artificial

Intelligence and Applications (IJAIA), 12(1).

[15] Dieber, J., & Kirrane, S. (2020). Why model why? Assessing the strengths and limitations of LIME. arXiv preprint arXiv:2012.00093.