A general model for fair and explainable recommendation in the loan domain

Giandomenico Cornacchia[1], Fedelucio Narducci[1] and Azzurra Ragone[2]

[1]Politecnico di Bari – Via E. Orabona 4, Bari (I-70125), Italy
[2]EY Business and Technology solution – Via Oberdan 40, Bari (I-70125), Italy

Abstract

Recommender systems have been widely used in the Financial Services domain and can play a crucial role in personal loan comparison platforms. However, the use of AI in this domain has brought to light many opportunities as well as new ethical and legal risks. The customers can trust the suggestions of these systems only if the recommendation process is Interpretable, Understandable, and Fair 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 ensure customer loyalty. To this end, in this paper, we propose a model for generating natural language and counterfactual explanations for a loan recommender system with the aim of providing fairer and more transparent suggestions.

Keywords

Politecnico di Bari – Via E. Orabona 4, Bari (I-70125), Italy Trustworthy AI, Financial Services, Loan recommender systems, Fairness, Explainability, Human-centered computing, Conversational systems

1. Introduction

As stated by the World Economic Forum's Global Future Council on Artificial Intelligence for Humanity:"*Artificial Intelligence (AI) is the engine of the Fourth Industrial Revolution. It holds the promise of solving some of society's most pressing issues, including repowering economies reeling from lockdowns, but requires thoughtful design, development, and deployment to mitigate potential risks*"[1].

These risks are related to the fact that AI applications are becoming more and more pervasive, and, most of the time, users often interact with such systems without even knowing that life-changing decisions like mortgage grants, job offers, patients screenings are in the hand of AI-based systems [1]. Moreover, such AI decisions may sometimes result arbitrary, inconsistent, or discriminatory, which cannot be allowed in highly regulated environments such as Financial Services. As these applications have became key enablers and more deeply embedded in processes, financial services organizations need to 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 most significant barrier to AI adoption and acceptance by users. In fact, AI systems often amplify social and ethical issues such as gender and demographic discrimination [2, 3], and they lack interpretability and explainability.

As sales activities of financial products require expert knowledge, recommender systems can offer significant benefits to financial services supporting the client in choosing the best option among the many financial products offered by different banks. However, compared to the subject of conventional recommender systems, their application in financial domains is a challenging task: there is the need to adhere to the regulation, follow specific fairness criteria, and providing, at the same time, an explanation of your decisions (black-box approaches are not allowed).

In this paper, we focus on the case of loan recommendation. In this domain, the recommendation problem is modeled as finding the right product of the lender company for the borrower, which, at the same time, satisfies their financial needs and will be likely to be paid back by the borrower.

In the last years, several online platforms for personal loan comparison[2] have emerged to help individual borrowers analyze different loans proposed by third-party lenders and suggest the best option. These platforms simplify the process of shopping for a personal loan, showing the users all the loans that are pre-approved for, so they can compare offers and make a conscious choice. In order to recommend the best loan for the user, on one side, these platforms usually ask several questions to profile

3rd Edition of Knowledge-aware and Conversational Recommender Systems (KaRS) & 5th Edition of Recommendation in Complex Environments (ComplexRec) Joint Workshop @ RecSys 2021, September 27–1 October 2021, Amsterdam, Netherlands

giandomenico.cornacchia@poliba.it (G. Cornacchia); fedelucio.narducci@poliba.it (F. Narducci); azzurra.ragone@it.ey.com (A. Ragone)
0000-0001-5448-9970 (G. Cornacchia); 0000-0002-9255-3256 (F. Narducci); 0000-0002-3537-7663 (A. Ragone)

[1] https://www.weforum.org/communities/gfc-on-artificial-intelligence-for-humanity

[2] To cite a few: https://www.creditkarma.com/, https://borrowell.com/, www.nerdwallet.com, www.meilleurtaux.com/, https://www.habito.com/, https://www.bankbazaar.com/

the client, like personal information (e.g., address, date of birth, Tax ID number), basic financial information (e.g., rent/mortgage payment, other major bills), requested loan amount and ideal term length. On the other side, to fill out the list of the best loans, the platforms have to evaluate several lenders, looking at key factors like interest rates, fees, loan amounts, and term lengths offered, customer service, and how fast you can get your funds.

In this paper, we propose an approach to model a personal loan recommender system that comply with the present European regulation (Section 2), guarantee fairness criteria (Section 3), provide a meaningful explanation of the decision of the algorithm (Section 4), and is able to provide a user-based explanation. In particular, Section 4 focuses the attention on defining a general model for generating natural language explanation in the aforementioned context of loan recommendations. In our opinion, this explanation model can be easily integrated in a conversational recommender system able to interact with the user by exchanging natural language messages. Furthermore, we enhance the power of explanations by providing also a counterfactual analysis and explanation (Section 5). In this way, we can provide more insightful explanations to make the interaction with the client more efficient, compliant with regulations, and, at the same time, reinforce customer trust in the system.

2. Regulation compliance

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"*[4]. Moreover, in EU the GDPR sets 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 [3] for a trustworthy AI where it is underline that users should not only understand AI outcomes but also be able to challenge them.

On 21 April 2021, the European Commission presented the "Proposal for a Regulation laying down harmonized rules on artificial intelligence" [4] a proposal law that could enter into force in the second half of 2022 in a transitional period. This proposal remarks on the importance of monitoring the deployed AI systems based on a scale of risk. The risk-based approach splits AI systems in four different categories, *unacceptable risk*, *high risk*, *limited risk*, *minimal risk* depending on the risk of the use case. AI systems intended to be used to evaluate the creditworthiness of natural persons or establish their credit score are placed in the *high risk* categories.

Furthermore, any application of artificial intelligence must be designed with responsibility and compliance to standards required by law. In the financial sector, this is not an easy task to solve. 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.

Generally speaking, every time a risk review of an AI system is performed, it is required to show how an outcome has been reached and whether it was fair and unbiased. This is not a one-time effort and should involve the contribution of different stakeholders: data scientists, business people, audit and compliance functions, ethicists, to name a few.

In the following, we will show how to cope with these requirements.

3. Fairness

The regulations of financial services do not start with the recent laws of artificial intelligence. Rather, the latter are a derivation of the steps taken by governments on financial and social regulations between the 1960s and 1980s. Indeed, governments have addressed discrimination against unprivileged groups as regulatory compliance requirements since the 1960s [5], [6], [7]. In USA, the Fair Housing Act (FHA) and Equal Credit Opportunity Act (ECOA), which protect consumers by prohibiting unfair and discriminatory practices, have focused on ensuring a quality of service that is independent of sensitive characteristics such as gender, race, age, disability, etc., avoiding discrimination against minorities.

These principles can be condensed into the definition of fairness, where fairness, accordingly to Mehrabi et al. [8], can be seen as *"the absence of any prejudice or favoritism toward an individual or a group based on their inherent or acquired characteristics"*. Contextualising it in the use of an AI system in financial services, it should allocate opportunities, resources, or information fairly, thus avoiding social or historical biases. However, this definition of fairness is independent of the technical concepts that arise when using any classifier, and that is why the definitions of fairness are different and various.

Since those norms were not set to prevent discrimination in not-human decision making (as in the case of ML

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

[4] https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52021PC0206

algorithms), ”Ethics guidelines for a Trustworthy AI” [4] and ”The White Paper” [9] were released to give guidelines for ethical and safe use of AI. Some critical keys requirements are ”equity, diversity and not-discrimination” enclosed in the concept of fairness. More recently, with the ”Proposal for a Regulation laying down harmonized rules on artificial intelligence” credit scoring applications, including loan recommender systems, are classified in the high-risk domain. Before deploying any AI system, the Financial Institution has to pass different conformity steps, and one of these concerns with Fairness.

In our analysis, we refer to personal loan recommender systems that suggest to each customer a personalized list of potential loan products based on their profile. We use this case study since for personal loan the concept of equal opportunity is crucial, and it lies very often in the hands of ML algorithms with a high risk that they discriminate without the awareness of both the financial institution and the client.

As these automated decision-making systems are increasingly used, they must guarantee these principles of fairness. In the case under consideration, the recommender system that suggests different offers based on the characteristics of the credit requested and the user’s profile must ensure that each offer has been processed through fair algorithms on the provider side.

Going deeper with this analysis, the concept of fairness in provider-side algorithms of a personal loan recommendation could be linked to one or more of these three statistical criteria [10]: (i) *Independence* [11], (ii) *Separation*[12], 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 recommended loans characteristics have been generated by fair-provider algorithms? In the next section we introduce another important requirements of the loan recommendation platform, the *explanation*. The platform and the loan provider, should be able to explain the outcome to the customer guaranteeing that the outcome is achieved under fairness constraint. Nowadays, this is often a step that is left out as AI systems already suggest loans to the customers but without giving in response the rationale behind the decision. However, following a black-box approach could lead to severe reputation damages for the financial institutions, as in the case of Apple and Goldman Sachs [13].

4. Explainability

For many years, research on ML and, more generally, AI algorithms has been focused on improving accuracy metrics such as precision, recall, etc. Recently, new laws and regulations [14] have introduced the need for those algorithms to show explanation capabilities in particular in a sensitive domain such as the financial one [15].

The ML algorithms belong to two main classes: interpretable and uninterpretable. More specifically, the former implement a *white-box* model design, the latter a *black-box* one. On this perspective, Sharma et al. [16] distinguish *model-agnostic* and *model-specific* explanations. Model-agnostic methods provide an explanation that is not dependent on the ML model adopted and are generally used for *black-box* models. A *surrogate* model is thus implemented with the aim of *simulating* the behavior of the original algorithm.

Several methods have been proposed to explain black-box models. In this paper we focus on SHAP [17]. SHAP 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). Considering the original theory, we have to compute all the possible combinations with the other sets of features. This choice is, first of all, impractical but, above all, computationally inefficient. Therefore, 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 the most to the less to the outcome. However, the explanation provided by this method 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, in general, the 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 SHAP in a natural language sentence. Figure 1 represents our proposed workflow for generating an explanation and a counterfactual explanation in order to recommend also corrective actions to the user. For the sake of simplicity, here we show the pipeline focusing on a single decision taken from the ML algorithm of a given lender. Naturally, the loan recommender will receive this information from all the lender services invoked. Let us

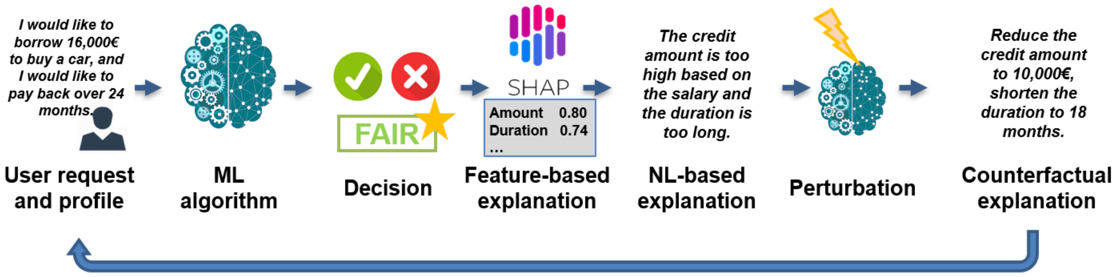
I would like to borrow 16,000€ to buy a car, and I would like to pay back over 24 months.
User request and profile
ML algorithm
FAIR
Decision
SHAP
Amount 0.80
Duration 0.74
...
Feature-based explanation
The credit amount is too high based on the salary and the duration is too long.
NL-based explanation
Perturbation
Reduce the credit amount to 10,000€, shorten the duration to 18 months.
Counterfactual explanation

Figure 1: Workflow for generating explanation and counterfactual explanation for loan application

suppose that the user asks for a personal loan through the following message: *"I would like to borrow 16,000€ to buy a car, and I would like to pay back over 24 months"*. Then the platform will ask to provide personal information such as age, income, etc., to be sent to the lender services. Once received the different proposals from the lender platforms, a list is ranked according to one or more criteria (e.g., rate, decision, etc.) and proposed to the user. Let us assume that each algorithm respects fairness criteria with regulatory bodies' labels as proof of compliance with that criteria. Each proposal (i.e., accepted or denied) is provided with a feature-based SHAP explanation that shows how the ML algorithm has produced that result. Next, those SHAP values are transformed in a natural language explanation like: e.g., *"The credit amount is too high based on the salary and the duration is too long."*.

A further interesting contribution in this direction is provided by a counterfactual analysis obtained by a feature perturbation step (see Section 5.1). This explanation shows how to modify the the loan request for getting the loan accepted [18]. For example, the system can add: *Reduce the credit amount to 10,000€, shorten the duration to 18 months, ..., and the loan request will probably be accepted.*

But how can we generate this kind of natural language explanation? In the next section, we propose a template-based formal model able to transform the SHAP values into a natural language sentence.

5. A model for generating NL explanation

The model we designed for generating Natural Language explanations is inspired by Musto et al. [19].

The principal insight is that our natural language explanation can be generated by exploiting a template composed of some slots that can be filled with features, adverbs, and adjectives according to the the output produced by SHAP. We remember that the SHAP output consists of a set of couples *<feature,score>* (e.g., <income, 0.8>).

Let us consider the example in Figure 1: *The credit amount is too high based on the salary and the duration is too long*. In that case the template for the explanation is: *<feature> <verb> <adverb> <adjective> <motivation>* followed by a new set of *<feature> <verb> <adverb> adjective>* without motivation. The problem is to properly fill each slot and compose the whole explanation.

In the above mentioned example, the number of features taken into account for generating the explanation are three: *the credit amount*, *the salary*, and *the duration* each of which associated to adverbs and/or adjectives (e.g., too high, too long, etc.). The number of features used for generating the explanation can be set as desired. However, since the explanation has to be as useful as possible, too much features can, in some cases, losing effectiveness and efficiency.

In our model, the generation of the natural language explanation exploits a set of rewriting rules using the Back-Naur Form (BNF) as described in the following. Even though these templates and rules can be exploited also in other domains, the terminal symbols (e.g., the credit amount, the duration, long, short, etc.) are specific for a loan application.

<explanation> ::= <sentence> | <explanation> <conjunction> <sentence>
<sentence> ::= <feature> <verb> <adverb> <adjective>
<sentence> ::= <sentence> <motivation>
<motivation> ::= <motivation> <conjunction> <motivation>
<motivation> ::= <adverbial phrase> <feature>
<adverbial phrase> ::= ‘based on’ | (etc.)
<adverb> ::= ‘too’ | ’so’ | ’few’ | ’almost’ | ’enough’ (etc.)
<adjective> ::= ‘high’ | ’long’ | ’short’ | ’little’ | (etc.)
<conjunction> ::= ‘and’ | ’but’ | , |(etc.)
<feature> ::= ‘the credit amount’ | ’the duration’ | ’the salary’ | (etc.)
<verb> ::= ‘is’ | ’are’ | ’has’ | ’have’ | ’is not’| (etc.)

These rewriting rules can be applied for generating, for example, the explanation *The credit amount is too high based on the salary and the duration is too long*.

A further problem is the choice of adverbs and adjectives. For the adverbs, we defined a matching between value intervals and the *intensity* of the adverb. As an example, if the SHAP value of a feature is 0.8 (the highest interval)[5], the corresponding <*adverb*> will be 'too' emphasizing how this feature has a strong impact on the loan application decision. Obviously, the association between the <*feature*> and the type of <*adjective*> is not arbitrary, but it depends on the type of <*feature*> is considered. Therefore, for each feature we defined a vocabulary of compatible adjectives.

5.1. Counterfactual explanation

In the previous subsection, we have described how a loan recommendation platform can generate the explanation for each decision given by a provider.

To make our explanation more effective, we propose to the user some indications useful for revising her request and getting the loan application accepted. This is obtained through a *counterfactual explanation.*

The counterfactual explanation consists of a set corrective actions to the characteristics of the requested loan, based on the results of a counterfactual analysis. Providing a counterfactual explanation is an opportunity for the loan provider that results in an additional service to enhance customer satisfaction and make the customer aware of his or her chances of getting a loan. This service will result in a Responsible and Trustworthy use of AI systems towards customers.

The counterfactual analysis performs a *perturbation* on the feature space of the customer's loan application. The perturbation will generate a new sample that will be considered as a new application. Subsequently, the counterfactual analysis will detect the new nearest sample to the original one that will be accepted by the ML algorithm. The result of this analysis will consist in detecting the change in the loan's characteristics of the customer and recommending corrective actions.

The approach we adopted for generating the counterfactual explanation is the same described in the previous section, namely a set of BNF rewriting rules.

Following the previous example, a counterfactual explanation can be: *"Reduce the credit amount to 10,000€, shorten the duration to 18 months."*.
The BNF template is:

<counterfactualexplanation>::= <sentence>|<counterfactualexplanation> <conjunction> <sentence>

<sentence>::= <action><feature><value>

<action> ::= 'reduce'|'expand'|'shorten'|etc.
<feature> ::= 'the credit amount'|'the duration'|etc.
<value> ::= '10,000€'|'18 months'|
<conjunction> ::= 'and' | 'but' | , |(etc.)

The counterfactual explanation has a small set of rules, in fact it includes a feature, the corrective actions, and optionally the desirable new feature value. Since the counterfactual analysis works by perturbing all the features of a determined instance, the recommended actions should impact the minimum set of features that allow to change the algorithm decision.

The action is chosen according to the relation between the old and the new feature value. For example, if the old value for the feature *duration* was 24 and the new value after the perturbation is 18, the verb (action) chosen will be *reduce.* Regarding the values, if the new value is equal to the original one, the respective feature will not be included in the explanation since there is no corrective action to be done, otherwise the new perturbed value will be shown in the explanation.

6. Conclusion and future research directions

This work proposes a model to generate natural language explanation for ML decisions in the context of loan recommendation platforms. In the first part of the paper, we analyzed which fairness metrics can be used for evaluating the ML model. Next, for improving the system transparency, financial platforms must understand the causality of the learned representations, and explain their decisions through visualization tools or natural language. Shapley values could help understand more on what features influence the outcome, however it is not very human friendly. For this reason, a model for generating NL explanations from Shapley values has been proposed.

Another contribution is the definition of a counterfactual explanation based on the result of a counterfactual analysis, This results in a set of corrective actions to be performed by the user.

The defined model finds a straightforward application in a scenario of conversational recommender system. The user expresses her request in natural language, the platform compares the different offers and provides an explanation for each of them. The user can thus ask for help on how to modify her request for getting the loan. Eventually, the platform, thanks to the counterfactual analysis and explanation, can provide a set of actions for getting the application accepted. However, the conversational system should preserve from discovering the complete set of decision criteria avoiding adverse action from unfair users.

[5] Please remember that the SHAP values are between 0 and 1

In the future work, first of all, the whole pipeline and conversational environment will be implemented (e.g, intent recognizer, entity recognizer, sentiment analyzer, NL generator, etc.). Then, extensive experimental evaluations and user studies have to be carried out for assessing the effectiveness of the model both in terms of the capability of generating NL explanations and in terms of improved user experience.

References

[1] S. Barocas, M. Hardt, A. Narayanan, Fairness and Machine Learning, fairmlbook.org, 2019.

[2] L. Cohen, Z. C. Lipton, Y. Mansour, Efficient candidate screening under multiple tests and implications for fairness, in: FORC, volume 156 of *LIPIcs*, Schloss Dagstuhl - Leibniz-Zentrum für Informatik, 2020, pp. 1:1–1:20.

[3] A. Chouldechova, Fair prediction with disparate impact: A study of bias in recidivism prediction instruments, Big data 5 (2017) 153–163.

[4] High-Level Expert Group on AI, Ethics guidelines for trustworthy AI, Report, European Commission, Brussels, 2019.

[5] Federal Reserve Board, The truth in lending act, 1968.

[6] Congress of the United States, Fair housing act, 1968.

[7] Federal Trade Commission, Equal credit opportunity act, 1974.

[8] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, A. Galstyan, A survey on bias and fairness in machine learning, 2019. arXiv:1908.09635.

[9] White Paper on Artificial Intelligence: Public consultation towards a European approach for excellence and trust, CONSULTATION RESULTS, European Commission, Brussels, 2020. URL: https://wayback.archive-it.org/12090/20210726215107/https://ec.europa.eu/digital-single-market/en/news/white-paper-artificial-intelligence-public-consultation-towards-european-approach-excellence.

[10] N. Kozodoi, J. Jacob, S. Lessmann, Fairness in credit scoring: Assessment, implementation and profit implications, arXiv preprint arXiv:2103.01907 (2021).

[11] C. Dwork, M. Hardt, T. Pitassi, O. Reingold, R. Zemel, Fairness through awareness, in: ITCS, 2012, pp. 214–226.

[12] M. Hardt, E. Price, N. Srebro, Equality of opportunity in supervised learning, in: NIPS, 2016, pp. 3315–3323.

[13] R. P. Bartlett, A. Morse, N. Wallace, R. Stanton, Algorithmic discrimination and input accountability under the civil rights acts, Available at SSRN 3674665 (2020).

[14] K. Croxson, P. Bracke, C. Jung, Explaining why the computer says ‘no’, FCA 5 (2019) 31.

[15] N. Bussmann, P. Giudici, D. Marinelli, J. Papenbrock, Explainable machine learning in credit risk management, Computational Economics 57 (2021).

[16] R. Sharma, C. Schommer, N. Vivarelli, Building up explainability in multi-layer perceptrons for credit risk modeling, in: DSAA, IEEE, 2020, pp. 761–762.

[17] S. M. Lundberg, S. Lee, A unified approach to interpreting model predictions, in: NIPS, 2017, pp. 4765–4774.

[18] I. Stepin, J. M. Alonso, A. Catala, M. Pereira-Fariña, A survey of contrastive and counterfactual explanation generation methods for explainable artificial intelligence, IEEE Access 9 (2021) 11974–12001.

[19] C. Musto, F. Narducci, P. Lops, M. De Gemmis, G. Semeraro, Explod: A framework for explaining recommendations based on the linked open data cloud, in: Proceedings of the 10th ACM Conference on Recommender Systems, RecSys ’16, Association for Computing Machinery, New York, NY, USA, 2016, p. 151–154. URL: https://doi.org/10.1145/2959100.2959173. doi:10.1145/2959100.2959173.