NOVA IMS Information Management School

MGI

Master Degree Program in

Information Management

ADVANCING MODEL TRANSPARENCY IN NATURAL LANGUAGE PROCESSING

A case study of Explainable AI at Banco de Portugal

Volodymyr Mykhayliv

Project Work

presented as partial requirement for obtaining the Master Degree in Information Management

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

ADVANCING MODEL TRANSPARENCY IN NATURAL LANGUAGE PROCESSING

A case study of Explainable AI at Banco de Portugal

by

Volodymyr Mykhayliv

Project Work presented as partial requirement for obtaining the Master's degree in Information Management, with a specialization in Knowledge Management and Business Intelligence.

Supervised by

Carlos Tam Chuem Vai, PhD, NOVA Information Management School

July, 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, July 2024

ACKNOWLEDGEMENTS

I would like to express my deepest gratitude to all those who have supported me during my master's project journey. All the guidance, time, encouragement, and support have been invaluable for the conclusion of this work.

I am especially grateful to my supervisor, Professor Carlos Tam Chuem Vai, whose knowledge, patience, and insightful feedback have been instrumental in the completion of this project. His continuous support and guidance have greatly contributed to the quality of my work.

I would also like to extend my sincere thanks to my colleagues at Banco de Portugal. Their suggestions on the project's topic, along with their support and availability, have been crucial in shaping the direction and outcome of this research. I am particularly grateful to Sara Vaz Cândido, Ricardo Carvalho, Luís Marcos, Elisabete Santos, and Natalina Ribeiro for their valuable insights and encouragement throughout this process. Additionally, I would like to thank my team at Banco de Portugal, mainly Baltasar Cordeiro, Filipe Fernandes, Cláudia Pinto, Alexandra Gonçalves, and Pedro Dantas for their patience and support. Their understanding and cooperation provided me with a great environment necessary to develop this work effectively.

Finally, I want to express my gratitude to my family and friends, namely my girlfriend, Sofia Nunes, for your amazing support. To my sister, Liliya Mykhayliv, for your help, encouragement, and motivation. And my friend, Ricardo Correia da Silva, for your assistance and invaluable feedback.

ABSTRACT

The efficiency of Machine Learning (ML) algorithms and projects presents a tremendous opportunity for better decision-making in the activity of financial institutions. However, the growing complexity of ML algorithms proves to be a considerable obstacle to the widespread adoption of these technologies due to the lack of transparency and interpretability, deterring users from understanding the reasoning for the decisions made by the model. Despite the clear value and success of these ML models, uncertainty and concerns from end users intensify this issue, forcing them to blindly trust the outcomes, especially in the context of Natural Language Processing (NLP) projects. At Banco de Portugal, the implementation of classification models for information requests has been very successful. However, the lack of explainability of some results poses an obstacle to the full adoption of some models by the business, resulting in constant skepticism about the results and fear of misinformation. We propose a framework that systematically addresses these challenges, with the final goal of overcoming these obstacles is a crucial step to develop trust, reduce doubt and ensure a smooth integration of ML technologies in the decision-making practices of financial institutions.

KEYWORDS

Explainable artificial intelligence; natural language processing; evaluation methods; trustworthy artificial intelligence; explainability

Sustainable Development Goals (SDG):

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS AND ACRONYMS

AI	Artificial Intelligence
AIX360	AI Explainability 360
BERT	Bidirectional Encoder Representations from Transformers
CNN	Convolutional Neural Network
DNN	Deep Neural Network
ML	Machine Learning
MEMS	Micro Electro-Mechanical Systems
NLP	Natural Language Processing
XAI	Explainable AI
LIME	Local Interpretable Model-agnostic Explanations
SHAP	SHapley Additive exPlanations
PDPs	Partial Dependence Plots
PSO	Particle Swarm Optimization
ELI5	Explain Like I'm 5
XGBoost	Extreme Gradient Boosting

1. INTRODUCTION

One of the pillars of the modern computing revolution is Artificial Intelligence (AI), a set of technologies that, in a simple manner, empowers computers to execute relatively easy or extremely sophisticated tasks by imitating human intelligence (Love et al., 2023). Many common AI tasks consist of language understanding, data analysis and visualization, shopping cart recommendations, visual perception, health, cyber-security and even food fields (Gnanasankaran et al., 2022). Over the last few years, it became clear that AI can be a driving power to unlock the hidden potential of individuals or businesses, serving as a backbone to innovation, creativity, and automation (Haque et al., 2023). For example, understanding the sentiments of individuals or clients has become an essential step of decision-making process, and sentiment analysis, a branch of NLP allows us to analyze and quantify a positive, negative, or neutral sentiment based on a body of text, using computational methods while being where human-to-machine communication is crucial (Abonizio et al., 2022).

Some algorithms are extremely user-friendly, thus display a much higher level of clearness and interpretability for the end users, potentially becoming a more favorable option independently of the provided results. However, the focus of current AI research clearly prioritizes ML models based on statistical performance, and not purely preference, one of the greatest examples being neural network algorithms, that simply surpass trivial ML algorithms in terms of performance. While these ML algorithms are very helpful in decision-making processes, the backbone of their actions are still hidden or not explained to the end-users, missing the explanation for the made decisions (Q. Wang et al., 2020).

There are different approaches and techniques to evaluate the performance of an algorithm, such as accuracy, precision, recall or F1 score, each serving a different objective in understanding it's value and painting a different picture about the algorithm's ability to correctly classify instances (Doulah, 2023). These methods allow us to understand algorithm's strengths and limitations in different applications, but it's not always easy to tell us which ML algorithm's result is more transparent and easier to figure out by the end user, that are eager to be informed about advice provided by the deployed models (Lehmann et al., 2022).

Lack of transparency in AI models is a big challenge and a threat to their widespread adoption and acceptance. If algorithms create "black boxes", that undermine trust by hiding the logic behind their decisions, stakeholders quickly feel the need of comprehending what is the reason for specific results (Lehmann et al., 2022). This lack of trust is especially noticeable in fields such as finance, healthcare, or even criminal justice, where decisions have extreme consequences, need to be made quickly and without a shadow of doubt (Zahoor et al., 2024).

In hopes of addressing this issue, the field of explainable AI (XAI) has been rising exponentially with hopes of enhancing the clarity of AI models and bridging the existing gap between complex algorithms and user comprehension to create trust and deploy AI technologies in a responsible way (Weber et al., 2023). How to find the right tools and frameworks to

understand ML models? XAI is a field in a constant transformation that through techniques like rule-based systems or feature importance analysis tries to demystify complex algorithms, allowing users to grasp the logic behind each prediction (Dieber & Kirrane, 2022).

The relevance of the research reported in this study was concluded from the observation of the constant growth of the field of XAI and its consequent evolution and importance in decision-making within financial institutions (Quinn, 2023). The way different stakeholders utilize and interact with the different ML models to achieve organizational goals motivated this research. But the application of XAI techniques without transparency and easy interpretability creates implications for the traditional approaches (Deshpande & Ambatkar, 2023).

The contributions of this study benefit not only financial institutions (from IT professionals to decision-makers across the board), but also AI and ML professionals. If managers can understand and trust the outputs of ML models, it will improve the confidence in decision-making and task planning that is appropriate for their specific teams, according to their level of performance and maturity, which will therefore increase the motivation of the members of those teams. On the other hand, stakeholders, and teams, more precisely their members, benefit in terms of productivity and satisfaction, resulting from enhancements in transparency and interpretability is processes supported by these models. Lastly, projects, both the simplest as well as the most complex ones, have a greater probability of success, resulting in the consumption of less resources to achieve the established goals. The constant emphasis on transparency, interpretability and resource optimization indicates the potential efficiency obtained through strategic and well-defined integration of XAI techniques in pursuing quick and effective outcomes while minimizing resource expenses. This study proposes to answer the following research question (RQ):

RQ: How can XAI methodologies improve the interpretability of ML models and provide trustworthy outcomes?

For this matter, the suggested question will be approached from various perspectives, allowing to investigate several aspects that can be later combined in meaningful manner. Initially, this study will try to identify adequate XAI techniques to improve the interpretability and transparency of ML models in the context of different projects and/or institutions. Afterwards, the impact of these techniques will be observed and assessed, focusing on their influence on decision-making processes, trustworthiness of results and outputs confidence. Finally, this dissertation will try to build the necessary steps to develop a trustworthy and easily interpretable methodology that can be successfully replicated in the future.

The goals of this study consist of a) providing a comprehensive definition of XAI and its relevance and importance in ML projects, by exploring the existing libraries and solutions; b) creating a framework/methodology of a practical application of XAI in NLP project at Banco de Portugal; c) teaching how to facilitate, from an operational point of view, the application

of XAI to boost trust and interpretability; and finally, d) summarizing key recommendations and reflections about the challenges faced during the project.

The remainder of this manuscript is organized as follows: ***Section 2*** presents a comprehensive Literature Review on XAI, diving into many key concepts including transparency, interpretability and a variety of XAI techniques that are valuable for this research. This section also explores the problem of misinformation, the adaptability of XAI, and the description of important evaluation metrics for explainability while reviewing related work to establish a context for this study.

Following this, ***Section 3*** outlines the deployed methodology, beginning with the conceptual model that frames the theoretical foundation. Then, this section describes the design and definition of the proposed framework for enhancing XAI, detailing steps such as aligning with business context and strategy, assessing AI capabilities, selecting and preparing data, and explaining model implementation through visualization techniques and performance evaluation.

Section 4 describes the practical set-up and implementation of the experiment, including an operational snapshot of Banco de Portugal, the integration of AI within the institution, and the meticulous process of data selection and preparation. It also provides a detailed explanation of the model, the application of various visualization techniques, and the evaluation metrics used to assess model performance.

In ***Section 5***, the results of the study are presented and discussed, examining both theoretical and practical implications, the benefits and limitations of XAI integration, and insights gained from misclassified cases, alongside the impact of feedback integration on continuous improvement.

Finally, ***Section 6*** concludes the study, summarizing the key findings and contributions, offering recommendations for future research, and emphasizing the importance of transparency and interpretability in AI models for enhancing trust and adoption in financial institutions.

2. LITERATURE REVIEW

XAI has become an increasingly significant topic of research in the field of ML, with a big emphasis on different concepts, taxonomies, utilization opportunities, and daily challenges that bring us closer to the responsible utilization of AI (Barredo Arrieta et al., 2020). XAI is a leading field of study withing ML world that focuses on enhancing transparency and interpretability of AI models (Laato et al., 2022; Saraswat et al., 2022). Key concepts in XAI comprise transparency, interpretability and explainability in AI systems. XAI attempts to provide easy, human-understandable explanations, closing the gap between human comprehension and complex algorithms.

2.1. XAI concepts

In the context of explainability, it is important to understand the notions of global and local explainability, and situations in which each might be a more appropriate solution (Shen, 2022). This raises the question of how explainability can be achieved in AI models. An answer to this question emerged over time through model-specific and model-agnostic approaches, by providing different standpoints on how to transform AI systems into more transparent ones. As an example, local interpretable model-agnostic explanations have been suggested as a solution to visualize the impact of combining relationship between different features and the interpretation process (Z. Chen et al., 2023).

In the field of NLP, relevant for this study, the concepts of readability, plausibility, and faithfulness are essential in developing faithful and understandable NLP systems (Lanham et al., 2023) and understanding them was indispensable for ensuring that NLP explanations were clear and faithful to the model's reasoning process. The complex field of XAI is extensive and multi-layered, with a big assortment of methodologies and techniques for achieving transparency, interpretability, and faithfulness. To build trustworthy AI systems, we need to study the trade-offs between different techniques, broaden our understanding of Human-AI interactions and apply our knowledge to NLP (Barredo Arrieta et al., 2020; Qu & Hullman, 2016; Sood & Craven, 2021).

2.2. The problem of misinformation and threats of explainability

With the rise of AI classifications, wrong results and the challenges faced by regulatory frameworks of addressing the need for transparency and accountability has become a growing concern too (Elkhatat et al., 2023). Fields such as journalism and archives have seen a massive increase in fact-checking initiatives, driven by the emergence of dozens, if not hundreds, of active fact-checking interventions globally (Amazeen, 2017). However, the scalability and the necessity to combat the increase of misinformation on the Internet continues to be a massive challenge. To combat this, we can observe a growing interest on leveraging NLP to partially or fully automate fact-checking processes (Glockner et al., 2022). One of the key fields of study has extensively focused on studying AI detection tools in differentiating between AI-generated

and human text. It became clear that AI detection tools are much more proficient in identifying specific pieces of content, such as generated by GPT3.5, while struggling with GPT 4.5 (Elkhatat et al., 2023). Additionally, it was suggested that fact-checking initiatives might be spreading as a response to challenges and perceived weakness of democratic institutions (Amazeen, 2017). Lastly, some research indicates that current NLP fact-checking technologies are struggling to combat real-world misinformation and it is safe to assume that future research will undoubtedly explore the connection of AI, misinformation and undergoing efforts of promoting transparency in information systems (Glockner et al., 2022).

Unfortunately, XAI is not only good news, in the field of XAI systems, there is a growing fear of potential misuses of XAI tools such as LIME (Local Interpretable Model-agnostic Explanations) and SHAP (SHapley Additive exPlanations) in different fields, as an example, adversarial attacks. It was clearly shown that SHAP values can be exploited to influence model predictions, leveraging immediate security questions in applications such as fake news detection or even automated adversarial sample generation during conflicts or elections (Slack et al., 2020). The development of different detection methods that rely on SHAP further contribute to this domain, in efforts to distinguish normal and adversarial inputs. By applying these methods on diverse state-of-the-art attaches, researchers show the potential of XAI methodologies on detection of adversarial examples (Fidel et al., 2020). We can already see the dual nature of XAI, as a tool for interpretability but also a potential protection against ongoing threats. Additionally, the literature suggests a cross-domain strategy that leverages different XAI models such as Anchor, ELI5, LIME and SHAP to design a new approach for detecting wrong predictions in different fields (Kanneganti, 2023). This is a clear demonstration of multidisciplinary application of XAI techniques in addressing misinformation challenges (Dobrovolskis et al., 2023; Tiwari, 2023).

2.3. XAI flexibility and adaptability

AI models continue to evolve, making the field of explainability in AI increasingly important but complex. The necessity of transparency in AI systems resulted in the exploration of fundamentally interpretable models such as decision trees, rule-based models, and linear models. By having the facility to provide explanations for their predictions, important aspects in gaining trust and acceptance in various fields, these models became the focus of many researchers(Arenas et al., 2022).

As an example, Garvin et al. (2020) used novel spatiotemporal and XAI models to find possibly adaptive mutations of the SARS-CoV-2 virus (Garvin et al., 2020). Despite focusing on biological data, we can clearly see the importance of XAI in understanding complex fields while utilizing decision trees and rule-based models. Another example presented by Akula et al. demonstrates the capacities of XAI in explaining user queries, focusing on transparent communication (Akula et al., 2019). On the other hand, Ferigo et al. (2023) explored the quality-diversity optimization to improve the interpretability of decision trees for reinforcement learning, by showing the potential of enhanced decision trees (Ferigo et al.,

2023). More in line with our study, Berka explored sentiment analysis using rule-based reasoning which aligns with understanding and interpreting sentiment in text data (Berka, 2020). All these studies contributed collectively to our understanding of the potential of interpretable models in AI systems by shedding some light on strengths and restrictions of these models in different fields (Mohammed & Shehu, 2023). As AI continues to grow and gets new applications, the need for explainability and transparency will only continue to grow, making the research in this field progressively more important (Haresamudram et al., 2023; Patel et al., 2024).

2.4. Evaluation metrics for explainability

As XAI systems become increasingly prevalent, their assessment for effectiveness and reliability has emerged as a critical research focus (Hsiao et al., 2021). Currently, various essential evaluation metrics are used to assess the quality of AI models. Vast literature demonstrates the application of these metrics in real-world scenarios (Bodria et al., 2023; Brandt et al., 2023) such as medical applications and image classification models. Several studies introduced robust evaluation metrics included in the AI Explainability 360 toolkit (Arya et al., 2022). LI et al. (2022) stressed the evaluation of explanation quality based on interpretability and fidelity attributes (LI et al., 2022) and Nauta et al. emphasized the need for quantitative evaluation methods to guarantee robust outcomes (Nauta et al., 2023). Additionally, DoX metric was introduced to quantitatively evaluate textual information's degree of explainability (Sovrano & Vitali, 2022). These studies, and many more, collectively display the importance of objective evaluation metrics to evaluate the performance of XAI methods across different domains (R. R. Hoffman et al., 2023).

2.5. XAI techniques

Several methods and tools aimed at addressing the necessity of interpretability in ML models have emerged over the last years. Some of the most prominent techniques in this field include LIME, SHAP, AIX360 (AI Explainability 360), Partial Dependence Plots (PDPs, Decision Trees and Rule-based Models, Integrated Gradients, Anchors, ELI5, and Graph LIME (Alabi et al., 2023; Baghbani et al., 2023; Hussain, 2023; Laatifi et al., 2023; Lundstrom et al., 2022, 2022; Marcinkevics & Vogt, 2023; Uddin et al., 2023; Van Quan, 2023).

The literature clearly showcases the diverse array of XAI methodologies and techniques and their applications across various fields, emphasizing the importance of interpretability in ML models. These studies greatly contribute to the global understanding and development of explainable ML methods, answering the growing need for transparent and interpretable AI systems. For our study, and before moving into related work, we explored some benefits and disadvantages of the most common XAI techniques (Kırboğa et al., 2023).

The landscape of XAI is rich with methodologies addressing the interpretability of ML models. A critical analysis reveals techniques such as LIME that distinguishes itself for its model-agnostic nature, allowing it to function across various ML models and data types. This

versatility makes it a favored choice for interpreting complex models, as it generates simpler, more understandable surrogate models. However, its reliance on random sampling can introduce significant unpredictability, leading to questions about the consistency and reliability of its explanations. This aspect becomes especially critical when dealing with critical decisions, where explanation fidelity is preeminent (Biecek et al., 2021; Zhang et al., 2019).

On the other hand, SHAP is a concept in cooperative game theory. SHAP assigns each feature an importance value for a particular prediction, considering all possible combinations of features. This comprehensive approach ensures unbiased feature attribution, which is a key factor in interpretability. Despite its robust theoretical foundation, SHAP can often face computational challenges, predominantly with large-scale, high-dimensional data, where the computation of SHAP values becomes resource-intensive. This constraint is particularly pertinent in real-time analysis scenarios (Bifarin, 2022; Hamilton & Papadopoulos, 2024; Parsa et al., 2020).

Additionally, AIX360 is an open-source toolkit for transparency enhancement in AI models that is part of IBM Research's AI Fairness 360 initiative. The big benefit is the possibility it gives for the integration of multiple explanation methods and user choice of the method to use, depending on the situation. Moreover, it is highly documented and has tutorials, allowing novices and experts to understand and use the toolkit. However, since AIX360 is very broad in the kind of things it does, it could become quite complex to navigate, while mastering it could become a great advantage in the XAI field (Arya et al., 2021).

Moving into more well-known methods, decision trees provide a straightforward, hierarchical structure that visually represents decision-making processes, making them intrinsically interpretable. Each node in a decision tree corresponds to a feature-based decision, offering a clear trail of the model's reasoning. In a similar way, rule-based models offer a direct, human-readable set of rules for each outcome. Contrasting LIME and SHAP, these models are designed with interpretability as a core feature, rather than as an accessory, supporting closely the principles of XAI. However, their simplicity can be a disadvantage for complex datasets, where they may not capture subtle patterns as effectively as more refined models (Costa & Pedreira, 2023; Fürnkranz et al., 2018; Izza et al., 2022).

Anchor-based explainers are another personalized method for complex, black-box models. As the name suggests, anchors strive to find minimal sets of conditions or 'anchors' that, when met, lead to consistent and reliable predictions. This method is predominantly useful for providing insight into individual predictions, offering a clear and concise explanation in cases where models make critical decisions. However, similarly to other local explanation methods, Anchors focus on individual instances and may not provide an all-encompassing understanding of the model's overall behavior (Lopardo et al., 2023).

On the other hand, ELI5 (Explain Like I'm 5) is designed to clarify complex ML models, offering simplified explanations that are comprehensible to non-experts. Its primary strength stands

in its user-friendly interface and ability to work with many ML models. Nevertheless, its simplicity can be a double-edged sword; while it makes ML more approachable, it may oversimplify complex model behaviors, leading to incomplete or deceptive interpretations. This is particularly evident in models that involve complicated interactions between features or when working with highly complex data structures (Sharma et al., 2023).

Lastly, Graph LIME, an extension of LIME, focuses in explaining predictions for graph-structured data, such as the ones used in social network analysis, bioinformatics, or recommendation systems. It adjusts the principles of LIME to fit the specific characteristics of graph data, including the relationships and interdependencies between nodes. Despite its specialized approach, Graph LIME also comes with the limitations of LIME, which include the potential instability occurring from its sampling process. This aspect can be of great challenge in graphs with large, complex structures, where ensuring representative and stable sampling is crucial (Costi et al., 2024).

The array of methodologies within XAI highlights the field's dynamic nature and the ongoing efforts to make ML models more transparent and understandable. Each of the methods brings a unique perspective to the table, contributing to a more comprehensive and multifaceted approach to explainability in AI (Khani et al., 2024). While techniques like LIME and SHAP offer robust model-agnostic explanations, they often are faced with challenges such as unpredictability and computational intensity. On the other hand, tools like AIX360 provide flexibility with multiple methods but can quickly become complex to navigate. Decision trees and rule-based models are amazing in terms interpretability but may struggle with complex datasets. Anchor-based explainers and ELI5 offer simplified explanations but can be limited in scope. Graph LIME addresses graph-structured data but faces the same instability issues as LIME (Chaddad et al., 2023; Dwivedi et al., 2023).

Overall, the diverse XAI landscape demonstrates the importance of selecting appropriate techniques based on specific needs and contexts, emphasizing the continuous evolution and refinement of explainability methods in AI research and practice (Tiwari, 2023).

2.6. Related work

For the last two decades, text classification and sentiment analysis have been the focal point of computational linguistics, leading to the development of sophisticated AI systems with remarkable accuracy. Despite these advancements, the challenge to make these systems transparent and user-friendly persists (Miller, 2019). Integration of XAI into the text classification and sentiment analysis process significantly enhances the interpretability of AI systems. To further advance their trust and usability, researchers in this area combine findings from social sciences, deep learning, and other techniques to pave the way for AI applications to be more transparent and user-friendly in the future (Al-Khazaleh et al., 2023).

In this section, we succinctly highlight the summary of pivotal research in this domain, that bridges the gap using XAI in text classification, sentiment analysis and other similar domains. These studies combine insights to build a way for more transparent and user-friendly applications, capable of figuring the generation of natural language explanations (Valentino & Freitas, 2022).

The exciting results and methodologies presented in ***Table 1*** are not only a testament to the field, but also a backbone and inspiration for this project that will try to contribute and deliver another perspective to XAI world.

Table 1 - Overview of the related studies

Study	XAI Technique	NLP Model/Application	Key Findings
(Liu & Liu, 2022)	SHAP and Particle Swarm Optimization (PSO)	Geoscience (Permeability Prediction in Tight Sandstone Reservoirs)	Integrated Explainable ML (XGBoost) and PSO with SHAP for interpretability, enhancing prediction accuracy in reservoir evaluation compared to traditional methods.
(Kurasinski & Mihailescu, 2020)	Attention Mechanism in CNN and BERT	Fake news detection in text classification	Investigated two deep learning architectures, BiDir-LSTM-CNN and Bidirectional Encoder Representations from Transformers (BERT), for fake news detection, accenting the need for explainability in model selection and assessment.
(Blanco-Justicia & Domingo-Ferrer, 2019)	Decision Trees	Black-box ML models	Focused on using decision trees as a surrogate model to explain black-box ML models. It aims to find a balance between comprehensibility and privacy of subjects in the training data.
(Manoharan et al., 2023)	XAI with micro electro-mechanical systems (MEMS)	Commercial communication systems	Observed the effectiveness of integrating XAI with MEMS in banking and finance applications, enhancing transparency and data security.
(Golizadeh Akhlaghi et al., 2021)	SHAP, Deep Neural Network (DNN)	Optimization in energy systems	Employed SHAP in DNNs to interpret feature contributions for optimizing dew point cooler performance.
(Kozik et al., 2024)	SHAP	Fake news detection	Showcased how SHAP can be used in adversarial attacks to manipulate fake news detection systems, revealing a dual nature of XAI in security and transparency.
(Zahoor et al., 2024)	LIME	Text classification and sentiment analysis	Demonstrated the necessity of explainability in evaluating the fairness, correctness, and reliability of text classification and sentiment analysis models using LIME.
(Y.-C. Wang & Chen, 2024)	Decision Tree-Based Interpretation, Dynamic Transition and Contribution Diagrams, Improved Bar Charts	Genetic algorithms in job scheduling	Proposed new XAI techniques to address the challenges of explaining genetic algorithms in job scheduling, focusing on high-dimensional data handling and visualization.

2.7. Challenges and future directions

There's no doubt that the field of XAI is being heavily researched and developed, while trying to combat different challenges of providing transparent and understandable AI systems (R. Hoffman et al., 2018) and integrating existing digital infrastructures. In addition, growing concerns over data privacy and security need well defined measures to protect customers' and companies' information while leveraging AI capabilities from the operational point of view. The threat of algorithmic bias further complicates the big picture, where the potential misuse can destroy public trust in new technological advancements (Nyberg et al., 2024).

The process of integration of AI technologies in existing infrastructures is a challenge because it needs to be flawless, avoiding any potential disruptions for ongoing operations. These difficulties still extend to the world of explainability, where tools like LIME and AIX360 aim to clarify decision-making processes, yet true transparency still looks miles away and just an illusion (Adamson, 2021).

Moving forward, the future of AI and XAI is filled with opportunities for innovation, creativity, and collaboration. Improved AI models promise to leverage edge-cutting technology and their applications need to move forward into areas like fraud detection and market analysis in combination with robust governance frameworks to ensure accountability, fairness, and ethical usage (Li et al., 2022; Zhou et al., 2020).

3. METHODOLOGY

3.1. CONCEPTUAL MODEL

Figure 1 represents the conceptual model serves as a theoretical scheme that outlines the relationships between different components within the study. It helps in understanding the abstract structure and high-level functions of the analyzed ML model. This includes the objectives, inputs, AI/ML models, data sources, and expected outcomes and provides a visual and narrative guide to how the system is expected to work (Maass & Storey, 2021).

Objective: Enhance understanding and trust in AI-driven classification of banking complaints using XAI techniques.

Data Input: Digital images containing advertisement content.

ML Model: Banco de Portugal's automatic classification system for advertising practices.

XAI Tools: Combination of resources that provide a comprehensive understanding of the model's decisions, enhancing transparency and trust in AI systems by addressing both overall and case-specific reasoning behind model outputs.

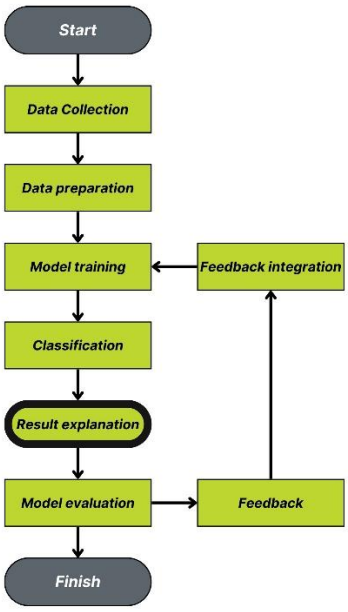

Figure 1 - Conceptual model representation

Expected Outcomes: Implementing a clear, user-friendly ML model for advertisement classification can lead to improved efficiency in handling and classifying data. Prioritizing transparency in the AI decision making process can also improve the external image and lastly, using ML can greatly contribute to reduced response times.

3.2. DESIGNING THE FRAMEWORK FOR ENHANCING XAI

In this section, we will provide a refined framework that aims at improving the transparency of AI systems. By incorporating explanatory tools such as LIME and AIX360, along with different visualization techniques, this framework is structured to provide clear insights into ML model decision-making process by following key concepts presented in ***Table 2***. Our approach is designed to significantly enhance the framework's adaptability across diverse applications, currently in use or still in development, ensuring that stakeholders can trust and understand AI-driven decisions. The framework translates the conceptual model into actionable steps and provides detailed guidelines on how to implement the model.

Table 2 - Proposed framework's key components

Framework's key components	
Interpretability techniques	Our primary objective is to enhance the transparency and comprehension of analyzed ML models. The selected interpretability techniques such as LIME and AI360, which are successfully used for unraveling the influence of individual features on model predictions and can be applied to different ML models (Ribeiro et al., 2016).
Visualization techniques	Data visualization is a key step in making the unpredictable and often complex behaviors of ML models accessible and clear to users. By highlighting areas of significance in data input and output it is expected to visually illustrate how decisions are made (Karran et al., 2022).
Model agnosticism	Our framework is designed to be model agnostic, supporting an extensive range of deep learning architectures, including convolutional neural networks (CNNs), widely used in image processing, but also transformers, applied in NLP. This flexibility ensures compatibility with emerging technologies without major adjustments.
Explanations	The key aspect of our framework is the generation of detailed, user centered explanations. It is expected to produce explanations that accommodate different levels of expertise, from non-technical end-users to field experts by ensuring that explanations are not only informative but also easy to understand.
Model evaluation	To ensure the effectiveness of our explanations, it is expected to implement both quantitative and qualitative assessments, by measuring the explanation fidelity, understandability, and practical utility.

3.3. FRAMEWORK DEFINITION AND OVERVIEW

Our framework is an outcome of a comprehensive literature review that guided us to best practices, based on previous findings within the field. It seeks to combine transparency and strategy into the development and deployment of internal AI systems, ensuring that these technologies not only produce good results, but are also both understandable and effectively integrated. This flowchart (***Figure 2***) presents an overview of the proposed framework designed to enhance the explainability of ML models. By synthesizing the concepts and strategies into a structured visual representation, this flowchart serves as a practical guide for implementing transparency and interpretability:

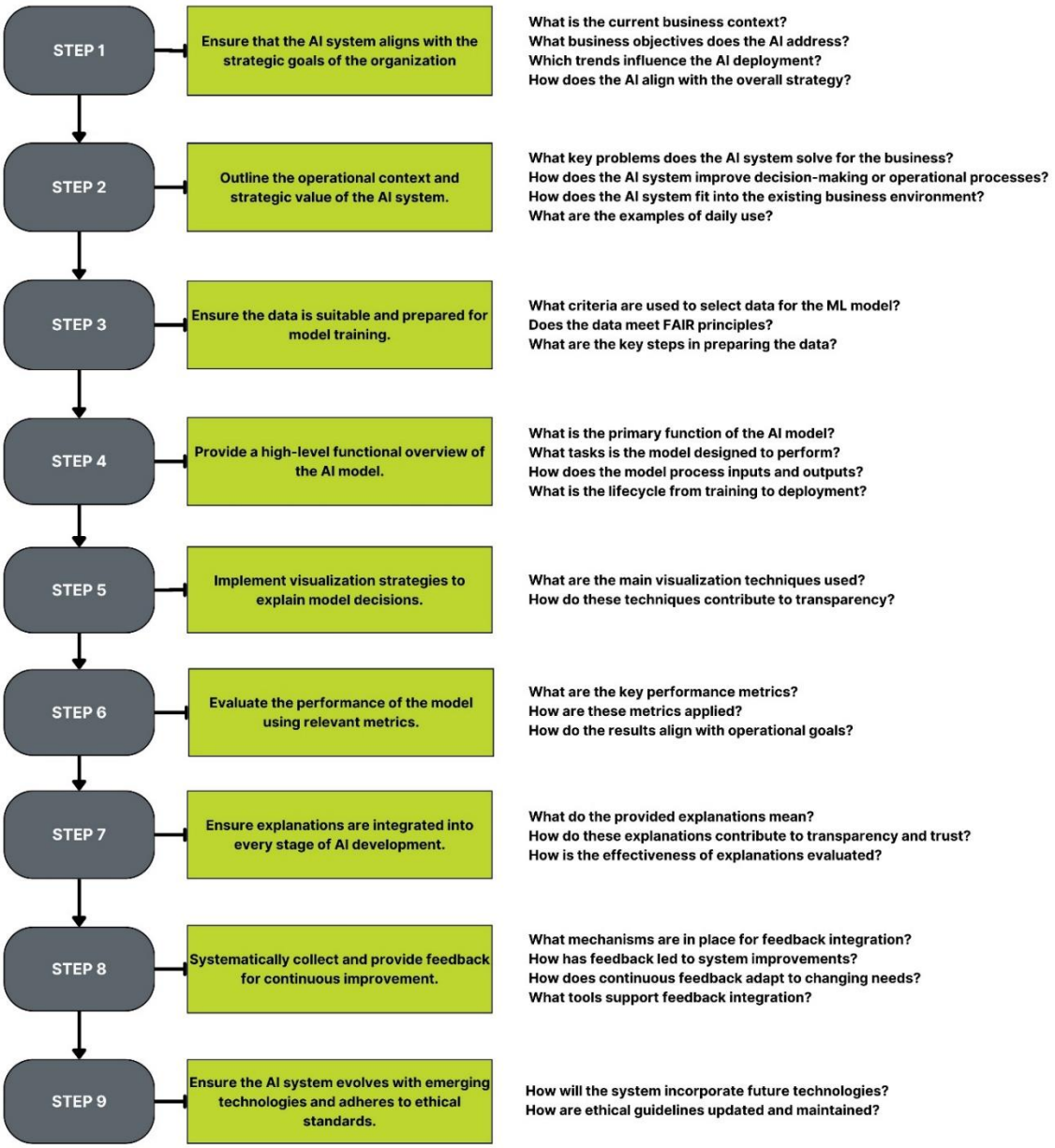

Figure 2 - Flowchart of the proposed framework

STEP 1: Business context and strategic alignment

Objective: Ensure that the AI system aligns with the strategic goals of the organization

In the first step, the industry or sector where target ML model will be deployed will be investigated, analyzing its unique challenges and opportunities to support the AI adoption. This breakdown will include an examination of current industry trends such as digital transformation, business necessities and advancements in ML.

This section will detail the business objectives supported by the ML system, being them enhancing customer engagement, increasing operational efficiency, or improving business strategies. Additionally, here will be defined the objectives to justify the system's development. In this section, by focusing on the business strategy and its role in achieving long-term goals, it should be possible to answer the following questions:

- **What is the current business context?**
- **What business objectives does the AI address?**
- **Which trends influence the AI deployment?**
- **How does the AI align with the overall strategy?**

STEP 2: AI capabilities overview

Objective: Outline the operational context and strategic value of the AI system.

The second step of the proposed framework explores the strategic significance and potential impact of the AI system within its business environment. It aims to describe and clarify the operational context and strategic value in the analyzed industry.

To address this, the motivation behind the AI system's development will be discussed, exploring the existing challenges or needs. This includes outlining the problems it aims to solve and highlighting expected benefits and development objectives. Additionally, real-world application scenarios should be indicated, detailing user interactions and operational use cases to provide a clear understanding of daily activities.

This structured approach aims to lay a solid foundation revealing not only the technical workings but also the rationale and practical utility of the ML system. By carefully addressing what the system is, why it was built, and how it is used, the framework provides comprehensive insight into its design and effectiveness in achieving operational goals.

Overall, this section of the framework will offer a structured and analytical overview of the practical applications of the system and aims to answer the following questions:

- **What key problems does the AI system solve for the business?**
- **How does the AI system improve decision-making or operational processes?**

- **How does the AI system fit into the existing business environment?**
- **What are the examples of daily use?**

STEP 3: Data selection and preparation

Objective: Ensure the data is suitable and prepared for model training.

In this section of the framework, the focus will fall on exploring the critical steps involved in choosing and preparing the appropriate dataset necessary for training the model.

The selection criteria for the data will be detailed, focusing on the importance of abiding by the FAIR principles—ensuring that the data is findable, accessible, interoperable, and reusable. These conditions are vital to guarantee that datasets not only meet the technical requirements of the model but also comply with high data governance standards (Barker et al., 2022).

Next, data sourcing strategies will be explored. That include identifying potential data sources, evaluating their credibility and relevance, and outlining the methods for acquiring data while ensuring it respects privacy and ethical guidelines.

Lastly, the data preparation processes will be covered by describing the steps involved in data cleaning, normalization, and transformation to convert raw data into a format that is suitable for model training. By doing that, the goal is to provide detailed insights into useful techniques for handling missing data, correcting anomalies, and enhancing data quality.

Overall, this section aims to provide a comprehensive guide on how to effectively select, source, prepare, and preprocess data, ensuring it is ready for use in AI model training. By addressing each of these aspects, the prepared dataset maximizes the potential of the analyzed ML system and expect to answer the following questions:

- **What criteria are used to select data for the ML model?**
- **Does the data meet FAIR principles?**
- **What are the key steps in preparing the data?**
- **How do these processes enhance quality and reliability?**

STEP 4: Model description

Objective: Provide a high-level functional overview of the AI model.

In this section of the framework, a brief overview of the model's architecture will be provided, while approaching it as a black box to focus on its application rather than its internal complexities and developments. This approach is useful for anyone who's mainly interested in the model's capabilities and outputs without the need for programming expertise and deep diving into the technical specifics.

The purpose and functionality of the model will be outlined, describing what it is designed to do and the tasks it is made to perform, such as image recognition, data prediction, or another specific purpose.

Additionally, the input-output mechanisms of the model will be discussed, focusing on how data is received and what outputs or results are produced. To complement this information, the general lifecycle of the model will be explored, moving from training to deployment, reviewing the key stages without delving into the technical methodologies.

In conclusion, this step aims to convey a clear and concise understanding of the AI model from a functional perspective, ensuring stakeholders understand its utility and operational integration without getting confused by the underlying technical details by answering the following questions:

- **What is the primary function of the AI model?**
- **What tasks is the model designed to perform?**
- **How does the model process inputs and outputs?**
- **What is the lifecycle from training to deployment?**

STEP 5: Visualization techniques implementation

Objective: Implement visualization strategies to explain model decisions.

This section of the framework explores how different visualization strategies are essential for explaining the decision-making processes of a model. Here it is expected to deploy different data visualization techniques to make AI operations transparent and visually effective in interpreting model behavior more efficiently.

Different types of visualization techniques will be described while providing examples of how these visualization strategies are applied within real-world scenarios. By answering the following questions, the objective is to demystify AI operations, creating a transparent environment where users can confidently interact with and rely on AI systems:

- **What are the main visualization techniques used?**
- **How do these techniques contribute to transparency?**

STEP 6: Model performance and evaluation

Objective: Evaluate the performance of the model using relevant metrics.

In the performance evaluation section of the framework, performance assessment of the model will be done, ensuring that it meets the required standards of accuracy and reliability essential for its expected applications.

Key performance metrics, used to evaluate the model, will be explored while providing a clear explanation of each one of the used metrics. This includes common metrics such as accuracy, precision, recall, and the F1 score. By doing this, stakeholders will understand the meaning and relevance of these metrics in assessing the effectiveness of the model.

The results of these evaluations will be explained and interpreted in relation to the model's operational goals. This involves aligning performance outcomes with business objectives or operational requirements, providing a comprehensive view of how well the model meets its intended purposes.

Overall, this step intends to present a structured and detailed approach to model performance evaluation, ensuring all stakeholders have a clear understanding of how the AI model's effectiveness is measured and maintained by answering these questions:

- **What are the key performance metrics?**
- **How are these metrics applied?**
- **How do the results align with operational goals?**

STEP 7: Model explanation

Objective: Ensure explanations are integrated into every stage of AI development.

Our framework promotes a holistic approach to explainability, integrating explanations into every stage of the AI development cycle. This approach is designed to be interactive and user-centric, breaking down the model's operations into components that are easy to understand and analyze. This ensures transparency and builds a deeper trust in the ML systems, as stakeholders can see and understand why and how decisions were made.

In this step it will be explained how the results of these evaluations are interpreted in relation to the model's operational goals. This involves aligning performance outcomes with business objectives or operational requirements, providing a comprehensive view of how well the model meets its intended purposes. Overall, this section aims to present a structured and detailed approach to model performance evaluation, ensuring all stakeholders have a clear understanding of how the AI model's effectiveness is measured and maintained by answering the following questions:

- **What do the provided explanations mean?**
- **How do these explanations contribute to transparency and trust?**
- **How is the effectiveness of explanations evaluated?**

STEP 8: Feedback integration and continuous improvement

Objective: Systematically collect and provide feedback for continuous improvement.

Feedback should be systematically collected, analyzed, and utilized to encourage ongoing enhancements in the AI system. This section is crucial for ensuring the system remains adaptive and responsive to user needs and operational demands.

Overall, this step aims to demonstrate a structured and effective approach to maintaining a high-performing, user-centric ML system. By continuously integrating user feedback into the development cycle, the system not only meets its current operational goals but also adapts to future challenges and user expectations. In this step, the following questions should be answered:

- **What mechanisms are in place for feedback integration?**
- **How has feedback led to system improvements?**
- **How does continuous feedback adapt to changing needs?**
- **What tools support feedback integration?**

STEP 9: Future-proofing the AI System

Objective: Ensure the AI system evolves with emerging technologies and adheres to ethical standards.

In the last step of the framework, the AI's system preparation to evolve with emerging technologies and adapt to changing regulatory environments will be explained, by exploring potential future technologies that could impact the AI system and possibilities of incorporating them as they become viable, ensuring the system remains at the forefront of technological innovation.

In addition, it is crucial to ensure that the system operates within established ethical boundaries and incorporates principles such as fairness, transparency, and accountability, particularly as it scales and evolves. This section will outline the mechanisms for reviewing and updating the ethical guidelines in response to new challenges and societal expectations. By planning for technological evolution and ensuring rigorous adherence to regulatory and ethical standards, this framework sets the stage for a resilient, adaptable, and responsible AI system that is prepared to meet future challenges and opportunities.

- **How will the system incorporate future technologies?**
- **How are ethical guidelines updated and maintained?**

4. EXPERIMENT SET-UP AND IMPLEMENTATION

4.1. OPERATIONAL SNAPSHOT OF BANCO DE PORTUGAL: MARKET CONDUCT SUPERVISION

Banco de Portugal is entrusted with the role of market conduct supervision for a wide range of financial entities, including credit institutions, financial companies, payment institutions, electronic money institutions, and credit intermediaries. The Banking Conduct Supervision Department's intervention in the advertising process is a crucial part of Banco de Portugal's mission to ensure that these entities comply with the regulations when advertising retail banking products and services. Banco de Portugal shares its findings from the investigation of advertising practices within its jurisdiction with the relevant stakeholders. In its communications, Banco de Portugal conveys the outcomes of its inquiry into the advertising conduct of the supervised entities, ensuring transparency and adherence to legal standards.

Key Statistics and Insights from the Supervision Process in 2023 (Conduct Supervision report, 2023):

- **Volume of advertising supervised**: Banco de Portugal reviewed 16,338 advertising supports from 56 financial institutions, indicating a comprehensive oversight scope.
- **Compliance issues**: Of these, 2.2% exhibited irregularities, up from 1.8% the previous year, highlighting a tightening in regulatory oversight. The most significant non-compliance rates were in mortgage credit (11.4%) and bank deposits (8.1%).
- **Focus on consumer credit**: Consumer credit products dominated the advertising, with credit cards accounting for 47% of such promotions.
- **Impact of interest rates**: Advertising related to bank deposits and housing credit saw dramatic increases, by 92.6% and 40.8% respectively, likely driven by rising interest rates.
- **Digital transition**: The shift toward digital channels was notable, with a 11.6% increase in digital advertising, now constituting nearly half (47.8%) of all analyzed supports.

4.2. AI AT BANCO DE PORTUGAL

AI now sits at the forefront of Banco de Portugal's digital transformation plan. As seen throughout this study, AI, and ML in particular, represent both an opportunity and a challenge for central banks and financial institutions by offering a significant opportunity as they provide new and powerful tools for managing the growing volume of data central banks work with. However, they also pose a challenge in ensuring their reliability, since AI and ML have been helpful in relieving teams from repetitive and discouraging tasks, allowing them to focus on higher-value activities, by this means enhancing efficiency.

To meet Banco de Portugal's needs, internally developed ML models are becoming essential companions in how the bank processes and analyzes document information. New tools enable the creation of summaries, identification of keywords, recognition of entities, and extraction

and classification of information, allowing for the near-instantaneous analysis of extensive documents.

By using AI applications, Banco de Portugal facilitates the process of managing sanction measures, information requests, and address complaints from customers concerning banking institutions and credit intermediaries. In line with this study, we will evaluate the model that is being developed to automatically classifying advertising practices, with big potential of becoming a crucial tool that Banco de Portugal uses to ensure financial institutions' compliance with the regulations governing their advertising activities.

By leveraging AI, Banco de Portugal aims to enhance its capability to oversee advertising practices effectively, aligning with the ever-evolving digital landscape and regulatory complexities. This approach not only helps in managing the current volume and complexity of advertising but also ensures that financial entities operate within the established legal framework, thereby protecting consumer interests and maintaining market integrity.

This study aims to evaluate the advertising ML model following the proposed framework and provide feedback with expectations to enhance Banco de Portugal's efficiency in handling and classifying issues related to advertising of banking products and services. The expected benefits include enhanced trust in the AI system by implementing XAI techniques, which provide clear, understandable reasons for the AI's decisions, fostering confidence among regulators, financial institutions, and the public, faster processing of advertising reviews, reduced response time to queries from banking customers and stakeholders and improvement of Banco de Portugal's ML tool implementation.

To achieve these results, we will follow the proposed framework outlined in the project. The framework is designed to ensure that the AI system's decision-making is both transparent and understandable, thus providing XAI insights into the classification of advertising reviewed by Banco de Portugal. By closely following the structured approach proposed in the project, the model will enhance internal transparency and foster trust in the AI's decision-making process among regulators, financial institutions, and the public. This adherence to the framework ensures that the AI system not only meets regulatory compliance standards but also aligns with best practices in XAI, making its operations auditable and its decisions justifiable.

4.3. DATA SELECTION AND PREPARATION

Strong emphasis was laid on image quality and the relevance of content for proper selection of the data, that is, only those images that are clear, legible, and relevant to the content of the advertisement are selected for the next stage. To this effect, we carefully chose 48 images, with some examples in ***Figure 3***, to make a heterogeneous dataset from which we used three images for each of 16 distinct publicity labels. This ensures that all the labels are equal, which balances the dataset for testing the model's performance unfailingly through unbiased classification and pointing out any possible pitfalls to work upon.

Figure 3 - Selected data preview

The following are some of the important key processes applied in the data preparation phase:

1. **Text Extraction**: images were converted into textual data by utilizing sophisticated optical character recognition (OCR) tools and libraries such as EasyOCR, which support multiple languages and fonts.

2. **Data Cleaning, Normalization, and Transformation:** Added lowercasing for consistency, especially in cleaning links, punctuation, special characters, and all numbers. In addition, the cleaning process recognized the language and translated it to both Portuguese and English for consistency across the dataset. The tokenization and encoding of the textual data were achieved using the transformers library by Hugging Face for preprocessing.

4.4. MODEL DESCRIPTION

In this section, we provide a brief overview of the model's architecture, emphasizing its application rather than delving into internal complexities and developments. This is helpful for those needing to understand what the model is capable of and its outputs, without the need to acquire the ability to code or deep technical knowledge.

The primary goal of the AI model is to better supervise the way advertising is carried out in the financial sector. It is, therefore, designed to automatically analyze, classify, and ensure compliance with the required content of the advertisement by the regulations, making it easy to supervise the market conduct.

The ML model is developed to perform several key tasks central to its main purpose. It uses OCR in technologies for text extraction, making the image of text to be machine-encoded text. Advanced techniques in NLP, such as regular expression matching and ML through BERT, are used for classification of the text according to predefined advertising categories. The model concludes with the checking of compliance through the analysis of the classified data in flagging possible non-compliant advertising content. This makes use of the texts extracted and classified to specified advertising categories to check on the compliance.

Input-Output Mechanisms

This model primarily uses data inputs such as digital images and documents that contain textual content of an advertisement. These inputs are fed into the OCR. First, their pre-processing is done by converting the images into text. They are then classified using regular expressions and classifications using the BERT models. And the results of this analysis are used for creating actionable reports for regulatory purposes. Inputs primarily include receiving the scanned or digital images, and the mechanisms produce the classified categories of advertisements, flags for non-compliance, and detailed reports summarizing the findings for regulatory review.

Role Within the Larger System and Contribution to Project Goals

Within the broader regulatory framework of Banco de Portugal, the analyzed ML model is expected to fill a pivotal role in automating the detection and reporting of non-compliant advertising practices. This automation significantly enhances the bank's ability to manage large volumes of data, reduce human error, and respond more swiftly to violations. The model's capabilities support the bank's strategic goals of maintaining high standards of market conduct, protecting consumer interests, and promoting transparency and trust in the financial sector.

4.5. VISUALIZATION TECHNIQUES

We have incorporated several techniques for visualization so that, in the case of the case study, a decision by an AI model can be transparent and interpretable. Some of the key visualization techniques include feature importance visualizations that using tools like LIME generate visual representation of the most important features in terms of model's decisions by highlighting the words and phrases in the text with significant influence on the classification outcome, providing clear justification for each prediction. We use a confusion matrix to represent the performance of the model by showcasing the correct and predicted classifications, aiding in identifying patterns of misclassifications and giving insights about model accuracy across different classes. For misclassified cases, word clouds are an excellent visualization tool for the most common words, helping to effectively highlight text that led to wrong classifications. Additionally, bar plots of feature importance will be used as a great adaptable way of comparing the importance of features between correct and incorrect predictions, providing a clear impact on the model's decision-making. Lastly, for more granular analysis, heatmaps are an effective way of representing how specific text elements contribute to the final decision across different stages of the classification process.

By utilizing different visualization techniques, we expect to simplify and visualize the decisions of the ML model, making it easier for users to understand the process and trust the outcomes.

4.6. MODEL PERFORMANCE EVALUATION

In evaluating the AI model's performance, we used several common performance measures such as Accuracy, Precision, Recall (sensitivity) and F1 score, as shown in ***Figure 4***. These metrics are applied in real-world testing environments by comparing the model's predictions to actual labels. The performance metrics are calculated using the results from these comparisons, providing insights into how well the model performs under realistic conditions.

```
Accuracy: 0.0625 (95% CI: (0.0, 0.125))
F1 Score: 0.0080 (95% CI: (0.0, 0.03926282051282051))
Precision: 0.9418 (95% CI: (0.8779069767441859, 1.0))
Recall: 0.0625 (95% CI: (0.0, 0.14583333333333334))

Classification Report:
                                      precision    recall  f1-score   support

                        Cartão de crédito       1.00      0.00      0.00         3
                         Cartão de débito       1.00      0.00      0.00         3
                           Conta ordenado       1.00      0.00      0.00         3
                Conta pacote - particulares       1.00      0.00      0.00         3
Conversão cambial e transferência de fundos       1.00      0.00      0.00         3
                       Crédito a empresas       1.00      0.00      0.00         3
                        Crédito automóvel       1.00      0.00      0.00         3
                     Crédito consolidado       1.00      0.00      0.00         3
                         Crédito pessoal       1.00      0.00      0.00         3
                     Crédito à habitação       1.00      0.00      0.00         3
                        Depósito a prazo       1.00      0.00      0.00         3
                           Institucional       0.07      1.00      0.13         3
                        Linha de crédito       1.00      0.00      0.00         3
                            Multiproduto       1.00      0.00      0.00         3
               Outros serviços financeiros       1.00      0.00      0.00         3
                     Serviços de Pagamento       1.00      0.00      0.00         3
                        cartão de crédito       0.00      1.00      0.00         0
                        crédito automóvel       0.00      1.00      0.00         0
                             multiproduto       0.00      1.00      0.00         0

                                 accuracy                           0.06        48
                                macro avg       0.79      0.21      0.01        48
                             weighted avg       0.94      0.06      0.01        48
```

Figure 4 – Model performance metrics

The **accuracy** of a classifier is found by calculating the ratio of correctly classified instances to the total instances. The overall model accuracy is very low, at 6.25%, which indicates that in the model, only a very small proportion of the instances get classified correctly.

Precision: This is a calculation of the number of true positive predictions divided by all the positive predictions made. It gives an insight into the correctness of positive predictions the model makes. Average weighted precision is 94.18%. In general, it is shown that if a model is very confident when it makes a prediction of a class, then it has high precision. However, in this model, the high precision is one in which the other metrics completely balance it, which is a bit concerning.

Recall (Sensitivity): Recall measures the proportion of true positive predictions out of all actual positive instances. It reflects the model's ability to identify positive instances. The weighted average recall is 6.25%. This low recall indicates that the model fails to identify a significant number of actual positive instances, meaning it misses many true cases.

F1 Score: The F1 score is the harmonic mean of precision and recall, providing a balance between the two metrics. It is particularly useful when the class distribution is imbalanced. The weighted average F1 score is 8.08%. The F1 score, being the harmonic mean of precision and recall, is low, reflecting the imbalance between high precision and low recall.

The above analysis gives a rough look at the model's performance on even basic tasks. This was highly imbalanced, presenting a high precision but coupled with extremely low recall. That is to say, the model is being extremely conservative in its predictions and doing well to avoid false positives but at the expense of missing very many true positives. Such an imbalanced model would raise operational risks in instances where the failure to detect true positives carries severe consequences. Say, a fraud detection system, where a missed fraudulent

activity is often assumed to have resulted in huge losses (low recall). However, the wide confidence intervals around estimates of accuracy, precision, recall, and the F1 score indicate very little certainty in these estimates—most likely because of the sample size—increasing the dataset size may yield more reliable performance estimates.

4.7. MODEL EXPLANATION

In this step, we evaluate the provided explanations using metrics and methodologies to ensure transparency and build trust in the AI system combined with the visualization techniques mentioned previously. The explanations generated using LIME and AIX360 aim to clarify the influence of various features on the model's predictions.

LIME approximates the model locally with an interpretable model to explain individual predictions. For each instance, LIME perturbs the data slightly and observes how the model's predictions change, identifying which features are most influential. AIX360 provides a set of algorithms to explain the predictions of ML models. These explanations complement LIME by offering different perspectives on feature importance and interactions.

4.7.1. Single sample explanations

Let's consider a randomly selected image (***Figure 5***) from our dataset. The selected image should be classified as ***crédito automóvel*** (auto loan), however it was misclassified as ***institucional*** with a relatively low confidence of 13.5%.

Figure 5 - Case study image

We can use LIME to check the probability distribution of the model's prediction using each instance (***Figure 6***). The model shows that the probability for the ***institucional*** class is 50%, that is, according to the model, this instance has a very strong likelihood of being in the indicated class. Here, it is important to underline that this prediction is wrong for the example at hand.

The probability of the instance belonging to the ***crédito automóvel*** class is also very low, at 18%. This is what the actual label for this instance is, but the model's confidence on that class is much lower than for ***institucional*** (institutional) —a clear sign of biasing towards the class for this given instance.

The true class, ***crédito automóvel***, has a relatively low probability of 18%, showing that the model is less confident about this correct classification. This misclassification suggests that the features relevant to ***crédito automóvel*** are not as strongly weighted or prominent in this instance. The probabilities spread for the other classes (***crédito pessoal*** (personal loan)***, crédito à habitação*** (mortgage loan)***, and others***) are signs that the model does consider these classes but with much less confidence.

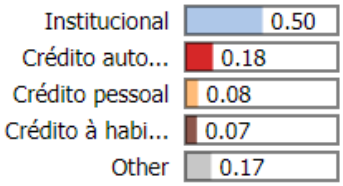

Figure 6 - Prediction probabilities for the case study

The explanations from LIME and AIX360 in ***Figure 7*** highlight the most influential features for this prediction:

```
{'finality': 'Institucional', 'confidence': 0.13525964319705963, 'true_label': 'Crédito automóvel'}
LIME Explanation:
[('garantia', -0.23570277284684887), ('Campanha', 0.12759765789943536), ('Inspeções', -0.10557266335561635), ('comprovadas', 0.09837
05819866513), ('KADJAR', 0.05147428836336767 6), ('COMEÇAR', 0.04626123148466989), ('USADOS', 0.03833085913021532), ('MENSALIDADE',
0.03677481629130958), ('1ª', 0.03328461223138816), ('120', 0.028351981869415097), ('29', -0.024589232711784287), ('_', -0.0243222821
69165522), ('UMA', 0.0225251816034629), ('2024', -0.02158478385193683), ('3', -0.02081664596475562), ('PRAZO', 0.01968964058555045
3), ('91', -0.01923141001743877), ('903', 0.018873296343721042), ('fevereiro', -0.01695683263441987), ('2021', -0.01510136309251508
4)]
AIX360 Explanation:
[('garantia', -0.23636170880910712), ('Inspeções', -0.10458898178879714), ('comprovadas', 0.09671253584368272), ('Campanha', 0.09316
0627774295), ('OFERTA', 0.05058506717792908), ('KADJAR', 0.046725326044253924), ('91', -0.041955458912435885), ('MENSALIDADE', 0.037
42302643875197), ('MONTANTE', -0.02883961732311213), ('fevereiro', -0.0263558496875372), ('CARRO', 0.025362038061820306), ('PRAZO',
0.023557280859086545), ('COMEÇAR', 0.023205323654047517), ('INICIAL', 0.022864033740074695), ('usApos', 0.02175192967463501), ('HIST
ÓRIA', 0.020465763523851987), ('3', -0.02019324862903346), ('MESES', 0.01964725575820077 5), ('ENTRADA', 0.019394169114957758), ('90
3', 0.01617246283955584)]
```

Figure 7 - LIME vs. AI360 explanations

Both LIME and AIX360 find importance in features such as ***garantia*** (guarantee), ***campanha*** (campaign), ***inspeções*** (inspections), and ***comprovadas*** (verified) as significant. This consistency suggests that these features are indeed influential in the model's decision-making process.

The feature ***garantia*** has a strong negative weight in both explanations, indicating it strongly influenced the prediction towards ***institucional*** rather than ***crédito automóvel***. We see that both LIME and AIX360 give different answers on features' weights and refer to other important features. So, LIME raises the feature ***mensalidade*** (monthly payment) as important, and AIX360 raises the features ***inicial*** (initial) and ***1ª*** (1st). This way they become great complements to each other, showing different points of the text now influencing the model. In particular, the negative weights of features like garantia and ***inspeções*** give the impression that the presence of these words in the lexicon pulls the model towards the class ***Institucional*** rather than ***Crédito automóvel,*** as shown in ***Figure 8***.

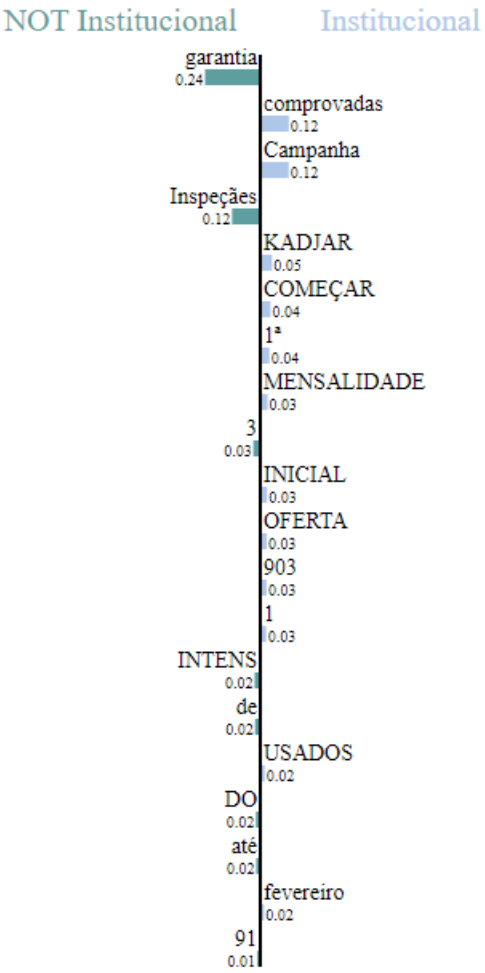

Figure 8 - Feature weight comparison.

Figure 9 shows a snippet of the text with some of the words highlighted. Again, the highlighted words are a nice example of how the visualizations represent the terms that the model picked up on as influential to the prediction. Despite the incorrect classification, the emphasis on terms related to promotions, inspections, and warranties indicates a strong contextual influence from automotive and financial sectors, which should ideally align with the correct class ***Crédito automóvel***.

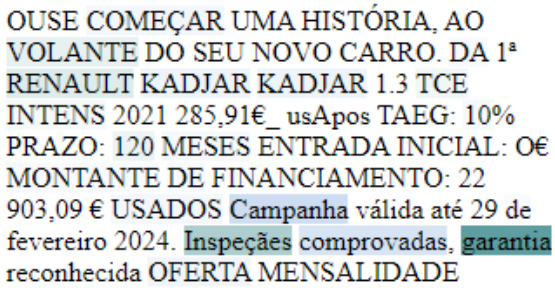

Figure 9 - Highlighted text example

It might be beneficial to ensure that terms like ***garantia*** and ***inspeções*** are adequately represented and correctly associated with ***Crédito automóvel*** in the training data.

The provided explanations from LIME and AIX360 show a coherent picture of which features influenced the model's prediction. They reveal that certain features strongly contributed to the misclassification, and these insights can guide further refinement of the model and data. By addressing the identified issues, we can improve the model's accuracy and reliability for future predictions.

4.7.2. Single class explanations

When evaluating the performance of ML models, it is crucial to consider not only the accuracy of individual predictions but also the overall, or global, performance of the model. This view not only makes sure a model is reliable, fair, and interpretable across different scenarios but, for instance, that one can use an individual prediction to be able to describe in which concrete cases the model is good or bad and hence also indicate the strengths and potential causes of failures of the model. On the other hand, the global performance of the model is described by the analysis of all the metrics and feature importances holistically and describes how the model works and how consistently it shows the described behavior.

Following the individual example above, let's compare the feature importance for the class ***Crédito automóvel***. A comparison of the 10 highest feature importances between the ***Crédito automóvel*** classification label, as determined by two different explanation methods is shown in ***Figure 10***.

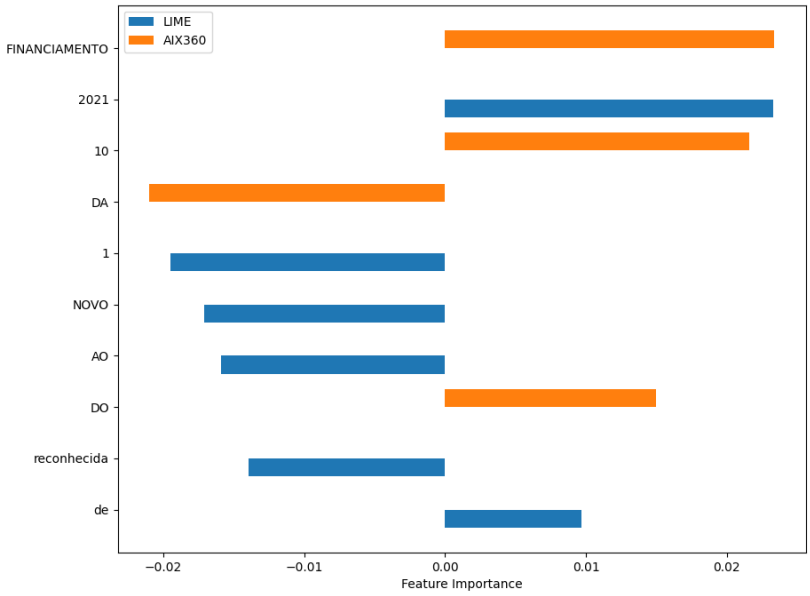

Figure 10 - Feature importance comparison

Features such as ***financiamento*** (financing), ***2021***, and ***reconhecida*** (acknowledged) have positive values of importance, indicating their significant role in the model's classification decisions. Conversely, some features exhibit negative importance values, suggesting they may contribute negatively towards the label.

Notably, there are differences in feature importance between both methods. For instance, according to AIX360, ***da*** (of) is of high positive importance, but that feature does not appear in LIME. Similarly, ***financiamento*** and ***reconhecida*** are given positive importance by LIME but not by AIX360, and features like ***10*** and ***ao*** (to the) have differing importance values between the two methods.

Using multiple explanation methods, such as LIME and AIX360, allows understanding the model's behavior from different perspectives, highlighting that each method might capture different aspects of the model's decision-making process. Features with consistently high importance across both methods, such as ***2021*** and ***financiamento,*** are likely significant for the classification of ***Crédito automóvel***. However, features with high importance in only one method should be further investigated to understand their role and relevance.

Combining insights from both LIME and AIX360 provides a more comprehensive understanding of the model, potentially leading to better trust and transparency in its predictions. In terms of data preparation and model training, attention should be paid to features identified as important by both methods. Additional pre-processing or engineering might be required for features that are important in only one method. Furthermore, it is essential to validate the model performance using features identified as significant by both methods to ensure they contribute positively to the model's predictive power.

Incorporating multiple explanation methods regularly during the model development process can help refine feature selection and improve model accuracy. These explanations can guide

targeted improvements, enhance the robustness and transparency of the AI system, and ultimately contribute to better decision-making and trust in the model's outputs.

4.7.3. Global explanations

Figure 11 shows the ten most positive and negative global importances. The insights from this might provide better judgment on what features are working to contribute positively to the model in giving better recommendations and which ones are impeding it.

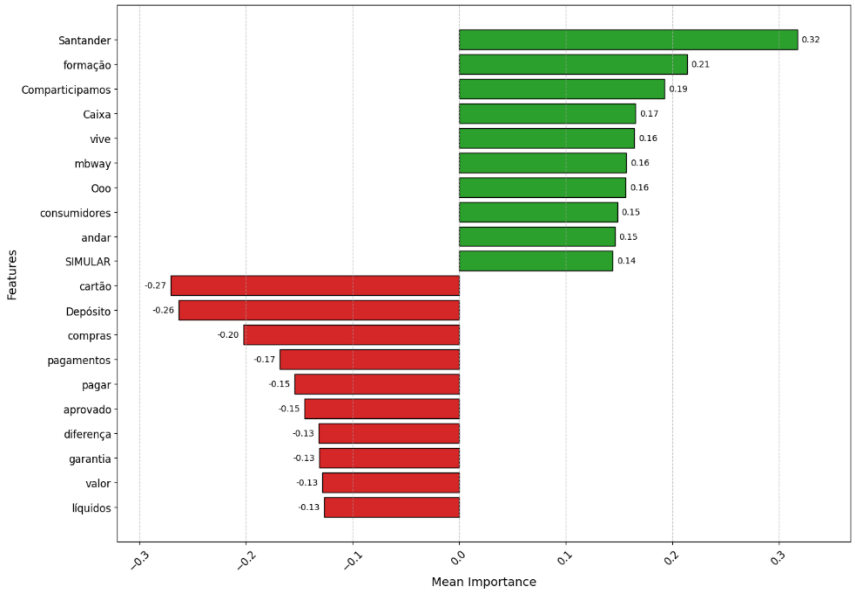

Figure 11 - Global feature importance

The feature importances identified in the model play an important role in influencing the predictions made by the ML model. The feature ***Santander*** (name of a bank) holds the highest positive importance, indicating its significant influence on the model's accurate predictions, especially in financial-related classifications. This aligns with the understanding that specific entities, such as banks, can strongly impact predictive outcomes in financial scenarios. Additionally, features like ***formação*** (training), ***comparticipamos*** (participate), and ***Caixa*** (name of a bank) also contribute positively to enhancing the model's accuracy, further emphasizing their importance in the prediction process.

On the contrary, the feature ***cartão*** (card) exhibits the highest negative importance, significantly reducing the model's predictive power. Similarly, ***depósito*** (deposit), ***compras*** (purchases), and ***pagamentos*** (payments) negatively impact the model's predictions, suggesting that these features might require further investigation or mitigation strategies to improve the model's performance.

Understanding global feature importances in ML models offers valuable insights that can benefit various aspects of model development and deployment. By identifying significant positive and negative impacts of features, organizations can interpret the behavior of the model and understand the factors driving its decision-making process (Fang et al., 2022). Features with high negative importance, as mentioned in the context of the model's negative feature importances, may require re-evaluation or transformation to mitigate their adverse impact. The use of feature importance scores in feature engineering is crucial to address bias and improve the interpretability of results based on individual input features (Sahoo et al., 2022). Lastly, enhancing features with high positive importance and addressing issues related to negatively impactful features can lead to model refinement. This process can significantly improve the overall performance and reliability of the model, by emphasizing the importance of feature selection in improving the generalizability of ML models (Pudjihartono et al., 2022).

4.7.4. Feature importance distribution

The distribution of feature importance weights in ***Figure 12*** shows the influence of the factors over final model predictions. The knowledge of this distribution allows one to purposefully choose opportunities for feature engineering, selection, and reduction.

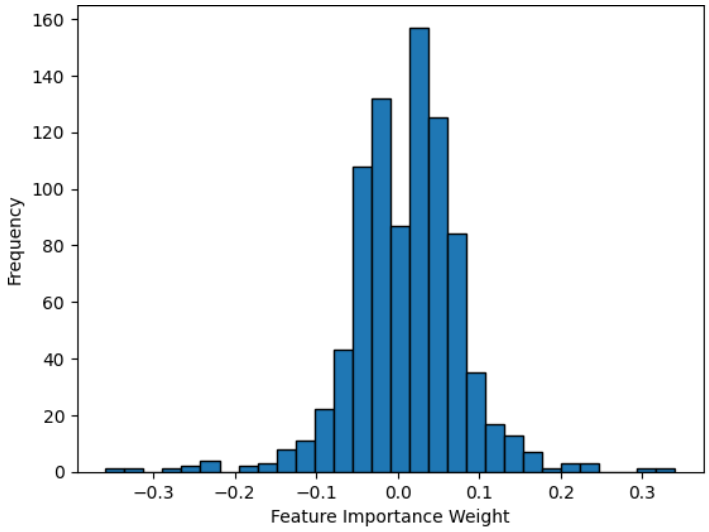

Figure 12 - Feature importance distribution

The observations and insights resulting from analyzing feature importances in a ML model provide valuable guidance for optimizing model performance and interpretability. These insights can inform decisions regarding feature selection, model refinement, and overall model simplification.

As shown in ***Figure 12*** the majority of feature importance weights cluster around 0, indicating that many features have minimal influence on the model's predictions. This suggests that only a subset of features significantly impacts the model's decision-making process. The symmetrical distribution of feature importances around the mean implies a balance between positive and negative contributions from features, showing that the model considers both beneficial and detrimental features in its predictions. The concentration of feature

importances around zero highlights that only a few features play a substantial role in influencing the model's predictions. This concentration simplifies the interpretability of the model by focusing on the most influential features.

By leveraging the information obtained from analyzing feature importances, organizations can make informed decisions regarding feature selection, model refinement, and dimensionality reduction. These actions contribute to building a more robust, transparent, and efficient AI system that maintains high standards of fairness, accuracy, and interpretability across diverse scenarios.

4.7.5. Misclassified cases analysis

The analysis of misclassified cases in the model provides valuable insights into its limitations and areas of struggle, offering opportunities for refinement and optimization. By understanding the characteristics and patterns of misclassifications, organizations can identify potential issues in data representation, feature selection, and model assumptions, leading to enhanced accuracy and robustness across diverse scenarios.

The total number of misclassified samples is 45, with the predominant error being the misclassification of various financial product labels as ***Institucional***. The presence of generic financial terms in misclassified samples, such as ***crédito***, ***para***, and ***em***, suggests potential sources of confusion leading to misclassifications. The model's tendency to over-predict the ***Institucional*** label across various financial products underscores the need for more granular feature engineering and dataset enhancement to differentiate between closely related categories.

The word cloud visualization in ***Figure 13*** and the corresponding bar chart in ***Figure 14*** highlight prevalent textual patterns that challenge the model's correct classification. The predominant misclassification error of various financial product labels as ***Institucional*** suggests a bias towards this category, particularly when faced with diverse financial terms and contexts. The detailed breakdown reveals consistent struggles in distinguishing specific financial products from the general ***Institucional*** category, indicating areas for improvement in model differentiation.

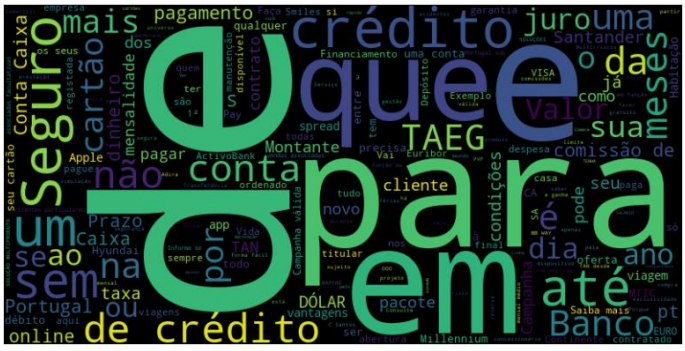

Figure 13 - Word cloud for misclassified samples

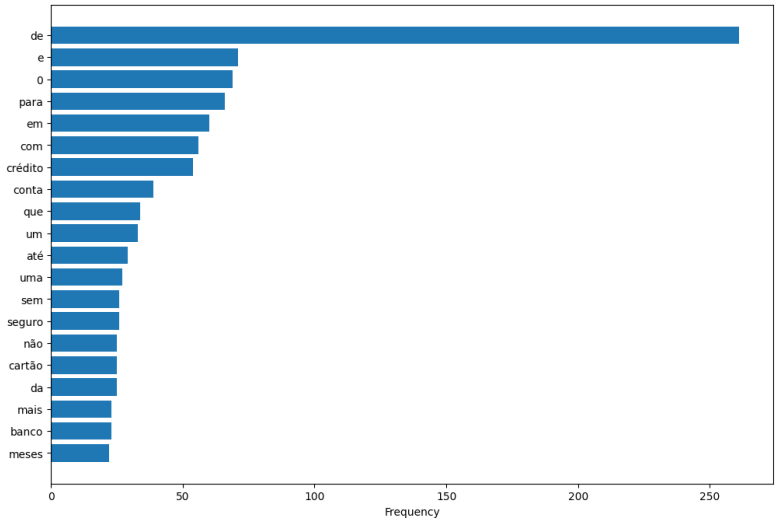

Figure 14 - Most common words in misclassified samples

The insights gained from analyzing misclassified cases can guide model refinement efforts, potentially through targeted feature engineering, enhanced training data, or improved model architecture. Addressing these misclassification patterns can lead to a more accurate and reliable model, ultimately improving its performance in real-world applications. In conclusion, the thorough analysis of misclassified cases provides a roadmap for refining the model, enhancing its accuracy, and ensuring robust performance across diverse scenarios, thereby advancing the effectiveness and reliability of the AI system.

4.7.6. Misclassified text analysis

The analysis of text length statistics for misclassified cases provides valuable insights into the nature of these instances. The mean length of 88.18 words indicates that, on average, misclassified texts consist of approximately 88 words. This average is influenced by the presence of significantly longer texts, as evidenced by the maximum length of 337 words, showcasing the existence of lengthy misclassified cases. Contrarywise, the minimum length of 13 words demonstrates that even very short texts can be misclassified, emphasizing the diversity in text lengths among these instances.

By understanding these statistics, we can gain a better perspective on the nature of the misclassified cases and consider strategies for improving the model's handling of texts of varying lengths. For instance, ensuring that the model is equally proficient at processing both short and long texts could help reduce misclassification rates.

The identification of uncommon words in misclassified texts, such as ***regime***, ***número***, ***mesmo*** and ***registo***, shows a presence of domain-specific or context-specific terms that are rare in the dataset. These terms, appearing only once, signify their infrequency and potential challenge for the model in accurately interpreting their meaning and context. The rarity of these words can be attributed to their specificity to certain domains, like finance or law, making them crucial for accurate classification in specialized contexts. Additionally, terms like ***https*** and ***site*** hint at references to websites or online resources, indicating potential out-of-context references that may confuse the model if not adequately trained based on such instances.

The Flesch Reading Ease score of 43.53 categorizes the text as difficult to read, suitable for individuals with a college education level. This score implies the presence of complex sentences, advanced vocabulary, and specialized jargon that can pose challenges for NLP models. The complexity of the text may contribute to higher misclassification rates, as the model may struggle to comprehend the intricate language structures present in the text. On the contrary, the Gunning Fog index of 11.64 suggests that the text is relatively straightforward to read, corresponding to an 11th-grade reading level. While the Fog index implies that the text should be comprehensible to the average reader, the inclusion of uncommon words and domain-specific terminology can still impede the model's accurate classification of such text.

The combination of uncommon words and readability scores stresses the challenges posed by complex and specific terms in the text data. The presence of rare vocabulary, coupled with the text's difficulty in readability, presents obstacles for the model in accurately classifying misclassified cases. The rarity of certain words indicates potential gaps in the model's training data, suggesting a need for enrichment with domain-specific texts to enhance familiarity with uncommon terms and specialized vocabulary. Feature engineering techniques, such as word embeddings or domain-specific lexicons, can aid in capturing the context and significance of rare terms, thereby improving the model's performance on such instances. In conclusion, addressing the challenges posed by uncommon words, complex language structures, and varying text lengths is crucial for enhancing the model's performance on misclassified cases. By enriching the training data with domain-specific texts, employing feature engineering strategies, and simplifying text preprocessing, the model can better handle specialized vocabulary, rare terms, and complex language patterns, leading to more accurate predictions in real-world applications (Ree et al., 2022).

4.8. FEEDBACK INTEGRATION AND CONTINUOUS IMPROVEMENT

Feedback integration is a critical component in maintaining and improving the performance and reliability of any AI system. In this step we are interested in systematically collecting, analyzing, and utilizing feedback to encourage ongoing enhancements. By implementing robust feedback mechanisms, organizations guarantee that the AI systems remain adaptive and responsive to user needs and operational demands.

The analysis of misclassified cases and feature importances highlights several critical areas where feedback can significantly enhance the AI model's performance. First and foremost, user feedback on the misclassification patterns can help identify the specific contexts and nuances that the model currently fails to capture. For instance, the tendency to over-predict the ***Institucional*** category indicates a need for more granular feature engineering and dataset enhancement. Users can provide valuable insights into additional features or contextual information that should be considered during training, helping to refine the model's ability to distinguish between closely related advertising categories.

The model's high confidence in incorrect classifications suggests a potential misalignment between these features and the actual categories they should represent. By collecting feedback on these specific features from domain experts, we can adjust their weights or redefine their associations within the model. This will not only improve the model's accuracy but also enhance its interpretability, as stakeholders will better understand how and why certain features influence the model's decisions.

Additionally, visualization techniques, such as LIME and AIX360, can be used to enhance the transparency and trustworthiness of the AI system. Users can comment on the clarity and usefulness of the visual explanations provided, suggesting ways to make these explanations more intuitive and actionable. By integrating this feedback into the continuous improvement cycle, we can ensure that the model remains not only accurate but also transparent and user-friendly, fostering greater trust and confidence among regulators, financial institutions, and the public.

4.9. FUTUREPROOFING THE AI SYSTEM

To ensure the AI system remains robust and relevant in the face of evolving technologies and regulatory landscapes, it is essential to incorporate mechanisms for futureproofing. One of the primary strategies involves continuous monitoring and adoption of emerging AI and ML technologies. By staying abreast of advancements in areas such as transfer learning, federated learning, and XAI, Banco de Portugal can integrate these innovations into the system, enhancing its capabilities and maintaining its competitive edge. Regularly scheduled reviews and updates to the model architecture and algorithms will ensure the AI system leverages the latest technological advancements.

Maintaining ethical standards and compliance with evolving regulations is another critical aspect of futureproofing. Establishing a proactive engagement strategy with regulatory bodies allows for early identification of potential regulatory changes and the development of compliance roadmaps. By implementing flexible policies and procedures, the AI system can swiftly adapt to new regulatory requirements. Additionally, incorporating ethical guidelines into the AI development process, such as fairness, transparency, and accountability, ensures that the system operates within established ethical boundaries and builds trust with stakeholders.

Preparing for the future also involves enhancing the AI system's scalability and flexibility. Designing the system with a modular architecture enables easy integration of new technologies and components as they become available. Investing in scalable infrastructure ensures the AI system can handle increasing data volumes and computational demands, supporting the bank's strategic goals of maintaining high standards of market conduct and protecting consumer interests.

Finally, fostering a culture of continuous improvement and innovation within the organization is crucial. Encouraging ongoing training and education for staff on the latest AI and ML developments, as well as on ethical AI practices, ensures that the team remains knowledgeable and adept at managing the AI system. By creating a dynamic feedback loop where user insights and technological advancements inform iterative improvements, Banco de Portugal can ensure that its AI system remains resilient, adaptable, and capable of meeting future challenges and opportunities.

5. RESULTS AND DISCUSSION

The challenges faced by financial institutions, such as Banco de Portugal, in adopting different ML algorithms, especially in NLP, originate from the lack of transparency and interpretability of these extremely complex models. The vagueness of generated results has created skepticism and some averseness among different stakeholders to fully embrace these emerging technologies.

The findings of this study clearly show that introducing XAI techniques can greatly enhance the transparency of ML models by utilizing tools like LIME and AIX360, that are becoming instrumental in making the decision-making process clear. By implementing these methodologies, Banco de Portugal was able to provide users with clearer image of different factors that influence the analyzed ML model decisions, improving the underlying technology while fostering trust and confidence.

The demand for transparent AI systems is exceptionally clear in critical fields like healthcare, finance, and autonomous systems, and integrating XAI in ML models can drastically improve adoption by enhancing the understanding of models' predictions.

5.1. THEORETICAL IMPLICATIONS

The findings of this study align with the existing theories greatly emphasizing the significance of transparency and interpretability in AI systems, especially in very sensitive fields such as finance. This study supports the idea complex AI models combined with lack of transparency are massive block to their adoption and trust among users. By integrating tools such as LIME and AIX360, this study shows that it is possible to enhance transparency and bridge the gap between users and technology.

Moreover, this study provides new and useful insights by showcasing real practical applications of XAI techniques in real-world scenario, filling the existing gap in empirical evidence of XAI's impact in real-world operation environments. We highlight the necessity of combining multiple XAI instead of relying on single methodology to gain comprehensive understanding of model's behavior and factors that influence individual and global predictions.

In practical terms, this study offers a framework for financial institutions to enhance the transparency of their AI models, being in development or in use, enabling improved decision-making, regulatory compliance, and stakeholder trust. By showcasing that even complex models can be made interpretable with simple and strategic XAI approach, the study challenges the existing ideas suggesting a need for a trade-off between model complexity and interpretability. Even more sophisticated models can achieve both high performance and transparency.

5.2. PRACTICAL IMPLICATIONS

Our study provides several practical implications that emphasize the real-world applications and benefits of integrating XAI techniques into ML models.

The transparency brought by XAI tools enables better decision-making within Banco de Portugal. For instance, when classifying financial advertisements, the ability to understand the reasoning behind each classification allows regulatory bodies to assess compliance more accurately. This not only accelerates the review process but also reduces the risk of oversight or misinterpretation, ensuring more consistent and reliable outcomes.

By automating the classification and review of advertising content, Banco de Portugal can significantly reduce the time and resources required for the task. The use of XAI techniques ensures that the automated processes remain transparent and trustworthy, thus minimizing the need for extensive manual verification. This leads to increased operational efficiency, allowing users to allocate their focus to more complex and strategic activities rather than routine checks.

The incorporation of XAI in ML models addresses ethical concerns related to AI usage in decision-making processes. By ensuring that AI systems are transparent, and their decisions are explainable, Banco de Portugal promotes responsible AI usage. This approach mitigates the risks associated with "black box" AI systems, where the lack of transparency can lead to unintended biases and dilemmas.

For end-users, the transparency of AI-driven decisions can improve their overall experience. When users understand how and why certain decisions are made, their confidence in the institution's processes is strengthened. This transparency can lead to greater user satisfaction and trust, which are crucial for maintaining strong customer relationships and institutional reputation.

In conclusion, the practical implications of this study demonstrate many tangible benefits of integrating XAI techniques into ML models used by financial institutions. These benefits span enhanced trust and adoption, improved decision-making, increased operational efficiency, facilitated compliance, promoted ethical AI usage, and enhanced user experience. By addressing these practical aspects, the study demonstrates the significant impact of XAI on the real-world application of AI technologies in the financial sector.

5.3. LIMITATIONS AND FUTURE RESEARCH

The conducted study on application of XAI at Banco de Portugal has showed several limitations and potential areas for future research. One of the biggest limitations is the generalization of results, while the study showcases great results and the potential of XAI, its direct replicability to other sectors and institutions with different operational dynamics remains unclear.

Moreover, the reliability of the study's results is heavily impacted by data constraints, resulting from a limited dataset. The lack of access to an extensive and heterogenous data types and volumes hinders the ability to confirm the scalability of the proposed framework and by relying on a smaller dataset, the study may not fully capture the complexity of financial data encountered in other institutions.

The methodological approach was also constrained by the scope and resources availability. The selection of XAI techniques such as LIME and AIX360 was based on their popularity and availability, potentially excluding other effective methods and the reliance on specific visualization tools and performance metrics might limit the comprehensiveness of the evaluation.

Looking towards future research directions, expanding the dataset and context is crucial to establish the applicability and effectiveness of the proposed XAI framework across different financial institutions and sectors. Including a wider variety of financial data and case studies from multiple institutions will help validate the framework's generalizability and robustness.

Future studies should also explore additional XAI techniques beyond LIME and AIX360, such as SHAP and Anchors, to enhance the explainability of AI models. Comparative studies evaluating the performance and interpretability of these techniques in various contexts would be valuable for advancing XAI research.

Interdisciplinary approaches that combine insights from AI, human-computer interaction, and organizational behavior should be considered for developing holistic XAI frameworks. Collaboration with experts in psychology, sociology, and ethics can help address the multifaceted challenges of AI explainability and trust.

In conclusion, while the study at Banco de Portugal has demonstrated the benefits of XAI in financial institutions, addressing the identified limitations and pursuing the outlined avenues for future research will be instrumental in advancing transparent, interpretable, and trustworthy AI systems.

6. CONCLUSION

In conclusion, this Project Work contributes to the growing literature on XAI in NLP projects within financial institutions, emphasizing the practical benefits of XAI in enhancing decision-making processes, regulatory compliance, and stakeholder trust. The study's results and implications underscore the importance of continued research and development of XAI techniques to ensure the successful deployment of AI models in critical domains like finance.

Emerging XAI techniques have been shown to significantly enhance model transparency and trust within the banking sector, particularly in NLP projects. The practical deployment of these techniques in banks, like Banco de Portugal, has improved decision-making processes by providing clear and understandable explanations of model outputs. This transparency not only aids in making better-informed decisions but also contributes to operational efficiency.

Moreover, the study highlights the critical role of XAI techniques in ensuring regulatory compliance and mitigating legal risks within financial institutions. Transparent AI models can automate the detection of non-compliant practices, reducing legal and reputational risks significantly. Additionally, the use of XAI techniques leads to model outcomes that are understandable and justified, thereby enhancing trust and satisfaction among model users. This increased trust can foster greater cooperation and collaboration among stakeholders.

The adaptable framework developed for XAI in NLP projects, as demonstrated in the study, can be extended to other sectors beyond finance, showcasing the diversified impacts of these techniques. However, the study also acknowledges limitations related to data quality constraints and suggests that future research should focus on training and testing NLP models on larger datasets with more variations for better generalization of findings. Furthermore, there is a call for improving AI models in terms of accuracy and recall through the exploration of advanced techniques and algorithms.

BIBLIOGRAPHICAL REFERENCES

Abonizio, H. Q., Paraiso, E. C., & Barbon, S. (2022). Toward Text Data Augmentation for Sentiment Analysis. *IEEE Transactions on Artificial Intelligence*, *3*(5), 657–668. https://doi.org/10.1109/TAI.2021.3114390

Adamson, G. (2021). Explainable Artificial Intelligence (XAI): A reason to believe? *Law in Context. A Socio-Legal Journal*, *37*(3). https://doi.org/10.26826/law-in-context.v37i3.177

Akula, A., Todorovic, S., Chai, J., & Zhu, S. (2019). *Natural Language Interaction with Explainable AI Models*.

Alabi, R., Elmusrati, M., Leivo, I., Almangush, A. A. ا. ا., & Mäkitie, A. (2023). Machine learning explainability in nasopharyngeal cancer survival using LIME and SHAP. *Scientific Reports*, *13*. https://doi.org/10.1038/s41598-023-35795-0

Al-Khazaleh, M., Alian, M., Biltawi, M., & Al-Hazaimeh, B. (2023). Sentiment Analysis for People's Opinions about COVID-19 Using LSTM and CNN Models. *International Journal of Online and Biomedical Engineering (iJOE)*, *19*(01), Article 01. https://doi.org/10.3991/ijoe.v19i01.35645

Amazeen, M. (2017). Journalistic interventions: The structural factors affecting the global emergence of fact-checking. *Journalism: Theory, Practice & Criticism*, *21*, 146488491773021. https://doi.org/10.1177/1464884917730217

Aparicio, M., Bação, F., & Oliveira, T. (2016). An e-Learning Theoretical Framework. *Journal of Educational Technology Systems*, *19*, 292–307.

Arenas, M., Barceló, P., Romero, M., & Subercaseaux, B. (2022). *On Computing Probabilistic Explanations for Decision Trees* (arXiv:2207.12213). arXiv. https://doi.org/10.48550/arXiv.2207.12213

Arya, V., Bellamy, R., Chen, P.-Y., Dhurandhar, A., Hind, M., Hoffman, S., Houde, S., Liao, V., Luss, R., Mojsilovic, A., Mourad, S., Pedemonte, P., Raghavendra, R., Richards, J., Sattigeri, P., Shanmugam, K., Singh, M., Varshney, K., Wei, D., & Zhang, Y. (2021). *AI Explainability 360: Impact and Design*.

Arya, V., Bellamy, R. K. E., Chen, P.-Y., Dhurandhar, A., Hind, M., Hoffman, S. C., Houde, S., Liao, Q. V., Luss, R., Mojsilović, A., Mourad, S., Pedemonte, P., Raghavendra, R., Richards, J., Sattigeri, P., Shanmugam, K., Singh, M., Varshney, K. R., Wei, D., & Zhang, Y. (2022). AI Explainability 360: Impact and Design. *Proceedings of the AAAI Conference on Artificial Intelligence*, *36*(11), Article 11. https://doi.org/10.1609/aaai.v36i11.21540

Baghbani, A.-N., Soltani, A., Kiany, K., & Daghistani, F. (2023). Predicting the Strength Performance of Hydrated-Lime Activated Rice Husk Ash-Treated Soil Using Two Grey-Box Machine Learning Models. *Geotechnics*, *3*. https://doi.org/10.3390/geotechnics3030048

Barker, M., Chue Hong, N., Katz, D., Lamprecht, A.-L., Martinez Ortiz, C., Psomopoulos, F., Harrow, J., Castro, L., Gruenpeter, M., Martinez, P., & Honeyman, T. (2022). Introducing the FAIR Principles for research software. *Scientific Data*, *9*. https://doi.org/10.1038/s41597-022-01710-x

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*, 82–115. https://doi.org/10.1016/j.inffus.2019.12.012

Berka, P. (2020). Sentiment analysis using rule-based and case-based reasoning. *Journal of Intelligent Information Systems*, *55*(1), 51–66. https://doi.org/10.1007/s10844-019-00591-8

Biecek, P., Burzykowski, T., Biecek, P., & Burzykowski, T. (2021). *Local Interpretable Model-agnostic Explanations (LIME)*. 107–123. https://doi.org/10.1201/9780429027192-11

Bifarin, O. O. (2022). *Interpretable machine learning with tree-based shapley additive explanations: Application to metabolomics datasets for binary classification*. https://doi.org/10.1101/2022.09.19.508550

Blanco-Justicia, A., & Domingo-Ferrer, J. (2019). Machine Learning Explainability Through Comprehensible Decision Trees. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *11713 LNCS*, 15–26. Scopus. https://doi.org/10.1007/978-3-030-29726-8_2

Bodria, F., Giannotti, F., Guidotti, R., Naretto, F., Pedreschi, D., & Rinzivillo, S. (2023). Benchmarking and survey of explanation methods for black box models. *Data Mining and Knowledge Discovery*, *37*(5), 1719–1778. https://doi.org/10.1007/s10618-023-00933-9

Brandt, R., Raatjens, D., & Gaydadjiev, G. (2023). *Precise Benchmarking of Explainable AI Attribution Methods*.

Chaddad, A., Peng, J., Xu, J., & Bouridane, A. (2023). Survey of Explainable AI Techniques in Healthcare. *Sensors*, *23*(2), Article 2. https://doi.org/10.3390/s23020634

Chen, H., Gomez, C., Huang, C.-M., & Unberath, M. (2022). Explainable medical imaging AI needs human-centered design: Guidelines and evidence from a systematic review. *Npj Digital Medicine*, *5*(1), 1–15. https://doi.org/10.1038/s41746-022-00699-2

Chen, Z., Lian, Z., & Xu, Z. (2023). Interpretable Model-Agnostic Explanations Based on Feature Relationships for High-Performance Computing. *Axioms*, *12*, 997. https://doi.org/10.3390/axioms12100997

Costa, V. G., & Pedreira, C. E. (2023). Recent advances in decision trees: An updated survey. *Artificial Intelligence Review*, *56*(5), 4765–4800. https://doi.org/10.1007/s10462-022-10275-5

Costi, F., Onchis, D., Hogea, E., & Istin, C. (2024). *Predictive Modeling for Diabetes Using GraphLIME* (p. 2024.03.14.24304281). medRxiv. https://doi.org/10.1101/2024.03.14.24304281

Deshpande, R. S., & Ambatkar, P. V. (2023). Interpretable Deep Learning Models: Enhancing Transparency and Trustworthiness in Explainable AI. *Proceeding International Conference on Science and Engineering*, *11*(1), Article 1. https://doi.org/10.52783/cienceng.v11i1.286

Dieber, J., & Kirrane, S. (2022). A novel model usability evaluation framework (MUsE) for explainable artificial intelligence. *Information Fusion*, *81*, 143–153. https://doi.org/10.1016/j.inffus.2021.11.017

Dobrovolskis, A., Kazanavičius, E., & Kižauskienė, L. (2023). Building XAI-Based Agents for IoT Systems. *Applied Sciences*, *13*(6), Article 6. https://doi.org/10.3390/app13064040

Doulah, Md. S. (2023). *Performance Evaluation of Machine Learning Algorithm in Various Datasets*. *3*, 14–32. https://doi.org/10.55529/jaimlnn.32.14.32

Dwivedi, R., Dave, D., Naik, H., Singhal, S., Omer, R., Patel, P., Qian, B., Wen, Z., Shah, T., Morgan, G., & Ranjan, R. (2023). Explainable AI (XAI): Core Ideas, Techniques, and Solutions. *ACM Computing Surveys*, *55*(9), 194:1-194:33. https://doi.org/10.1145/3561048

Elkhatat, A., Elsaid, K., & Almeer, S. (2023). Evaluating the efficacy of AI content detection tools in differentiating between human and AI-generated text. *International Journal for Educational Integrity*, *19*. https://doi.org/10.1007/s40979-023-00140-5

Fang, L., Jin, J., Segers, A., Lin, H. X., Pang, M., Xiao, C., Deng, T., & Liao, H. (2022). Development of a regional feature selection-based machine learning system (RFSML v1.0) for air pollution forecasting over China. *Geoscientific Model Development*, *15*(20), 7791–7807. https://doi.org/10.5194/gmd-15-7791-2022

Ferigo, A., Custode, L. L., & Iacca, G. (2023). Quality–diversity optimization of decision trees for interpretable reinforcement learning. *Neural Computing and Applications*. https://doi.org/10.1007/s00521-023-09124-5

Fidel, G., Bitton, R., & Shabtai, A. (2020). *When Explainability Meets Adversarial Learning: Detecting Adversarial Examples using SHAP Signatures*. 1–8. https://doi.org/10.1109/IJCNN48605.2020.9207637

Fürnkranz, J., Kliegr, T., & Paulheim, H. (2018). On Cognitive Preferences and the Interpretability of Rule-based Models. *ArXiv*. https://www.semanticscholar.org/paper/On-

Cognitive-Preferences-and-the-Interpretability-F%C3%BCrnkranz-Kliegr/1ae1a73d88361dca313fdf049c2d2a1bf728ec7d

Garvin, M., Prates, E., Pavicic, M., Jones, P., Amos, B., Geiger, A., Shah, M., Streich, J., Gazolla, J., Kainer, D., Cliff, A., Romero, J., Keith, N., Brown, J., & Jacobson, D. (2020). Potentially adaptive SARS-CoV-2 mutations discovered with novel spatiotemporal and explainable AI models. *Genome Biology*, *21*. https://doi.org/10.1186/s13059-020-02191-0

Glockner, M., Hou, Y., & Gurevych, I. (2022). *Missing Counter-Evidence Renders NLP Fact-Checking Unrealistic for Misinformation* (arXiv:2210.13865). arXiv. https://doi.org/10.48550/arXiv.2210.13865

Gnanasankaran, N., Rakesh, G., & Manikumar, T. (2022). *Multidisciplinary Applications of Machine Learning* (pp. 41–57). https://doi.org/10.1201/9781003240310-3

Golizadeh Akhlaghi, Y., Aslansefat, K., Zhao, X., Sadati, S., Badiei, A., Xiao, X., Shittu, S., Fan, Y., & Ma, X. (2021). Hourly performance forecast of a dew point cooler using explainable Artificial Intelligence and evolutionary optimisations by 2050. *Applied Energy*, *281*, 116062. https://doi.org/10.1016/j.apenergy.2020.116062

Hamilton, R. I., & Papadopoulos, P. N. (2024). Using SHAP Values and Machine Learning to Understand Trends in the Transient Stability Limit. *IEEE Transactions on Power Systems*, *39*(1), 1384–1397. https://doi.org/10.1109/TPWRS.2023.3248941

Haque, A. B., Islam, A. K. M. N., & Mikalef, P. (2023). Explainable Artificial Intelligence (XAI) from a user perspective: A synthesis of prior literature and problematizing avenues for future research. *Technological Forecasting and Social Change*, *186*, 122120. https://doi.org/10.1016/j.techfore.2022.122120

Haresamudram, K., Larsson, S., & Heintz, F. (2023). Three Levels of AI Transparency. *Computer*, *56*(2), 93–100. https://doi.org/10.1109/MC.2022.3213181

Hoffman, R., Mueller, S. T., Klein, G., & Litman, J. (2018). Metrics for Explainable AI: Challenges and Prospects. *ArXiv*. https://www.semanticscholar.org/paper/Metrics-for-Explainable-AI%3A-Challenges-and-Hoffman-Mueller/be711f681580d3a02c8bc4c4dab0c7a043f4e1d2

Hoffman, R. R., Mueller, S. T., Klein, G., & Litman, J. (2023). Measures for explainable AI: Explanation goodness, user satisfaction, mental models, curiosity, trust, and human-AI performance. *Frontiers in Computer Science*, *5*. https://doi.org/10.3389/fcomp.2023.1096257

Hsiao, J., Ngai, H. H. T., Luyu, Q., Yang, Y., & Cao, C. (2021). *Roadmap of Designing Cognitive Metrics for Explainable Artificial Intelligence (XAI)*.

Hussain, I. (2023). An Explainable EEG-Based Human Activity Recognition Model Using Machine-Learning Approach and LIME. *Sensors*, *23*. https://doi.org/10.3390/s23177452

Izza, Y., Ignatiev, A., & Marques-Silva, J. (2022). On Tackling Explanation Redundancy in Decision Trees. *Journal of Artificial Intelligence Research*, *75*, 261–321. https://doi.org/10.1613/jair.1.13575

Kanneganti, D. (2023). *A New cross-domain strategy based XAI models for fake news detection* (arXiv:2302.02122). arXiv. https://doi.org/10.48550/arXiv.2302.02122

Karran, A. J., Demazure, T., Hudon, A., Senecal, S., & Léger, P.-M. (2022). Designing for Confidence: The Impact of Visualizing Artificial Intelligence Decisions. *Frontiers in Neuroscience*, *16*. https://doi.org/10.3389/fnins.2022.883385

Khani, P., Moeinaddini, E., Abnavi, N. D., & Shahraki, A. (2024). Explainable artificial intelligence for feature selection in network traffic classification: A comparative study. *Transactions on Emerging Telecommunications Technologies*, *35*(4), e4970. https://doi.org/10.1002/ett.4970

Kırboğa, K. K., Abbasi, S., & Küçüksille, E. U. (2023). Explainability and white box in drug discovery. *Chemical Biology & Drug Design*, *102*(1), 217–233. https://doi.org/10.1111/cbdd.14262

Klerings, I., Robalino, S., Booth, A., Escobar-Liquitay, C. M., Sommer, I., Gartlehner, G., Devane, D., & Waffenschmidt, S. (2023). Rapid reviews methods series: Guidance on literature search. *BMJ Evidence-Based Medicine*, *28*(6), 412–417. https://doi.org/10.1136/bmjebm-2022-112079

Kozik, R., Ficco, M., Pawlicka, A., Pawlicki, M., Palmieri, F., & Choraś, M. (2024). When explainability turns into a threat—Using xAI to fool a fake news detection method. *Computers and Security*, *137*. Scopus. https://doi.org/10.1016/j.cose.2023.103599

Kurasinski, L., & Mihailescu, R.-C. (2020). *Towards Machine Learning Explainability in Text Classification for Fake News Detection*. 775–781. Scopus. https://doi.org/10.1109/ICMLA51294.2020.00127

Laatifi, M., Douzi, S., Ezzine, H., el Asry, C., Naya, A., Bouklouze, A., Younes, Z., & Naciri, M. (2023). Explanatory predictive model for COVID-19 severity risk employing machine learning, Shapley addition, and LIME. *Scientific Reports*, *13*. https://doi.org/10.1038/s41598-023-31542-7

Laato, S., Tiainen, M., Najmul Islam, A. K. M., & Mäntymäki, M. (2022). How to explain AI systems to end users: A systematic literature review and research agenda. *Internet Research*, *32*(7), 1–31. https://doi.org/10.1108/INTR-08-2021-0600

Lanham, T., Chen, A., Radhakrishnan, A., Steiner, B., Denison, C., Hernandez, D., Li, D., Durmus, E., Hubinger, E., Kernion, J., Lukošiūtė, K., Nguyen, K., Cheng, N., Joseph, N., Schiefer, N.,

Rausch, O., Larson, R., McCandlish, S., Kundu, S., & Perez, E. (2023). *Measuring Faithfulness in Chain-of-Thought Reasoning*.

Lehmann, C., Haubitz, C., Fügener, A., & Thonemann, U. (2022). The risk of algorithm transparency: How algorithm complexity drives the effects on use of advice. *Production and Operations Management*, *31*. https://doi.org/10.1111/poms.13770

Li, B., Qi, P., Liu, B., Di, S., Liu, J., Pei, J., Yi, J., & Zhou, B. (2022). *Trustworthy AI: From Principles to Practices* (arXiv:2110.01167). arXiv. https://doi.org/10.48550/arXiv.2110.01167

Ll, D., Liu, Y., Huang, J., & Wang, Z. (2022). *A Trustworthy View on XAI Method Evaluation*. https://doi.org/10.36227/techrxiv.21067438

Liu, J.-J., & Liu, J.-C. (2022). Permeability Predictions for Tight Sandstone Reservoir Using Explainable Machine Learning and Particle Swarm Optimization. *Geofluids*, *2022*, e2263329. https://doi.org/10.1155/2022/2263329

Liyanagamage, N., & Fernando, M. (2023). Machiavellian leadership in organisations: A review of theory and research. *Leadership & Organization Development Journal*, *44*(6), 791–811. https://doi.org/10.1108/LODJ-07-2022-0309

Lopardo, G., Precioso, F., & Garreau, D. (2023). *Understanding Post-hoc Explainers: The Case of Anchors*. https://doi.org/10.48550/ARXIV.2303.08806

Love, P. E. D., Fang, W., Matthews, J., Porter, S., Luo, H., & Ding, L. (2023). Explainable artificial intelligence (XAI): Precepts, models, and opportunities for research in construction. *Advanced Engineering Informatics*, *57*, 102024. https://doi.org/10.1016/j.aei.2023.102024

Lundstrom, D., Huang, T., & Razaviyayn, M. (2022). *A Rigorous Study of Integrated Gradients Method and Extensions to Internal Neuron Attributions*.

Maass, W., & Storey, V. C. (2021). Pairing conceptual modeling with machine learning. *Data & Knowledge Engineering*, *134*, 101909. https://doi.org/10.1016/j.datak.2021.101909

Manoharan, H., Yuvaraja, T., Kuppusamy, R., & Radhakrishnan, A. (2023). Implementation of explainable artificial intelligence in commercial communication systems using micro systems. *Science Progress*, *106*(3). Scopus. https://doi.org/10.1177/00368504231191657

Marcinkevics, R., & Vogt, J. (2023). Interpretable and explainable machine learning: A methods-centric overview with concrete examples. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *13*. https://doi.org/10.1002/widm.1493

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

Mohammed, K., & Shehu, A. (2023). A REVIEW OF ARTIFICIAL INTELLIGENCE (AI) CHALLENGES AND FUTURE PROSPECTS OF EXPLAINABLE AI IN MAJOR FIELDS: A CASE STUDY OF NIGERIA. *Open Journal of Physical Science (ISSN: 2734-2123)*, *4*(1), Article 1. https://doi.org/10.52417/ojps.v4i1.458

Nauta, M., Trienes, J., Pathak, S., Nguyen, E., Peters, M., Schmitt, Y., Schlötterer, J., van Keulen, M., & Seifert, C. (2023). From Anecdotal Evidence to Quantitative Evaluation Methods: A Systematic Review on Evaluating Explainable AI. *ACM Computing Surveys*, *55*(13s), 1–42. https://doi.org/10.1145/3583558

Nyberg, A. J., Cragun, O. R., Conroy, S. A., & Weller, I. (2024). Artificial Intelligence and Pay Information Disclosure: Changing How Pay is Communicated. *Compensation & Benefits Review*, *56*(2), 58–75. https://doi.org/10.1177/08863687231195477

Parsa, A. B., Movahedi, A., Taghipour, H., Derrible, S., & Mohammadian, A. (Kouros). (2020). Toward safer highways, application of XGBoost and SHAP for real-time accident detection and feature analysis. *Accident Analysis & Prevention*, *136*, 105405. https://doi.org/10.1016/j.aap.2019.105405

Patel, A. U., Gu, Q., Esper, R., Maeser, D., & Maeser, N. (2024). The Crucial Role of Interdisciplinary Conferences in Advancing Explainable AI in Healthcare. *BioMedInformatics*, *4*(2), Article 2. https://doi.org/10.3390/biomedinformatics4020075

Pudjihartono, N., Fadason, T., Kempa-Liehr, A. W., & O'Sullivan, J. M. (2022). A Review of Feature Selection Methods for Machine Learning-Based Disease Risk Prediction. *Frontiers in Bioinformatics*, *2*. https://doi.org/10.3389/fbinf.2022.927312

Qu, Z., & Hullman, J. (2016). *Evaluating Visualization Sets: Trade-offs Between Local Effectiveness and Global Consistency*. 44–52. https://doi.org/10.1145/2993901.2993910

Quinn, B. (2023). *Explaining AI in Finance: Past, Present, Prospects* (arXiv:2306.02773). arXiv. https://doi.org/10.48550/arXiv.2306.02773

Ree, E., Wiig, S., Seljemo, C., Wibe, T., & Lyng, H. B. (2022). Managers' strategies in handling the COVID-19 pandemic in Norwegian nursing homes and homecare services. *Leadership in Health Services*, *36*(2), 200–218. https://doi.org/10.1108/LHS-05-2022-0052

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why Should I Trust You?": Explaining the Predictions of Any Classifier. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1135–1144. https://doi.org/10.1145/2939672.2939778

Sahoo, S. S., Kobow, K., Zhang, J., Buchhalter, J., Dayyani, M., Upadhyaya, D. P., Prantzalos, K., Bhattacharjee, M., Blumcke, I., Wiebe, S., & Lhatoo, S. D. (2022). Ontology-based feature

engineering in machine learning workflows for heterogeneous epilepsy patient records. *Scientific Reports*, *12*(1), 19430. https://doi.org/10.1038/s41598-022-23101-3

Saraswat, D., Bhattacharya, P., Verma, A., Prasad, V. K., Tanwar, S., Sharma, G., Bokoro, P. N., & Sharma, R. (2022). Explainable AI for Healthcare 5.0: Opportunities and Challenges. *IEEE Access*, *10*, 84486–84517. https://doi.org/10.1109/ACCESS.2022.3197671

Sharma, J., Mittal, M. L., & Soni, G. (2023). *Explainable artificial intelligence (XAI) enabled anomaly detection and fault classification of an industrial asset*. https://doi.org/10.21203/rs.3.rs-2780708/v1

Shen, M. (2022). *Trust in AI: Interpretability is not necessary or sufficient, while black-box interaction is necessary and sufficient*.

Slack, D., Hilgard, S., Jia, E., Singh, S., & Lakkaraju, H. (2020). *Fooling LIME and SHAP: Adversarial Attacks on Post hoc Explanation Methods* (arXiv:1911.02508). arXiv. https://doi.org/10.48550/arXiv.1911.02508

Sood, A., & Craven, M. (2021). *Feature Importance Explanations for Temporal Black-Box Models*.

Sovrano, F., & Vitali, F. (2022). *How to Quantify the Degree of Explainability: Experiments and Practical Implications*. 1–9. https://doi.org/10.1109/FUZZ-IEEE55066.2022.9882574

Tiwari, R. (2023). Explainable AI (XAI) and its Applications in Building Trust and Understanding in AI Decision Making. *INTERANTIONAL JOURNAL OF SCIENTIFIC RESEARCH IN ENGINEERING AND MANAGEMENT*, *07*. https://doi.org/10.55041/IJSREM17592

Uddin, M., Li, L., Deng, B., & Ye, J. (2023). Interpretable XGBoost–SHAP machine learning technique to predict the compressive strength of environment-friendly rice husk ash concrete. *Innovative Infrastructure Solutions*, *8*. https://doi.org/10.1007/s41062-023-01122-9

Valentino, M., & Freitas, A. (2022). *Scientific Explanation and Natural Language: A Unified Epistemological-Linguistic Perspective for Explainable AI* (arXiv:2205.01809). arXiv. https://doi.org/10.48550/arXiv.2205.01809

Van Quan, T. (2023). Selection of single machine learning model for designing compressive strength of stabilized soil containing lime, cement and bitumen. *Journal of Intelligent & Fuzzy Systems*, *45*, 1–18. https://doi.org/10.3233/JIFS-222899

Wang, Q., Huang, Y., Jasin, S., & Singh, P. V. (2020). *Algorithmic Transparency with Strategic Users* (SSRN Scholarly Paper 3652656). https://doi.org/10.2139/ssrn.3652656

Wang, Y.-C., & Chen, T. (2024). Adapted techniques of explainable artificial intelligence for explaining genetic algorithms on the example of job scheduling. *Expert Systems with Applications*, *237*. Scopus. https://doi.org/10.1016/j.eswa.2023.121369

Weber, L., Lapuschkin, S., Binder, A., & Samek, W. (2023). Beyond explaining: Opportunities and challenges of XAI-based model improvement. *Information Fusion*, *92*, 154–176. https://doi.org/10.1016/j.inffus.2022.11.013

Zahoor, K., Bawany, N. Z., & Qamar, T. (2024). Evaluating text classification with explainable artificial intelligence. *IAES International Journal of Artificial Intelligence*, *13*(1), 278–286. Scopus. https://doi.org/10.11591/ijai.v13.i1.pp278-286

Zerilli, J. (2022). Explaining Machine Learning Decisions. *Philosophy of Science*, *89*(1), 1–19. https://doi.org/10.1017/psa.2021.13

Zhang, Y., Song, K., Sun, Y., Tan, S., & Udell, M. (2019). "Why Should You Trust My Explanation?" Understanding Uncertainty in LIME Explanations. *arXiv: Learning*. https://www.semanticscholar.org/paper/%22Why-Should-You-Trust-My-Explanation%22-Understanding-Zhang-Song/468d0b4bfcbdf0d9c14f49efda196945b3e28cef

Zhou, J., Chen, F., Berry, A., Reed, M., Zhang, S., & Savage, S. (2020). A Survey on Ethical Principles of AI and Implementations. *2020 IEEE Symposium Series on Computational Intelligence (SSCI)*, 3010–3017. https://doi.org/10.1109/SSCI47803.2020.9308437

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

Universidade Nova de Lisboa