Explaining AI Decisions to Bank Customers: A Systematic Literature Review

Xanti Lizanzu
x.l.lizanzu@student.utwente.nl
University of Twente
Enschede, Netherlands

ABSTRACT

The increasing usage of black-box Artificial Intelligence (AI) models in banking has caused a rise in demand for Explainable AI (XAI) methods. Bank customers, an important target audience for XAI methods, are in need of XAI methods that explain decisions made by AI models. This study focuses on reviewing model-agnostic methods which can be applied to the model of any bank. Through a systematic literature review, this study examines studies on the application of model-agnostic XAI methods in credit risk assessment and customer segmentation. The results include showcasing the methods used, categorising them into classes, and indicating their level of globality. The study found that there is some existing literature on applying model-agnostic methods to explain decisions on credit risk assessment and customer segmentation, mostly feature-based.

KEYWORDS

Systematic Literature Review, Bank Customers, Credit Risk, Customer Segmentation, Model-agnostic, XAI

ACM Reference Format:
Xanti Lizanzu. 2024. Explaining AI Decisions to Bank Customers: A Systematic Literature Review. In *41st Twente Student Conference on IT, July 05, 2024, Enschede, the Netherlands.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

Banks have adopted AI systems for various applications [22]. Among these applications, different types of AI tools have been explored for assessing credit risk of bank customers [9] [43] [8]. Credit risk is defined as the risk of financial loss when a loan borrower fails to repay the loan. Financial institutions conduct credit risk assessments using methods such as classification (predicting defaults) and regression (predicting credit scores) [9]. Both methods use information about the loan applicant and the application to output a prediction, but output either a qualitative or quantitative risk, respectively. Machine learning (ML) techniques for both classification and regression exist, and they are also applied in the domain of credit risk assessment [9]. Bhatore et al. [9] reviewed 136 papers on machine learning techniques for credit risk evaluation, including research from Zhao et al. [55] who applied a neural network to predict defaults and Huang et al. [28] who applied a support vector machine (SVM) to credit scoring.

Another application of ML in the banking sector is customer segmentation [22]. Fares et al. [22] mention for example the study from Smeureanu et al. [46], which implements ML to segment customers because segmentation is important to gain new customers and get higher value out of the established customers.

However, the increasing use of ML techniques, which are black-box models, caused a demand for explainability of these techniques [35]. XAI addresses this issue by developing a range of methods to explain AI decisions.

The importance of XAI is also highlighted by the new EU's AI Act [21], approved by the Council of the EU in May 2024 [15]. The act follows a risk-based approach, where AI methods of higher risk have more restricting rules [15]. Credit risk assessment is considered to be a high-risk AI system, which is the highest risk excluding the systems with such high risk that they are prohibited, meaning that these systems must adhere to strict explainability [21]. For customer segmentation, which is not mentioned as high-risk in the act, explainability is still important because it might individually grant or not grant access to services.

XAI has various target audiences, as can be seen in Figure 1. This study focuses on bank customers affected by model decisions in credit risk assessment and customer segmentation. These customers mostly want to understand their situation and verify whether a decision is fair [6]. Understanding a situation can be achieved by locally explaining a model, opposing that is explaining a model globally (see Figure 2).

Although Figure 2 addresses interpretability methods, the correct term according to Barredo Arrieta et al. [6] would be understandability methods. This study follows the definitions from Barredo Arrieta et al. [6]:

- **Understandability**: the characteristic of a model to make a human understand how a model works without understanding its structure or underlying algorithm;
- **Interpretability**: the ability to explain or to provide the meaning in understandable terms to a human;
- and **Explainability**: an interface between humans and a decision maker.

Here one can see interpretability and explainability both contribute to understandability, but interpretability is a characteristic of a model and explainability covers the techniques used to make a non-interpretable model into an explainable one. The purpose of this study thus is explaining black-box and complex models through posthoc analysis (see Figure 2).

TScIT41, July 05, 2024, Enschede, the Netherlands

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

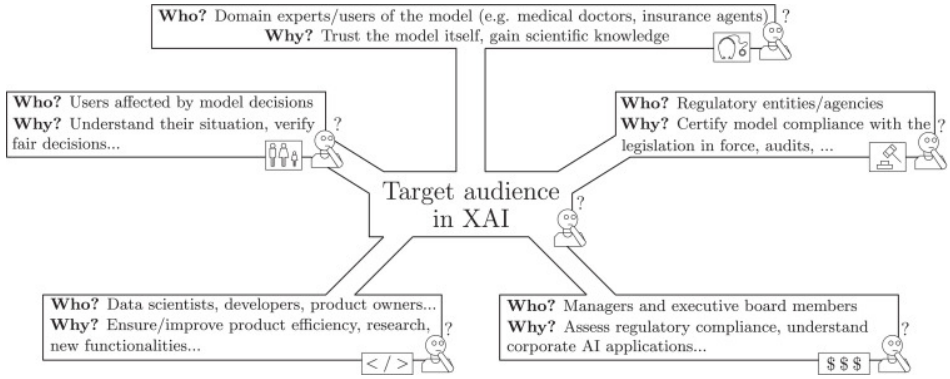


Figure 1: Target audiences in XAI [6]

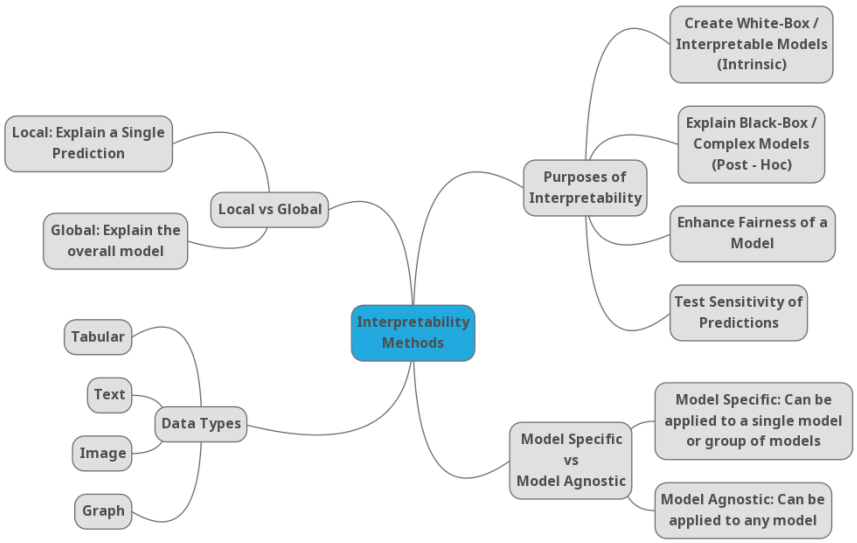


Figure 2: Overview of XAI landscape [30]

Figure 2 categorises explanation methods into model-specific and model-agnostic methods. This study will only research model-agnostic methods, since these methods are not dependent on the model of a bank.

To get an overview of XAI in finance, this study conducts a systematic literature review (SLR) of existing literature on XAI in credit risk assessment and bank customer segmentation. It aims to discover model-agnostic methods, relevant to bank customers. Applications of and future research into these methods is much needed with the newly approved EU AI Act. This study will answer the research question (RQ): *what model-agnostic XAI method(s) exist for bank customers*? Since both credit risk assessment and bank customer segmentation are researched the following two sub-RQs (SRQs) will be answered to answer the RQ:

- **SRQ1**: what model-agnostic XAI research exists on credit risk assessment?
- **SRQ2**: What model-agnostic XAI research exists on bank customer segmentation?

2 METHODOLOGY

This SLR follows the framework described by Varsha et al. [50]. The research issue was defined as the two SRQs. For SRQ2, the scope was extended from bank customer segmentation to customer segmentation due to limited results.

Scopus was selected as the database, for its easy-to-use support of Boolean operators, exporting results to CSV format. It is preferred by Varsha et al. [50] over other databases since it has more formal indexing criteria. Filters within Scopus were applied to filter the type and language of the publication.

Initially, the search included articles from 2016 onwards. However, including these results produced too much literature to review, and thus articles published before 2019 were excluded. Any literature that could not be accessed freely was excluded. Any article that did not discuss applying XAI methods to credit risk or customer segmentation, or did only discuss applying interpretable methods was excluded.

Table 1: Criteria for research

Search terms in Title, Keywords, or Abstract	Inclusion
Of type article	Inclusion
In English	Inclusion
Published in 2018 or earlier	Exclusion
Duplicate in findings	Exclusion
Not relevant to the topic	Exclusion
About interpretable models	Exclusion
Not freely available	Exclusion

Two distinct queries were composed with the keywords that can be seen in Table 2. The domains were joined by the AND operator like (Understandability) AND (AI) AND (Credit risk) and (Understandability) AND (AI) AND (Customer segmentation).

The understandability keywords are the same as the "Explainability" keywords in [51]. The AI keywords were derived from Bhatore et al. [9]. For credit risk, the keywords are straightforward and for customer segmentation inspiration was taken from Amato et al. [3]. It could not be found if Scopus applies stemming, so wildcards were used.

Table 2: Search keywords per domain

Domain	Keywords
Understandability	“Transparen*” OR “explain*” OR "explanat*" OR “interpret*” OR “black box” OR “white box”
AI	“AI” OR “artifcial intelligen*” OR “ML” OR “machine learning” OR “classification” OR “supervised” OR “unsupervised” OR “deep learning” OR “neural network*” OR “radial basis function networks” OR “SVM*” OR “support vector machine*” OR “decision tree*” OR “discriminant analysis” OR “naive bayes” OR “nearest neighbor*” OR “random forest*” OR “hidden markov” OR “markov chain*” OR “regression” OR “fuzzy logic” OR “expert system*”
Credit risk	"Credit risk" OR "credit scor*"
Customer segmentation	"Customer segment*" OR "cluster*" OR "credit portfolio"

The query results[1] were exported to CSV files and processed by a Python script to be formatted correctly to an Excel file. The results included the authors, document title, year, source title, citation count, abstract, and author keywords.

Dwivedi et al. [20] classifies model-agnostic techniques as either feature- or example-based. Since this is a clear classification and there is a lack of a better way to classify techniques, it was chosen for analysis in this article.

To produce data for the results, Excel was used to count techniques and the globality of the techniques. If an article includes multiple techniques or explains decisions at varying levels, all were counted.

[1]Both queries were executed at May 16, 2024.

3 RESULTS

The detailed list of papers can be found in Table B.1 and Table B.2 for credit risk and customer segmentation, respectively. Figure 3 shows how many papers were returned through the query and eventually used in the analysis.

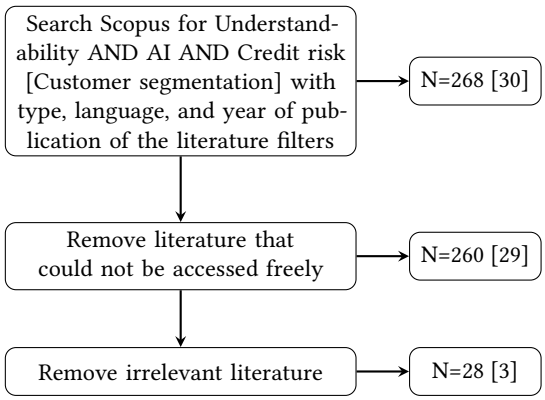


Figure 3: Literature selection process

Figure 4 shows the trend of publications. While overall an incline through the years can be observed, a peak in 2022 occurs for credit risk, while the low results in 2024 can be explained because the year is only halfway through. With only 28 articles researched the peak is not significant. Regarding customer segmentation, no significant results can be observed.

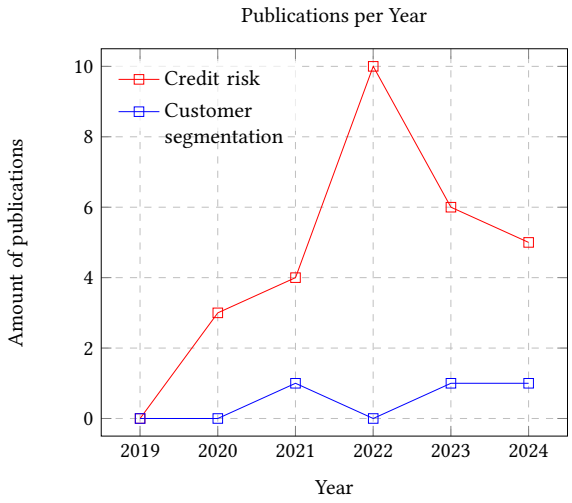
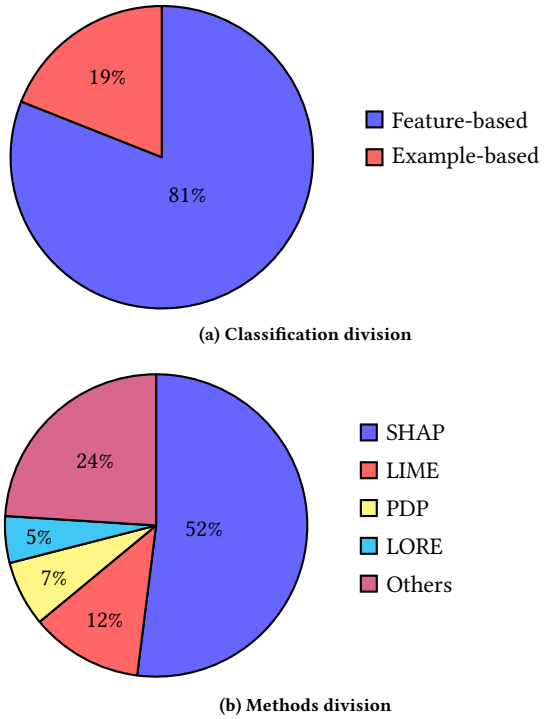


Figure 4: XAI Publications per Year

In response to the EU's AI act, Chen et al., Alonso Robisco and Carbó Martínez, Hamon et al., and de Lange et al. [2, 13, 18, 24] all mentioned the importance of XAI research regarding this act.

3.1 Answer to SRQ1: XAI research on credit risk



(a) Classification division

(b) Methods division

Figure 5: XAI methods in credit risk

3.1.1 Methods. A broad range of techniques were found used in credit risk assessment. The majority of techniques are feature-based (see Figure 5a). Among these, are SHapely Additive SHapley Additive exPlanations (SHAP) [34] and Local Interpretable Model-agnostic Explanations (LIME) [40], making up for 64% of the methods used (Figure 5b). Both methods calculate feature attribution: LIME creates an interpretable linear model around a data point it explains and from this model, the feature attributions can be retrieved [40], and SHAP calculates SHAP values using either Shapley sampling values, known from game theory, or KernelSHAP which uses LIME to create a linear model around the data point it explains and calculates SHAP values with that model [34].

Several studies explored both LIME and SHAP methods to explain models. Nallakaruppan et al. [37] use both to explain a Random Forest model. Chen et al. [13] evaluate the stability of both methods on datasets of progressively increased class imbalance. Moscato et al. [36] propose a benchmarking study comparing five XAI methods including LIME and SHAP. Additionally, Dastile et al. [16] convert tabular data into 2D images on which then CNNs are applied and explained among other things by LIME and SHAP. The one study to use only LIME is from Aljadani et al. [1], using it to explain instances out of 3 datasets.

SHAP, accounting for 52% of the methods used, has been applied in various studies, with some explaining things locally. Hjelkrem and Lange [26] apply a deep learning model to textual descriptions of transactions from customers to predict whether the customer defaulted in the following 12 months, explain globally the most important features and include some examples of local explanations. Do et al. [19] include a local explanation of a logistic regression model, repeated random subsampling estimation of feature SHAP values for all its models, and global feature importance for all models. Liu et al. [32] display a wide range of plots using SHAP, including an absolute mean SHAP ranking for each feature, SHAP LZ value ranking for each feature, the overall distribution of SHAP values for each feature, distribution of SHAP values for each feature, and a local explanation for both a defaulted and non-defaulted class. de Lange et al. [18] include a SHAP variable importance ranking, an overall distribution of SHAP values for each feature, and local explanations. Liu et al. [31] use the variation of SHAP specifically designed for tree-based models, TreeSHAP[2] [33], to create explanations globally and two decisions locally. TreeSHAP makes use of the tree structure, making it computationally more efficient [33].

Other methods simply use SHAP to give a global overview of the most important features. Talaat et al., Wen et al., Onari et al., and Bastos and Matos [7, 39, 48, 52] simply give feature rankings of their model. Xia et al. [53] do so as well but use TreeSHAP. Bueff et al. [11] use it on a dataset and the augmented version of the dataset, which was created to check robustness. Xia et al. [54] use SHAP on two models on two different datasets. Nwafor and Nwafor [38] give this overview by comparing 6 different models, Alonso Robisco and Carbó Martínez [2] for 3 different models, and Ariza-Garzon et al. [5] explains 2 models. Hamon et al. [24] explain in a radar chart the two most important features for 4 models and each model for 4 subsets of the dataset.

In addition to LIME and SHAP, feature attribution was also calculated using a permutation-based variable importance method, from Breiman [10], in the paper by Hu et al. [27]. This permutation-based technique measures how random permutations of all variables, except the one being tested, impact prediction accuracy, with great difference in impact indicating a high importance of a variable.

Furthermore, Dastile and Celik [16], which also use LIME and SHAP, explains features with the convolution explanation methods Gradient weighted Class Activation Map (Grad-CAM) [42] and Saliency maps [44]. GRAD-Cam uses gradients to visualise and localise important regions of input images [42], and saliency maps, generated through a single back-propagation pass in a neural network, identifies where a specific class is spatially supported within an image [44].

Some techniques look at the influence of feature values on predictions. The most used method for this is visualising the influence in Partial Dependence Plots (PDPs), comprising 7% (see Figure 5b) of the used methods. Nallakaruppan et al. [37] use these plots to both show the influence of variables individually in a 2D plot and their combined influence in a 3D plot. Szepannek and Lübke [47] use 2D PDPs with a trellis visualisation to visualise the interaction of two variables simultaneously. Hu et al. [27] use both 2D PDPs

[2]TreeSHAP is classified as SHAP in this study

and 3D PDPs, making 3D plots through multiple line plots in a single graph.

Accumulated Local Effects (ALE) plots are another way to visualise feature effects [4]. ALE plots were invented in response to PDPs which are considered faulty if the features are strongly correlated. They calculate the local effect of small changes in feature value on the model's prediction and accumulate these effects on the variable's range to provide an overall feature influence plot. Bastos and Matos [7] use these ALE plots in a trellis visualisation to explain two variables at the same time.

Example-based methods account for 19% of the found methods (see Figure 5a). The method most used is the counterfactual generator Local Rule-Based Explanations (LORE), which trains a decision tree on data generated by a genetic algorithm and extracts a rule for the decision and several counterfactual rules from the interpretable decision tree. This method is used in the benchmarking study by Moscato et al. [36] and in the study by Bueff et al. [11] to assess its robustness as an XAI method.

An unnamed method similar to LORE is proposed by Dastile et al. [17], which uses an optimisation-based method to generate sparse counterfactuals by minimising changes to the input features while achieving the desired outcome [17]. In the paper of invention the method is also applied to credit risk data and compared to other counterfactual rule generating methods [17].

Another method for creating counterfactual rules is Multi-Objective Game-based Counterfactual Explanation (MOGCE) [39]. MOGCE, applied in its paper to credit risk data, finds the closest data point with reversed prediction to a certain point through Multi-Objective Particle Swarm Optimization (MOPSO), and incorporates the Prisoner's Dilemma during optimisation [39]. The paper filters out non-actionable features for customers such that the counterfactual explanations produced are actionable.

Building on counterfactual rules, Directive Explanations (DEs) are a subset of counterfactual rules where the explanations are ones that the individual could perform [45]. DEs are proposed by Singh et al. [45] and applied to the credit risk domain. The study uses the Markov Decision Process to get from counterfactual rules to DEs [45].

Some methods create rules as explanations using IF, AND, and THEN, which can be visualised as a decision tree. Hayashi and Takano [25] create these rules by applying Recursive-Rule Extraction (Re-RX) with a J48graft decision tree. Moscato et al. [36] use Balanced English Explanations of Forecasts (BEEF) [23][3] and Anchors [41] in their benchmarking study. Anchors, aiming to improve precision in comparison to LIME, use a beam-search algorithm to iteratively find and optimise rules [41].

All in all numerous studies on XAI on credit risk were reviewed. Feature-based examples are the most popular type of methods used, with LIME and SHAP being the most prevalent methods used, while SHAP is used over 4 times more than LIME.

3.1.2 Globality. Both local and global explanations were found, with the majority, around 65%, being a global explanation (see Figure 6). Papers including local explanations are those using LIME [1, 13, 16, 37] and using counterfactual explanations [11, 17, 25, 39], with only Moscato et al. [36] using both. All other explanations were only global.

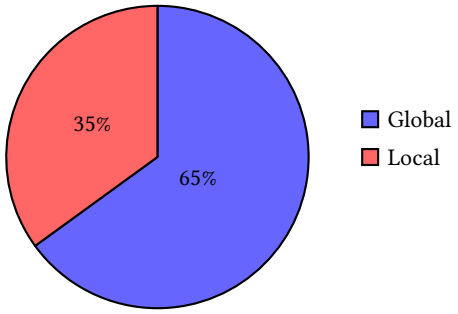


Figure 6: Globality division of XAI methods

3.2 Answer to SRQ2: XAI research on customer segmentation

For customer segmentation, no charts were produced, since only 3 papers were retrieved. Choi et al. [14] use TreeSHAP to explain the most globally significant features and in the same plot for which clusters they are important. Lee et al. [29] come up with there own XAI method. They train a Fuzzy Decision Tree (FDT) along with their segmenting models, and then rules are derived from the FDT [29], which explain the model globally. Talaat et al. [48] use the method DeepLimeSeq which can explain locally for a customer which features were the most important for being assigned a certain segment.

4 DISCUSSION

We found that feature-based explanations are the dominant class of methods used in both credit risk (see Figure 5a) assessment and customer segmentation. For both, SHAP and LIME [1, 2, 5, 7, 11–14, 16, 18, 19, 24, 26, 31, 32, 36–39, 48, 49, 52–54] are the most used methods for feature-based explanations. Other feature-based methods used in credit risk assessment are PDP [37, 47], ALE [7], Grad-CAM [17], Saliency maps [17], and permutation-based variable importance [27]. PDP and ALE stand out to be the only methods that explain feature effects.

For example-based methods, every method was either based on producing IF-THEN rules or counterfactual rules. In credit risk assessment, LORE was the only example-based method used more than once [11, 36]. Other example-based methods used in credit risk assessment are DE [45], MOGCE [39], Anchors [36], BEEF [36], GA [16], and J48graft with Re-RX [25]. The latter is similar to the FDT, the only example-based method used for customer segmentation where rules are derived from a self-developed Decision Tree [29].

5 CONCLUSION

Since an increasing amount of black-box models are deployed by banks the need for XAI methods to explain the decisions is great. Model-agnostic methods are a solution to this problem.

This study has shown there is already literature on applying model-agnostic methods to explain decisions on credit risk assessment and customer segmentation. Feature-based explanations are

[3]This paper could not be accessed so the algorithms behind BEEF are not known.

the more popular class of explanations, but example-based explanations exist as well. SHAP is the most popular method in credit risk assessment and is used for creating feature-based explanations by calculating feature importance. In total 15 different methods were found. Global explanations to explain the model are more in use than local explanations, but local explanations which are important for explaining decisions to bank customers are applied.

5.1 Limitations

5.1.1 Consideration of Interpretable Methods. This paper only focuses on model-agnostic XAI methods, because these methods can explain any model used. However, the query results included a lot of papers about interpretable models. Due to time constraints, they are not part of this study. It was however hard to determine for some models if they are interpretable models or XAI methods. For example, the Fuzzy Decision Tree from Lee et al. [29] is a model but in this case an XAI method.

5.1.2 Customers not explicitly Bank Customers. As can be seen in Figure 3, only 3 papers were retrieved for customer segmentation [14, 29, 48]. Due to this already low amount of results, it was decided not to focus on bank customers explicitly. When reproducing this study in the future one might look into only bank customers but due to a lack of results, this is not done in this study. The results of customer segmentation thus should not be used to draw any further conclusions. It was decided to still include them in this paper such that the found literature can be used to research XAI for bank customer segmentation.

5.1.3 Search Terms and Terminology. As can be seen in Figure 3, a lot of papers were excluded. This was due to a lot of papers not being about the domains credit risk or customer segmentation, and a lot being on interpretable models. The latter could be because of the keywords "transparen*", "interpret*", and "white box". If this research were to be executed again, one would need to better research if these terms should be included. As mentioned in the introduction, Barredo Arrieta et al. [6] make a clear distinction between terms in understandable AI, because there is a lot of wrong use of the terms interpretability and explainability. In this study, this misuse was also found so including the described keywords might have not been a mistake. This makes a systematic literature search into XAI methods more difficult.

5.1.4 Quality of Analysis and Reproducibility. Although much attention was given to the analysis, it could be methods were overlooked and when looking if the datasets used were about loan applicants errors were made. The classification of methods is also hard since there is no good consensus on which XAI methods exist and to which classes they belong. Classifying TreeSHAP as SHAP but not generalising J48graft with Re-RX was done because there is a lack of a framework. In addition, determining whether something is local, global, or both is ambiguous. Lee et al. [29] made an FDT to create rules for clusters, which is a global explanation. For a single instance, the rules can be used to explain the decision locally, but this was not done in the study. Thus for the FDT, it was decided to classify it as a global method.

In addition to this, since there is no consensus about the field of XAI, it is not always clear what counts as an XAI method and what does not. This research relied on studies mentioning terms regarding understandable AI, so it is unknown if relevant papers were not included.

A way to improve the quality of the analysis was to have this be done by different people, but this study did not have the resources to do so. The method of study was described to make it reproducible so that this analysis could be redone the same.

5.2 Future Research

This study only serves as a basis but lays the groundwork for more XAI research to make explanations understandable to bank customers. Future research could delve into many different directions, some recommendations are listed here.

5.2.1 Establishing the XAI Methods Landscape. As mentioned in Section 2 Methodology, there is no clear framework for classifying XAI methods. Since the field is still emerging, this is hard to establish, but much needed when researching XAI. With a framework in place, research into XAI can be more structured.

5.2.2 Researching Broader Literature. The domain could be broadened compared to this study to also include model-specific methods and interpretable models. Also, different databases can be used instead of only Scopus to find more papers to review. This paper could only research papers from 2019 until May 16, 2024, further research could take into account more years.

5.2.3 Executing More Extensive Literature Reviews. With all the found papers, more extensive reviews can be done. This paper researched what methods exist, but did not produce quantitative results about how methods were applied or the actual results of explanation methods. An example of research to produce quantitative results about how methods are applied could be the different types of plots in which SHAP is used.

5.2.4 Comparing Methods. Different methods should also be compared to each other. Similar methods should be compared to see for which use cases they are applicable. An example of this is to compare local explanations by LIME and SHAP. Also combinations of completely different methods, for example, an example-based method with a feature-based method should be reviewed to see what is more effective and useful for bank customers.

5.2.5 Interacting with End Users. A broad range of XAI methods already exist, but little research has been done into how explanation methods are understood by end users. Regarding the previous point, different methods should also be compared in their understandability by end users. In the banking domain, bank customers can be confronted with different methods to see what helps most in their understanding of decisions.

5.2.6 Discovering new XAI Methods and Improving Existing Methods. Although a broad range of methods already exist, more research should be conducted into developing new XAI methods and improving existing methods. The methods found can create powerful explanations to see which features are important and how to reverse a decision by counterfactual rules, but they do not explain exactly how a model reasons.

REFERENCES

[1] Abdussalam Aljadani, Bshair Alharthi, Mohammed A. Farsi, Hossam Magdy Balaha, Mahmoud Badawy, and Mostafa A. Elhosseini. 2023. Mathematical Modeling and Analysis of Credit Scoring Using the LIME Explainer: A Comprehensive Approach. *Mathematics* 11, 19 (2023). https://doi.org/10.3390/math11194055 Cited by: 4; All Open Access, Gold Open Access.

[2] Andrés Alonso Robisco and José Manuel Carbó Martínez. 2022. Measuring the model risk-adjusted performance of machine learning algorithms in credit default prediction. *Financial Innovation* 8, 1 (2022). https://doi.org/10.1186/s40854-022-00366-1 Cited by: 15; All Open Access, Gold Open Access.

[3] Alessandra Amato, Joerg R. Osterrieder, and Marcos R. Machado. 2024. How can artificial intelligence help customer intelligence for credit portfolio management? A systematic literature review. *International Journal of Information Management Data Insights* 4, 2 (2024), 100234. https://doi.org/10.1016/j.jjimei.2024.100234

[4] Daniel W. Apley and Jingyu Zhu. 2020. Visualizing the effects of predictor variables in black box supervised learning models. *Journal of the Royal Statistical Society. Series B: Statistical Methodology* 82, 4 (2020), 1059 – 1086. https://doi.org/10.1111/rssb.12377 Cited by: 493; All Open Access, Green Open Access.

[5] Miller Janny Ariza-Garzon, Javier Arroyo, Antonio Caparrini, and Maria-Jesus Segovia-Vargas. 2020. Explainability of a Machine Learning Granting Scoring Model in Peer-to-Peer Lending. *IEEE Access* 8 (2020), 64873 â€“ 64890. https://doi.org/10.1109/ACCESS.2020.2984412 Cited by: 71; All Open Access, Gold Open Access.

[6] Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador Garcia, Sergio Gil-Lopez, Daniel Molina, Richard Benjamins, Raja Chatila, and Francisco Herrera. 2020. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion* 58 (2020), 82–115. https://doi.org/10.1016/j.inffus.2019.12.012

[7] JoÃ£o A. Bastos and Sara M. Matos. 2022. Explainable models of credit losses. *European Journal of Operational Research* 301, 1 (2022), 386 â€“ 394. https://doi.org/10.1016/j.ejor.2021.11.009 Cited by: 17; All Open Access, Green Open Access.

[8] Imane Rhzioual Berrada, Fatima Zohra Barramou, and Omar Bachir Alami. 2022. A review of Artificial Intelligence approach for credit risk assessment. In *2022 2nd International Conference on Artificial Intelligence and Signal Processing (AISP)*. IEEE, 1–5.

[9] Siddharth Bhatore, Lalit Mohan, and Y Raghu Reddy. 2020. Machine learning techniques for credit risk evaluation: a systematic literature review. *Journal of Banking and Financial Technology* 4, 1 (2020), 111–138.

[10] Leo Breiman. 1996. Bagging predictors. *Machine Learning* 24, 2 (1996), 123 – 140. https://doi.org/10.1023/A:1018054314350 Cited by: 18241; All Open Access, Bronze Open Access.

[11] Andreas C. Bueff, Mateusz CytryÅ„ski, Raffaella Calabrese, Matthew Jones, John Roberts, Jonathon Moore, and Iain Brown. 2022. Machine learning interpretability for a stress scenario generation in credit scoring based on counterfactuals. *Expert Systems with Applications* 202 (2022). https://doi.org/10.1016/j.eswa.2022.117271 Cited by: 13; All Open Access, Green Open Access.

[12] Niklas Bussmann, Paolo Giudici, Dimitri Marinelli, and Jochen Papenbrock. 2020. Explainable AI in Fintech Risk Management. *Frontiers in Artificial Intelligence* 3 (2020). https://doi.org/10.3389/frai.2020.00026 Cited by: 83; All Open Access, Gold Open Access, Green Open Access.

[13] Yujia Chen, Raffaella Calabrese, and Belen Martin-Barragan. 2024. Interpretable machine learning for imbalanced credit scoring datasets. *European Journal of Operational Research* 312, 1 (2024), 357 â€“ 372. https://doi.org/10.1016/j.ejor.2023.06.036 Cited by: 9; All Open Access, Green Open Access, Hybrid Gold Open Access.

[14] Insu Choi, Woosung Koh, Bonwoo Koo, and Woo Chang Kim. 2024. Network-based exploratory data analysis and explainable three-stage deep clustering for financial customer profiling. *Engineering Applications of Artificial Intelligence* 128 (2024). https://doi.org/10.1016/j.engappai.2023.107378 Cited by: 1.

[15] Council of the European Union. 2024. Artificial intelligence (AI) act: Council gives final green light to the first worldwide rules on AI. https://www.consilium.europa.eu/en/press/press-releases/2024/05/21/artificial-intelligence-ai-act-council-gives-final-green-light-to-the-first-worldwide-rules-on-ai/ Accessed: 2024-06-21.

[16] Xolani Dastile and Turgay Celik. 2021. Making Deep Learning-Based Predictions for Credit Scoring Explainable. *IEEE Access* 9 (2021), 50426 â€“ 50440. https://doi.org/10.1109/ACCESS.2021.3068854 Cited by: 29; All Open Access, Gold Open Access.

[17] Xolani Dastile, Turgay Celik, and Hans Vandierendonck. 2022. Model-Agnostic Counterfactual Explanations in Credit Scoring. *IEEE Access* 10 (2022), 69543 â€“ 69554. https://doi.org/10.1109/ACCESS.2022.3177783 Cited by: 9; All Open Access, Gold Open Access, Green Open Access.

[18] Petter Eilif de Lange, Borger Melsom, Christian Bakke Vennerød, and Sjur Westgaard. 2022. Explainable AI for Credit Assessment in Banks. *Journal of Risk and Financial Management* 15, 12 (2022). https://doi.org/10.3390/jrfm15120556 Cited by: 12; All Open Access, Gold Open Access.

[19] Thanh Thuy Do, Golnoosh Babaei, and Paolo Pagnottoni. 2023. Explainable Machine Learning for Credit Risk Management When Features are Dependent. *Measurement* (2023). https://doi.org/10.1080/15366367.2023.2261186 Cited by: 0.

[20] Rudresh Dwivedi, Devam Dave, Het Naik, Smiti Singhal, Rana Omer, Pankesh Patel, Bin Qian, Zhenyu Wen, Tejal Shah, Graham Morgan, and Rajiv Ranjan. 2023. Explainable AI (XAI): Core Ideas, Techniques, and Solutions. *Comput. Surveys* 55, 9 (2023). https://doi.org/10.1145/3561048 Cited by: 102; All Open Access, Green Open Access.

[21] European Union. 2024. Regulation of the European Parliament and of the Council laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act). https://data.consilium.europa.eu/doc/document/PE-24-2024-INIT/en/pdf Accessed: 2024-06-21.

[22] Omar H Fares, Irfan Butt, and Seung Hwan Mark Lee. 2023. Utilization of artificial intelligence in the banking sector: a systematic literature review. *Journal of Financial Services Marketing* 28, 4 (2023), 835–852.

[23] Sachin Grover, Chiara Pulice, Gerardo I. Simari, and V.S. Subrahmanian. 2019. BEEF: Balanced English Explanations of Forecasts. *IEEE Transactions on Computational Social Systems* 6, 2 (2019), 350 – 364. https://doi.org/10.1109/TCSS.2019.2902490 Cited by: 19.

[24] Ronan Hamon, Henrik Junklewitz, Ignacio Sanchez, Gianclaudio Malgieri, and Paul De Hert. 2022. Bridging the Gap between AI and Explainability in the GDPR: Towards Trustworthiness-by-Design in Automated Decision-Making. *IEEE Computational Intelligence Magazine* 17, 1 (2022), 72 â€“ 85. https://doi.org/10.1109/MCI.2021.3129960 Cited by: 35.

[25] Yoichi Hayashi and Naoki Takano. 2020. One-dimensional convolutional neural networks with feature selection for highly concise rule extraction from credit scoring datasets with heterogeneous attributes. *Electronics (Switzerland)* 9, 8 (2020), 1 â€“ 15. https://doi.org/10.3390/electronics9081318 Cited by: 11; All Open Access, Gold Open Access.

[26] Lars Ole Hjelkrem and Petter Eilif de Lange. 2023. Explaining Deep Learning Models for Credit Scoring with SHAP: A Case Study Using Open Banking Data. *Journal of Risk and Financial Management* 16, 4 (2023). https://doi.org/10.3390/jrfm16040221 Cited by: 1; All Open Access, Gold Open Access.

[27] Linwei Hu, Jie Chen, Joel Vaughan, Soroush Aramideh, Hanyu Yang, Kelly Wang, Agus Sudjianto, and Vijayan N. Nair. 2021. Supervised Machine Learning Techniques: An Overview with Applications to Banking. *International Statistical Review* 89, 3 (2021), 573 â€“ 604. https://doi.org/10.1111/insr.12448 Cited by: 4; All Open Access, Green Open Access.

[28] Cheng-Lung Huang, Mu-Chen Chen, and Chieh-Jen Wang. 2007. Credit scoring with a data mining approach based on support vector machines. *Expert Systems with Applications* 33, 4 (2007), 847–856. https://doi.org/10.1016/j.eswa.2006.07.007

[29] Zne-Jung Lee, Chou-Yuan Lee, Li-Yun Chang, and Natsuki Sano. 2021. Clustering and classification based on distributed automatic feature engineering for customer segmentation. *Symmetry* 13, 9 (2021). https://doi.org/10.3390/sym13091557 Cited by: 9; All Open Access, Gold Open Access.

[30] Pantelis Linardatos, Vasilis Papastefanopoulos, and Sotiris Kotsiantis. 2021. Explainable ai: A review of machine learning interpretability methods. *Entropy* 23, 1 (2021), 1 – 45. https://doi.org/10.3390/e23010018 Cited by: 1143; All Open Access, Gold Open Access, Green Open Access.

[31] Wanan Liu, Hong Fan, and Meng Xia. 2022. Credit scoring based on tree-enhanced gradient boosting decision trees. *Expert Systems with Applications* 189 (2022). https://doi.org/10.1016/j.eswa.2021.116034 Cited by: 51.

[32] Yang Liu, Fei Huang, Lili Ma, Qingguo Zeng, and Jiale Shi. 2024. Credit scoring prediction leveraging interpretable ensemble learning. *Journal of Forecasting* 43, 2 (2024), 286 â€“ 308. https://doi.org/10.1002/for.3033 Cited by: 3.

[33] Scott M. Lundberg, Gabriel Erion, Hugh Chen, Alex DeGrave, Jordan M. Prutkin, Bala Nair, Ronit Katz, Jonathan Himmelfarb, Nisha Bansal, and Su-In Lee. 2020. From local explanations to global understanding with explainable AI for trees. *Nature Machine Intelligence* 2, 1 (2020), 56 – 67. https://doi.org/10.1038/s42256-019-0138-9 Cited by: 2928; All Open Access, Green Open Access.

[34] Scott M Lundberg and Su-In Lee. 2017. A Unified Approach to Interpreting Model Predictions. In *Advances in Neural Information Processing Systems*, I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Eds.), Vol. 30. Curran Associates, Inc.

[35] Branka Hadji Misheva, Joerg Osterrieder, Ali Hirsa, Onkar Kulkarni, and Stephen Fung Lin. 2021. Explainable AI in credit risk management. *arXiv preprint arXiv:2103.00949* (2021).

[36] Vincenzo Moscato, Antonio Picariello, and Giancarlo SperlÃ. 2021. A benchmark of machine learning approaches for credit score prediction. *Expert Systems with Applications* 165 (2021). https://doi.org/10.1016/j.eswa.2020.113986 Cited by: 124.

[37] M.K. Nallakaruppan, Balamurugan Balusamy, M. Lawanya Shri, V. Malathi, and Siddhartha Bhattacharyya. 2024. An Explainable AI framework for credit evaluation and analysis. *Applied Soft Computing* 153 (2024). https://doi.org/10.1016/j.asoc.2024.111307 Cited by: 0.

[38] Chioma Ngozi Nwafor and Obumneme Zimuzor Nwafor. 2023. Determinants of non-performing loans: An explainable ensemble and deep neural network approach. *Finance Research Letters* 56 (2023). https://doi.org/10.1016/j.frl.2023.104084 Cited by: 3; All Open Access, Hybrid Gold Open Access.

[39] Mohsen Abbaspour Onari, Mustafa Jahangoshai Rezaee, Morteza Saberi, and Marco S. Nobile. 2024. An explainable data-driven decision support framework for strategic customer development. *Knowledge-Based Systems* 295 (2024). https://doi.org/10.1016/j.knosys.2024.111761 Cited by: 0.

[40] Marco Túlio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. *CoRR* abs/1602.04938 (2016). arXiv:1602.04938

[41] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2018. Anchors: High-precision model-agnostic explanations. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 32.

[42] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. 2017. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. *Proceedings of the IEEE International Conference on Computer Vision* 2017-October (2017), 618 – 626. https://doi.org/10.1109/ICCV.2017.74 Cited by: 11521; All Open Access, Green Open Access.

[43] Si Shi, Rita Tse, Wuman Luo, Stefano D'Addona, and Giovanni Pau. 2022. Machine learning-driven credit risk: a systemic review. *Neural Computing and Applications* 34, 17 (2022), 14327–14339.

[44] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. 2014. Deep inside convolutional networks: Visualising image classification models and saliency maps. *2nd International Conference on Learning Representations, ICLR 2014 - Workshop Track Proceedings* (2014). https://www.scopus.com/inward/record.uri?eid=2-s2.0-85083953896&partnerID=40&md5=5897b0590b10086cbf0bd356292a0908 Cited by: 1680.

[45] Ronal Singh, Tim Miller, Henrietta Lyons, Liz Sonenberg, Eduardo Velloso, Frank Vetere, Piers Howe, and Paul Dourish. 2023. Directive Explanations for Actionable Explainability in Machine Learning Applications. *ACM Transactions on Interactive Intelligent Systems* 13, 4 (2023). https://doi.org/10.1145/3579363 Cited by: 3; All Open Access, Green Open Access.

[46] Ion Smeureanu, Gheorghe Ruxanda, and Laura Maria Badea. 2013. Customer segmentation in private banking sector using machine learning techniques. *Journal of Business Economics and Management* 14, 5 (2013), 923 – 939. https://doi.org/10.3846/16111699.2012.749807 Cited by: 22; All Open Access, Gold Open Access.

[47] Gero Szepannek and Karsten Lübke. 2023. How much do we see? On the explainability of partial dependence plots for credit risk scoring. *Argumenta Oeconomica* 2023, 1 (2023), 137 â€“ 150. https://doi.org/10.15611/aoe.2023.1.07 Cited by: 1.

[48] Fatma M. Talaat, Abdussalam Aljadani, Bshair Alharthi, Mohammed A. Farsi, Mahmoud Badawy, and Mostafa Elhosseini. 2023. A Mathematical Model for Customer Segmentation Leveraging Deep Learning, Explainable AI, and RFM Analysis in Targeted Marketing. *Mathematics* 11, 18 (2023). https://doi.org/10.3390/math11183930 Cited by: 2; All Open Access, Gold Open Access.

[49] Fatma M. Talaat, Abdussalam Aljadani, Mahmoud Badawy, and Mostafa Elhosseini. 2024. Toward interpretable credit scoring: integrating explainable artificial intelligence with deep learning for credit card default prediction. *Neural Computing and Applications* 36, 9 (2024), 4847 â€“ 4865. https://doi.org/10.1007/s00521-023-09232-2 Cited by: 2.

[50] PS Varsha, Amrita Chakraborty, and Arpan Kumar Kar. 2024. How to Undertake an Impactful Literature Review: Understanding Review Approaches and Guidelines for High-impact Systematic Literature Reviews. *South Asian Journal of Business and Management Cases* (2024), 22779779241227654.

[51] Patrick Weber, K Valerie Carl, and Oliver Hinz. 2023. Applications of explainable artificial intelligence in finance—a systematic review of finance, information systems, and computer science literature. *Management Review Quarterly* (2023), 1–41.

[52] Hongmei Wen, Xin Sui, and Shaopeng Lu. 2022. Study on Effect of Consumer Information in Personal Credit Risk Evaluation. *Complexity* 2022 (2022). https://doi.org/10.1155/2022/7340010 Cited by: 1; All Open Access, Gold Open Access.

[53] Yufei Xia, Yinguo Li, Lingyun He, Yixin Xu, and Yiqun Meng. 2021. Incorporating multilevel macroeconomic variables into credit scoring for online consumer lending. *Electronic Commerce Research and Applications* 49 (2021). https://doi.org/10.1016/j.elerap.2021.101095 Cited by: 16.

[54] Yufei Xia, Zijun Liao, Jun Xu, and Yinguo Li. 2022. FROM CREDIT SCORING TO REGULATORY SCORING: COMPARING CREDIT SCORING MODELS FROM A REGULATORY PERSPECTIVE. *Technological and Economic Development of Economy* 28, 6 (2022), 1954 â€“ 1990. https://doi.org/10.3846/tede.2022.17045 Cited by: 1; All Open Access, Gold Open Access.

[55] Zhenyu Zhao, Wei Zhang, and Yayue Zhou. 2011. Notice of Retraction: National student loans credit risk assessment based on GABP algorithm of neural network. *2011 2nd International Conference on Artificial Intelligence, Management Science and Electronic Commerce, AIMSEC 2011 - Proceedings* (2011), 2196 – 2199. https://doi.org/10.1109/AIMSEC.2011.6010906 Cited by: 3.

A USE OF AI

During the preparation of this work the author(s) used Grammarly in order to check for spelling and grammar mistakes. After using this tool/service, the author(s) reviewed and edited the content as needed and take(s) full responsibility for the content of the work.

B SUMMARY OF PAPERS

Table B.1: Credit risk papers

Authors	Title	Year	Cited by	XAI method	Globality
Nallakaruppan et al. [37]	An Explainable AI framework for credit evaluation and analysis	2024	0	LIME; PDP; SHAP	Global; Local
Hjelkrem and Lange [26]	Explaining Deep Learning Models for Credit Scoring with SHAP: A Case Study Using Open Banking Data	2023	1	SHAP	Global; Local
Do et al. [19]	Explainable Machine Learning for Credit Risk Management When Features are Dependent	2023	0	SHAP	Global; Local
Chen et al. [13]	Interpretable machine learning for imbalanced credit scoring datasets	2024	8	LIME; SHAP	Global
Liu et al. [32]	Credit scoring prediction leveraging interpretable ensemble learning	2024	3	SHAP	Global; Local
Talaat et al. [49]	Toward interpretable credit scoring: integrating explainable artificial intelligence with deep learning for credit card default prediction	2024	1	SHAP	Global
Aljadani et al. [1]	Mathematical Modeling and Analysis of Credit Scoring Using the LIME Explainer: A Comprehensive Approach	2023	3	LIME	Local
Nwafor and Nwafor [38]	Determinants of non-performing loans: An explainable ensemble and deep neural network approach	2023	3	SHAP	Global
Singh et al. [45]	Directive Explanations for Actionable Explainability in Machine Learning Applications	2023	3	DE	Local
Onari et al. [39]	An explainable data-driven decision support framework for strategic customer development	2024	0	MOGCE; SHAP	Global; Local
Alonso Robisco and Carbó Martínez [2]	Measuring the model risk-adjusted performance of machine learning algorithms in credit default prediction	2022	15	SHAP	Global
Bastos and Matos [7]	Explainable models of credit losses	2022	17	ALE; SHAP	Global
Szepannek and Lübke [47]	How much do we see? On the explainability of partial dependence plots for credit risk scoring	2023	1	PDP	Global
Xia et al. [53]	Incorporating multilevel macroeconomic variables into credit scoring for online consumer lending	2021	14	SHAP	Global
Xia et al. [54]	FROM CREDIT SCORING TO REGULATORY SCORING: COMPARING CREDIT SCORING MODELS FROM A REGULATORY PERSPECTIVE	2022	1	TreeSHAP	Global
Hamon et al. [24]	Bridging the Gap between AI and Explainability in the GDPR: Towards Trustworthiness-by-Design in Automated Decision-Making	2022	33	SHAP	Global
de Lange et al. [18]	Explainable AI for Credit Assessment in Banks	2022	11	TreeSHAP	Global; Local
Moscato et al. [36]	A benchmark of machine learning approaches for credit score prediction	2021	121	Anchors; BEEF; LIME; LORE; SHAP	Global; Local

Authors	Title	Year	Cited by	XAI method	Globality
Bueff et al. [11]	Machine learning interpretability for a stress scenario generation in credit scoring based on counterfactuals	2022	13	LORE; SHAP	Global; Local
Hu et al. [27]	Supervised Machine Learning Techniques: An Overview with Applications to Banking	2021	4	Permutation-based variable importance; PDP	Global
Wen et al. [52]	Study on Effect of Consumer Information in Personal Credit Risk Evaluation	2022	0	SHAP	Global
Dastile et al. [17]	Model-Agnostic Counterfactual Explanations in Credit Scoring	2022	9	GA-based Counterfactual Explanations	Local
Liu et al. [31]	Credit scoring based on tree-enhanced gradient boosting decision trees	2022	51	TreeSHAP	Global; Local
Dastile and Celik [16]	Making Deep Learning-Based Predictions for Credit Scoring Explainable	2021	26	Grad-CAM; LIME; Saliency maps; SHAP	Local
Hayashi and Takano [25]		2020	11	J48graft with Re-RX	Global
Bussmann et al. [12]	Explainable AI in Fintech Risk Management	2020	82	SHAP	Local
Ariza-Garzon et al. [5]	Explainability of a Machine Learning Granting Scoring Model in Peer-to-Peer Lending	2020	71	SHAP	Global

Table B.2: Customer segmentation papers

Authors	Title	Year	Cited by	XAI method	Globality
Talaat et al. [49]	A Mathematical Model for Customer Segmentation Leveraging Deep Learning, Explainable AI, and RFM Analysis in Targeted Marketing	2023	2	DeepLimeSeq	Local
Lee et al. [29]	Clustering and classification based on distributed automatic feature engineering for customer segmentation	2021	9	Fuzzy Decision Tree	Global
Choi et al. [14]	Network-based exploratory data analysis and explainable three-stage deep clustering for financial customer profiling	2024	1	TreeSHAP	Global