Artificial Intelligence Ethics and Applications (CIS4057-N-FJ1-2020) Assignment #2

Karen Kaur Premajit Singh
A0243311
11 May 2021

Explainable AI in the Transparency of Credit Risk Management: A Mixed Method Research

Word count: 1994

ABSTRACT

As we leap into the dawning digital revolution, the implementation of Artificial Intelligence across various industries is increasing at as fast a pace as is the formation of data. However, as we progress towards the ethical management of said data, we find ourselves facing a limitation to the use of AI-based systems: transparency. The nature of previous and current machine learning models are black-box based which limits our access to understanding how these models arrive to their results. Thus, both providers and consumers are blindsighted as to how the data garnered weighs into said results.

This is where the discussion of Explainable Artificial Intelligence emerges that promises an increase in the trust and transparency of Artificial Intelligence systems. Trust is key to building confidence and acceptance of Artificial Intelligence-based decision-making systems regardless of the industry it is implemented in. This study explores a survey on Explainable Artificial Intelligence to understand how it promotes transparency and will be followed by a replication of an empirical analysis of SMEs requesting peer to peer lending using an Explainable Artificial Intelligence model to understand how their financial characteristics can be used to predict future behaviour.

1. INTRODUCTION

FinTech – or financial technology – is a new financial industry that applies technology to improve financial services. The innovation of fintech was driven by the progress in e-finance and mobile technology for financial companies after the global financial crisis in 2009. This development is characterised by the integration of internet technology, artificial intelligence (AI) and big analytic data. This has challenged traditional financial institutions such as banks to update their business models to adapt to the increase in data collection and usage.

One of the current uses of AI in FinTech is analysing a client risk profile – an example of which is peer to peer lending approval. Banks and other financial institutions are required to decide whether an individual or business is creditworthy to offer a suitable level of credit that is priced accordingly.

A focus on fairness in AI decision making is becoming more prevalent alongside the discussion around ethical decision making. An important question to be addressed is the need for greater transparency and best practices around model interpretability.

The EU General Data Protection Regulation (GDPR) introduced a right of explanation for individuals to have "meaningful information of the logic involved" when an automated decision-making takes place with "legal or similar relevant effect" on individuals.

The current implementation of black box machine learning models provide little access to understanding how variables are combined to make its predictions. An important step in constructing a machine learning model is the explanation of its logic that is expressed in a comprehensible, human-readable format that highlights the biases that are learned by the model. As such, black box AI has proved unsuitable in regulated financial services because both designers and users have little access to understanding how variables contribute to the predictions made. In addition, there may exist possible biases inherited by the algorithms from human prejudices embedded in the training data that can lead to unfair or even wrong decisions [11].

Relating to the ethical aspect, black box AI models challenges the ethical guidelines of transparency. In a presentation by the European Commission High Level Expert Group on AI on their Ethics Guidelines for Trustworthy AI in April 2019, AI systems and their decisions should be explained in a manner that is adapted to the relevant stakeholder [12]. Extending this to accountability, AI systems should develop mechanisms for auditability, assessment of algorithms, data and design processes. This then allows AI developers and the relevant stakeholders to validate their decision rationale. As per the Bank of England's definition of explainability, interested stakeholders should be able to comprehend the main drivers of a model driven decision.

2. LITERATURE REVIEW

2.1. TRANSPARENCY

Kurniawan (2019) states that there is a significant effect that trust has on a customer's attitude towards using a peer to peer lending service. He had studied the factors that were influencing the decision to implement peer to peer lending service platforms. He finds that the transparency of the information provided encourages a positive attitude towards a service. [3]
Transparency is a fundamental principle when processing data under the EU General Data Protection Regulation (GDPR) with its focus being on the provision of information and explanation. However, this is at odds with the reality of what is being practiced as written by Felzmann et al (2019). The authors had concluded that research is needed into the implementation of legal transparency requirements into technical systems. [4]

2.2. EXPLAINABLE AI

As a result of the growing need for transparency in automated decision making, there has been a growing acceptance for Explainable AI models that allows both highly accurate and predictive results. A proposed methodology by Bussman et al. (2020) utilises a gradient boosting machine learning algorithm alongside Shapley values to predict credit risk of enterprises that have applied for credit. [5].
The term Explainable Artificial Intelligence (XAI) relates to the initiatives and efforts carried out in response to AI transparency and trust concerns rather than to a formal technical concept [9]. As per DARPA [10], XAI aim to "produce more explainable models, while maintaining a high level of learning performance (prediction accuracy); and enable human users to understand, appropriately trust and effectively manage the emerging generation of artificially intelligent partners". Some of key concepts of XAI are interpretable machine learning, responsible AI and accurate AI which all addresses the greater need for transparency in decision-making models such as the credit scoring model in this study.

This mixed method research is based on this written paper by Bussmann et al (2020) to raise the key research question which is as follows:
How can transparency be encouraged using Explainable AI in credit risk management?

3. RESEARCH METHODOLOGY

This study uses a mixed methods research methodology and attempts to conduct parallel quantitative and qualitative research. As such the central research question proposed as per paragraph **2.2**. It is then broken down into individual specific research questions that include quantitative and qualitative sub-questions that will be integrated back to the main research question.

The first sub-question is to be addressed using the qualitative component of the study and is as below:
RQ1: How does Explainable Artificial Intelligence increase transparency?

The component of the study utilises an existing publication on a survey on Explainable AI. [9] By answering this research question, the qualitative study seeks to explore how Explainable AI has been utilised to promote transparency.

The second sub-question is to be answered by the quantitative component of this study, specified as follows:

RQ2: How can Explainable Artificial Intelligence be used to implement the transparency of decision making within peer to peer lending?

As this study uses a mixed methods research, the quantitative and qualitative components should be connected with each other rather than be studied separately. The connection between the quantitative and qualitative studies begins at the research design – the stage of constructing the research questions.

4. METHODOLOGY

4.1 CREDIT RISK MODELS

A credit risk model is used to model and predict the probability of an individual firm defaulting. These models are based on machine learning approaches used for financial analysis and decision-making tasks. These models extract non-linear relationships from financial information found in the balance sheets and each model is chosen with the aim of optimising the predictive accuracy.

Consider *N* firms that observes *T* different variables such as balance-sheet measures or financial ratios. For each firm *n*, a variable γ_n that will indicate whether the firm has defaulted on its loan or not where $\gamma_n = 1$ if the firm defaults and $\gamma_n = 0$ otherwise. A credit risk model is used to develop a relationship between the explanatory variables, T and the dependent variable $\gamma_.$

As per the study by Bussmann et al. (2020), the most common method utilised for a credit scoring model is the logistic regression model that aims to classify the dependent variable into two groups (ie. whether or not a firm has defaulted in this scenario). The following model is used:

$$ln(\frac{p_n}{1-p_n}) = \alpha + \sum_{t=1}^{T} \beta_t x_{nt}$$

Image 1. Logistic regression model sourced from Bussmann et al. (2020)

The variable p_n represents the probability of a firm *n* defaulting, x_i represents the T-dimensional vector of explanatory variables specific to the individual firm, the parameter α is the model's intercept and β_t represents the t-th regression coefficient.

As such, the probability of an individual firm defaulting is as follows:

$$p_n = (1 + exp(\alpha + \sum_{t=1}^{T} \beta_t x_{nt}))^{-1}$$

Image 2. Probability of default sourced from Bussmann et al. (2020)

5. ANALYSIS APPROACH

5.1. QUALITATIVE ANALYSIS

This study refers to an existing publication on a survey on Explainable AI [9] to answer the qualitative research question of how Explainable AI promotes transparency which will be discussed below.

5.2. QUANTITATIVE ANALYSIS

The data considered was sourced from a study by Giudici, P. et al [7] that was initially supplied by the European External Credit Assessment Institution (ECAI). They specialise in credit scoring for peer to peer (P2P) lending platforms focused on SME commercial lending. The dataset is made up of official financial information on 4,514 Italian SMEs that represent the target of P2P lending platforms. The information about the status where 0 = active and 1 = defaulted is also included. The dataset shows that 11.03% of the firms have defaulted.

This study aims to recreate the XGBoost model that was proposed by Bussmann et al (2020) that was able to increase the performance of prediction of an existing logistic regression scoring model constructed by Giudici (2018), Ahelegby et al. (2019), and Giudici et al. (2019a,b). The results from the model will then be explained by employing Shapley values.

The combination of both the qualitative and quantitative analysis will be used to show how the application of Explainable Artificial Intelligence can increase the transparency of decision making models.

6. RESULTS

The Python system was used for this study. The dataset from the ECAI was split into a training set of 80% and test set of 20%. A logistic regression model was first estimated on the training set then applied to the test set. This was used as a basis for comparison with the gradient boosting model XGBoost. The XGBoost model was then also estimated on the training set to be applied to the test set. The ROC curves of the two models are in **Image 3** as follows.

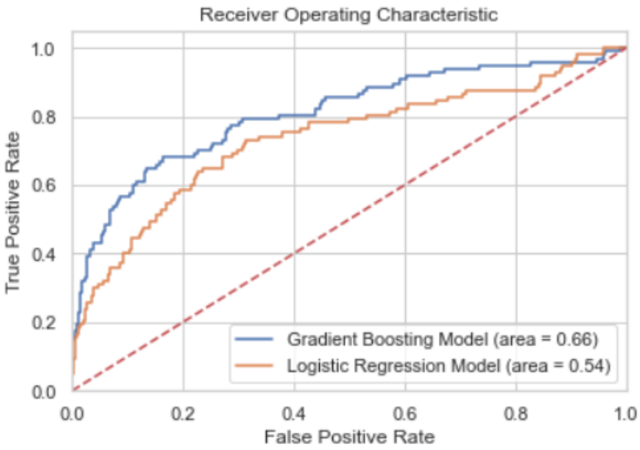


Image 3. Receiver Operating Characteristic (ROC) curves for the logistic credit risk model and for the XGboost model. The results relating to the XGBoost model is shown in blue and the logistic model's results is in orange.

Note that the XGBoost improves the predictive accuracy in comparison to the logistic regression model, where the graph indicates the increase from 0.54 to 0.66.

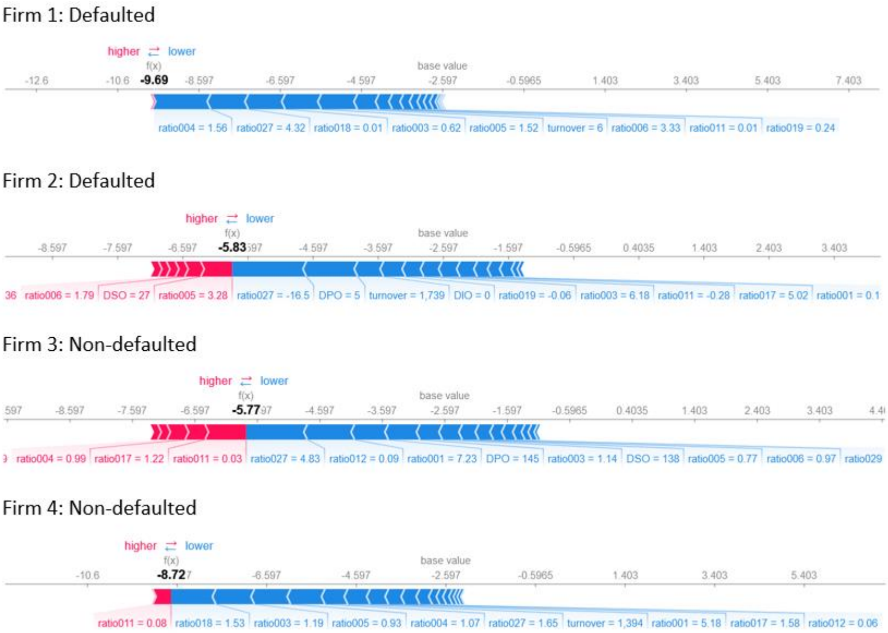
Shapley values were calculated for four randomly selected firms in the data set to indicate which variables contributed towards the predictions of whether or not that firm would default. **Image 4** in the discussions shows the usefulness of indicating the variables that largely contribute to a prediction of individual firms, as per feature selection models.

7. DISCUSSION

The XGBoost model combined with the interpretation of results by Shapley values is a suitable alternative with respect to the lack of transparency in black-box decision-making models. The concept of black-box model have been exploited by technological companies in order to protect their intellectual property to conserve competitiveness. This contradicts the need for responsible AI whose three pillar are accountability, responsibility and transparency.

The use of XAI systems generally requires justifications for a particular outcome, especially in cases where unexpected decisions are made. This allows for auditable ways to support algorithmic decisions to encourage fair and ethical decisions that will build trust. In addition to justifying a decision, there is also a deeper understanding to unknown vulnerabilities that can assist debugging.

As discussed in the survey by Adadi (2018), AI tools in the financial services poses queries around fair lending despite the fact that the financial industry is heavily regulated and are responsible for making fair decisions. XAI implementation provides an outlet to provide explanations around credit application decisions.



Firm 1: Defaulted

Firm 2: Defaulted

Firm 3: Non-defaulted

Firm 4: Non-defaulted

Image 4. Contribution of each explanatory variable according to the Shapley's decomposition for four prediction, two of which have defaulted and the other have not. The red colour is an indication of low variable importance and a blue colour indicates a high variable importance.

The above plots shows each feature that contributes to pushing the model output from the base value - average model output over the training dataset - to the model output. The features that push the prediction higher are shown in red.

8. CONCLUSION

The replication of the study by Bussmann et al (2020) shows that the use of explainable artificial intelligence, in this case the XGBoost model alongside interpretation by Shapley values, reaffirms that the use of explainable AI can indeed increase transparency in decision making models. This meets the current need for high predictive accuracy alongside high interpretability in this era of transparent data management.

Future studies would propose to explore alternative algorithms that can further improve predictive accuracy whilst maintaining or even improving the model interpretability. As also proposed by Bussmann et al. (2020), an alternative would also be to improve model development by utilising the Shapley values to identify key features that can improve model predictions.

Stricter implementation of legal responsibilities should also be put in place and monitored. This responsibility lies within the organisations utilising the data for their decision making systems. Relating this to the current EU GDPR regulations, it remains to be quite general with respect to the aritificial intelligence systems. More research should be carried out to explore possible technical and theoretical checks to be put in place to regulate the ethical use of data within decision making.

9. REFERENCES

[1] Stevens, A., Deruyck, P., Veldhoven, Z.V., Vanthienen, J. (2020) *Explainability and Fairness in Machine Learning: Improve Fair End-to-end Lending for Kiva*, 2020 IEEE Symposium Series on Computational Intelligence (SSCI), pp. 1241-1248

[2] Paul, L.R. and Sadath, L. (2021) *A Systematic Analysis on FinTech and Its Applications,* 2021 International Conference on Innovative Practices in Technology and Management (ICIPTM), pp. 131-136

[3] Kurniawan, R. (2019) *Examination of the Factors Contributing to Financial Technology Adoption in Indonesia using Technology Acceptance Model: Case Study of Peer to Peer Lending Service Platform*, 2019 International Conference on Information Management and Technology (ICIM Tech)

[4] Felzmann, H., Villaronga, E.F., Lutz, C., Tamo-Larrieux, A. (2019) *Transparency you can trust: Transparency requirements for artificial intelligence between legal norms and contextual concerns,* Big Data & Society

[5] Bussmann, N., Paolo, G., Dimitri, M., Jochen, P. (2020) *Explainable AI in FinTech Risk Management*

[6] Giudici, P., Hadji-Misheva, B., and Spelta, A. (2019a), *Correlation network models to improve P2P credit risk management.* Artif. Intell. Finance

[7] Giudici, P., Hadji-Misheva, B., and Spelta, A. (2019b), *Network based credit risk models*, Qual. Eng. 32, 199-211

[8] Giudici, P. (2018) *Financial data science* in Statistics and Probability Letters, Vol. 136 (Elsevier). 160-164

[9] Adadi, A. and Berrada, M. (2018) *Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI)*, IEEE Access, vol. 6, pp. 52138-52160

[10] Gunning, D. (2018) *Explainable artificial intelligence (XAI)*, Defense Advanced Research Projects Agency (DARPA). Accessed: 5 April 2021 [Online]
Available: https://www.darpa.mil/program/explainable-artificial-intelligence

[11] Grossi, V., Giannotti, F., Pedreschi, D. et al. (2021) *Data science: a game changer for science and innovation.* Int J Data Sci 11, 263-278

[12] European Commission High Level Expert Group on AI on their Ethics Guidelines for Trustworthy AI. Accessed: 5 April 2021 [Online]
Available: https://ec.europa.eu/futurium/en/ai-alliance-consultation

10. APPENDIX

Table A1
Description of variables included in the dataset.

ID	FORMULA	TYPE
RATIO001	(Total assets – Shareholders Funds) / Shareholders Funds	Continuous
RATIO002	(Long term debt + Loans) / Shareholders Funds	Continuous
RATIO003	Total assets / Total liabilities	Continuous
RATIO004	Current assets / Current liabilities	Continuous
RATIO005	(Current assets – Current assets: stocks) / Current liabilities	Continuous
RATIO006	(Shareholders funds + Non current liabilities) / Fixed assets	Continuous
RATIO008	EBIT / Interest paid	Continuous
RATIO011	(Profit (loss) before tax + Interest paid) / Total assets	Continuous
RATIO012	P/L after tax / Shareholders Funds	Continuous
RATIO013	Gross profit / Operating revenues	Continuous
RATIO017	Operating revenues / Total assets	Continuous
RATIO018	Sales / Total assets	Continuous
RATIO019	Interest paid / (Profit before taxes + Interest paid)	Continuous
RATIO027	EBITDA / Interest paid	Continuous
RATIO029	EBITDA / Operating revenues	Continuous
RATIO030	EBITDA / Sales	Continuous
RATIO036	Constraint EBIT	Dichotomous
RATIO037	Constraint EBIT	Dichotomous
RATIO039	Constraint PL before tax	Dichotomous
RATIO040	Constraint Financial PL	Dichotomous
DPO	Constraint P/L for period th EUR	Continuous
DSO	Trade payables / Operating revenues	Continuous
DIO	Inventories / Operating revenues	Continuous