Improving the user experience and the trustworthiness of Financial Services

Giandomenico Cornacchia[1][0000−0001−5448−9970], Fedelucio Narducci[1][0000−0002−9255−3256], and Azzurra Ragone[2]

[1] Politecnico di Bari, Italy - firstname.lastname@poliba.it
[2] EY, Italy - firstname.lastname@it.ey.com

Abstract. Decision-making systems have been widely used in the Financial Services domain. AI is bringing both many innovations and opportunities as well as new risks linked to ethical considerations. Customer trust is at the forefront of customer retention. To build trust, there is the need to make the decision process Interpretable, Understandable, and Trustworthy for the end-user. Since products offered within the banking sector are usually of an intangible nature, customer trust perception is crucial to maintain a long-standing relationship and to ensure customer loyalty. To this end, in this paper we propose more insightful and user-friendly explanations for decisions made by AI systems in the financial domain.

Keywords: Trustworthy AI · Financial Services · Credit Scoring · Fairness · Explainability · Human-centered computing.

1 Introduction

Artificial Intelligence (AI) has increasingly played a predominant role in the interaction with consumers, citizens, and patients in the last years. It has been capable of revolutionizing their daily lives and improving all those user-centered services that are expanding out from the back office into customer-facing applications[3]. However, users often interact with such systems without even knowing that life-changing decisions like mortgage grants, job offers, patients screenings, etc., are in the hand of an AI-based system. Sometimes, such decisions may result arbitrary or inconsistent and limit users' ability to access the opportunities for which they are indeed qualified [1]. In particular, AI adoption is growing rapidly in the financial sector. Financial firms have widely used it to monetize data assets, improve customer experience, customize product and service offerings, drive business growth, and enhance operational efficiencies.

On the other hand, AI has shown its weakness by emphasizing social and ethical issues such as gender and demographic discrimination [3, 4], and lack of interpretability and explainability. As these applications become key enablers and more deeply embedded in processes, financial services organizations need to

[3] https://www.fca.org.uk/publications/research/research-note-machine-learning-uk-financial-services

cope with AI applications' inherent risks. This is true both from a compliance point of view (regulatory and ethical norms) and because the lack of trust is the leading barrier to AI adoption and acceptance by users. Building trust requires transparency and clear communication with internal stakeholders and customers, who should know when, how, and why AI is being used. This is especially true in a sector heavily regulated, such as the financial services one.

AI-based systems are increasingly attracting the attention of regulatory agencies and society at large, as they can cause, although unintentionally, harm. Indeed as reported by the Ethics guidelines for trustworthy AI from the European Commission's High-Level Expert Group on AI: *"The development, deployment, and use of any AI solution should adhere to some fundamental ethical principles such as respect for human autonomy, prevention of harm, fairness, and explainability"* [10]. Moreover, the GDPR set off the *right to explanation*: users have the right to ask for an explanation about an algorithmic decision made about them. In the UK, the Financial Conduct Authority (FCA) requires firms to explain why a more expensive mortgage has been chosen if a cheaper option is available. The G20 has adopted the OECD AI Principles [4] for a trustworthy AI where it is underline that users should not only understand AI outcomes but also be able to challenge them. In the financial sector, this is not an easy task to solve. As on one side, it is required to show how an outcome has been reached and whether it was fair and unbiased. On the other, not all the rationales behind a decision can be disclosed to prevent users from gaming the system.

In this paper, we propose an approach to provide more insightful explanations to make the interaction more user-friendly and, at the same time, to reinforce customer trust in the system.

2 What does it mean to be fair?

There is no single, globally accepted definition of fairness in AI [17]. However, being fair implies ensuring the same quality of service to all people, avoiding discriminating against minorities, and using protected characteristics like gender, nationality, age. AI-based systems should allocate opportunities, resources, or information fairly, thus avoiding societal or historical biases. The definition provided by Mehrabi et al. [13] summarizes well these concepts: *the absence of any prejudice or favouritism toward an individual or a group based on their inherent or acquired characteristics.*

In our analysis, we refer to Credit Scoring (CS) systems that compute the probability of a customer to repay a loan. We use this case study since for credit loan the concept of equal opportunity is crucial, and it lies very often in the hands of ML algorithms. Indeed, governments have addressed demographic, gender, and racial discrimination as regulatory compliance requirements since the 1960s [7], [5], [8]. Since those norms were not set to prevent discrimination in not-human decision making (as in the case of ML algorithms), "Ethics guidelines for a Trustworthy AI" [10] and "The White Paper" [5] were released to give

[4] https://oecd.ai/ai-principles

[5] https://ec.europa.eu/digital-single-market/en/news/white-paper-artificial-intelligence-public-consultation-towards-european-approach-excellence

guidelines for ethical and safe use of AI. Some critical keys requirements are "equity, diversity and not-discrimination" enclosed in the concept of fairness.

Going deeper with this analysis, the concept of fairness in CS could linked to one or more of these three criteria [11]: (i) *Independence* [6], (ii) *Separation* [9], and (iii) *Sufficiency* [3]. The (i) Independence guarantees that the fraction of customer classified as good-risks is the same in each sensitive groups. Therefore, if the gender is considered as sensitive, both men and women should have the same percentage of good-risk classification. The (ii) Separation criterion is related to the concepts of misclassification. Accordingly, the errors in classifying will be the same both in sensitive and non-sensitive groups. Finally, the (iii) Sufficiency criterion states that the probability that an individual belonging to the good-risk class is classified as good-risk will be the same for both sensitive groups. In this case, if the algorithm shows a gender bias, for example, a woman that belongs to the good-risk customer could be classified in the bad-risk class.

Once defined the concept of fairness and described the dimensions it is based on, the next question is: how can the customer be sure that the decision made by the algorithm is fair? We introduce now the next step that allows the customer to realize that the decision is fair. In particular, the person accountable for the AI system should be able to *explain* their outcome to the customer. The following section will address what explanation means and how to reach this goal.

3 From model fairness to end-user explanation

Several definitions are provided in the literature on what *explainable* means when we talk about an ML algorithm. The most relevant one for our purpose is provided by Bracke et al. [2] *"explanations can answer different kinds of questions about a model's operation depending on the stakeholder they are addressed to"*. This definition introduces an interesting characteristic of the explanation that has to consider the point of view of a specific stakeholder. Accordingly, in a CS scenario, for example, the explanation for a given decision might be different if addressed to customers rather than to the risk management functions. From the customer's point of view, which is the most interesting in our analysis, the explanation should describe the motivations behind a decision in a way that is easy to understand. Naturally, as abovementioned, the decisions are made by algorithms thus, it is crucial to know how these algorithms work. The ML algorithms belong to two main classes: interpretable and uninterpretable. More specifically, the former implement a *white-box* model, the latter a *black-box* one. On this perspective, Sharma et al. [15] distinguish *model-agnostic* and *model-specific* explanations. Model-agnostic methods provide an explanation that is not dependent on the ML model adopted and are geneally used for *black-box* models. A *surrogate* model is thus implemented with the aim of *simulating* the behavior of the original algorithm.

Today, explaining how a black box model works is still a challenging task.However, several methods have been proposed to explain black-box models. Two of the most important are LIME and SHAP. LIME trains local surrogate models explaining single data [14]. It generates a perturbation of initial data creating a

Fig. 1. An example of generation of natural-language explanations

new dataset and observing how the prediction changes through training an interpretable model. The analysis of the outcome of the perturbated data allows to interpret the original model. SHAP [12] is inspired by the cooperative game theory based on the Shapley Values. Each feature is considered a player that contributes differently to the outcome (i.e., the algorithm decision). SHAP does not compute all the possible combinations between all the features but performs only a random set of combinations for efficiency constraints. SHAP provides a ranked list of the features that contributed to the outcome ordered from the most to the less important. However, this explanation probably is not so clear for a customer who does not have experience with how an algorithm works. For this reason, if we want to improve the user's trust and general user experience with the system, we need to make the explanation more understandable. In that direction, we guess that an effective solution could be to transform the output produced by software like LIME or SHAP in a natural language sentence. We propose the pipeline described in Figure 1. Customer characteristics are the input, then the algorithm makes a decision, e.g. the computation of the CS, and shows using SHAP the features that contributed the most to the decision. At this point, another module takes as input the decision and the SHAP output and generate a natural-language explanation: e.g. *Dear Giulio, your loan application has been rejected since you don't have an account with us, the credit amount you asked for is too high compared to your income, and the duration is too long.* An interesting opportunity in this context could be provided by a counterfactual explanation that explains how the output of the algorithm could be changed [16]. For example, the system can add: *In the case you decide to open an account with us, to reduce the credit amount to 10,000$, and to reduce the duration to 12 months, the application will be probably accepted.* Conversely, model-specific explanations are based on the analysis of the structural information and the internal components of the algorithm that should be interpretable natively. From a technical perspective, these algorithms are easier to explain, but in this case, as well, most users will not be able to understand them. Therefore, the scenario is quite similar to the previous one and also here the exploitation of natural language can improve comprehensibility.

4 Future directions and challenges

In the credit risk context, we analyzed which fairness metrics can better evaluate the ML model and which explanation tools can get a better insightful interpretation of the decision process. For decision-making systems, it is necessary to

understand the causality of learned representations, and visualization tools need to be human-centered through natural language. It turns out that Shapley values can help scientists to understand something more on what features have more influence on the outcome. However, no explanation has been developed specifically for the knowledge domain of the end-user. One of the possible future directions concerns the development of intelligent conversational systems that can adapt the explanation to the type of user they interface with, guaranteeing the fairness of the treatment received and proposing counterfactual analyses of their characteristics. The explanation needs to be more Human-centered and more user-friendly without disclosing all the financial institution's decision criteria, risking adverse actions from unfair users.

References

1. Barocas, S., Hardt, M., Narayanan, A.: Fairness and Machine Learning. fairmlbook.org (2019)
2. Bracke, P., Datta, A., Jung, C., Sen, S.: Machine learning explainability in finance: an application to default risk analysis (816) (Aug 2019)
3. Chouldechova, A.: Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. Big data **5**(2), 153–163 (2017)
4. Cohen, L., Lipton, Z.C., Mansour, Y.: Efficient candidate screening under multiple tests and implications for fairness. In: FORC. LIPIcs, vol. 156, pp. 1:1–1:20. Schloss Dagstuhl - Leibniz-Zentrum für Informatik (2020)
5. Congress of the United States: Fair housing act (1968)
6. Dwork, C., Hardt, M., Pitassi, T., Reingold, O., Zemel, R.: Fairness through awareness. In: ITCS. pp. 214–226 (2012)
7. Federal Reserve Board: The truth in lending act (1968)
8. Federal Trade Commission: Equal credit opportunity act (1974)
9. Hardt, M., Price, E., Srebro, N.: Equality of opportunity in supervised learning. In: NIPS. pp. 3315–3323 (2016)
10. High-Level Expert Group on AI: Ethics guidelines for trustworthy ai. Report, European Commission, Brussels (Apr 2019)
11. Kozodoi, N., Jacob, J., Lessmann, S.: Fairness in credit scoring: Assessment, implementation and profit implications. arXiv preprint arXiv:2103.01907 (2021)
12. Lundberg, S.M., Lee, S.: A unified approach to interpreting model predictions. In: NIPS. pp. 4765–4774 (2017)
13. Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., Galstyan, A.: A survey on bias and fairness in machine learning (2019)
14. Ribeiro, M.T., Singh, S., Guestrin, C.: "why should I trust you?": Explaining the predictions of any classifier. In: KDD. pp. 1135–1144. ACM (2016)
15. Sharma, R., Schommer, C., Vivarelli, N.: Building up explainability in multi-layer perceptrons for credit risk modeling. In: DSAA. pp. 761–762. IEEE (2020)
16. Stepin, I., Alonso, J.M., Catala, A., Pereira-Fariña, M.: A survey of contrastive and counterfactual explanation generation methods for explainable artificial intelligence. IEEE Access **9**, 11974–12001 (2021)
17. Verma, S., Rubin, J.: Fairness definitions explained. In: ICSE-FairWare. pp. 1–7 (2018)