Enhancing Explainability in Credit Risk Modelling

Submitted on: **07-03-2024**

Alina Baciu
alina.baciu@student.uva.nl
University of Amsterdam
Amsterdam, The Netherlands

Erman Acar
e.acar@uva.nl
University of Amsterdam
Amsterdam, The Netherlands

1 INTRODUCTION

Over the past decades, advancements in Artificial Intelligence (AI) and Machine Learning (ML) have had an important role in reshaping and developing the economy's branches, including the financial service sector [10]. This evolution has been characterised by the transformation of traditional business processes to automated tasks, leading to increased productivity, accuracy and business value [7]. Among the various applications of AI and ML in finance, credit risk modelling has received considerable recognition due to its positive impact on risk management practices [11]. Models developed using this tool can assist financial institutions (e.g banks, fintechs) in automatically evaluating the ability of an applicant to repay its requested loan, categorizing individuals as at risk of default or not [13]. Identifying real defaulters is essential on a global scale as failing to achieve that can have drastic consequences not only for the financial institution (lender) and customer (borrower) [2] but also on the stability of a country's economy [13]. For instance, Malaysia's central bank has experienced a considerable cumulative amount (RM31 billion) of non-performing loans until 2021, potentially affecting the country's economy [13]. As a result, a notable amount of research effort has been dedicated to building and improving the performance of the credit risk models, achieving significant accuracy scores [13].

However, despite these developments, certain institutions active in the finance area face a global and social challenge. While the implementation of complex models can result in high performance, these classifiers, often being referred to as "black box" models lack complete explainability [13]. This aspect can have legal consequences for the lender as it is argued that the General Data Protection Regulation (GDPR) requires a certain level of understanding of the steps of systems involving automated decision-making processes [2, 8]. Moreover, when applicants face unfavourable treatment, namely their requested loan is rejected, it can reduce transparency, clarity and trust in the financial institution, which can negatively affect the organisation's reputation and liquidity stability.

Although this challenge was commonly addressed by employing *explainable AI* (XAI) methods [8], the problem area remains open for exploration and improvement. Based on the conducted literature review, it is apparent that credit risk modelling research is limited regarding the assessment of the impact of *conformal prediction* (CP) on *explainable AI* (XAI) and *algorithmic recourse* (AR). By addressing this gap, this thesis explores how the stated approach can influence explainability and therefore the reliability (transparency, trust and clarity) of credit risk models globally , significantly contributing to the existing literature and new potential implementation practices in real-world scenario developments.

The dataset [6] chosen for this thesis is public and sourced from Home Credit, a non-banking international financial institution providing lending services. This source will serve as the foundation for building the credit risk model and for implementing and integrating the three stated sub-fields based on the classifier's predictions.

The main research question that this thesis will address is:

- What impact does *conformal prediction* have by means of *algorithmic recourse* and *explainable AI*?

The following sub-questions need be researched based on the main research question:

(1) How do mainstream explainable AI methods (e.g. SHAP, LIME) differ when used independently compared to their performance when integrated with *conformal prediction*?
(2) How does *conformal prediction* influence mainstream *algorithmic recourse* approaches?
(3) To what extent does the general public perception align with the integration of *conformal prediction* with *explainable AI* and *algorithmic recourse*?

2 RELATED WORK

The gap of understanding the impact of CP in relation to XAI and the AR in credit risk modelling will be addressed in this section by outlining the the latest methodologies applied in relevant studies to tackle each mentioned notion. Moreover, the main hypothesis of the case study will be stated.

Firstly, concerning the AR, Karimi, Barthe, Scholkopf and Valera [2] present an overview of the concept in connection with credit risk modelling, stating that the main objective of the process is to provide minimal consequential recommendations that result in (nearest) contrastive explanations to individuals who receive unfavorable treatment from automated decision-making systems and desire to improve their outcome. If achieving this goal is not possible, the second priority should represent generating solely (nearest) contrastive explanations, without associating them with recommendations. Both type of insights are considered counterfactual as they consider an alternative outcome after performing alterations to an entity's characteristics.

Regarding the finance subject of this case study, this implies providing the following insights to applicants who were denied loans: the profile characteristics that would have led to loan approval and the actions required to build such a profile. The authors of the paper also state that both explanations and recommendations are originally local, which imply that they are generated per individual outcome rather than collectively, for the entire dataset.

To fulfil the priority objective, we need to define the optimization task and its components: the objective function and general constraints (model; plausibility and actionability; diversity and sparsity). The optimization process should also include insights concerning the structural causal model (SCM) [5] of the involved features. Instead of presuming that independent feature alterations from generated explanations can represent accurate recommendations, SCM considers the dependency of variables, therefore preventing suboptimal solutions.

The recourse task can represent a complex challenge as the objective function and constraints can be non-linear/non-convex/non-differentiable. In addition, constraints can also have non- monotonic behavior. Due to its potential challenging nature, achieving desirable properties (optimality, perfect coverage, efficient runtime, and access) of the optimizer and its results at once is difficult, requiring trade-off approximate solutions. These approaches are gradient-, model-, search-, verification-, heuristic-based, the first and the third being the ones employed by the authors of the paper depending on the differentiability setting of both objective and constraints. Concerning the differentiability aspect, the gradient-based optimization should be used, applying methods such as FISTA for convex objective functions and constraints, to find global optimal solution and (L-) BFGS for non-convex nature, to detect local optima. Regarding the non-differentiability setting, the search-based optimization should be used, employing integer/mixed linear programming methods for linear objective and constraints.

To fulfil the second priority goal, the overview of Recourse Algorithms for Consequential Decision-Making Settings presented in [2] can be used as it is focused more on contrastive explanations rather than consequential recommendations. The suitable algorithm(s) can be selected based on available properties.

Secondly, concerning the XAI concept, several techniques exist, including counterfactual instances, LIME and SHAP [3], the first method referring to the (nearest) contrastive explanations generated by Algorithmic Recourse concept. Both LIME and SHAP are commonly applied techniques in finance research studies aimed at improving the reliability of credit risk classifiers. Each method offers different characteristics: SHAP is a game theoretic approach which uses Shapley values to measure how much each feature contributes to the credit risk model's prediction on a global or local level, illustrating the most important features ranking. On the other hand, LIME solely provides a local interpretable approximation of the complex classifier's behavior for a specific instance [13].

Furthermore, the concept of CP can be defined as a measure of uncertainty used to construct confidence intervals for the predictions of models [12]. It involves two main phases, calibration and prediction. The first one implies calculating nonconformity scores (e.g. least confidence score, smallest margin, ratio of distance to nearest neighbors) for each instance in a calibration set, which aim to measure how much each data point' predicted probability deviates from the expected patterns established by the model during training. Then, a cutoff point is selected (e.g. 95th percentile) from these nonconformity scores, entering the prediction phase. The nonconformity score for a new applicant data point is then computed and compared to the cutoff point to form prediction intervals [9] [1]. Referring to the binary task of this case study, these regions are: empty, default, non-default and default, non-default [4]. The first type is generated when the new score is higher than the cutoff point, outlining complete uncertainty. However, if the measure is lower, the intervals can include one or both classes, depending on significance level.

The main hypothesis that will be tested throughout this case study concerning the assessment of integrating CP with XAI and AR is: General public perception aligns with the outcome that CP can positively influence explanations generated by AR and XAI in credit risk modelling. Regardless of whether this hypothesis is practically validated, this thesis will contribute to filling the identified research gap.

3 METHODOLOGY

The existing methods described in the "Related work" section will serve as the baseline for this case study. In addition to these foundational techniques, other methodologies and approaches will be applied and evaluated with the aim to perform a strong analysis of the assignment's problem.

This research will commence by employing the CRISP-DM methodology, an established framework for structuring machine learning projects following five out of its six steps: Business Understanding, Data Understanding, Data Preparation, Modelling, Evaluation, to build a credit risk classifier. The chosen dataset [6] for this thesis is labelled, indicating the presence of the target variable, distinguishing between default and non-default cases. Since the imbalance of the target variable represents one of the main issues in the research field, the following tools will be used to assess the performance of the model: confusion matrix, its derived metrics and the ROC curve - AUC score. Concerning the methods for performance evaluation, the holdout technique will be used as the dataset of the case study is substantial (307.511 instances).

After the model is developed, several steps will be added to the stated baseline.The CP process will be conducted on the classifier's prediction results on the test set. As mentioned in section 2, the output of the concept can be categorised into four intervals. For a better understanding, these determined region values will be converted into more comprehensive labels based on their confidence degree (default and non-default to high; default,non-default to medium and empty to low). Then, the process of fulfilling the priority objective in AR will commence by attempting to generate local consequential recommendations and explanations based on data properties. Efforts will be made to integrate causal dependencies of SCM into the task's constraints, in this way, ensuring that the insights are consistent with known variable relationships, contributing to the process efficiency and accuracy. In the event that the properties of the data do not allow

this development, the second priority goal will commence, solely focusing on generating contrastive explanations based on the stated algorithm overview. Next, the application of both LIME and SHAP techniques will be performed on local level to maintain consistency with AR. The resulted explanations will be filtered to only include results that align across the two concepts (AR and XAI), potentially indicating more reliable insights.

Due to the individual nature of the explanations, a small random subset of instance results will be selected to take part in an online survey, aiming at assessing the general public's perception on the Conformal Prediction's impact on Algorithmic Recourse and XAI.

The survey will target the general public and it will include two sections: one for viewing the generated consistent insights along with their confidence intervals and the latter for significant queries addressing the trust, transparency, clarity and preference of the insights :

(1) To what extent do you trust AI models in generating explanations for credit risk predictions?
(2) How important is it for you to understand the factors that AI models take into account when assessing credit risk?
(3) Do you find the explanations generated by AI models for credit risk easy to understand?
(4) Would you be more likely to trust an AI model if it provided confidence levels for its explanations of credit risk predictions?

After obtaining the results of the survey, the main hypothesis will be tested, contributing to filling the stated research gap and therefore to the explainability challenge of credit risk systems.

4 RISK ASSESSMENT

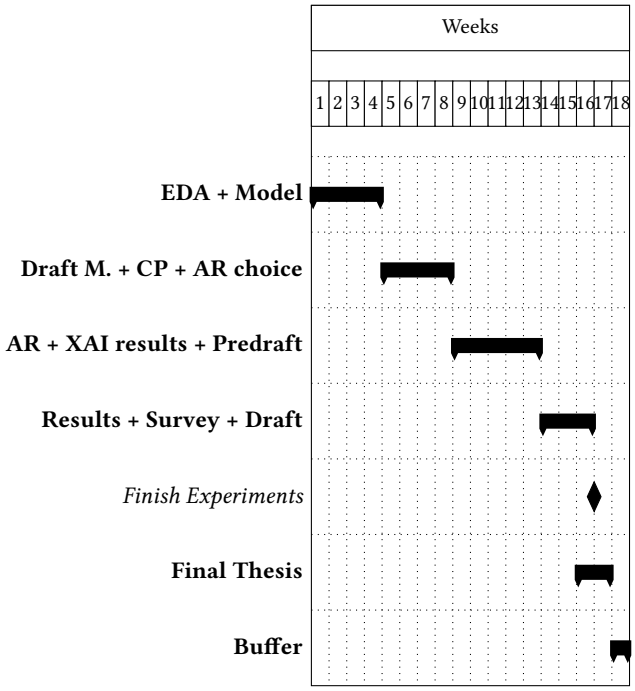
This case study can experience certain risks (Table 1). Firstly, it is possible that the student loses working code files due to potential technical issues. This possibility can be reduced by using Git as a version control system to keep track of latest version of developments. Secondly, the lack of knowledge concerning causality can represent another risk that can affect the fulfilment of the priority objective of Algorithmic Recourse. This issue will be tackled by obtaining and studying the necessary material from the Causal Data Science elective.

Risk	Measure
File loss	Use version control software
Causality uncertainty	Causal DS elective
Priority goal - Alg. Recourse	Second priority goal
Mathematical uncertainty	Meetings with supervisor

Table 1: Risk Assessment

Another risk regarding the priority objective of Algorithmic Recourse can represent the mismatch between the data properties expected by the mentioned methods and the Home Credit's dataset properties (e.g. non-differentiability and linearity of objective and constraints). In such a scenario, the second priority goal will commence instead using the stated algorithm overview, based on available settings. Furthermore any mathematical issues that can arise will be addressed by having meetings with the thesis supervisor.

5 PROJECT PLAN



REFERENCES

[1] Alphanome.AI. 2023. *Conformal Prediction: A Guide for Investors.* https://www.alphanome.ai/post/conformal-prediction-a-guide-for-investorsfbclid=IwAR1xXEAKC0DgF6jbZg3zOcSQNEX4Mtmdu54_bdyglbxaphxCv8wtsg-FlYk

[2] Bernhard Sholkopf Isabel Valera Amir-Hossein Karimi, Gilles Barthe. 2022. A Survey of Algorithmic Recourse: Contrastive Explanations and Consequential Recommendations. 55, 95 (2022), 1–29. https://doi.org/10.1145/3527848

[3] Francesco Dallanoce. 2022. *Explainable AI: A Comprehensive Review of the Main Methods.* https://medium.com/mlearning-ai/explainable-ai-a-complete-summary-of-the-main-methods-a28f9ab132f7

[4] Leo Dreyfus-Schmidt. 2020. *Measuring Models' Uncertainty: Conformal Prediction.* https://blog.dataiku.com/measuring-models-uncertainty-conformal-prediction

[5] Bruno Gonçalves. 2020. *Structural Causal Models.* https://medium.data4sci.com/causal-inference-part-iv-structural-causal-models-df10a83be580

[6] Home Credit Group. 2018. *Home Credit Default Risk.* https://www.kaggle.com/competitions/home-credit-default-risk/overview

[7] Oracle Netherlands. [n. d.]. *What is AI in finance?* https://www.oracle.com/nl/erp/ai-financials/what-is-ai-in-finance/

[8] Christian Bakke Vennerød Sjur Westgaard Petter Eilif de Lange, Borger Melsom. 2022. Explainable AI for Credit Assessment in Bank. 15, 12 (2022), 1–23. https://doi.org/10.3390/jrfm15120556

[9] Igor Radovanovic. 2023. *Conformal Prediction – A Practical Guide with MAPIE.* https://algotrading101.com/learn/conformal-prediction-guide/?fbclid=IwAR0P5r6EHqcPX3xUwxKuWZnzWeykzT47lMjqLrbCGlhwF-VF-az1lRLIKZU

[10] Kris Sharma. 2023. *Machine learning in finance: history, technologies and outlook.* https://ubuntu.com/blog/machine-learning-in-finance-history-technologies-and-outlook

[11] Prashant Singh. 2023. *Credit Risk Modelling: An Essential Tool for Financial Institutions.* https://www.linkedin.com/pulse/credit-risk-modelling-essential-tool-financial-prashant-singh/

[12] Kristof Slechten. 2023. *Conformal predictions.* https://kristofsl.medium.com/conformal-predictions-94506997429c
[13] Preethi Subramanian Yi Sheng Heng. 2022. A Systematic Review of Machine Learning and Explainable Artificial Intelligence (XAI) in Credit Risk Modelling. In *Proceedings of the Future Technologies Conference (FTC) 2022, Volume 1*, Kohei Arai (Ed.). Vancouver, Canada.