# Comparative Analysis of Interpretability and Accuracy between Gradient Boosting Machine and Explainable Boosting Machine on Default Credit domain

MSc Research Project
Fintech

Patricio Garay
Student ID: 21221154

School of Computing
National College of Ireland

Supervisor: Noel Cosgrave

**National College of Ireland**

**MSc Project Submission Sheet**

**School of Computing**

**Student Name:** …Patricio Andrés Garay Pacheco……………………………………………………………

**Student ID:** …21221154……………………………………………………………………………………………

**Programme:**…MSc in Fintech….……………………………… **Year:** …2023…………………

**Module:** …Research Project (MSCFTD1)…………………………………………………………………

**Supervisor:** … Noel Cosgrave……………………………………………………………………………….………

**Submission Due Date:** …13/08/23………………………………………………………………………………………….………

**Project Title:** … Comparative Analysis of Interpretability and Accuracy between Gradient Bosting Machine and Explainable Boosting Machine on Default Credit domain…………………………………………………………………………………………………….…

**Word Count:** …9008……………………….… **Page Count**…23……………………………………….…….…..

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

**Signature:**… Garay ………………………………………………………………………………………………

**Date:** …14/08/23……………………………………………………………………………………………………

**PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST**

| | |
|---|---|
| Attach a completed copy of this sheet to each project (including multiple copies) | □YES |
| **Attach a Moodle submission receipt of the online project submission,** to each project (including multiple copies). | □YES |
| **You must ensure that you retain a HARD COPY of the project**, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer. | □YES |

Assignments that are submitted to the Programme Coordinator Office must be placed into the assignment box located outside the office.

| **Office Use Only** | |
|---|---|
| Signature: | |
| Date: | |
| Penalty Applied (if applicable): | |

# Title Comparative Analysis of Interpretability and Accuracy between Gradient Bosting Machine and Explainable Boosting Machine on Default Credit domain

Patricio Garay
Student ID: 21221154

**Abstract**

Machine learning (ML) algorithms have gained ground in credit default modelling due to their great predictive power. Gradient Boosted Machines (GBM) have excellent prediction performance. However, their lack of interpretability challenges their use in the finance sector. This study investigates to what extent GBM can be made explainable in the context of credit default modelling. To assess GBMs, this study contrasts GBM with Explainable Boosted Machines (EBM), a glass box model. The evaluation examines accuracy, precision, recall, F1 score, AUC-ROC, and confusion matrix. The accuracy of both models was around 82%; the other performance metrics were similar.

LIME compared local interpretability, showing that GBM and EBM had a Default prediction of around 0.72. Also, LIME illustrated that PAY_0 (September payment records) contributes most to these predictions, followed by PAY_6 (April payment records) and PAY_2 (July payment records). However, LIME does not explain how GBM decides to reach the prediction. GBM can be used in the credit default domain with an acceptable accuracy level. Using the LIME technique helps to know what variables are relevant in the prediction, but understanding how GBM makes decisions through post-hoc method of local interpretability is not helpful. Future studies should investigate global interpretability techniques or other local interpretability methods. Further research into the relationship between outside variables like the Mid-Autumn Festival and the likelihood of credit default could yield insightful results. The study of the interpretability of GBMs in the credit default domain has relevant implications for financial institutions and regulators.

# 1 Introduction

## 1.1. Background on the research topic

### 1.1.1. The integration of ML algorithms in our society.

Since the development of computers and the internet, humanity has experienced exponential technological development, known as the digitisation process. This digitisation process has been integrated into different industries in the last decade. Many companies have adopted ML and artificial intelligence (AI) tools within their operations to improve their efficiency and obtain better results (Sakri, 2022). Some examples of algorithms created from ML and AI that we see present day by day for most people can be seen in product recommendations (Amazon), movie or series suggestions (Netflix), music recommendations (Spotify), social media friend suggestions (Facebook, Instagram). More examples of using these technologies are sectors such as finance (mortgage approval models, risk models), military, and legal

(Adadi and Berrada, 2018). An example of the healthcare sector could be an ML model used to provide a diagnosis to help doctors to choose a suitable treatment based on the patient's symptoms (Konstantinov and Utkin, 2021).

### 1.1.2. ML algorithms in the financial sector

The financial sector is highly competitive, made up of traditional banking and the emerging fintech sector, which has modernized traditional banking with the use of new technologies. Many financial companies, and banks have incorporated ML to create credit scoring models to improve their risk assessment for loan applicants. These algorithms are based on ML, which minimizes the risks of default by predicting default. The default prediction is a crucial activity to reduce the risks for these companies of getting loose when they lend money (Liu, Fan and Xia, 2022; Frost *et al*., 2020). Koh and Chan (2002) illustrate a range of potential data mining applications in the banking industry such as fraud detection, market basket analysis, segment customers. Fintech has created new opportunities for groups of people who previously could not be part of the financial system, because they did not qualify under the standards of traditional banking (Sakri, 2022; Frost *et al*., 2020).

### 1.1.3. Challenges in credit risk prediction and default risks

The client's default on credit obligations has multiple impacts, affecting their economic well-being, the credit institutions' stability, and the domestic and global economy. Local financial problems, such as those experienced in the United States, have the potential to trigger global economic instability, as seen during the global financial crisis in 2007 (Adadi and Berrada, 2018; Cecchetti and Schoenholtz, 2015), where credit risk management played a fundamental role. That is why financial default represents a significant challenge for financial institutions and governments. Today with the advent of big data, the traditional methods used by financial institutions to perform credit risk analysis have lost their effectiveness due to the large volume of information available for analysis. This has led to the search for new ML algorithms capable of handling big data more accurately and efficiently (Liu *et al*., 2022).

### 1.1.4. The need for interpretability in ML

Our society has increased its reliance on ML algorithms and the constant creation of ML algorithms have generated multiple challenges. One of these challenges is the ability to interpret the operation and results of the black box algorithms (Sakri, 2022; Adadi and Berrada, 2018). Both traditional financial institutions and emerging fintech companies can benefit from properly implementing credit scoring algorithms (Sakri, 2022).

ML algorithms are closely related to AI. Explainable Artificial Intelligence (XAI) has emerged as a field of study to develop interpretability in ML and provide more transparent AI. Increasing the interpretability of these algorithms has become a challenge for academics, government entities, and multinational companies (Carvalho, Pereira and Cardoso, 2019; Adadi and Berrada, 2018). Recent studies have shown that interpretability should not be sacrificed for better performance. Utilizing models considered black boxes, like EBM, can offer a good level of performance and interpretability at the same time (Liu and Sun, 2023; Jayasundara, Indika and Herath, 2022).

### 1.1.5. Interpretability

Interpretability is a crucial feature for ML, an acceptable level of interpretation facilitates the digital transformation process to more areas of our society. Interpretability facilitates algorithm comparisons, identification and improvement of unreliable models, social acceptance of machine learning, and model insights (Adadi and Berrada, 2018). Evaluating machine learning models in credit risk assessment requires careful consideration of several performance metrics, such as accuracy, misclassification rate, sensitivity, and specificity (Carvalho *et al.*, 2019). Baesens *et al.* (2003) say that the pre-processing techniques and selection of the ML algorithm are fundamental to getting the desired trade-offs between predictive accuracy and interpretability.

## 1.2. Justification for the research

The use of ML must be careful since the algorithms can present several problems and provide incorrect information that could seriously impact the area where they are used, such as the health area (Rudin, 2019). An appropriate level of interpretation helps to detect problems; once the problem is detected, it can be solved. One of the problems with algorithms is bias, which could occur when the model is trained using partial data, producing a biased result. The previous case could cause discrimination against some minority groups (Honegger, 2018; Varian, 2014). According to Rudin (2019), the use of black boxes should discourage, and policymakers should demand the utilization of ML models with a proper level of interpretability to prevent poor decisions in a wide range of sectors and industries.

Financial services require high interpretability level due to the robust government regulations that must be met for their operation (Rudin, 2019). An example of how a low level of interpretability can affect the financial sector is the case of a bank that receives a loan application from a client. This request is evaluated and rejected by an ML algorithm, which has a high level of accuracy but low interpretability. After the rejection, the client turns to the bank to find out the reason for the rejection. Because the algorithm is considered a black box, bank analysts cannot give such an explanation. A real-life example can be seen in the work of Varian (2014) based on the analysis of the data of Home Mortgage Disclosure Act (HMDA) (Munnell *et al.*, 1996).

The need for interpretability is due to "the problem is that a single metric, such as classification accuracy, is an incomplete description of most real-world tasks." In other words, to fully understand the data set, it is necessary to have good interpretability (Doshi-Velez and Kim, 2017). Even though the excellent performance of these models is considered black boxes, there is still work to be done and studies to be carried out. Sakri (2022) suggests the need to consider both the performance and the interpretability of the algorithms in credit risk assessment.

The objective of this research is to explore the local interpretability of GBM comparing GBM and EBM in terms of performance and interpretability.

## 1.3. What gap in the literature it seeks to fill

This research project seeks to explore the interpretability and accuracy of GBM model in a credit default domain through a comparison between GBM and EBM. GBM is considered a black box and EBM is considered a white box. Most ML research has been focused on developing new methods rather than improving existing ones or comparing and evaluating the interpretability properties of explanation methods (Carvalho *et al*., 2019; Rudin, 2019). According to Liu *et al*. (2022), there is a lack of exploration on interpretability of credit score model, this creates an opportunity to compare the GBM and EBM algorithms and determine the strengths and weaknesses of each one.

### 1.4. The main research question.

The research question that this work tries to answer is "To what extent can Gradient Boosted Machines be made explainable in the context of credit default modelling?".

### 1.5. Paper's structure

This research project is organized as follows. Section 1 introduces the background on the research topic, such as the integration of ML algorithms in our society, the finance sector, ML challenges, the concept of interpretability in ML, the research project and the justification project. Section 2 describes the literature review about ML techniques in credit default and interpretability in ML. Section 3 presents the research methodology applied, exploring the data set and a deeper understanding of local interpretability, the LIME method, GBM and EBM ML models, the methodology approach, and the research project limitations. Finally, section 4 concludes the study with the conclusion and discussion of this research project.

## 2 Literature Review

### 2.1 ML techniques in credit default

In recent years, ML techniques have gained significant popularity in credit risk assessment, where they have been widely used to predict the probability of customer default. The most widely used models in the context of credit default are Decision Trees (DT) and Logistic Regression (LR) due to their high level of interpretability (Yu, 2020).

DT offer a variety of approaches to analyse customer behaviour in credit risk assessment. This model type stands out for providing a graphical representation of the algorithm's decision-making process, making it understandable even for people without technical knowledge of ML (Koh and Chan, 2002). Similarly, LR has received much attention due to its prominent features of easy implementation, good interpretability, and handling imbalanced datasets (Sohn, Kim and Yoon, 2016).

Research by Sakri (2022) compares GBM and Deep Neural Network (DNN) algorithms using the Receiver Operating Characteristic of the Area Under the Curve (ROC AUC) to assess the accuracy of the predictive model. Although GBM and DNN are considered state-of-the-art algorithms for solving classification problems, the findings showed that GBM outperforms DNN regarding prediction accuracy. Another study applied GBM to a P2P dataset to predict the risk of peer-to-peer (P2P) credit loans coming due, yielding excellent accuracy, precision, recovery, and F1 score (Zhu and Chen, 2021).

The performance of different machine learning algorithms in credit risk assessment is still a matter of debate. Alam *et al*. (2020) conducted a study indicating that the artificial neural network (ANN) algorithm obtained the best performance compared with other ML algorithms. However, that study recognized that other algorithms could excel on different metrics. On the other hand, Srinath and Gururaja (2022) found that the random forest algorithm outperformed LR and DT algorithms performed better in predicting credit default. When looking for a higher level of precision, random forest models and a group of boosting ensemble algorithms known as black boxes offer excellent performance results. Despite their excellent results, boosting ensemble algorithms face a trade-off between performance and interpretability (Liu *et al.*, 2022).

GBM is considered one of the best algorithms in precision performance and versatility because the original model can be modified, improving the model's precision at each stage through error correction (Sakri, 2022). Liu *et al.* (2022) address the limitations of traditional Gradient Boosted Decision Tree (GBDT) algorithms in terms of the lack of diversity in individual classifiers and interpretability. This study proposes two tree-based augmented GBDTs (AugBoost-RFS and AugBoost-RFU) for credit scoring. The findings were that AugBoost-RFS/AugBoost-RFU outperformed GBDT in offering intrinsic global interpretability. GBM with randomized DT as a basic model can perform better than NN and interpretability than linear models (Kovalev, Utkin and Kasimov, 2020). Zhou and Feng (2017) explore a modification of GBMs called gcForest (multi-Grained Cas- cade Forest), which is a new type of DT ensemble method. Konstantinov and Utkin (2021) claim that GBMs can be applied to local and global interpretation problems.

In the need to improve the interpretation of the algorithms known as black boxes, EBM arises, which promises to provide excellent performance results, such as both global and local interpretability, without the need for methods (Liu and Sun, 2023). EBM has been applied in various domains, including environmental assessment (Liu, Shi, *et al.*, 2022), civil engineering (Liu and Sun, 2023), medical care (Sarica, Quattrone and Quattrone, 2022) and education (Jayasundara *et al.*, 2022).

## 2.2 Interpretability in ML

An investigation by Yang and Wang (2021) is focused on comparing the performance and interpreting the decision-making process between several machine learning algorithms that are used, such as GBDT, LR, DT, K-Nearest Neighbor Classifier (KNN), Multilayer Perceptron Classifier (MRF), SVM, and a new stacking classifier method. The study concludes that GBDT had the best performance. Also, it explores the visual interpretability of GBDT showing the relevance of each variable for the model.

Blakely and Granmo (2021) present an alternative method based on Tsetlin Machines (TMs). This study used different data sets to obtain an acceptable level of global and local interpretability compared to Shapley Additive Explanations (SHAP), obtaining similar interpretation results. Additionally, TMs had competitive precision performance against other black ML models such as XGBoost, explainable booster machines, and neural additive models. Huber *et al.* (2021) and Konstantinov and Utkin (2021) carried out studies focused on global and local interpretability. These studies state that there are few studies on global interpretability, which is hard to explain when the model is complex because of the trade-off between flexibility and interpretability, in addition, this type of explication often cannot explain the whole model (Ribeiro, Singh and Guestrin, 2016a). In contrast a local interpretability explanation would be easier to achieve Karlsson *et al.* (2020) explored the interpretability in time series, concluding that the explainable local algorithm that uses the forest unexpectedly outperforms the global KNN solution in terms of cost and compactness.

Improving interpretability in ML models is a challenge researchers and practitioners are actively addressing (Adadi and Berrada, 2018). Several techniques exist to interpret the black box models locally; some previous works that involve them will be reviewed below. One of the techniques that has gained significant relevance around local interpretability is the Model-Agnostic Local Interpretable Explanations (LIME). LIME explains the model's predictions locally without being able to model the decision-making process, LIME uses simple and easily understandable linear models to approximate the predictions of black box models locally (Ribeiro *et al*., 2016a; Ribeiro, Singh and Guestrin, 2016b); there are also many other modifications of the original LIME (Konstantinov and Utkin, 2021). Petsiuk, Das and Saenko (2018) compared LIME to other methods for estimating pixel salience using DNN, obtaining a good approximation locally, but not better than the proposed Random Input Sampling method for Explanation (RISE) that generates an importance map indicating how prominent each pixel is for the model's prediction.

Ribeiro *et al*. (2016a) have used SP-LIME to explain the predictions of any classifier of the RF and NN models in credit risk assessment. Shankaranarayana and Runje (2019) state that LIME has two problems; the explanation of the instability caused by the randomly generated data set, but if it is solved by generating a large set of point to train a local surrogate model, it will decrease the local fidelity. They proposed a successful solution to deal with this trade-off problem using an autoencoder-based approach for local interpretability (ALIME). According to Zafar and Khan (2019), the use of model-agnostic interpretable local deterministic explanations (DLIME) and hierarchical clustering (GC) can address the lack of "stability" present in the traditional LIME method, which is produced by generating data using random disturbances. Ribeiro, Singh and Guestrin (2018) proposed a novel model-independent system called Anchor LIME to explain the black box model of different dataset domains. Anchor LIME improved over the traditional LIME approach regarding interpretability, making this process more intuitive and easier to understand. Another variation of LIME is SurvLIME, which is an extension of the traditional LIME. In that research SurvLIME was used to work with survival models (Kovalev *et al*., 2020).

LIME's flexibility allows to analyse tabular, image, time series and text data. The use of the inductive logic programming system Aleph and LIME together called LIME-Aleph to analyse images by extracting symbol rules from the images. Visual explanations provide information about which parts of an image are relevant to a classifier's decision (Rabold *et al*., 2019). LIME has also served as a benchmark for comparing new interpretability methods. Ross, Hughes, and Doshi-Velez (2017) developed a method called "input gradients" that was compared against different types of data sets from different domains, from RGB images to Ines-cancer. The results obtained by the input gradients were contrasted with various versions of LIME depending on the data set used. After the study it was concluded that the new method is consistent with LIME, faster and even superior when working with continuous data.

Some interpretation research has looked at the implementation of SHAP, which assigns each feature an importance value for a particular prediction. Lundberg and Lee (2017) unified six existing methods to develop a new version that improved performance and interpretability. However, this model can still be improved by making fewer assumptions for faster estimation. Strumbelj and Kononenko (2010) developed a method based on coalition game theory, which explains the predictions of individual feature values, the method proposed makes the explanation intuitive for classification models.

Some other research focused on perturbation methods in ML image classification domain. Fong and Vedaldi (2017) propose a general framework for learning explanation for any black box algorithms, their contribution is specialized imagine classifier decision. This approach provides insight into the model's decision-making process and improves transparency, allowing users to understand and trust the model's predictions. Vu *et al.* (2019) proposed a framework called c-Eval to evaluate different interpretability methods in the context of machine learning image classifiers.

## 2.3 Conclusion Literature Review

The literature review gives us an overview of the main ML algorithms used in credit default, where DT and LR stand out due to their high level of interpretability. Many investigations consider metrics such as accuracy, precision, recall, f1-score and AUC-ROC for performance evaluation between ML models. In the reviewed papers, there is no consensus on the best model for credit default analysis; however, it is agreed that some models outperform others in specific metrics.

The models with a better performance level are known as black boxes. However, these models have a lower interpretability capacity than those considered glass boxes, such as DT, LR and EBM. GBM is considered an excellent black box model with a remarkable ability to adapt to the required problems. EBM is a glass model that promises to have a high level of performance together with a high level of global and local interpretability, without the need for additional interpretability methods.

Regarding interpretability, there has yet to be a consensus on its definition. However, most studies have a similar definition. Interpretability can be global or local. Some local interpretability methods were reviewed, such as LIME, which has many variations. LIME is a model that can be used as a benchmark for comparing the interpretability of different models. Another widely used interpretability method is SHAP.

# 3 Research Methodology

The modernization of the financial sector and the beginning of the Fintech industry, together with other types of technologies such as blockchain and IA that promise improvements in efficiency and productivity, have awakened the interest to apply ML algorithms to get better performance. As mentioned in the previous chapters, companies and financial institutions have opted for glass boxes due to the high interpretability standards that the financial sector is subjected to. However, the high level of precision of the algorithms considered as black boxes has motivated the study of interpretability characteristics to favour and promote the use of these algorithms in areas that need a high explanatory level of the decision-making process on the part of algorithms, such as credit card default analysis, crime prevention, support in medical diagnoses. This study intends to evaluate the interpretability of GBM (blass box) and EBM (glass box) in credit card default domain.

## 3.1. Local Interpretability

Although there is no consensus on the measurement of interpretability, most definitions revolve around the ability of humans to understand the decision-making process of algorithms (Molnar, 2022; Doshi-Velez and Kim, 2017; Bibal and Frénay, 2016). Some notions of the objectives sought by interpretability include precision, understandability, and efficiency (Carvalho *et al.*, 2019). Blakely and Granmo (2021) state that the ML literature says that a model is fully interpretable only if it is possible to understand how the underlying data characteristics affect the model's predictions, both from the whole model (global) perspective and at the individual sample level (local). The local ones try to explain a black box model locally around a test example, while the global methods derive interpretations from all or part of the data set (Konstantinov and Utkin, 2021; Kovalev *et al.*, 2020). Blakely and Granmo, (2021) used SHAP method to explain global interpretability. Furthermore, interpretability can be grouped into (i) model-agnostic (ii) intrinsic or post-hoc, and (iii) perturbation- or salience-based (Molnar, 2022; Shankaranarayana and Runje, 2019).

There are many methods for achieving local interpretability, Carvalho *et al.* (2019) applied a solution for a black box ML model using a surrogate model, which simplified the model through an approximation of the behaviour of the black box ML model. Other well-known methods are Feature Importance (Rajbahadur *et al.*, 2022), Rule Extraction (Chakraborty, Biswas, and Purkayastha, 2022), Attention Mechanisms (Mu *et al.*, 2023), and Influence Functions (Koh and Liang, 2017). Also, post-hoc methods explain the decisions made by the model once it has been built and used to make predictions (Honegger, 2018). These models do not show how the model has made the decisions; they focus on the subsequent interpretation of the results, applying to all types of glass and black box models without affecting their predictive power. One popular post-hoc approach to interpretation is approximating a black-box model with a linear model. Methods such as SHAP and LIME build a linear model around the instance to be explained (Liu *et al.*, 2022; Konstantinov and Utkin, 2021; Ribeiro *et al.*, 2016a).

Consequently, the evaluation of the local interpretability of the models was conducted using the LIME method proposed by Ribeiro *et al.* (2016a). This method has proven effective for interpreting tabular and image models. However, when assessing global interpretability, utilizing an agnostic model presents challenges due to the trade-off between flexibility and interpretability. Typically, the quality of global interpretation falls short as these models tend to oversimplify the complexity of the entire model (Ribeiro *et al.*, 2016b). In contrast, the EBM model offers a direct global interpretation. For this study, the comparison with the GBM model will not be included, as the focus lies solely on comparing the local interpretability of both models. According to Shankaranarayana and Runje (2019), there is not a suitable metric to compare two different interpretable models. Performing an interpretability comparison between different ML models as a framework can be challenging, particularly when comparing a black box model like GBM and a glass box model like EBM. However, an agnostic model (Nori *et al.*, 2019; Ribeiro *et al.*, 2016b) can be employed to make the comparison using the same criteria without impacting the operation or performance of the models due to LIME is considered a post-hoc method.

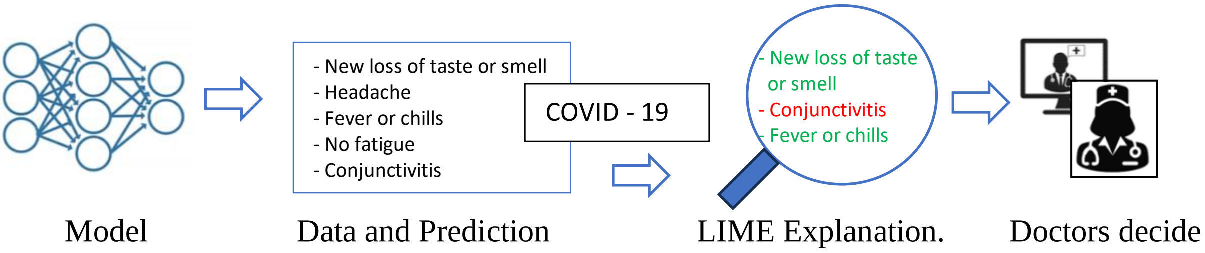
## 3.2. Brief Explanation of LIME

We can distinguish between two sub-types: Local interpretability for a single prediction, and local interpretability for a group of predictions (Molnar, 2022). According to Zafar and Khan (2019), LIME is used to increase the interpretability of black box ML algorithms, by generating an explanation for a single prediction made by any ML model by learning a simpler interpretable model (for example, a linear classifier) around the prediction and obtaining feature importance by using some type of feature selection. LIME can be used to interpretate tabular, images, time series and text data sets.

**L (Local):** Avoid making generalisations and concentrate on specific predictions or incidents.
**I (Interpretable):** Helps humans understand sophisticated machine learning models.
**M (Model-Agnostic):** It can be used with both glass boxes and black boxes in ML models.
**E (Explanations):** Produces a local-surrogate model that simulates the complex model's behaviour close to a particular instance.

LIME uses simple and easily understandable linear models to locally approximate the predictions of black-box models (Kovalev *et al.*, 2020). Figure 1 shows an explanation of how LIME helps humans to understand the models.

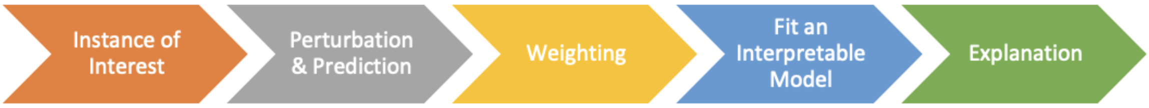
**Figure 1:** Explaining individual predictions (Ribeiro *et al.*, 2016a)



LIME highlights the symptoms in the patient's past that contributed to the model's prediction that the patient had COVID - 19. New loss of taste or smell and fever or chills are cited as factors supporting the "COVID - 19" hypothesis, whereas "Conjunctivitis" is cited as supporting the opposite hypothesis. Finally, a doctor can use these to decide whether to accept the model's forecast.

Because LIME is a model-agnostic model, it can be utilised to give local interpretability in GBM and EBM. Figure 2 represents the LIME's process.

**Figure 2:** LIME's process (Own elaboration)



Firstly, the user chooses an instance for which a black boxes model (GBMs) prediction explanation is required. Secondly, LIME generates new artificial data points by slightly perturbing the provided training dataset, then extracts the relevant predictions from the black box model used. In this step, LIME investigates what happens to the ML model's predictions when different versions of the training data are fed to it. Then, the new synthetic data points are weighted according to how close they are to our instance of interest (the closer they are, the higher the weight assigned to them). Later, fit a new interpretable model, such as a decision tree or a linear regression, using the newly constructed and weighted artificial dataset. Last, justify the relevant case using the recently trained interpretable model (Honegger, 2018).

Some advantages of LIME are that this method could be applied to glass and black boxes models, also it can be used to compared two models with the same benchmarks such as black box and glass box models (Zafar and Khan, 2019; Ribeiro *et al.*, 2016a). LIME has certain downsides despite its advantages. The first downside is that there is a problem with the *instability* of generated explanations. Repeating the LIME's random sampling process can result in different outcomes, which makes the output unpredictable. The accuracy of the estimation may also be impacted by LIME's use of simpler linear models to approximate the complex. Health experts have rejected its use in Computer-Aided Diagnosis (CAD) systems due to this drawback (Zafar, and Khan, 2019). The lack of stability is made worse by LIME's non-deterministic character, which is a desirable trait for an interpretable model. This instability is caused in part by the process of randomly perturbing spots. Finally, the computational complexity of LIME is another drawback, particularly when dealing with high-dimensional perturbed input instances (Kovalev *et al.*, 2020; Shankaranarayana, and Runje, 2019).

## 3.3. Dataset

The selection of the data set was on the credit default domain, which obtained on the machine learning repository https://archive.ics.uci.edu/ml/datasets/default+of+credit+card+clients. This dataset corresponds to the banking information of a Taiwan bank between the months of April and September 2005, which allows to visualize the clients that had credit problems in the payment of their financial obligations.

The dataset has 23 columns and 30,000 rows, of which 6636 (22.12%) were classified as default. The features of the columns or attributes are presented below. This dataset was selected because its characteristics meet the information volume requirements to use machine learning algorithms, it has a variable that can be used as a dependent (Y) named *Default payment next month*, and additionally, it has both categorical and continuous variables, which will be explained in more detail below:

**Default payment next moth:** Yes = 1; No = 0
**LIMIT_BAL (Amount of the given credit, NT dollar):** it includes both the individual consumer credit and his/her family (supplementary) credit.
**SEX (Gender):** 1 = male; 2 = female.
**EDUCATION (Education Level):** 1 = graduate school; 2 = university; 3 = high school; 4 = others.
**MARRIAGE (Marital status):** 1 = married; 2 = single; 3 = others.
**AGE (Age):** year.
**PAY_0 – PAY_6 (History of past payment):** Monthly payment records (PAY_0 = Sep; PAY_1 = Aug; …PAY_6 = Apr). The measurement scale for the repayment status is -1 = pay duty; 1 = payment delay for one month; 2 = payment delay for two months; ... 9 = payment delay for nine months and above.
**BILL_AMT1 - BILL_AMT6 (Amount of bill statement, NT dollar):** Amount of bill statement (BILL_AMT1= Sep; BILL_AMT2 = Aug; … BILL_AMT6 = Apr).
**PAY_AMT1 - PAY_AMT6 (Amount of previous payment, NT dollar):** Amount paid (PAY_AMT1 = Sep; PAY_AMT2 = Aug; … PAY_AMT6 = Apr).

This data set was subjected to data-cleaning techniques such as analysing missing values, outliers, and duplicates. However, the dataset had been cleaned already. Then, the dataset was randomly split with a static seed into training and test sets with a percentage of 80% and 20%, respectively.

## 3.4. Gradient Boosted Machines (GBM)

GBMs are an effective ensemble learning method that combines weak learners, usually decision trees, to produce a robust predictive model. In tasks involving credit default modelling, GBMs have demonstrated great predicted accuracy (Zhu *et al.*, 2023). According to Fernandez-Delgado *et al*. (2014), they can successfully capture complicated correlations and nonlinearities in credit data and produce reliable estimates of default risk. Also, GBMs offer metrics of feature relevance, enabling analysts to pinpoint the factors that have the greatest impact on the forecast of loan default. The drivers of default behaviour can be better understood thanks to this feature important study (Chen and Guestrin, 2016). However, the inherent lack of interpretability in GBMs poses several challenges, such as Complex Model Structure (Lundberg and Lee, 2017), Black Box Nature (Rudin, 2019), and Feature Importance Interpretation (Hooker, 2004).

## 3.5. Explainable Boosted Models (EBM)

EBM is a glass box ML algorithm, that has demonstrated its ability to handle this trade-off. (Nori *et al*., 2019). EBM is a generalised additive model (GAM) with the form showed in Figure 3.

**Figure 3:** GAM’s form of the EBM

$$g(E[y]) = \beta_0 + \sum f_j(x_j)$$

Where *g* is the link function, which allows the GAM to be adjusted for various situations like regression or classification.

EBM provides a few significant advantages over conventional GAMs. Firstly, EBM employs state-of-the-art machine learning methods such as bagging and gradient boosting to train each feature function fj. The boost method is severely restricted to training on a single function at a time in a turn-based fashion with a low learning rate, making the order of the functions irrelevant. To reduce the impacts of collinearity and train the best feature function for each feature, round-robin loops iterate through the features. Secondly, EBM can automatically detect and include pairwise interaction terms with the form in Figure 4 increasing accuracy and maintaining intelligibility (Nori *et al*., 2019).

**Figure 4:** GAM’s form of the EBM with pairwise interaction

$$g(E[y]) = \beta_0 + \sum f_j(x_j) + \sum f_{i_j}(x_i, x_j)$$

Because each feature's contribution to a final prediction can be seen and understood by charting fj, EBMs are very understandable. Due to the additive nature of EBM, each feature contributes to forecasts in a modular manner that facilitates understanding the role that each characteristic plays in the prediction.

EBM was selected to carry on this research because it is considered a glass box ML algorithm inside the GBM, which still offers many possibilities to study and get new knowledge to the interpretability for credit default domain (Jayasundara *et al*., 2022). EBM in contrast with the classic GBM has interpretability features that allow both global and local interpretability to be obtained without the need to sacrifice performance in its prediction. Due to its prominent level of interpretability, EBM does not require a post-hoc technique like LIME or SHAP.

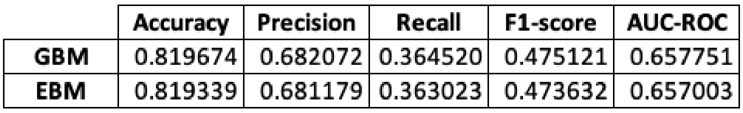
## 3.6. Methodological Approach

To carry out this Project, Google Colab was used together with the Python programming language after loading the data set and performing the data cleaning procedures mentioned in the previous chapter. First, the GBM model was applied, together with evaluation metrics, confusion matrix and the LIME method for local interpretability analysis. The same procedure was then repeated for the EBM model. The previously described procedures will be explained in greater detail below.

### 3.6.1. Evaluation Metrics

The performance metrics for GBM and EBM models are displayed in Table 1. After that, evaluate the performance using several assessment measures like accuracy, precision, recall, F1 score, and AUC-ROC. Due to their capacity to offer thorough insights into several facets of model performance and interpretability, these metrics are important for comparing GBM with EBM (Liu *et al.*, 2022).

**Table 1:** Performance Metrics (Own elaboration)



| | Accuracy | Precision | Recall | F1-score | AUC-ROC |
|---|---|---|---|---|---|
| GBM | 0.819674 | 0.682072 | 0.364520 | 0.475121 | 0.657751 |
| EBM | 0.819339 | 0.681179 | 0.363023 | 0.473632 | 0.657003 |

Both the GBM and EBM models had an elevated level of accuracy, with the GBM obtaining an accuracy of 0.81967 and the EBM's accuracy of 0.81933. While the GBM had a precision of 0.682072, the EBM had a precision of 0.681179, indicating that approximately 68% of the predictions made by each model were correct based on the test data.

Analysing the recall for the GBM was greater (0.364520) than for the EBM (0.363023), which suggests a lower rate of false negatives (FN). The F1-score for the GBM and the EBM, which both provided balanced measures of performance, were 0.475121 and 0.473632, respectively. Additionally, the GBM showed a slightly higher AUC-ROC value of 0.657751 compared to the EBM's AUC-ROC value of 0.657003, demonstrating the GBM's superior discrimination performance (Liu *et al.*, 2022). In conclusion, both models have similar performance in the terms assessed.

### 3.6.2. Confusion Matrix

To gain a deeper understanding of the model's performance confusion matrixes were evaluated for both models (see Figure 5). The confusion matrix provides insights into the distribution of true positive (TP), true negative (TN), false positive (FP), and false negative (FN) predictions.

**Figure 5:** Left: GBM’s Confusion Matrix.
Right: EBM’s Confusion Matrix.

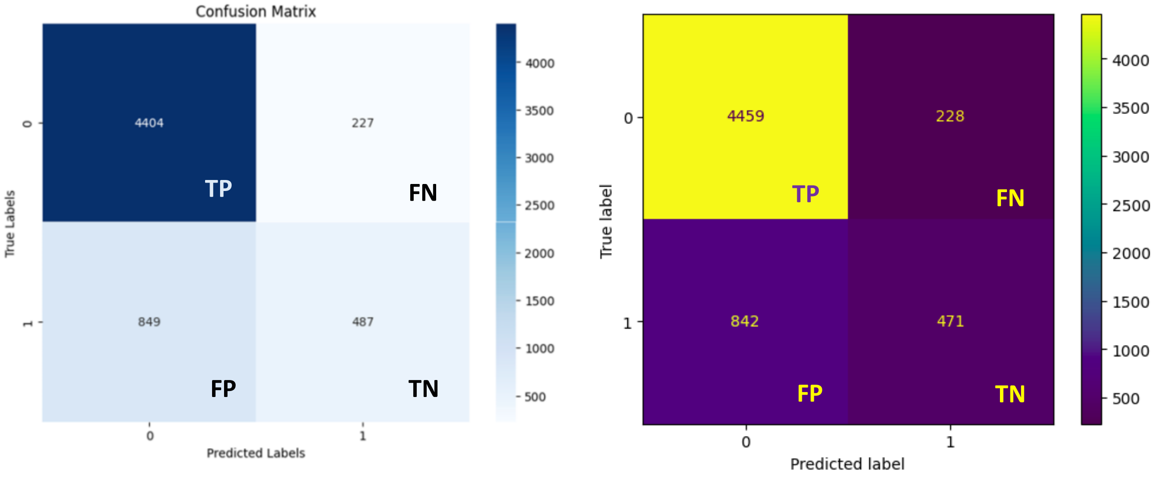
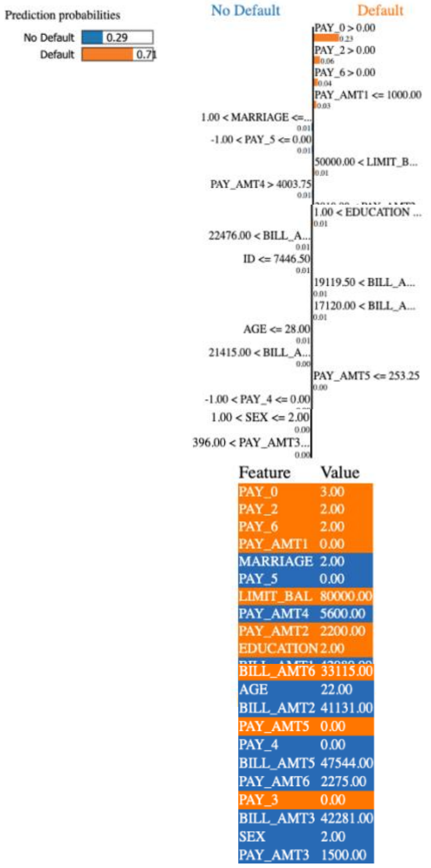


Figure 5 shows that their confusion matrices have very minor variations. Both models indicate comparable TP and FN counts, demonstrating their efficacy in detecting *Default payment*. The GBM model, in contrast to the EBM model, predicts Default payment next month more frequently when it is not the case, according to slightly higher FP counts. Additionally, the GBM model has a greater TN count, which indicates a higher accuracy in correctly forecasting *Non-Default payment*.

### 3.6.3. LIME

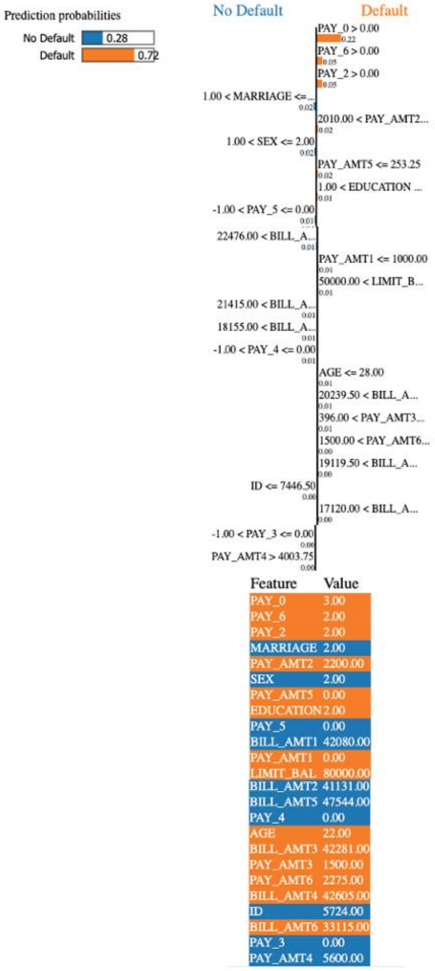
Applying LIME method to explain individual predictions of the GBM and EBM models. LIME provides local interpretability by approximating the GBM's decision-making process for a specific instance (Default payment next month). LIME produces feature importance and explanation plots (see Figure 6, and Figure 7) to shed light on the GBM model's decision for default payment.

**Figure 6:** GBM's LIME



| Feature | Value |
|---|---|
| PAY_0 | 3.00 |
| PAY_2 | 2.00 |
| PAY_6 | 2.00 |
| PAY_AMT1 | 0.00 |
| MARRIAGE | 2.00 |
| PAY_5 | 0.00 |
| LIMIT_BAL | 80000.00 |
| PAY_AMT4 | 5600.00 |
| PAY_AMT2 | 2200.00 |
| EDUCATION | 2.00 |
| BILL_AMT6 | 33115.00 |
| AGE | 22.00 |
| BILL_AMT2 | 41131.00 |
| PAY_AMT5 | 0.00 |
| PAY_4 | 0.00 |
| BILL_AMT5 | 47544.00 |
| PAY_AMT6 | 2275.00 |
| PAY_3 | 0.00 |
| BILL_AMT3 | 42281.00 |
| SEX | 2.00 |
| PAY_AMT3 | 1500.00 |

**Figure 7:** EBM's LIME



| Feature | Value |
|---|---|
| PAY_0 | 3.00 |
| PAY_6 | 2.00 |
| PAY_2 | 2.00 |
| MARRIAGE | 2.00 |
| PAY_AMT2 | 2200.00 |
| SEX | 2.00 |
| PAY_AMT5 | 0.00 |
| EDUCATION | 2.00 |
| PAY_5 | 0.00 |
| BILL_AMT1 | 42080.00 |
| PAY_AMT1 | 0.00 |
| LIMIT_BAL | 80000.00 |
| BILL_AMT2 | 41131.00 |
| BILL_AMT5 | 47544.00 |
| PAY_4 | 0.00 |
| AGE | 22.00 |
| BILL_AMT3 | 42281.00 |
| PAY_AMT3 | 1500.00 |
| PAY_AMT6 | 2275.00 |
| BILL_AMT4 | 42605.00 |
| ID | 5724.00 |
| BILL_AMT6 | 33115.00 |
| PAY_3 | 0.00 |
| PAY_AMT4 | 5600.00 |

Figure 6 shows that the GBM has a probability of 0.71 of getting default, quite like the EBM model (0.72). Also, LIME allows us to see the instances that impacts the prediction of No Default (blue colour) and Default (orange colour) for each model. The most crucial instance that impact in the prediction of default in the GBM model is the instance 0.23 for PAY_0 (September payment records), followed by 0.06 for PAY_2 (July payment records) and 0.04 for PAY_6 (April payment records). In contrast, the instances that impact the prediction of No default are Marriage and PAY_5 (May payment records). Figure 7 illustrates that also the most important instance that impacts the prediction of default in the EBM is PAY_0 (0.22), followed by PAY_6 (0.05) and PAY_2 (0.05), which this time share the same level of importance. EBM's LIME showed that MARRIAGE (0.02) and Sex (0.02) are essential for No Default prediction.

In conclusion, although the models present differences in the rankings of the instances that impact the default prediction, most of them coincide with the most critical values since both models obtained similar prediction models to determine if a person will default. Additionally, after the visual analysis of both plots of the LIME method for each model, it can be observed that both models agree that the essential variables for the prediction of "Default" was PAY_0, followed by PAY_2, PAY_6. It is relevant that both models agreed that Marriage is a relevant instance for the prediction of No default.

### 3.7. Limitations

Next, the limitations of this research project will be described. Firstly, in the cleaning preparation, no analysis was performed to identify potential bias issues. Secondly, this research project was focused on local interpretability only using the LIME methods; the project could be extended by analysing other local interpretability methods. As was mentioned before, one of the problems with the application of LIME is its instability. Additionally, a more challenging project could focus on comparing both models globally. Finally, precision, recall, F1 score, AUC-ROC and confusion matrix are valuable indicators for evaluating model performance. However, these metrics do not directly assess model interpretability.

# 4 Conclusion and Discussion

This research project aimed to investigate to what extent GBM can be made explainable in the context of credit default modelling. The first step was to identify which ML algorithms are used in the credit default domain. The finding showed that the most used algorithms are glass boxes, such as DT and LR, due to their elevated level of interpretability. However, DT and LR have already been extensively investigated. This research focused on analyzing the GBM, which is a black box ML model; in consequence, GBM has a high-performance level but a low level of interpretability. To contrast the result of GBM in terms of performance and interpretability, the EBM was selected, which is a glass box with excellent performance levels and global and local interpretability without the need for other methods.

It should be noted that EBM has an elevated level of interpretability since it directly provides local and global interpretability. GBM, for its part, does not have these characteristics, which is why it requires implementing some local interpretability method. LIME was selected for the comparison of the local interpretability of both models.

Then, interpretability in ML was investigated, finding two approaches to interpreting ML models: globally and locally. Global interpretability tries to explain the model in its entirety. Global interpretability is more challenging, and the results could be more appropriate sometimes. In contrast local interpretability focuses on the explanation of an instance. This study focused on the explanation of the instance named *Default payment next month* using a local interpretability method post-hoc named LIME.

After performing an analysis of performance metrics considering accuracy, precision, recall, F1-score, AUC-ROC and matrix confusion, it was determined that both models have similar performance in the metrics and confusion matrix applied. The accuracy levels were close to 82% for both models. The precision of the models was around 68%, which indicates that approximately 68% of the predictions made by each model were correct based on the test data. The analysis of the decision matrix showed that GBM and EBM had very minor variations, the results indicate comparable TP and FN counts, demonstrating their efficacy in detecting Default payment.

Regarding the interpretability analyses that are the ones sought in this project, LIME indicated that for both models, the default prediction was 0.71 GBM and 0.72 EBM, and that the variable that contributes the most to this prediction is PAY_0 (September payment records), which represents the September payment record. Also, LIME determined for both models, but with different rankings, that PAY_6 (April payment records) and PAY_2 (July payment records) are variables that impact the prediction of a *default* (Default payment next month). In contrast, LIME determined that the Marriage variable impacts the prediction that people will not default. The LIME method was used to compare both models, obtaining similar performance and local interpretability results.

After the analysis of the results, it was possible to conclude that GBM can be satisfactorily explained from the point of view of local interpretability using LIME, obtaining equivalent results in terms of the variables that impact the prediction of Default payment next month to those obtained by EBM using the LIME method also. These results provide information on the impact percentage, both positively and negatively, in the prediction of a specific instance or variable. In consequence, LIME can be used complementary with other analyses to understand the behaviour of an ML model. However, GBM's local interpretability does not provide information on how the ML algorithm makes decisions to arrive at a prediction. This study cannot provide information on whether it is feasible to understand how GBM works or how it makes decisions to reach the predictions it develops. Studies of GBM's global interpretability could provide insights into how GBM works internally to deliver predictions.

Additionally, since GBM is a black box, the literature review says that the global interpretability results are not optimal in terms of the results' quality and require more complex analysis. In future work, global interpretability could be addressed. Also, to broaden the local interpretability methods' scope compared to these two ML models. Additionally, finding or building specially designed metrics to assess model interpretability in credit default would benefit the financial sector. Since PAY_0 (September payment records) contributed the most to this prediction of default, it was analysed which event could produce an overconsumption in September that would trigger a credit default on the part of the clients of this bank in Taiwan.

# References

Adadi, A. and Berrada, M. (2018) ‘Peeking inside the black-box: A survey on explainable artificial intelligence (XAI)’, IEEE Access, 6, pp. 52138–52160. doi: 10.1109/access.2018.2870052.

Alam, T. M., Shaukat, K., Hameed, I. A., Luo, S., Sarwar, M. U., Shabbir, S., Li, J. and Khushi, M. (2020) ‘An investigation of credit card default prediction in the imbalanced datasets’, IEEE Access, 8, pp. 201173–201198. doi:10.1109/ACCESS.2020.3033784.

Baesens, B., Van Gestel, T., Viaene, S., Stepanova, M., Suykens, J. and Vanthienen, J. (2003) ‘Benchmarking state-of-the-art classification algorithms for credit scoring’, Journal of the Operational Research Society, 54(6), pp. 627–635. doi:10.1057/palgrave.jors.2601545.

Bibal, A. and Frénay, B. (2016) ‘Interpretability of machine learning models and representations: An introduction’, in Proceedings of the 24th European Symposium on Artificial Neural Networks ESANN, Bruges, Belgium, 27–29 April 2016, pp. 77–82. Available at: https://www.researchgate.net/profile/Adrien-Bibal-2/publication/326839249_Interpretability_of_Machine_Learning_Models_and_Representations_an_Introduction/links/5b6861caa6fdcc87df6d58e4/Interpretability-of-Machine-Learning-Models-and-Representations-an-Introduction.pdf [Accessed 16 June 2023].

Blakely, C. D. and Granmo, O.-C. (2021) ‘Closed-form expressions for global and local interpretation of Tsetlin’, in 34th International Conference on Industrial, Engineering and

Other Applications of Applied Intelligent Systems, IEA/AIE 2021, Kuala Lumpar, Malaysia, 26-29 July 2021, pp. 158–172. doi: 10.1007/978-3-030-79457-6_14.

Carvalho, D. V., Pereira, E. M. and Cardoso, J. S. (2019) ‘Machine learning interpretability: A survey on methods and metrics’, Electronics, 8(8), p. 832. doi: 10.3390/electronics8080832.

Cecchetti, S. and Schoenholtz, K. (2015) Money, banking, and financial markets. 4th edn. New York: McGraw-Hill Education.

Chakraborty, M., Biswas, S. K. and Purkayastha, B. (2022) ‘Rule extraction using ensemble of neural network ensembles’, Cognitive Systems Research, 75, pp. 36–52. doi:10.1016/j.cogsys.2022.07.004.

Chen, T. and Guestrin, C. (2016) ‘XGBoost: A scalable tree boosting system’, in Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, California, 13-17 August 2016, pp. 785-794. doi: 10.1145/2939672.2939785

Doshi-Velez, F. and Kim, B. (2017) Towards a rigorous science of interpretable machine learning. Available at: https://arxiv.org/pdf/1702.08608.pdf [Accessed 15 June 2023].

Fernández-Delgado, M., Cernadas, E., Barro, S. and Amorim, D. (2014) ‘Do we need hundreds of classifiers to solve real-world classification problems?’, Journal of Machine Learning Research, 15(1), pp. 3133-3181. Available at: https://www.jmlr.org/papers/volume15/delgado14a/delgado14a.pdf?source=post_page [Accessed 13 July 2023].

Fong, R. C. and Vedaldi, A. (2017) ‘Interpretable explanations of black boxes by meaningful perturbation’, in Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22-29 October 2017, pp. 3429–3437. doi: 10.1109/ICCV.2017.371.

Frost, J., Gambacorta, L., Huang, Y., Song Shin, H. and Zbinden, P. (2020) ‘BigTech and the changing structure of financial intermediation’, Economic Policy, 34(100), pp. 761-799. doi: 10.1093/epolic/eiaa003.

Hooker, G. (2004) ‘Discovering additive structure in black box functions’, in KDD '04: Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Seattle, WA, USA, 22-25 August 2014, pp. 575-580. doi: 10.1145/1014052.1014122

Honegger, M. (2018) Shedding light on black box machine learning algorithms: Development of an axiomatic framework to assess the quality of methods that explain individual predictions. MA thesis. Karlsruhe: Karlsruhe Institute of Technology. doi: 10.48550/arXiv.1808.05054.

Huber, T., Weitz, K., André, E. and Amir, O. (2021) ‘Local and global explanations of agent behavior: Integrating strategy summaries with saliency maps’, Artificial Intelligence, 301, p.103571. doi:10.1016/j.artint.2021.103571.

Jayasundara, S., Indika, A. and Herath, D. (2022) ‘Interpretable student performance prediction using explainable boosting machine for multi-class classification’, in 2022 2nd

International Conference on Advanced Research in Computing (ICARC). Belihuloya, Sri Lanka, 23-24 February 2022, pp. 391-396. doi: 10.1109/ICARC54489.2022.9753867.

Karlsson, I., Rebane, J., Papapetrou, P. and Gionis, A. (2020) ‘Locally and globally explainable time series tweaking’, Knowledge & Information Systems, 62(5), pp. 1671–1700, doi: 10.1007/s10115-019-01389-4.

Koh, H. C., and Chan, K. L. G. (2002) ‘Data mining and customer relationship marketing in the banking industry’, Singapore Management Review, 24(2), pp. 1–27. Available at: https://discovery.ebsco.com/linkprocessor/plink?id=63a931ca-ff48-39a6-b13e-bb159d0a7879 [Accessed 16 June 2023].

Koh, P. W. and Liang, P. (2017) ‘Understanding black-box predictions via influence functions’, in International Conference on Machine Learning (ICML), 70, pp. 1885-1894. Available at: http://proceedings.mlr.press/v70/koh17a/koh17a.pdf [Accessed 08 July 2023].

Konstantinov, A. V. and Utkin, L. V. (2021) ‘Interpretable machine learning with an ensemble of gradient boosting machines’, Knowledge-Based Systems, 222, p.106993. doi:10.1016/j.knosys.2021.106993.

Kovalev, M. S., Utkin, L. V. and Kasimov, E. M. (2020) ‘SurvLIME: A method for explaining machine learning survival models’, Knowledge-Based Systems, 203, p.106164. doi: 10.1016/j.knosys.2020.106164.

Liu, W., Fan, H. and Xia, M. (2022) ‘Credit scoring based on tree-enhanced gradient boosting decision trees’, Expert Systems with Applications, 189, p.116034. doi: 10.1016/j.eswa.2021.116034.

Liu, X., Shi, X.-Q., Li, X.-B. and Peng, Z.-R. (2022) ‘Quantification of multifactorial effects on particle distributions at urban neighbourhood scale using machine learning and unmanned aerial vehicle measurement’, Journal of Cleaner Production, 378, p.134494. doi: 10.1016/j.jclepro.2022.134494.

Liu, G. and Sun, B. (2023) ‘Concrete compressive strength prediction using an explainable boosting machine model’, Case Studies in Construction Materials, 18. doi: 10.1016/j.cscm.2023.e01845.

Lundberg, S. M. and Lee, S.-I. (2017) ‘A unified approach to interpreting model predictions’, in 31st Conference on Neural Information Processing Systems (NIPS 2017). Long Beach, USA. 2017. Available at: https://proceedings.neurips.cc/paper_files/paper/2017/file/8a20a8621978632d76c43dfd28b67767-Paper.pdf [Accessed 20 June 2023].

Molnar, C. (2022) Interpretable machine learning: A guide for making black box models explainable. 2nd edn. Available at: https://christophm.github.io/interpretable-ml-book/. [Accessed 16 June 2023].

Mu, Y., Ni, R., Fu, L., Luo, T., Feng, R., Li, J., Pan, H., Wang, Y., Sun, Y., Gong, H., Guo, Y., Hu, T., Bao, Y. and Li, S. (2023) ‘DenseNet weed recognition model combining local variance preprocessing and attention mechanism’, Frontiers in Plant Science, 13. doi: 10.3389/fpls.2022.1041510.

Munnell, A., Tootell, G., Browne, L., and McEneaney, J. (1996) ‘Mortgage lending in Boston: Interpreting HMDA data’, American Economic Review, 86 (1), pp. 25–53. Available at: http://cob.jmu.edu/doylejm/EC%20485docs/Munnell_HMDA_AER1996.pdf [Accessed 05 March 2023].

Nori, H., Jenkins, S., Koch, P., and Caruana, R. (2019) ‘InterpretML: A unified framework for machine learning interpretability. doi: 10.48550/arXiv.1909.09223

Petsiuk, V., Das, A. and Saenko, K. (2018) RISE: Randomized Input Sampling for Explanation of black-box models. Available at: https://arxiv.org/pdf/1806.07421.pdf [Accessed 20 June 2023].

Rabold, J., Deininger, H., Siebers, M. and Schmid, U. (2019) ‘Enriching visual with verbal explanations for relational concepts – Combining LIME with Aleph’, in International Workshops of ECML PKDD 2019, Würzburg, Germany, 16-20 September 2019, pp.180–192. doi: 10.1007/978-3-030-43823-4_16.

Rajbahadur, G. K., Wang, S., Oliva, G. A., Kamei, Y. and Hassan, A. E. (2022) ‘The impact of feature importance methods on the interpretation of defect classifiers’, IEEE Transactions on Software Engineering, 48(7), pp. 2245-2261. doi: 10.1109/tse.2021.3056941.

Ribeiro, M. T., Singh, S. and Guestrin, C. (2016a) ‘Why should I trust you? Explaining the predictions of any classifier’, in KDD '16: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, California, 13-17 August 2016, pp. 1135-1144. doi: 10.1145/2939672.2939778.

Ribeiro, M. T., Singh, S. and Guestrin, C. (2016b) ‘Model-agnostic interpretability of machine learning’, in 2016 ICML Workshop on Human Interpretability in Machine Learning (WHI 2016), New York, NY, 23 June 2016, pp. 91-95. doi: 10.48550/arXiv.1606.05386.

Ribeiro, M. T., Singh, S. and Guestrin, C. (2018) ‘Anchors: High-precision model-agnostic explanations’, in Proceedings of the AAAI Conference on Artificial Intelligence, 32(1), pp. 1527-1535. doi: 10.1609/aaai.v32i1.11491.

Ross, A. S., Hughes, M. C. and Doshi-Velez, F. (2017) ‘Right for the right reasons: Training differentiable models by constraining their explanations’ in Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence (IJCAI-17), Melbourne, Australia, 19-25 August 2017, pp. 2662-2670. doi: 10.24963/ijcai.2017/371.

Rudin, C. (2019) ‘Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead’, Nature Machine Intelligence, 1(5), pp. 206–215. doi: 10.1038/s42256-019-0048-x.

Sakri, S. (2022) ‘Assessment of deep neural network and gradient boosting machines for credit risk prediction accuracy’, in 14th IEEE International Conference on Computational Intelligence and Communication Networks (CICN). Al-Khobar, Saudi Arabia, 4 - 6 December 2022, pp. 170-176. doi: 10.1109/CICN56167.2022.10008264.

Sarica, A., Quattrone, A. and Quattrone, A. (2022) ‘Explainable machine learning with pairwise interactions for the classification of Parkinson’s disease and SWEDD from clinical and imaging features’, Brain Imaging and Behavior, 16(5), pp. 2188–2198. doi: 10.1007/s11682-022-00688-9.

Shankaranarayana, S. M. and Runje, D. (2019) ‘ALIME: Autoencoder Based Approach for Local Interpretability’, in 20th International Conference on Intelligent Data Engineering and Automated Learning – IDEAL 2019, Manchester, UK, 14-16 November 2019, pp. 454-463. doi: 10.1007/978-3-030-33607-3_49.

Srinath, T. and Gururaja, H. S. (2022) ‘Explainable machine learning in identifying credit card defaulters’, Global Transitions Proceedings, 3(1), pp. 119-126. doi: 10.1016/j.gltp.2022.04.025.

Sohn, S. Y., Kim, D. H. and Yoon, J. H. (2016) ‘Technology credit scoring model with fuzzy logistic regression’, Applied Soft Computing, 43, pp. 150–158. doi:10.1016/j.asoc.2016.02.025.

Strumbelj, E. and Kononenko, I. (2010) ‘An efficient explanation of individual classifications using game theory’, Journal of Machine Learning Research, 11, pp. 1–18. Available at: https://www.jmlr.org/papers/volume11/strumbelj10a/strumbelj10a.pdf?ref=https://githubhelp.com [Accessed 15 June 2023].

Varian, H. (2014) ‘Big data: New tricks for econometrics’, Journal of Economic Perspectives, 28(2), pp. 3–28. doi: 10.1257/jep.28.2.3.

Vu, M. N., Nguyen, T. D., Phan, N., Gera, R. and Thai, M. T. (2019) Evaluating explainers via perturbation. Available at: https://www.researchgate.net/profile/Nhat-Hai-Phan/publication/333640578_Evaluating_Explainers_via_Perturbation/links/5cfd8c28299bf13a384a47b4/Evaluating-Explainers-via-Perturbation.pdf [Accessed 20 June 2023].

Yang, J. and Wang, H. (2021) ‘Interpretability analysis of academic achievement prediction based on machine learning’, in 2021 11th International Conference on Information Technology in Medicine and Education (ITME). Wuyishan, Fujian, China, 19-21 November 2021, pp. 475-479. doi: 10.1109/ITME53901.2021.00101.

Yu, Y. (2020) ‘The application of machine learning algorithms in credit card default prediction’, in 2020 International Conference on Computing and Data Science (CDS), Stanford, CA, USA, 1-2 August 2020, pp. 212-218. doi: 10.1109/CDS49703.2020.00050.

Zafar, M. R. and Khan, N. M. (2019) ‘DLIME: a deterministic local interpretable model-agnostic explanations approach for computer-aided diagnosis systems’, in Proceedings of ACM SIGKDD Workshop on Explainable AI/ML (XAI) for Accountability, Fairness, and Transparency, Anchorage, AK, 4-8 August 2019, pp. 1-6. Available at: https://arxiv.org/abs/1906.10263 [Accessed 19 June 2023].

Zhou, Z.-H. and Feng, J. (2017) ‘Deep forest: Towards an alternative to deep neural networks’, in Proceedings of the 26th International Joint Conference on Artificial Intelligence (IJCAI’17), Melbourne, Australia, 19-25 August 2017, pp. 3553–3559. doi: 10.24963/ijcai.2017/497.

Zhu, X. and Chen, J. (2021) ‘Risk prediction of P2P credit loans overdue based on gradient boosting machine model’, in 2021 IEEE International Conference on Power, Intelligent Computing and Systems (ICPICS), Shenyang, China, 29-31 July 2021, pp. 212-216. doi: 10.1109/icpics52425.2021.9524127.

Zhu, J., Fang, S., Yang, Z., Qin, Y., and Chen, H. (2023) ‘Prediction of concrete strength based on Random Forest and Gradient Boosting Machine’ in 2023 IEEE 3rd International Conference on Power, Electronics and Computer Applications (ICPECA). Shenyang, China, 29-31 January 2023, pp. 306-312. doi: 10.1109/ICPECA56706.2023.10075839.