NOVA IMS Information Management School

MDSAA

Master Degree Program in

Data Science and Advanced Analytics

Transparent AI in Finance

A study on how Explainable AI can help financial institutions justify automated decisions

Vasco Miguel Inácio Brigas Bargas

Master Thesis

presented as partial requirement for obtaining a Master's Degree in Data Science and Advanced Analytics

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

NOVA Information Management School

Instituto Superior de Estatística e Gestão de Informação

Universidade Nova de Lisboa

Transparent AI in Finance

A study on how Explainable AI can help financial institutions justify automated decisions

by

Vasco Miguel Inácio Brigas Bargas

Master Thesis presented as partial requirement for obtaining the Master's degree in Data Science and Advanced Analytics, with a specialization in Data Science.

Supervised by

Professor Vitor Manuel Pereira Duarte dos Santos, PhD, Universidade Nova de Lisboa

June, 2024

STATEMENT OF INTEGRITY

I hereby declare having conducted this academic work with integrity. I confirm that I have not used plagiarism or any form of undue use of information or falsification of results along the process leading to its elaboration. I further declare that I have fully acknowledged the Rules of Conduct and Code of Honor from the NOVA Information Management School.

Lisbon, June 2024

ACKNOWLEDGEMENTS

I would like to thank my supervisor, professor Vitor Santos for his support and guidance throughout the development of my thesis. His continuous support and availability helped me enormously. I would also like to thank my family and friends for their assistance in this past year.

ABSTRACT

The adoption of Artificial Intelligence in the financial sector has brought numerous quality of life improvements on internal procedures and on decision making processes. This adoption means that AI systems actively have a say on decisions that can affect people's lives. Meanwhile, most AI systems are opaque, casting doubts and uncertainty regarding the usage of these systems to make impactful decisions. This thesis focuses on using Explainable Artificial Intelligence (XAI) techniques and algorithms to provide explanations to an AI system's decision and the steps it took to get there. Through a detailed literature review where the current state of XAI in the financial sector is revised and a practical use case where the LIME and SHAP algorithms are tested against an AI system developed to predict a person's credit risk, this study tests the applicability of XAI techniques for explaining an AI system. The results suggest that the implementation of XAI techniques can provide a satisfactory degree of explainability to a model, demystifying its decision making processes.

KEYWORDS

Explainable Artificial Intelligence; Machine Learning; Banking; Finance

Sustainable Development Goals (SDG):

TABLE OF CONTENTS

1. Introduction 1
   1.1. Background and problem identification 1
   1.2. Objectives 2
   1.3. Importance and Relevance 3
   1.4. Methodological outline 4
2. Literature review 6
   2.1. Explainable Artificial Intelligence 6
      2.1.1. XAI Terminology 7
      2.1.2. Related Concepts 8
      2.1.3. Approaches 9
      2.1.4. Algorithms 10
      2.1.5. Challenges 12
   2.2. Explainable Artificial Intelligence in Financial area 13
   2.3. Analysis 18
3. Methodology Review and Improvement 21
   3.1. Exploration phase 21
   3.2. Analytical Phase 22
   3.3. Conclusive Phase 22
4. Analysis 23
   4.1. Experimental design 23
   4.2. Execution 24
      4.2.1. Data Understanding 24
      4.2.2. Data Preparation/Feature Engineering 25
      4.2.3. Model Selection/Hyperparameter Tuning 26
      4.2.4. Model Evaluation 27
      4.2.5. XAI Algorithms Implementation 28
5. Results Evaluation and discussion 32
6. Conclusions 35
   6.1. Limitations 35
   6.2. Future Work 36
References 38

LIST OF FIGURES

Figure 1. Methodology phase proposal. 5
Figure 2. Article Selection Process. 15
Figure 3. Research Methodology Design. 21
Figure 4. Flow-chart representing development tasks 24
Figure 5. Confusion Matrix of GNB model | Confusion Matrix of XGB model 27
Figure 6. Instance with bad credit risk 28
Figure 7. Instance with good credit risk 29
Figure 8. SHAP output 30

LIST OF TABLES

Table 1. Systematic Literature Review Research Questions 13
Table 2. Keywords 14
Table 3. Inclusion and Exclusion Criteria 14
Table 4. Articles included in the Literature Review 16
Table 5. Dataset features 25
Table 6. ML models tested 27

LIST OF ABBREVIATIONS AND ACRONYMS

AI	Artificial Intelligence
CRISP-DM	Cross-Industry Standard Process for Data Mining
GNB	Gaussian Naïve Bayes
LGBM	Light Gradient Boosting Machine
LIME	Local Interpretable Model-Agnostic Explanations
ML	Machine Learning
MU	Monetary Units
PRISMA	Preferred Reporting Items for Systematic Reviews and Meta-Analyses
SHAP	SHapley Additive exPlanations
SLR	Systematic Literature Review
XAI	Explainable Artificial Intelligence
XGB	EXtreme Gradient Boosting

1. INTRODUCTION

This chapter is composed by the introductory steps of the study, where a background contextualization and problem identification is made, followed by the goal definition, the study's relevance and a proposed methodological outline is presented.

1.1. BACKGROUND AND PROBLEM IDENTIFICATION

The data and technology revolution opened lots of room for new and cutting-edge companies to enter even the most traditional and well-established sectors like the banking and financial industry (Alt et al., 2018). With the emergence in competitiveness, traditional banks were forced to make a decision: to change its processes and upgrade their technology stack in order to incorporate new functionalities or to keep following an outdated business plan doomed to eventually fail (Frame et al., 2018). To avoid falling behind, traditional banks and financial institutions started implementing practices derived from financial technology (FinTech) companies.

Among these practices and new technologies, artificial intelligence (AI) stands out as a great way to increase efficiency, discover new patterns, unlock information, and optimize a corporation's resources. AI's correct use can translate in better insights, better informed decisions and overall better results(Wamba-Taguimdje et al., 2020). This prompted companies to use AI models in every way possible, without looking too much into its flaws. As a result, lots of corporations make decisions that can impact our lives solely based on the output of a Machine Learning model (Shaw, 2019). There are AI programs that decide which applicant should get the job, whether a person is eligible for a loan, or even determine if a patient is sick based on an X-ray image. These decisions can deeply impact the life of the person in question, and there is a growing trend of letting this kind of life-altering decisions be decided based solely on an AI model's output (Shaw, 2019).

The growing concern of this modus operandi is that AI models are basically a black-box ("a complex system whose internal workings are hidden or not readily understood"), which means that the decisions are provided with no justification and with no way of knowing the steps the model took to arrive to a certain conclusion (Carabantes, 2020). Depending on the

magnitude of the situation, this leads to a decision that can severely jeopardize one's life, providing no context as to why that happened. AI models have been known to learn some biased behavior, sometimes discriminating individuals based on certain attributes that should not matter to the final decision (Nelson, 2019). The problem is that since we don't know what happens inside a model, we can't really figure out if the decisions are based on the right type of attributes or not.

Explainable Artificial Intelligence (XAI) is a field of AI that tries to provide some reasoning and explanation to a model's decision (Barredo Arrieta et al., 2020). It uses specific techniques and algorithms with the goal of providing a more transparent model and a way of tracking an AI model's thought process. This is the perfect solution to combat the black-box problem. With insights about the inside of a model, it is possible to check for unwanted biases, reach its root and modify whatever is necessary to improve the model's performance (Minh et al., 2022). It is also helpful to fine-tune AI models by identifying some gaps that can be found when evaluating a model's decision. Another great benefit of XAI is that, regarding the responsibility of AI models on making decisions that can impact a person's life, is that a justification can be provided (Barredo Arrieta et al., 2020). If a model, that is determining who qualifies for a bank loan, is functioning as it should and denies credit to someone, it can show its reasoning. This then allows the person whose credit was not granted, to know what the weakest points of their application were and improve them, rather than being completely in the dark about the possible next steps for a future application. Furthermore, XAI provides a unique opportunity for man and machine to work in harmony, since all information is being shared, making the role of a human shift from simply providing inputs and testing outputs, to a supervisor guiding the creation of an AI model in each step of the process (Adadi & Berrada, 2018).

1.2. OBJECTIVES

The goal of the research is to perform a study on the practical applicability of XAI algorithms in the financial area. In order to achieve this goal, the following intermediate objectives were defined:

- Framework XAI area;

- Make a comprehensive analysis of most prominent XAI algorithms;
- Create and execute use case to apply relevant algorithms in a real financial case;
- Analyze and discuss the results.

1.3. IMPORTANCE AND RELEVANCE

The financial industry has a very important role in a society's ecosystem. It deals with one of the most valuable assets an individual can have, which is money. Since exchanging money for a product or service is the standard way of doing business, it holds a lot of value in our society.

Financial institutions have to frequently make decisions regarding people's access to loans and credits, based on gathered data regarding the asking customer. Currently, a lot of these decisions are made by trusting an AI model's output (Cao, 2020). The power to deny someone money to buy a house, or to start a new business venture is too great, and an AI model lacks accountability and responsibility for its own actions (Carabantes, 2020).

In a financial institution scope, having this kind of information allows it to understand how the model calculates a client's risk level, which personal attributes the model sees as riskier, and it can even find and explain new trends unknown to man (Minh et al., 2022). Moreover, it translates in a better overall customer service, by simply informing the customers the reason of their bad credit score or why their loan application was declined, which allows them to actively try to improve the lagging points of their application.

As previously stated in the context section of this study, accessing a model's way of making decisions can also help determine whether it is performing as intended. In a black-box scenario, like most Machine Learning (ML) models, the output is composed of only the final decision, disregarding any kind of justification (Carabantes, 2020). The consequences of using these kinds of models without fully understanding them can significantly impact people's lives.

The importance of this research lies on being able to implement XAI algorithms to provide transparency and clarity to a currently murky subject: What goes on inside an AI model?

Answering this question will potentiate the use of AI models to a new level, by combatting the biggest concerns and objections people have against this technology (Dietterich & Horvitz,

2015). It will result in more informed decisions, an overall wider acceptation and more trust regarding the field of AI, elevate human-AI collaboration and it will help debug and audit AI systems more effectively (Barredo Arrieta et al., 2020).

The outcome of this project will hopefully shine a light on how an AI model makes decisions, allowing us humans to understand its thought process and to guarantee it makes fair and justifiable decisions.

Ultimately, successfully implementing XAI algorithms to work alongside AI models will result in a more effective, safe, and fair experience for everyone involved.

1.4. METHODOLOGICAL OUTLINE

The first draft of the planned methodology in this research, which may be subject to changes, was divided into 3 main phases:

- Exploration Phase, where a literature review is performed in order to establish a strong foundation to answer this study's research questions. Furthermore, it also helps identify gaps and possible challenges in the field, which allows for a better understanding of the context of this field. Afterwards, a methodology review and improvement is performed based on the new information gathered in the literature review.
- Analytical Phase, where an evaluation is made on the applicability of the Literature Review's recommended XAI algorithms (according to the context) in a model designed to predict a financial institution customer's credit risk, to test the performance and viability of XAI tools.
- Conclusive Phase, which consists of the presentation, interpretation, discussion and evaluation of the results that originated from the Analytical Phase.

In Figure 1 is presented a visual representation of the planned methodology phases:

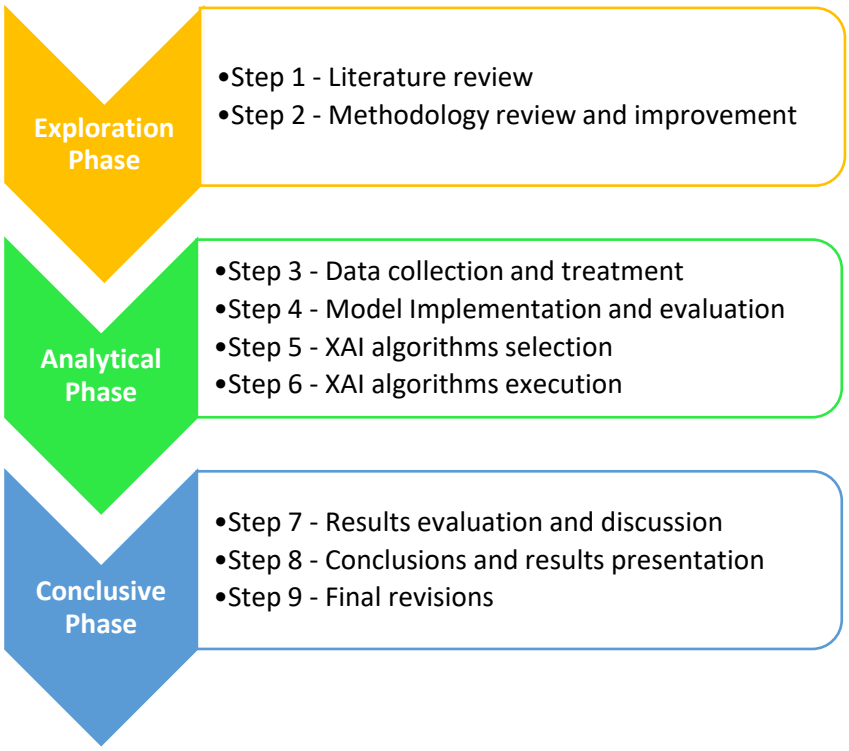

Figure 1. Methodology phase proposal.

2. LITERATURE REVIEW

The Literature Review is divided into sub sections based on the area of research. To start, a literature review on the Explainable Artificial Intelligence area is performed to gather the knowledge and technical terms required for the study. Then, a systematic literature review is performed to connect the area of Explainable Artificial Intelligence with the area of Finance, allowing a good understanding of the current state of research about these two topics.

2.1. EXPLAINABLE ARTIFICIAL INTELLIGENCE

Explainable Artificial Intelligence (XAI) is a relatively recent field in Artificial Intelligence (AI) and Machine Learning (ML) designed to make AI systems more comprehensible and understandable to humans. Nowadays, AI is used to assist and sometimes decide real-world situations and problem(Shukla Shubhendu & Vijay, 2013). With the growing size of available information to make decisions, a lot of these systems are extremely complex, analysing a lot of data in order to produce a final decision. This means that the human mind can't fully comprehend the dimension of the data, and it can't find dependencies and correlations as easily as an ML model (Meske et al., 2022). On top of that, AI systems are able to process and make a decision on new data entries at a speed that is impossible to match as a human being. To bridge the gap between the AI system's "knowledge", and the human's understanding of such AI system, XAI offers some techniques to add clarity to an ML model's decision-making process (Barredo Arrieta et al., 2020).

Most AI models used in real-world scenarios use deep learning to decipher problems and make decisions. A big concern regarding current AI systems is that its method of decision is described as a "black-box", meaning that there is no information regarding the process of decision-making (Carabantes, 2020). The user simply provides the input data and the AI model provides an output containing only the final decision. The usage of these systems on real-world situations can lead to undesirable results, since there is no kind of justification as to why a particular decision was made, which can lead to an unknowingly unfair outcome for a human being due to unnoticed bias and other irrelevant fields being given importance (Samek & Müller, 2019). Using the healthcare industry as an example, if an AI system designed to screen

medical patients based on X-ray images does not provide justification for its decisions, there could easily be some error in the decision-making process that wrongfully labels a specific set of patients. Even though the ML model was thoroughly tested before being used, the people developing it may have overlooked certain rare situations, not knowing that the system currently in use has flaws (Nazer et al., 2023). In this case, if XAI is applied with the model, it is possible to follow along its decision-making process. This allows us to easily find incorrect decisions, pinpoint where those decisions are being made and act accordingly.

XAI is also useful in other domains, not only on automating a task to increase efficiency like the example above. When used in combination with AI systems designed to make decisions based on huge amounts of data, it helps people understand dependencies and insights captured by the model, which otherwise would be invisible to the human eye (Ding et al., 2022). Having this kind of knowledge regarding a model's internal functionality allows us to comprehend its decisions, to attribute accountability, to fine-tune some undesirable conclusions and to justify its results to everyone involved (Barredo Arrieta et al., 2020).

All these benefits would push the use of AI even further across society, given that currently a lot of people still don't trust these systems to aid and make decisions (Lockey et al., 2021). With a good reasoning to support decisions, people will be more comfortable with the adoption of these technologies.

Business-wise, XAI offers the perfect solution to justify to shareholders the decisions made by AI systems (Zednik, 2021). By having a detailed explanation on a model's decision-making process, model results can be easily explained and, in cases where a business denies some sort of service to a client, it can even help customers figure out what is not favourable in their application so they can improve it and have better chances next time.

2.1.1. XAI Terminology

Below is a list of the most relevant XAI terms and expressions for this study, and its meaning.

- Post-hoc: Type of XAI techniques used after the creation of an AI model. These techniques aim to explain the decisions made by an AI model after its conception, meaning it won't affect the model's training and development (Bellucci et al., 2021).

- Transparent models: Machine Learning models that provide clear and understandable decision-making processes by default. Models such as Decision Trees, and Logistic and Linear Regressions have simple inner workings that are easily understandable by themselves (Bellucci et al., 2021).
- Feature relevance: Magnitude of a feature's importance regarding the output produced by an ML model. It reflects the impact of a feature on a model's decision-making process (Ali et al., 2023).
- Model-agnostic techniques: Techniques that don't rely on a model's characteristics to explain that model's prediction. Algorithms such as SHAP and LIME don't need to access a model's inner workings to successfully interpret its decisions (Bellucci et al., 2021).
- Model simplification: Process of creating a more easily interpretable version of a complex ML model. Models such as deep neural networks can be too complex to understand, but using model simplification techniques like feature selection or dimensionality reduction can untangle a model's complexity (Barredo Arrieta et al., 2020).
- Counterfactual Explanation: Framework that calculates and presents what changes in the input data would result in a change regarding the decision of an AI model (Bellucci et al., 2021).
- Anchor method: XAI approach which creates easy to follow explanations for the decisions of an AI model. It creates a simple set of rules (anchors) based on the features available that, if met, ensure a certain prediction from a model (S Band et al., 2023).

2.1.2. Related Concepts

There are other concepts that are related to the Explainable Artificial Intelligence field and that can be seen as relevant to XAI's objective:

- Responsible AI: Development and usage of AI systems prioritizing not only performance, but also ethical considerations, fairness, transparency and accountability of such systems. Its main goal is to make sure that AI is used to benefit society while minimizing possible risks and liabilities (Bellucci et al., 2021).

- Explainability: Ability of an AI system to provide understandable and transparent explanations for its decision-making process, focusing on making the inner workings of AI models clear to humans regardless of their previous knowledge in the area, allowing them to understand the reasoning behind specific results (Bellucci et al., 2021).
- Transparency: Refers to the accessibility and clarity of an AI system's decision-making process and overall functioning. It focuses on providing visibility into the internal mechanisms of an AI model to its stakeholders (Bellucci et al., 2021).
- Fairness: Capability of an AI system to provide ethical and equal treatment to everyone during the whole development process and consequent outcome, promoting AI systems to not discriminate against any particular subset of a population (Bellucci et al., 2021).
- Accountability: Represents the level of responsibility of an AI system and its developers regarding the decisions made by AI models. Its main objective is to be able to hold someone accountable when poor and unfair decisions are made (Bellucci et al., 2021).

2.1.3. Approaches

The two main approaches when it comes to XAI implementation can be distinguished by which part of the AI system they want to explain (Barredo Arrieta et al., 2020). A global approach aims to provide a general understanding of a model's behaviour, meaning it tries to assess feature importance and insights gathered from all instances, providing a macro level explanation to a model's decision-making process (Saleem et al., 2022). Examples of this approach include the SHAP (SHapley Additive exPlanations) algorithm and transparent models.

A local approach focuses on a single instance for its analysis. This approach provides a more personalized explanation, since it focuses on the decision-making process of only one decision at a time (Le et al., 2023). Examples of this approach include the LIME (Local Interpretable Model-Agnostic Explanations) algorithm, Counterfactual Explanations and Anchors.

2.1.4. Algorithms

This section will focus on the applicability and technical description of the most popular algorithms of each approach, Local Interpretable Model-Agnostic Explanations (LIME) and SHapley Additive exPlanations (SHAP) in a real world scenario, in order to see if these approaches have the potential to reliably justify and help understand AI systems' decisions.

LIME:

Local Interpretable Model-Agnostic Explanations, also known as LIME, is a XAI algorithm that has the capability of explaining the decisions of any ML model through a local approximation technique. This is a model agnostic method, meaning it is not required to have any information on the inner workings of a model to explain its reasoning. The algorithm works by selecting an input data point, slightly alter its feature values and create a sample of points based on those alterations. Afterwards, the altered data points are inputted in the black box AI model, which makes its predictions. Then, the LIME algorithm creates a transparent model, such as a regression or decision tree, with the intent of mimicking the black box model output. Now we have a transparent model that locally behaves in the same manner as our black box model. The transparent model can then be interpreted to find explanations as to why it chose that outcome. Through the transparent model we can assess the similarity of the altered data points with the original and weigh them accordingly. We can also calculate the feature importance of each feature used to make the prediction. The output of the LIME algorithm is a set of features that the algorithm deems as being the most impactful for a certain decision. For example, a black box model developed to predict if a bank customer will adhere to a new service that used LIME would go through the following process:

1. Select random data point;
2. Create subset of data points by slightly altering original data point. In this case, if the features used were account balance, age, number of transactions and date of account creation, the altered data points would be the result of applying small tweaks in the feature values;
3. Use black box model to make predictions on the altered data points;
4. Assign weights to the altered data points according to its proximity to the original point;

5. Train a transparent model with the altered data points and predictions made by the black box model;
6. Check feature importance to find out how much each feature is responsible for the black box model output.

If the most important features were account balance and number of transactions, we could conclude that the likelihood of a client joining a new service would heavily depend on these two features. This approach grants a level of transparency and interpretability otherwise inexistent when it comes to black box models.

SHAP:

SHapley Additive exPlanations, also known as SHAP, is an algorithm that uses the Shapley value concept from game theory to successfully explain a model's prediction. Like LIME, SHAP uses a model agnostic approach, meaning that the inner workings of a model are not needed in order to provide some reasoning behind its choices. SHAP works by evaluating the impact of a value change in a single feature while maintaining every other feature constant, repeating this process for all features. Having previous knowledge regarding the output of a random data point inputted in a black box model, the algorithm proceeds to tweak one feature and maintain all other features, thus creating a set of new data points. This set is then inputted in the model and its predictions allow the algorithm to quantify the importance of a feature in the decision making process. The importance is calculated by checking the difference between the new data points' results and the result of the original data point. This process is repeated throughout all features used and then tested using multiple instances of the dataset. Using the same example as above (black box model developed to predict if a bank customer will adhere to a new service, using the features account balance, age, number of transactions and date of account creation), the steps would be:

1. Select random data point (for example, [1000; 45; 50; 01/04/2015]);
2. Generate multiple data points by tweaking a feature value. For instance, the altered data points would see an alteration of the account balance feature, generating vectors such as [1050; 45; 50; 01/04/2015], [900; 45; 50; 01/04/2015], [1200; 45; 50; 01/04/2015], etc.;
3. Use the black box model to predict the outcomes of the altered data points;

4. Compare the predictions from the altered data points with the prediction of the original data point to assess feature importance;
5. Repeat steps 2 through 4 for every feature used;
6. Repeat steps 1 through 5 for multiple data points;
7. Final output consists of a set of values, one per feature, with the overall importance of each feature for the black box model's decision making process.

By analysing each feature's impact individually, SHAP makes sure that there is no oversight or negligence over any variable, providing a fair attribution of feature importance.

2.1.5. Challenges

The ultimate goal of mass adoption of Explainable Artificial Intelligence is not without its problems and challenges. There are certain aspects of AI systems that will inevitably pay the price for greater transparency and explainability. Generally, the so-called Transparent Models have the downside of performing slightly worse than "black-box" models when complex datasets are involved (Adadi & Berrada, 2018). Attempts to increase the transparency of such models may require a trade-off in terms of performance. In extremely complex models, even with the application of several XAI tools, there is the possibility of these tools being unable to properly explain and allow humans to interpret the full extent of AI models (Adadi & Berrada, 2018). In these cases, it would be required to simplify the model in question, which would probably result in a poorer performance.

With the increase of data generation, gathering and subsequent use for developing AI systems, the resources necessary to create explanations regarding a model's decisions are also increasing. The scalability of XAI when analyzing big data problems using complex models can present an efficiency problem (Saeed & Omlin, 2023).

The dynamic nature of AI systems can also present a challenge to XAI. Models are constantly evolving, being trained with newer data in order to keep up with the current trends and overall situation. These model changes can impact previous decisions, since the model can adjust the importance of features to better suit the current situation (Watson, 2022). In this event, the

previous model predictions can become obsolete and give room to an argument about the validity of previous decisions.

The intellectual property aspect is also a pertinent topic while discussing XAI challenges. When organizations rely on its developed AI systems to generate value in their service or product, they may not want to disclose the inner workings of their systems because that jeopardizes their competitive edge against competitors (Spartalis et al., 2024). Having to publicly disclose key information about one's tool can deter them from rallying in favor of a mandatory XAI adoption.

The number of stakeholders and interested parties regarding Artificial Intelligence raises the issue of finding the baseline for model explainability. Interested people on this topic will have different levels of knowledge about the field, so the level of detail and complexity of an explanation required is totally different for a developer and a person with no previous contact with AI (Doshi-Velez & Kim, 2017). Creating the perfect middle ground for wide understanding of decision-making processes in AI systems can be challenging due to the different levels of familiarity among stakeholders.

2.2. EXPLAINABLE ARTIFICIAL INTELLIGENCE IN FINANCIAL AREA

The literature review performed regarding Explainable Artificial Intelligence provided an introduction to this field and some explanations on its purpose, benefits and challenges. To complement this study, a systematic literature review will be conducted in order to identify all relevant papers to our study, which connects the XAI field with the Financial field. The objective of this review is to understand what kind of progress these two areas have achieved together and to answer the following research questions:

Table 1. Systematic Literature Review Research Questions

SLRQ1	What is the current state of research in this area?
SLRQ2	What kind of XAI techniques are currently useful in this area?
SLRQ3	What are the advantages and disadvantages of applying XAI techniques in this field?

According to the PRISMA methodology, to find the most relevant papers regarding this set of topics, a set of keywords relevant to this study has to be defined. To perform this step and the remaining ones, it was decided that the search would only include English keywords and, consequently, only English-written papers published after 2020. The chosen keywords are represented in table 2.

Table 2. Keywords

Explainable Artificial Intelligence	**Finance**
Explainable Artificial Intelligence	Finance
Neural Networks	Banking
Machine Learning	Loan
Transparent Models	Credit
Explainability	Credit Score

To search for these keywords in scientific documents databases, a search query was created containing the keywords and some conditions to guarantee the presence of keywords related to both fields pertinent to the study. The search query is as follows:

("Explainable Artificial Intelligence" OR "XAI" OR "Transparent Models" OR "SHAP" OR "LIME") AND ("Finance" OR "Banking" OR "Credit")

The search query was used in two scientific research databases: Scopus (https://www.scopus.com/) and Web of Science (https://www.webofscience.com/wos/). To further filter the quality and relevance of the research articles, some inclusion and exclusion criteria was defined:

Table 3. Inclusion and Exclusion Criteria

Inclusion Criteria	**Exclusion Criteria**
XAI techniques used in financial domain	Non-English publications
Publications between 2021 and 2024	Published date before 2021
Theoretical approaches relevant to this study	Articles with unrelated content
	Articles with irrelevant methodology

Upon inserting the search query in the scientific databases, a total of 718 articles were discovered, completing the Identification phase of the PRISMA methodology. In the Screening phase, 125 duplicate documents were removed and other 197 articles were excluded because their publication date was prior to 2021. Advancing to the Eligibility phase, a total of 396 articles were assessed according to their and keywords, where 339 were ultimately excluded from the literature review. To conclude the PRISMA flowchart a deeper inspection of the remaining 57 articles occurred to confidently assess which publications would provide more value to this research. This step removed 26 articles, leaving a total of 31 articles to be included in this research. Below is a visual representation of the whole process:

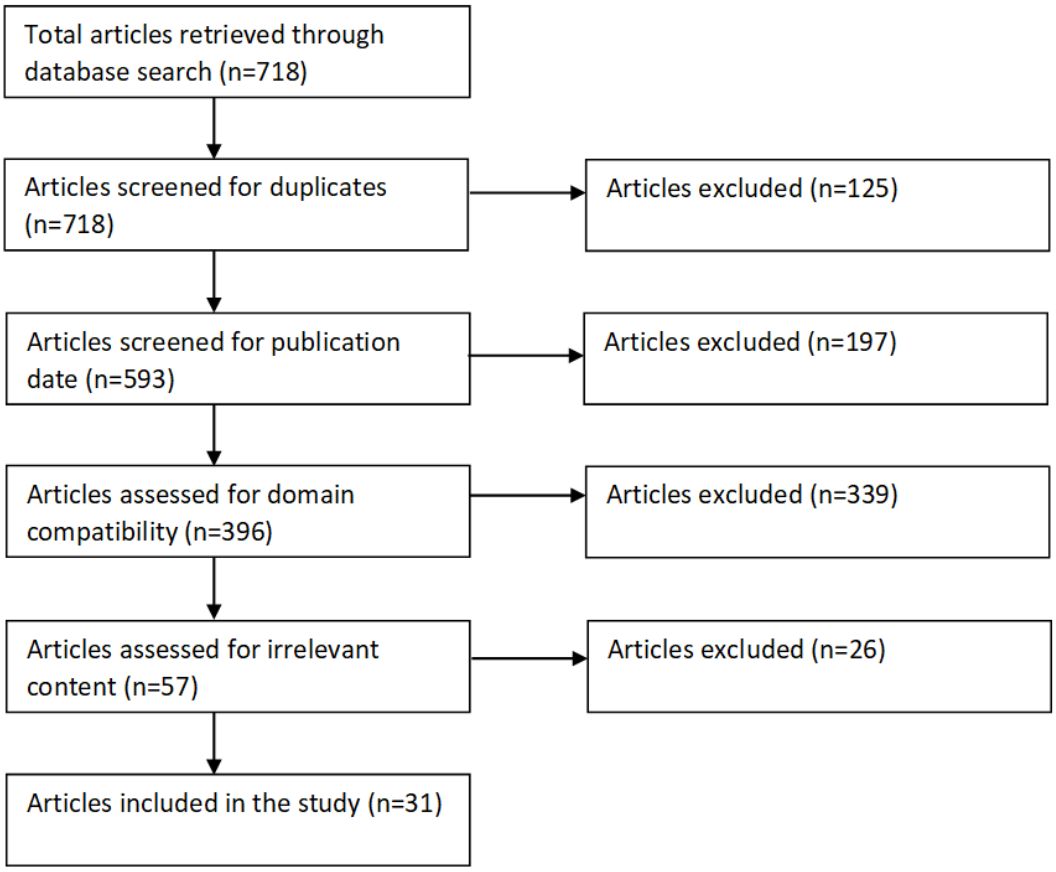

***Figure 2**. Article Selection Process.*

Below is a table containing all the articles included in this systematic literature review:

Table 4. Articles included in the Literature Review

#	Authors	Title
1	Moscato, V., Picariello, A., & Sperlí, G. (Moscato et al., 2021)	A benchmark of machine learning approaches for credit score prediction
2	Wu, H. D., & Han, L. (Wu & Han, 2021)	A novel reasoning model for credit investigation system based on Fuzzy Bayesian Network
3	Heng, Y. S., & Subramanian, P. (Heng & Subramanian, 2023)	A Systematic Review of Machine Learning and Explainable Artificial Intelligence (XAI) in Credit Risk Modelling
4	Hall, P., Cox, B., Dickerson, S., Ravi Kannan, A., Kulkarni, R., & Schmidt, N. (Hall et al., 2021)	A United States Fair Lending Perspective on Machine Learning
5	Sriram, A., Gorti, S. S., Amin, E. G., & Kumar, A. (Sriram et al., 2022)	Analyzing Banking Services Applicability Using Explainable Artificial Intelligence
6	Weber, P., Carl, K. V., & Hinz, O. (Weber et al., 2024)	Applications of Explainable Artificial Intelligence in Finance—a systematic review of Finance, Information Systems, and Computer Science literature
7	Hadji Misheva, B., Jaggi, D., Posth, J. A., Gramespacher, T., & Osterrieder, J. (Hadji Misheva et al., 2021)	Audience-Dependent Explanations for AI-Based Risk Management Tools: A Survey
8	Xu, R., Meng, H., Lin, Z., Xu, Y., Cui, L., & Lin, J. (Xu et al., 2021)	Credit Default Prediction via Explainable Ensemble Learning
9	Ponsam, J. G., Gracia, S. J. B., Geetha, G., Karpaselvi, S., & Nimala, K. (Ponsam et al., 2021)	Credit Risk Analysis using LightGBM and a comparative study of popular algorithms
10	Amato, F., Ferraro, A., Galli, A., Moscato, F., Moscato, V., & Sperlí, G. (Amato et al., 2022)	Credit Score Prediction Relying on Machine Learning
11	Liu, Y., Huang, F., Ma, L., Zeng, Q., & Shi, J. (Liu et al., 2024)	Credit scoring prediction leveraging interpretable ensemble learning

12	Nwafor, C. N., & Nwafor, O. Z. (Nwafor & Nwafor, 2023)	Determinants of non-performing loans: An explainable ensemble and deep neural network approach
13	Hadji Misheva, B., & Papenbrock, J. (Hadji Misheva & Papenbrock, 2022)	Editorial: Explainable, Trustworthy, and Responsible AI for the Financial Service Industry
14	Gramespacher, T., & Posth, J. A. (Gramespacher & Posth, 2021)	Employing Explainable AI to Optimize the Return Target Function of a Loan Portfolio
15	Rizinski, M., Peshov, H., Mishev, K., Chitkushev, L. T., Vodenska, I., & Trajanov, D. (Rizinski et al., 2022)	Ethically Responsible Machine Learning in Fintech
16	De Lange, P. E., Melsom, B., Vennerød, C. B., & Westgaard, S. (de Lange et al., 2022)	Explainable AI for Credit Assessment in Banks
17	Bussmann, N., Giudici, P., Marinelli, D., & Papenbrock, J. (Bussmann et al., 2021)	Explainable Machine Learning in Credit Risk Management
18	Cardenas-Ruiz, C., Mendez-Vazquez, A., & Ramirez-Solis, L. M. (Cardenas-Ruiz et al., 2022)	Explainable Model of Credit Risk Assessment Based on Convolutional Neural Networks
19	Hjelkrem, L. O., & Lange, P. E. D. (Hjelkrem & Lange, 2023)	Explaining Deep Learning Models for Credit Scoring with SHAP: A Case Study Using Open Banking Data
20	Kuiper, O., van den Berg, M., van der Burgt, J., & Leijnen, S. (Kuiper et al., 2022)	Exploring Explainable AI in the Financial Sector: Perspectives of Banks and Supervisory Authorities
21	Boardman, J., Alam, M. S., Huang, X., & Xie, Y. (Boardman et al., 2022)	Integrated Gradients is a Nonlinear Generalization of the Industry Standard Approach to Variable Attribution for Credit Risk Models
22	Walambe, R., Kolhatkar, A., Ojha, M., Kademani, A., Pandya, M., Kathote, S., & Kotecha, K. (Walambe et al., 2021)	Integration of Explainable AI and Blockchain for Secure Storage of Human Readable Justifications for Credit Risk Assessment

23	Chen, Y., Calabrese, R., & Martin-Barragan, B. (Chen et al., 2024)	Interpretable machine learning for imbalanced credit scoring datasets
24	Lusinga, M., Mokoena, T., Modupe, A., & Mariate, V. (Lusinga et al., 2021)	Investigating statistical and machine learning techniques to improve the credit approval process in developing countries
25	Dastile, X., & Celik, T. (Dastile & Celik, 2021)	Making Deep Learning-Based Predictions for Credit Scoring Explainable
26	Aljadani, A., Alharthi, B., Farsi, M. A., Balaha, H. M., Badawy, M., & Elhosseini, M. A. (Aljadani et al., 2023)	Mathematical Modeling and Analysis of Credit Scoring Using the LIME Explainer: A Comprehensive Approach
27	Sathe, S. S., & Mahalle, P. (Sathe & Mahalle, 2023)	Predictive Analytics in Financial Services Using Explainable AI
28	Gramegna, A., & Giudici, P. (Gramegna & Giudici, 2021)	SHAP and LIME: An Evaluation of Discriminative Power in Credit Risk
29	Wen, H., Sui, X., & Lu, S.	Study on Effect of Consumer Information in Personal Credit Risk Evaluation
30	Bücker, M., Szepannek, G., Gosiewska, A., & Biecek, P. (Bücker et al., 2022)	Transparency, auditability, and explainability of machine learning models in credit scoring
31	Zhang, L., Wang, J., & Liu, Z. (Zhang et al., 2023)	What should lenders be more concerned about? Developing a profit-driven loan default prediction model

2.3. ANALYSIS

Upon gathering the most pertinent articles following the PRISMA methodology, one will thoroughly analyse the full article list to get the information needed to answer the Systematic Literature Review Research Questions.

SLRQ1: What is the current state of research in this area?

The financial sector is no stranger to the implementation of AI systems to aid and optimize processes and services. Regarding Explainable AI, even though there are lots of articles combining these two fields, the recency of XAI research in general makes it challenging to find robust and heavily studied methods to use in financial cases at such an early stage. Practical studies are usually conducted by using machine learning models combined with XAI algorithms to assess a person's credit risk (Cardenas-Ruiz et al., 2022). Some studies try a more "macro" approach, analysing or using XAI techniques to provide explainability to ML models predicting credit risk and loan default on enterprises or countries (Hadji Misheva et al., 2021). Theoretical studies in this area revolve around bridging the knowledge gap between the financial field and the AI field (Bücker et al., 2022), with some touching on the ethical and legal concerns hovering around the AI topic (Heng & Subramanian, 2023).

SLRQ2: What kind of XAI techniques are currently useful in this area?

The main techniques that currently provide some explainability to AI systems are the usage of XAI algorithms like LIME and SHAP (Chen et al., 2024), which involves implementing a "black-box" model and implementing one of these algorithms to explain local and global decisions, and the usage of transparent models (Boardman et al., 2022), which are models that natively provide decision making explanations.

SLRQ3: What are the advantages and disadvantages of applying XAI techniques in this field?

The adoption of XAI techniques in the financial industry has its pros and cons. The advantages brought on by the usage of XAI techniques are the offer of explainability and transparency to AI systems, which will result in a better understanding of the decision making processes occurring inside such systems (Bussmann et al., 2021). XAI techniques will also assist financial institutions to act according to regulatory compliances (Kuiper et al., 2022), will improve interaction and collaboration between humans and AI systems (Gramespacher & Posth, 2021), and will offer clear justifications regarding its decisions to stakeholders and other interested parties (Sriram et al., 2022).

On the other hand, implementing XAI techniques can lead to a decrease in an AI system's performance. Due to an ML model's innate complexity, explaining its decisions in a way that is comprehensible by humans is difficult (Rizinski et al., 2022). Even with XAI techniques that

offer the capability to interpret “black-box” models, there are cases where the complexity is too great to understand (Dastile & Celik, 2021). In those cases, the need to understand a certain model requires a simplified version of itself, ultimately trading performance for explainability (Gramegna & Giudici, 2021). This trade-off can result in a worse model performance, causing financial institutions’ AI systems to wrongfully classify more customers, but is a necessary step towards a harmonious relationship between AI systems and society.

3. METHODOLOGY REVIEW AND IMPROVEMENT

Upon conducting a literature review, and using it as support, a methodology review and adjustment is performed in order to better represent the required steps and phases for this study. The analytical phase of this study was slightly adjusted to better fit the standard process of testing XAI algorithms in ML models. The analytical phase is now divided in four parts: one where an overview of the whole phase is designed, one regarding the selection and preparation of the data, followed by the implementation and evaluation of the ML model and finally the XAI algorithms execution. In addition, the implementation of the ML model also follows the CRISP-DM methodology. A detailed explanation of each methodology phase is also provided in this section.

In Figure 3 is presented a visual representation of the research methodology phases:

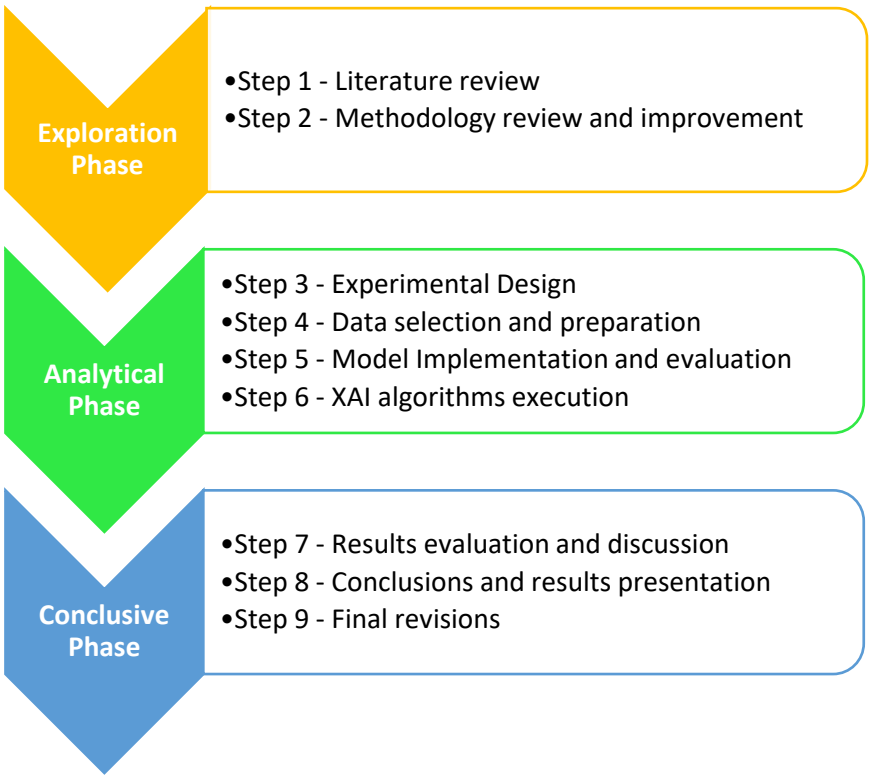

***Figure 3**. Research Methodology Design.*

3.1. EXPLORATION PHASE

In the exploration phase, a literature review is conducted to set a solid foundation for further exploration of this paper's research question and to create a vast knowledge base regarding

the topic of Explainable AI in Finance. Afterwards, a methodology review is performed to assess its effectiveness given the new knowledge gathered in the literature review.

3.2. ANALYTICAL PHASE

In the analytical phase, a dataset containing a bank's customer information is selected with the intent to create a machine learning model to predict if a customer will default on its credit/loan. The next step is to process and prepare the data to be inputted in the ML model. Afterwards the model implementation, selection and evaluation is performed. To finalize the analytical phase, one will evaluate the applicability of the most prominent XAI algorithms, LIME (Local Interpretable Model-Agnostic Explanations) and SHAP (SHapley Additive exPlanations) when paired with the previously created ML model.

3.3. CONCLUSIVE PHASE

The conclusive phase is the final phase of this study, where an evaluation and discussion of the results obtained is performed, to assess whether the ML model's decisions are successfully interpreted. After the result's discussion and evaluation, the conclusion and results are presented. Finally, revisions are made to ensure the quality and coherence of the whole study, making way to possible changes in the thesis.

4. ANALYSIS

The analytical phase of this study consists of the implementation of a practical use case where XAI algorithms are used to provide explainability to a Machine Learning model trained to classify an individual's credit risk. Having the literature review as reference, an experimental design is created in the format of a flow chart to represent the needed steps to achieve the desired results.

4.1. EXPERIMENTAL DESIGN

The experimental design is divided into 3 main categories:

- Data-related tasks: This category is composed by all tasks regarding data ingestion, data analysis and data preparation. It is in this category that the data is prepped to be inputted into an ML model.
- Modeling tasks: This category is divided into 3 steps, which are the model selection, hyperparameter tuning and model evaluation. The final output of this set of tasks is the model that will be analysed using XAI tools.
- XAI-related tasks: The final category represents the implementation of XAI techniques and algorithms and the analysis of the results obtained regarding the model's explainability.

A more detailed explanation of each task is given further ahead. Below is represented a flow chart (Figure 4) displaying all tasks to be performed and its order.

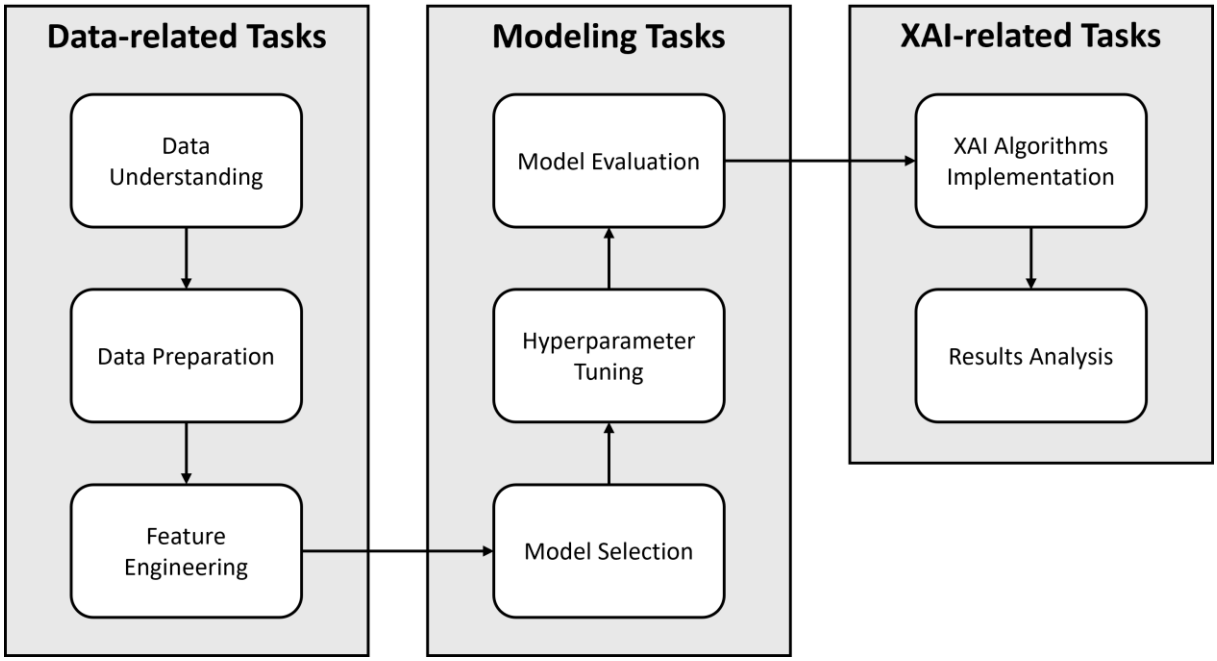

Figure 4. Flow-chart representing development tasks

4.2. EXECUTION

The Execution section is divided into 5 parts, representing the main development stages of the practical portion of this study.

4.2.1. Data Understanding

The dataset selected to perform the use case is named *German Credit Data* and is from UCI Machine Learning Repository (Dataset Link). It is composed of 1000 instances of individuals and 21 variables, as shown in table 5, containing personal information that can be useful for classifying credit risk, of which 20 are features and one is the target. The target is a binary feature where the value "0" represents a low default probability/good credit risk and the value "1" represents a high default probability/bad credit risk. The dataset has no missing values and no abnormal values that can indicate the occurrence of typos or other errors on the data. The target variable is not balanced, since 700 instances represent a low default probability/good credit risk and only 300 instances represent a high default probability/bad credit risk. This distribution makes sense when thinking about a real-life scenario, where the majority of people are able to pay off its loans, resulting on a lower number of defaults than the other way around.

Table 5. Dataset features

Variable Name	Variable Type
Checking Account Status	Categorical
Credit Duration (months)	Numerical
Credit History	Categorical
Credit Purpose	Categorical
Credit Amount	Numerical
Savings Account Status	Categorical
Current Employment Duration (years)	Categorical
Installment rate in % of disposable income	Numerical
Personal Status and Gender	Categorical
Other Debtors/Guarantors	Categorical
Current Residence Duration (years)	Numerical
Owned Property	Categorical
Age	Numerical
Other Installment Plans	Categorical
Housing Situation	Categorical
Number of Existing Credits	Numerical
Profession	Categorical
Number of Dependents	Numerical
Telephone	Binary
Foreign Worker	Binary
Credit Default	Binary

4.2.2. Data Preparation/Feature Engineering

To guarantee that the data was suitable for further analysis and for modeling purposes, some transformations on the original data were made. Since there were no missing values, no imputing technique was required to fill missing values. Due to the scarcity of instances and the overall quality of the data provided, no outliers were removed, as excluding data from a dataset of such small size could mean a lot of information loss. Categorical features were transformed using one-hot encoding, a method that converts all possible values from a feature into a binary feature. For example, the feature *Housing Situation* contains the possible values *rent*, *own* and *free*. Using one-hot encoding, each value is now a binary feature, and for each data instance, one of the features will have a value of "1" and the others a value of "0". The categorical features that represent an order between values were transformed using an

ordinal encoder. After these transformations the final dataframe was composed of 45 features, due to the high number of possible values in some categorical features. The data was then split into train and test sets, with a distribution of 75% for training and 25% for testing. Before the modeling phase, the numerical features were scaled using a MinMax scaler with the interval set from 0 to 1. With these operations, all features were using the same scale, with all values being within the defined interval [0,1].

4.2.3. Model Selection/Hyperparameter Tuning

To select the most suitable model for this use case, several models were tested using its default hyperparameters. The ones with the best performance (highlighted in table 6) were selected to perform hyperparameter tuning. Due to the imbalanced nature of this dataset, the most accurate model, which was an eXtreme Gradient Boosting (XGB) classifier, was misclassifying a significant part of the underrepresented class (high credit risk), as it is shown when analysing the Recall metric. This phenomenon applied to most of the models evaluated. For that reason, the models chosen were the one with best overall accuracy (XGB) and the one with the greatest capability to classify correctly high credit risk individuals. The latter model is the Gaussian Naïve Bayes (GNB) classifier, that despite having a slightly worse overall performance than XGB, had a more balanced correct classification rate between classes. To perform the tuning of hyperparameters in order to further improve the model's performance, a grid search was conducted to test several combinations of hyperparameters. The hyperparameter tuning did not result in a significantly better performance regarding the XGB classifier. In GNB's case, its lack of hyperparameters make a grid search pointless, so no tuning was performed.

Table 6. ML models tested

Model Name	Accuracy	Recall	Precision	F1 Score
XGB Classifier	0.788	0.542	0.661	0.595
Gaussian Naive Bayes	0.712	0.681	0.500	0.576
LGBM Classifier	0.772	0.444	0.653	0.529
Logistic Regression	0.768	0.500	0.621	0.554
Support Vector Classifier	0.756	0.361	0.634	0.460
Random Forest Classifier	0.748	0.333	0.615	0.432
AdaBoost Classifier	0.724	0.431	0.525	0.473

4.2.4. Model Evaluation

To evaluate both models, the Accuracy, Precision, Recall and F1 score metrics were analysed to determine the best overall performance. Due to the imbalance present in the target variable, the confusion matrix (shown in Figures 5 and 6) was also considered when selecting the best model. It is important to note that, in a practical scenario, the ability to identify high credit risk customers is very important for decision making. Due to that, the GNB model was chosen because its ability to correctly classify high credit risk individuals was far superior when compared to XGB classifier, as seen in the confusion matrices below.

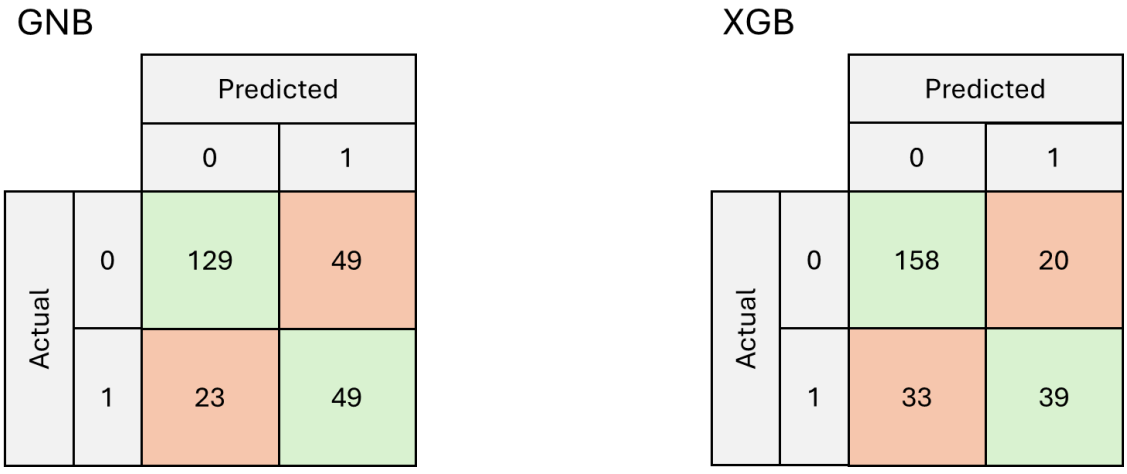

Figure 5. Confusion Matrix of GNB model | Confusion Matrix of XGB model

4.2.5. XAI Algorithms Implementation

The XAI algorithms tested on the GNB model were the LIME and SHAP. The goal was to test both a local explainability technique (LIME) and a global explainability technique (SHAP). To show the LIME algorithm execution, one correctly classified instance of each class (Figure 7 is an instance with good credit risk and Figure 6 is an instance with bad credit risk) from the test set was randomly chosen. The output of the algorithm (as depicted in figures 6 and 7), contains the probability of the instance belonging to each class, the importance that each feature has to the prediction probability and to which class it favours, and the respective feature value of the instance being analysed.

Figure 6. Instance with bad credit risk

Legend referring to Figures 6 and 7:

- credit_purpose_A410 -> Credit Purpose: Others
- credit_purpose_A44 -> Credit Purpose: Domestic Appliances
- credit_history_A31 -> Credit History: All previous credits paid back duly
- credit_purpose_A46 -> Credit Purpose: Education
- credit_purpose_A48 -> Credit Purpose: Retraining
- savings_account_status_A64 -> Savings Account: Over 1000 M.U. (Monetary Units)
- other_installment_plans_A142 -> Other Installment Plans: Stores
- savings_account_status_A63 -> Savings Account: Between 500 and 1000 M.U.
- credit_purpose_A45 -> Credit Purpose: Repairs
- checking_account_status_A13 -> Checking Account: Over 200 M.U.
- other_guarantors_A103 -> Other Guarantors: Guarantor

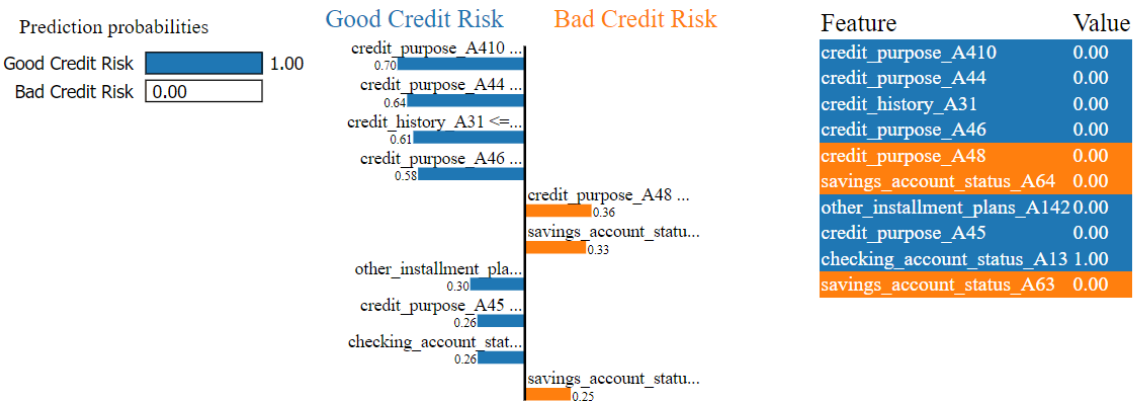

Figure 7. Instance with good credit risk

The SHAP algorithm implementation represented in Figure 8 provides an output where the top 20 features are presented in order of importance. The SHAP value determines the impact of each feature in the model decision and the points represent each instance, where the color red indicates a high feature value and the color blue indicates a low feature value.

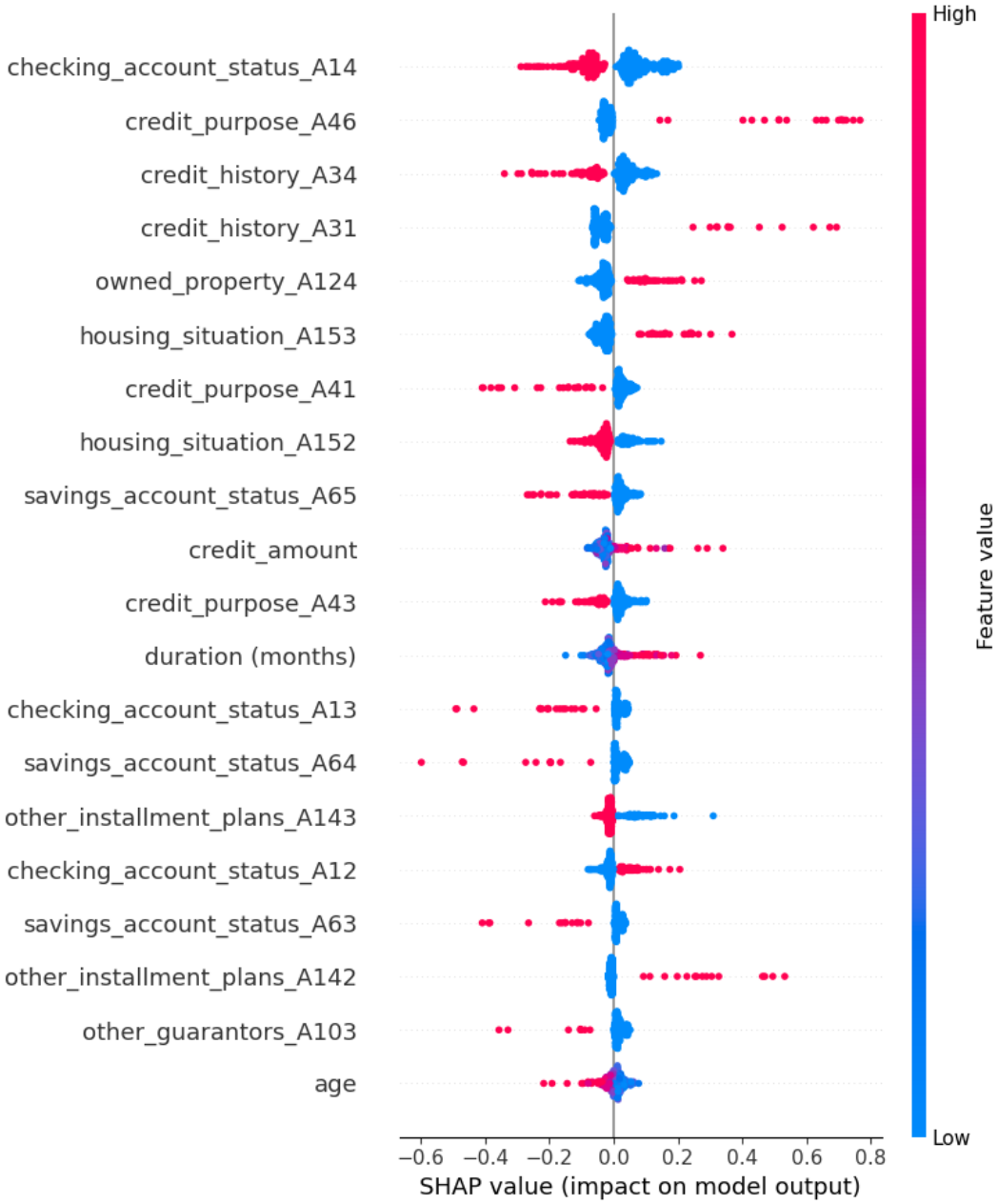

Figure 8. SHAP output

<u>Legend referring to Figure 8:</u>

- checking_account_status_A14 -> Checking Account: No checking account
- credit_purpose_A46 -> Credit Purpose: Education
- credit_history_A34 -> Credit History: Other credits existing
- credit_history_A31 -> Credit History: All previous credits paid back duly
- owned_property_A124 -> Owned Property: Unknown/No Property
- housing_situation_A153 -> Housing Situation: For Free

- credit_purpose_A41 -> Credit Purpose: Used Car
- housing_situation_A152 -> Housing Situation: Owner
- savings_account_status_A65 -> Savings Account: Unknown/No savings account
- credit_amount -> Credit Amount
- credit_purpose_A43 -> Credit Purpose: Electronic Devices
- duration (months) -> Duration in months
- checking_account_status_A13 -> Checking Account: Over 200 M.U. (Monetary Units)
- savings_account_status_A64 -> Savings Account: Over 1000 M.U.
- other_installment_plans_A143 -> Other Installment Plans: None
- checking_account_status_A12 -> Checking Account: Between 0 and 200 M.U.
- savings_account_status_A63 -> Savings Account: Between 500 and 1000 M.U.
- other_installment_plans_A142 -> Other Installment Plans: Stores
- other_guarantors_A103 -> Other Guarantors: Guarantor
- age -> Age

5. RESULTS EVALUATION AND DISCUSSION

The LIME algorithm defined, in both cases presented, the same 10 features as the most influential to classify an instance, with only the order of importance suffering a change. On the instance classified as a bad credit risk (Figure 6), the fact that the customer's credit purpose was not to purchase domestic appliances or other unidentified products is seen as a positive sign. The customer's credit history being in pristine condition, with all previous credits being paid back in time is also a positive factor when weighing the model's decision. The lack of store related instalment plans also has a favorable impact on the customer's credit request. Despite all these reasons, the credit was correctly classified as a bad credit risk. The credit purpose being for education was seen as a negative sign. The reason for that might be that an individual that is looking to pursue another degree of education most likely does not have a full-time job. It also impacted the decision negatively the fact that the credit purpose was not for retraining purposes (retraining might mean learning a new set of skills to further improve an individual's professional career, and is easier to do while already having a full-time job). Regarding the savings account status, since it wasn't between the interval of 500 to 1000 or above, it reflected poorly on the model's decision process. Adding to this individual's lack of monetary units on the savings account, its checking account also had less than 200 monetary units. These factors may tell us that the individual requesting the loan doesn't have a reliable "safety net" in case things don't go as expected. Finally, having no guarantor to guarantee the repayment of the loan in case of the individual's default made the model's final decision to classify this individual as a bad credit risk.

On the instance classified as good credit risk (Figure 7), the first three features are equal to the previous instance. The purpose not being domestic appliances, repairs or other unidentified products, and having a pristine credit history contributed positively for the decision made. In this case, the fact that the purpose was not education also had a favorable impact on the model. Having no store related instalment plans also helps this individual's case. This individual's checking account being over 200 monetary units (can indicate the possibility of repayment) was the final positive factor of its application. In this instance's credit risk assessment, the fact that its savings account was below 500 monetary units weighed negatively on the decision. It also weighed negatively, such as in the previous instance, the objective of the loan not being for retraining purposes. Despite these negative factors, the

overall assessment had a positive outcome, with the model correctly classifying this instance as a good credit risk.

The SHAP algorithm showcased the 20 most important features and its values among instances, providing a more general and big picture explanation regarding the decisions made. In this particular case, a negative SHAP value means that the feature contributes to a good credit risk classification and a positive SHAP value means the opposite. Following this principle, Figure 8 shows us that if the features representing not having a checking account, having other active credits, using the loan to acquire a used vehicle, owning a house, not having a savings account, using the loan to acquire electronic devices, having over 200 monetary units on a checking account, having over 500 monetary units on a savings account, having no other instalment plans, having a guarantor and age have a high value (on binary features, a high value is 1 and a low value is 0), its SHAP value will be negative, leaning the decision towards a good credit risk. In the opposite cases, these features' SHAP values will be positive, contributing for a bad credit risk decision. The features representing using the loan to pursue an education, having all credits paid back in time, not owning property/unknown, living somewhere for free, the credit amount, the credit duration, having between 0 and 200 monetary units in a checking account and having other instalment plans in stores have the opposite effect, meaning a low value will return a negative SHAP value. The distribution of instances shown on each feature row dictates the overall impact on the model. For example, the features "Other Installment Plans: None" and "Checking Account: Between 0 and 200 M.U." have a lower impact on decision making, since their SHAP values are mostly between -0.1 and 0.2. Lots of insights can be gathered from the output of the SHAP algorithm. It is possible to gather that, for this model, the older a person is, the more likely it is for the loan to be paid, probably because an older person tends to have more financial capabilities than a younger person. Or the fact that owning a house will increase the chances of being classified as a good credit risk, because in case of default, there are assets that can be used to repay the loan.

The implementation of both XAI algorithms provided indeed a better understanding of the AI model's thought process and decision-making reasoning. The way both techniques operate allows us to understand and identify key features, which results tend to favor one class over the other and the degree of importance for each feature, allowing us to assess its impact in

the overall situation. The downside of this use case is that the XAI algorithms interpret the model's decisions, meaning that a flawed model will present flawed explanations and thought processes. Some features' interpretations were inconsistent with what is witnessed in real life (e.g. in Figure 8, an individual with a pristine credit history (feature *credit_history_A31*) being more likely to be classified as a bad credit risk and an individual with a critical credit history (feature *credit_history_A34*) being more likely be classified as a good credit risk), but these situations are not caused by a poor performance on the XAI algorithms' side.

6. CONCLUSIONS

Combining these two perspectives (LIME and SHAP), it is possible to analyse an AI model's inner workings on a macro level and on an instance level, allowing the gathering of more insights by combining the findings of both techniques. Using both LIME and SHAP can complement the overall explainability of a model, since one algorithm may fill a knowledge gap that the other was unable to capture.

On a practical level, if a financial institution were to adopt these XAI techniques, it would gain the ability to identify riskier variables and common patterns between variables, allowing employees and stakeholders to better recognize good and bad credit risks. These insights can bring value to other services or actions within an organization, transferring the knowledge gathered from credit risk classification to other domains. These techniques would also allow a financial institution to provide a clear justification to a customer whose credit request has been denied. By using SHAP to understand the general thought process of the model, and LIME to see the exact request classification process, the financial institution can provide a set of justifiable reasons for the credit denial, allowing the customer to know the lacking points of its credit application. This adoption would ultimately result in a system capable of providing a level of reasoning similar to a human but with the efficiency of a computer program.

These tools can also be used to monitor a model's performance and guarantee its compliance with ethical guidelines and company policies. If, on the explanation provided, there is a deviation on how the model should work and how the model actually works, the output of XAI algorithms can help find these undesired thought processes to start working on how to improve the overall model functionality, making these techniques powerful monitoring and evaluation tools.

6.1. LIMITATIONS

This use case and the general adoption of these techniques is not without some limitations. The dataset used in this study presents some limitations that make it more challenging for a model to accurately capture the relationships between features:

- Its target is imbalanced. The dataset is composed of 1000 instances, where only 300 represent the class of high credit risk. Although this value makes sense if we think

about real life, where most people will not default on their loans, this discrepancy when training an AI model can introduce biases, skew feature importance and disregard important predictors for the smallest class, which, in this case, is the most important class to correctly predict.

- Most of its features are categorical. The original dataset (before data transformation) has 20 features, where 13 of them are categorical. When converting these features to binary using the one-hot encoding technique, the total number of features more than doubles its initial size. Having an overwhelmingly large number of binary variables can dilute the importance of other features that also provide a lot of information gain. Separating categorical features' values into independent variables also introduces the problem of skewness towards less distributed values, possibly causing the model to misjudge certain features' importance.

On a more general level, the results obtained with the usage of XAI algorithms heavily depend on the model they are being used on, so if a model has poor performance, XAI algorithms will also produce sub optimal outputs. Model complexity can also be a limiting factor. Current XAI algorithms find it difficult to accurately translate decision making processes of very complex models, since all nuances of an AI system may not be captured in the output available for humans to understand. Another limitation is the level of knowledge required to fully understand the XAI algorithms' output. Even though these techniques are very useful for client-facing processes, the employees that usually have the required knowledge to successfully manage these algorithms and extract insights are not directly involved in the before mentioned processes, creating a knowledge gap between the ones who implement and know how to deal with these systems and the ones who actively use them.

Despite these factors, the results presented accurately show the reasoning of the model's decision through the impact of a feature on the general model and on specific instances.

6.2. FUTURE WORK

To further improve this study, a real-world application of these techniques must ensue, to test the applicability of these outputs in a real scenario. This would compare the XAI algorithms' outputs with the selection criteria of a financial institution, allowing one to see if the valued features are the same between the two classification methods. With XAI currently being an

extremely hot topic, new algorithms and ways to explain models are surging at a fast pace. A broader experimentation of techniques could find better solutions to solve the black-box problem in AI.

REFERENCES

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

Ali, S., Abuhmed, T., El-Sappagh, S., Muhammad, K., Alonso-Moral, J. M., Confalonieri, R., Guidotti, R., Del Ser, J., Díaz-Rodríguez, N., & Herrera, F. (2023). Explainable Artificial Intelligence (XAI): What we know and what is left to attain Trustworthy Artificial Intelligence. *Information Fusion*, *99*. https://doi.org/10.1016/j.inffus.2023.101805

Aljadani, A., Alharthi, B., Farsi, M. A., Balaha, H. M., Badawy, M., & Elhosseini, M. A. (2023). Mathematical Modeling and Analysis of Credit Scoring Using the LIME Explainer: A Comprehensive Approach. *Mathematics*, *11*(19). https://doi.org/10.3390/math11194055

Alt, R., Beck, R., & Smits, M. T. (2018). FinTech and the transformation of the financial industry. In *Electronic Markets* (Vol. 28, Issue 3). https://doi.org/10.1007/s12525-018-0310-9

Amato, F., Ferraro, A., Galli, A., Moscato, F., Moscato, V., & Sperlí, G. (2022). Credit Score Prediction Relying on Machine Learning. *CEUR Workshop Proceedings*, *3194*.

Barredo Arrieta, A., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., Garcia, S., Gil-Lopez, S., Molina, D., Benjamins, R., Chatila, R., & Herrera, F. (2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*, *58*. https://doi.org/10.1016/j.inffus.2019.12.012

Bellucci, M., Delestre, N., Malandain, N., & Zanni-Merk, C. (2021). Towards a terminology for a fully contextualized XAI. *Procedia Computer Science*, *192*. https://doi.org/10.1016/j.procs.2021.08.025

Boardman, J., Alam, M. S., Huang, X., & Xie, Y. (2022). Integrated Gradients is a Nonlinear Generalization of the Industry Standard Approach to Variable Attribution for Credit Risk Models. *Proceedings - 2022 IEEE International Conference on Big Data, Big Data 2022*. https://doi.org/10.1109/BigData55660.2022.10020687

Bücker, M., Szepannek, G., Gosiewska, A., & Biecek, P. (2022). Transparency, auditability, and explainability of machine learning models in credit scoring. *Journal of the Operational Research Society*, *73*(1). https://doi.org/10.1080/01605682.2021.1922098

Bussmann, N., Giudici, P., Marinelli, D., & Papenbrock, J. (2021). Explainable Machine Learning in Credit Risk Management. *Computational Economics*, *57*(1). https://doi.org/10.1007/s10614-020-10042-0

Cao, L. (2020). AI in Finance: A Review. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3647625

Carabantes, M. (2020). Black-box artificial intelligence: an epistemological and critical analysis. *AI and Society*, *35*(2). https://doi.org/10.1007/s00146-019-00888-w

Cardenas-Ruiz, C., Mendez-Vazquez, A., & Ramirez-Solis, L. M. (2022). Explainable Model of Credit Risk Assessment Based on Convolutional Neural Networks. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *13612 LNAI*. https://doi.org/10.1007/978-3-031-19493-1_7

Chen, Y., Calabrese, R., & Martin-Barragan, B. (2024). Interpretable machine learning for imbalanced credit scoring datasets. *European Journal of Operational Research*, *312*(1). https://doi.org/10.1016/j.ejor.2023.06.036

Dastile, X., & Celik, T. (2021). Making Deep Learning-Based Predictions for Credit Scoring Explainable. *IEEE Access*, *9*. https://doi.org/10.1109/ACCESS.2021.3068854

de Lange, P. E., Melsom, B., Vennerød, C. B., & Westgaard, S. (2022). Explainable AI for Credit Assessment in Banks. *Journal of Risk and Financial Management*, *15*(12). https://doi.org/10.3390/jrfm15120556

Dietterich, T. G., & Horvitz, E. J. (2015). Viewpoint: Rise of concerns about AI: Reflections and directions. In *Communications of the ACM* (Vol. 58, Issue 10). https://doi.org/10.1145/2770869

Ding, W., Abdel-Basset, M., Hawash, H., & Ali, A. M. (2022). Explainability of artificial intelligence methods, applications and challenges: A comprehensive survey. *Information Sciences*, *615*. https://doi.org/10.1016/j.ins.2022.10.013

Doshi-Velez, F., & Kim, B. (2017). A Roadmap for a Rigorous Science of Interpretability. *ArXiv Preprint ArXiv:1702.08608v1*.

Frame, W. S., Wall, L., & White, L. J. (2018). Technological change and financial innovation in banking. Some implications for Fintech. In *Federal Reserve Bank of Atlanta, Working Papers*.

Gramegna, A., & Giudici, P. (2021). SHAP and LIME: An Evaluation of Discriminative Power in Credit Risk. *Frontiers in Artificial Intelligence*, *4*. https://doi.org/10.3389/frai.2021.752558

Gramespacher, T., & Posth, J. A. (2021). Employing Explainable AI to Optimize the Return Target Function of a Loan Portfolio. *Frontiers in Artificial Intelligence*, *4*. https://doi.org/10.3389/frai.2021.693022

Hadji Misheva, B., Jaggi, D., Posth, J. A., Gramespacher, T., & Osterrieder, J. (2021). Audience-Dependent Explanations for AI-Based Risk Management Tools: A Survey. *Frontiers in Artificial Intelligence*, *4*. https://doi.org/10.3389/frai.2021.794996

Hadji Misheva, B., & Papenbrock, J. (2022). Editorial: Explainable, Trustworthy, and Responsible AI for the Financial Service Industry. In *Frontiers in Artificial Intelligence* (Vol. 5). https://doi.org/10.3389/frai.2022.902519

Hall, P., Cox, B., Dickerson, S., Ravi Kannan, A., Kulkarni, R., & Schmidt, N. (2021). A United States Fair Lending Perspective on Machine Learning. In *Frontiers in Artificial Intelligence* (Vol. 4). https://doi.org/10.3389/frai.2021.695301

Heng, Y. S., & Subramanian, P. (2023). A Systematic Review of Machine Learning and Explainable Artificial Intelligence (XAI) in Credit Risk Modelling. *Lecture Notes in Networks and Systems*, *559 LNNS*. https://doi.org/10.1007/978-3-031-18461-1_39

Hjelkrem, L. O., & Lange, P. E. de. (2023). Explaining Deep Learning Models for Credit Scoring with SHAP: A Case Study Using Open Banking Data. *Journal of Risk and Financial Management*, *16*(4). https://doi.org/10.3390/jrfm16040221

Kuiper, O., van den Berg, M., van der Burgt, J., & Leijnen, S. (2022). Exploring Explainable AI in the Financial Sector: Perspectives of Banks and Supervisory Authorities. *Communications in Computer and Information Science*, *1530 CCIS*. https://doi.org/10.1007/978-3-030-93842-0_6

Le, T. T. H., Prihatno, A. T., Oktian, Y. E., Kang, H., & Kim, H. (2023). Exploring Local Explanation of Practical Industrial AI Applications: A Systematic Literature Review. In *Applied Sciences (Switzerland)* (Vol. 13, Issue 9). https://doi.org/10.3390/app13095809

Liu, Y., Huang, F., Ma, L., Zeng, Q., & Shi, J. (2024). Credit scoring prediction leveraging interpretable ensemble learning. *Journal of Forecasting*, *43*(2). https://doi.org/10.1002/for.3033

Lockey, S., Gillespie, N., Holm, D., & Someh, I. A. (2021). A review of trust in artificial intelligence: Challenges, vulnerabilities and future directions. *Proceedings of the Annual Hawaii International Conference on System Sciences*, *2020-January*. https://doi.org/10.24251/hicss.2021.664

Lusinga, M., Mokoena, T., Modupe, A., & Mariate, V. (2021). Investigating statistical and machine learning techniques to improve the credit approval process in developing countries. *IEEE AFRICON Conference*, *2021-September*. https://doi.org/10.1109/AFRICON51333.2021.9570906

Meske, C., Bunde, E., Schneider, J., & Gersch, M. (2022). Explainable Artificial Intelligence: Objectives, Stakeholders, and Future Research Opportunities. *Information Systems Management*, *39*(1). https://doi.org/10.1080/10580530.2020.1849465

Minh, D., Wang, H. X., Li, Y. F., & Nguyen, T. N. (2022). Explainable artificial intelligence: a comprehensive review. *Artificial Intelligence Review*, *55*(5). https://doi.org/10.1007/s10462-021-10088-y

Moscato, V., Picariello, A., & Sperlí, G. (2021). A benchmark of machine learning approaches for credit score prediction. *Expert Systems with Applications*, *165*. https://doi.org/10.1016/j.eswa.2020.113986

Nazer, L. H., Zatarah, R., Waldrip, S., Ke, J. X. C., Moukheiber, M., Khanna, A. K., Hicklen, R. S., Moukheiber, L., Moukheiber, D., Ma, H., & Mathur, P. (2023). Bias in artificial intelligence algorithms and recommendations for mitigation. *PLOS Digital Health*, *2*(6). https://doi.org/10.1371/journal.pdig.0000278

Nelson, G. S. (2019). Bias in Artificial Intelligence. *North Carolina Medical Journal*, *80*(4). https://doi.org/10.18043/ncm.80.4.220

Nwafor, C. N., & Nwafor, O. Z. (2023). Determinants of non-performing loans: An explainable ensemble and deep neural network approach. *Finance Research Letters*, *56*. https://doi.org/10.1016/j.frl.2023.104084

Ponsam, J. G., Bella Gracia, S. V. J., Geetha, G., Karpaselvi, S., & Nimala, K. (2021). Credit Risk Analysis using LightGBM and a comparative study of popular algorithms. *Proceedings of the 2021 4th International Conference on Computing and Communications Technologies, ICCCT 2021*. https://doi.org/10.1109/ICCCT53315.2021.9711896

Rizinski, M., Peshov, H., Mishev, K., Chitkushev, L. T., Vodenska, I., & Trajanov, D. (2022). Ethically Responsible Machine Learning in Fintech. *IEEE Access*, *10*. https://doi.org/10.1109/ACCESS.2022.3202889

S Band, S., Yarahmadi, A., Hsu, C. C., Biyari, M., Sookhak, M., Ameri, R., Dehzangi, I., Chronopoulos, A. T., & Liang, H. W. (2023). Application of explainable artificial intelligence in medical health: A systematic review of interpretability methods. *Informatics in Medicine Unlocked*, *40*. https://doi.org/10.1016/j.imu.2023.101286

Saeed, W., & Omlin, C. (2023). Explainable AI (XAI): A systematic meta-survey of current challenges and future opportunities. *Knowledge-Based Systems*, *263*. https://doi.org/10.1016/j.knosys.2023.110273

Saleem, R., Yuan, B., Kurugollu, F., Anjum, A., & Liu, L. (2022). Explaining deep neural networks: A survey on the global interpretation methods. In *Neurocomputing* (Vol. 513). https://doi.org/10.1016/j.neucom.2022.09.129

Samek, W., & Müller, K. R. (2019). Towards Explainable Artificial Intelligence. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and*

Lecture Notes in Bioinformatics): Vol. 11700 LNCS. https://doi.org/10.1007/978-3-030-28954-6_1

Sathe, S. S., & Mahalle, P. (2023). Predictive Analytics in Financial Services Using Explainable AI. *Lecture Notes in Networks and Systems*, *641 LNNS*. https://doi.org/10.1007/978-981-99-0483-9_35

Shaw, J. (2019). Artificial Intelligence & Ethics. *Harvard Magazine*.

Shukla Shubhendu, S., & Vijay, J. (2013). Applicability of Artificial Intelligence in Different Fields of Life. *International Journal of Scientific Engineering and Research (IJSER)*, *1*(1).

Spartalis, C. N., Semertzidis, T., & Daras, P. (2024). *Balancing XAI with Privacy and Security Considerations*. https://doi.org/10.1007/978-3-031-54129-2_7

Sriram, A., Gorti, S. S., Amin, E. G., & Kumar, A. (2022). Analyzing Banking Services Applicability Using Explainable Artificial Intelligence. *ACM International Conference Proceeding Series*. https://doi.org/10.1145/3549206.3549259

Walambe, R., Kolhatkar, A., Ojha, M., Kademani, A., Pandya, M., Kathote, S., & Kotecha, K. (2021). Integration of Explainable AI and Blockchain for Secure Storage of Human Readable Justifications for Credit Risk Assessment. *Communications in Computer and Information Science*, *1368*. https://doi.org/10.1007/978-981-16-0404-1_5

Wamba-Taguimdje, S. L., Fosso Wamba, S., Kala Kamdjoug, J. R., & Tchatchouang Wanko, C. E. (2020). Influence of artificial intelligence (AI) on firm performance: the business value of AI-based transformation projects. *Business Process Management Journal*, *26*(7). https://doi.org/10.1108/BPMJ-10-2019-0411

Watson, D. S. (2022). Conceptual challenges for interpretable machine learning. *Synthese*, *200*(1). https://doi.org/10.1007/s11229-022-03485-5

Weber, P., Carl, K. V., & Hinz, O. (2024). Applications of Explainable Artificial Intelligence in Finance—a systematic review of Finance, Information Systems, and Computer Science literature. *Management Review Quarterly*, *74*(2). https://doi.org/10.1007/s11301-023-00320-0

Wu, H. D., & Han, L. (2021). A novel reasoning model for credit investigation system based on Fuzzy Bayesian Network. *Procedia Computer Science*, *183*. https://doi.org/10.1016/j.procs.2021.02.060

Xu, R., Meng, H., Lin, Z., Xu, Y., Cui, L., & Lin, J. (2021). Credit Default Prediction via Explainable Ensemble Learning. *ACM International Conference Proceeding Series*. https://doi.org/10.1145/3503181.3503195

Zednik, C. (2021). Solving the Black Box Problem: A Normative Framework for Explainable Artificial Intelligence. *Philosophy and Technology*, *34*(2). https://doi.org/10.1007/s13347-019-00382-7

Zhang, L., Wang, J., & Liu, Z. (2023). What should lenders be more concerned about? Developing a profit-driven loan default prediction model. *Expert Systems with Applications*, *213*. https://doi.org/10.1016/j.eswa.2022.118938

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação

Universidade Nova de Lisboa