

TDK-dolgozat

Mócza-Muráncsik Erika Mária

Debreceni Egyetem
Informatikai Kar
Adattudomány és Vizualizáció Tanszék

Sustainable AI-based Credit Scoring:
A Way to Shift from Responsible AI to Sustainable AI

Témavezető:

Dr. Szathmáry László, egyetemi docens
Információ Technológia Tanszék

Szerző:

Mócza-Muráncsik Erika Mária
Adattudomány MSc 2. évfolyam

Lezárás dátuma: 2024.10.28.

Table of Contents

1. Introduction	5
2. From Responsible to Sustainable AI.....	6
2.1. Introduction to Sustainable Development	6
2.2 Gaps in Responsible AI Framework.....	6
3. Implementation of Sustainable AI in Credit Scoring.....	8
3.1. Credit Scoring	8
3.2. Sustainable AI Framework – Credit Scoring Approach	10
3.2.1. AI Risk Assessment.....	10
3.2.2. Fairness – Data and Model Bias.....	12
3.2.3. Explainability and Transparency	13
3.2.4. Environmental Impact	14
3.2.5. Privacy and Security.....	15
3.3. Operationalize Sustainable AI in Credit Scoring – Case Study	15
3.3.1. Dataset and Preprocessing.....	15
3.3.2. Machine Learning Models Used in Training	16
3.3.3. Fairness Analysis	18
3.3.4. Explainability Analysis.....	23
3.3.5. Environmental Impact	27
4. Conclusions	28
5. References.....	29
6. Tools and Libraries Used.....	33
7. Author’s Own Contribution	34
8. Acknowledgements	34

1. Introduction

With the development of artificial intelligence (AI), a variety of opportunities have emerged to implement machine learning solutions to real world problems. However, there is also an even increasing need to make sure that these techniques perform as intended so different industries where trust plays an important role could also benefit from machine learning solutions.

Responsible AI, as a new paradigm tries to fill this gap. However, still less attention has been put to investigate the completeness of its layers. Recently, some studies started to focus on the environmental impact of machine learning algorithms that shall also be part of this AI framework.

Finance is a heavily regulated industry where reliability is critical. The new EU AI Act – the first comprehensive regulation on AI – has identified machine learning-based credit scoring models that evaluate creditworthiness as a high-risk use case in financial sector.

In this research we try to investigate the potential gaps in the current responsible AI paradigm and propose a sustainable AI framework on a machine learning-based credit scoring case study. We focus on the bias, explainability and environmental impact of these algorithms in one analysis that was mainly investigated separately in other research papers. It will help groups in society to get access to credit that would have been otherwise ignored by traditional credit scoring firms, increase transparency as it will not consider AI as a black box and measure environmental impact that could also highlight inefficiencies while achieving climate change goals. This will hopefully give a new insight into AI risk management.

2. From Responsible to Sustainable AI

2.1. Introduction to Sustainable Development

In 1987, the United Nations Brundtland Commission defined sustainable development as “meeting the needs of the present without compromising the ability of future generations to meet their own needs.” (UN sustainability, 2024) Sustainable development is a holistic approach that considers the interaction of environmental, social and economic values. For any organization planning to implement this approach should consider the harmonization of these three pillars and integrate them into their core business. The latter is essential because if we think of sustainability related goals separately from the company’s business strategy, it might not be authentic, and those actions might have a risk to become greenwashing, sponsoring activities or even temporary causes that might not be successful in the long run.

To be able to integrate these goals to the company’s operation, Edward Freeman’s stakeholder theory could help to put this in context. In his book “Strategic Management: A Stakeholder Approach” he defines a stakeholder as “any group or individual who can affect or is affected by the achievement of the organization’s objectives” (Stakeholder, 2024). Examples of stakeholders could be customers, employees, managers, competitors, government, non-governmental organizations, the environment or even future generations. If organizations look after their stakeholders as part of their business strategy, sustainability could be achieved at company level because groups in society and even environment are considered as stakeholders. Now that we have the foundation of what responsibility and sustainability means in the context of company management, we could start to apply these concepts to AI management.

2.2 Gaps in Responsible AI Framework

There is no single definition of responsible AI (or trustworthy AI). Each institution has their own policy that determines what they mean under this concept. However, the main idea behind responsible AI is that it functions as a risk management framework for machine learning solutions.

I chose IBM’s definition since it is a comprehensive one that includes all its major aspects: “Responsible artificial intelligence (AI) is a set of principles that help guide the design, development, deployment and use of AI—building trust in AI solutions that have the potential to empower organizations and their stakeholders. Responsible AI involves the consideration of a broader societal impact of AI systems and the measures required to align these technologies

with stakeholder values, legal standards and ethical principles. Responsible AI aims to embed such ethical principles into AI applications and workflows to mitigate risks and negative outcomes associated with the use of AI, while maximizing positive outcomes” (IBM Responsible AI, 2024).

The key in this definition is **trust** because it is the fundamental element of any systems to be able to succeed in the long run: if customers of AI systems lose their trust, it might have an impact not just on the company’s performance but on the whole AI community as well and once it has been lost, it is very hard to earn back.

IBM considers the following layers of responsible AI:

- Explainability
- Fairness
- Robustness
- Transparency
- Privacy

Google also identified similar Responsible AI pillars as IBM: fairness, interpretability, safety and security (Google RAI, 2024).

As we could see, to be able to establish trust, we should promote AI solutions that are transparent, have necessary tools to understand its decision-making process while not considering it as a black box. With these AI systems we should move towards fairness via addressing bias and address privacy, information security concerns.

If we consider what this definition means in terms of sustainability, we could see that there is an aim to integrate these principles into the company’s core AI strategy (“principles that help guide the design, development, deployment and use of AI”) that is essential to foster responsibility.

Social values are also incorporated into the definition, however environmental impacts have not been emphasized yet. Although the major AI contributors, like Google have sustainability strategies but still not considering it as part of responsible AI. That might be because when the AI community started to look at sustainability as a new area in machine learning, they started to focus on what AI could do to achieve sustainability goals. This is AI for Sustainability when AI is used to tackle climate change related issues. However, it shall not be confused with Sustainable AI, where the latter is about to assess the sustainability impact of machine learning

models: e.g. measuring the carbon footprint of training algorithms (van Wynsberghe, 2021). That indicates a gap that shall be filled to aim for sustainable AI.

Stakeholder theory also has been included in responsible AI definition that could put sustainable frameworks into context. This could also align with the human-centered AI approach because it not just fosters how AI could augment human capabilities but helps to “understand how people engage with and trust AI systems. There is also a need to explain the operation of AI models and improve peoples’ understanding of how AI systems operate. A human-centered AI strategy must also gauge the negative consequences or potential misuse of AI systems, include ways to mitigate human and AI biases, and be able to measure peoples’ perceptions (or misperceptions) of AI systems” (IBM HCAI, 2024).

What the current responsible AI framework still doesn’t address is the interaction and potential conflict of its pillars (Sadek et al, 2024). This is another area that is also part of sustainable development concepts: we should not only see these layers in silos but evaluate if interaction of these values would cause any synergies or conflicts. In that case companies adopting AI systems shall assess those impacts and understand how to manage them.

One of the other critics that have been raised regarding trustworthy AI is that these are high level principles and there is no “one-fit all” solution. To address that we should define these in a use case specific way. That would lead us to “operationalizing trustworthy AI” (Sadek et al, 2024).

In the next chapter, I start to elaborate on how we could operationalize sustainable AI and use credit scoring machine learning algorithms to demonstrate that.

3. Implementation of Sustainable AI in Credit Scoring

3.1. Credit Scoring

If someone wants to apply for credit in most countries, financial institutions should determine if credit shall be granted to that person based on their creditworthiness analysis. To accelerate this process, banks needed an automated way to estimate the potential credit risks associated with the applicants.

We could see an evolution how these risk assessment techniques emerged. Most financial institutions outsource that activity to Fintech companies (for example FICO¹, the dominant

¹ <https://www.fico.com/en>

player in that sector) so they can use their technology expertise to analyse large customer datasets. Up until now mainly credit scorecard methods are applied to access creditworthiness, so we use the name credit scoring when talking about this analysis.

Credit score is a number that gives an estimate on how likely you can pay your loan back. It is ranging from 300 to 850 where the highest number indicates the highest creditworthiness. These scorecards are based on the applicant’s credit history.

FICO Score, the widely used credit score, is calculated based on the above five categories:

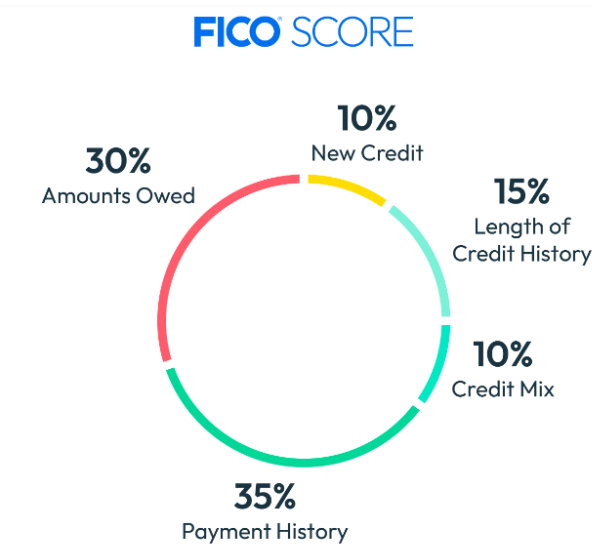


Figure 1: Fico Credit Score (FICO 2, 2024)

As we can see, all the information that credit rating firms use, are based on our credit history: if we manage to pay our credit bill back on time, what the actual amount of credit is that we own and how long we have been managing credit accounts. (FICO 1, 2024)

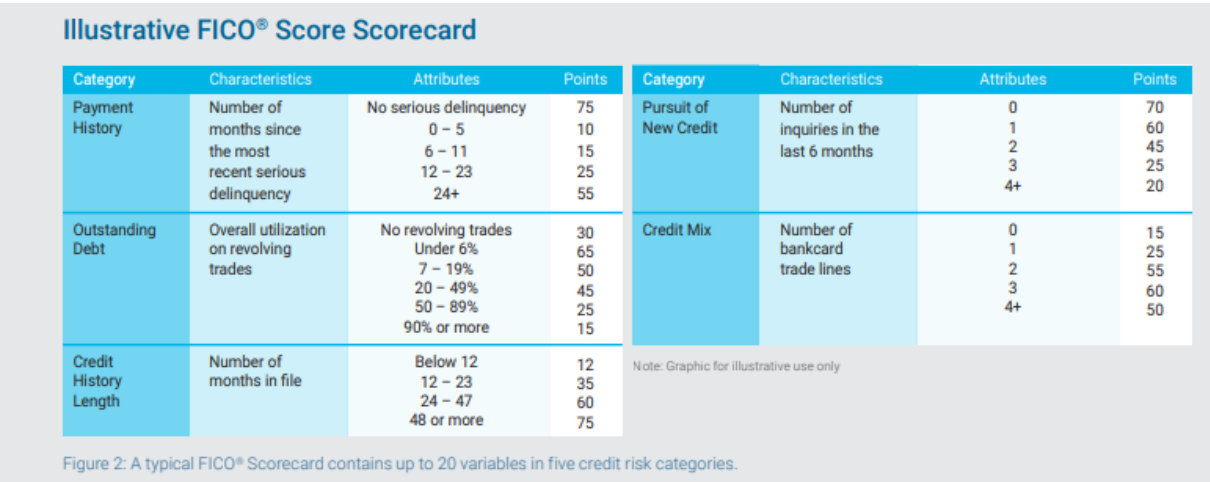


Figure 2: A typical FICO® Scorecard contains up to 20 variables in five credit risk categories.

Figure 2: FICO Scorecard (FICO 1, 2024)

Thirty years ago, Fintech companies, like FICO started to work on machine learning solutions as a new way to determine creditworthiness in a lab environment however given the lack of understanding of these black box models and biases, they still have not fully implemented them yet (FICO White Paper, 2024). To addition, with the rise of new regulations, like EU AI Act², credit scoring algorithms have been determined as high risk so these machine learning models could not be implemented in production without proper risk management.

3.2. Sustainable AI Framework – Credit Scoring Approach

To successfully implement Sustainable AI, we would need to see it as a strategy and risk management framework that shall be in the cornerstone of the company's AI operation that affect the whole AI lifecycle.

As described in previous sections, sustainability is a mindset and to be able to adopt this, we shall construct a use-case specific methodology and with that we could gradually start to operationalize this approach. In this research our use case is credit scoring machine learning algorithms.

3.2.1. AI Risk Assessment

As a first step, firms shall perform a risk assessment to determine the level of risk their AI applications have and understand their rules and regulations they operate in. The below figure shows the EU AI Act risk pyramid on AI application (EU AI Act and N. Díaz-Rodríguez et al. 2023)

² <https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai>

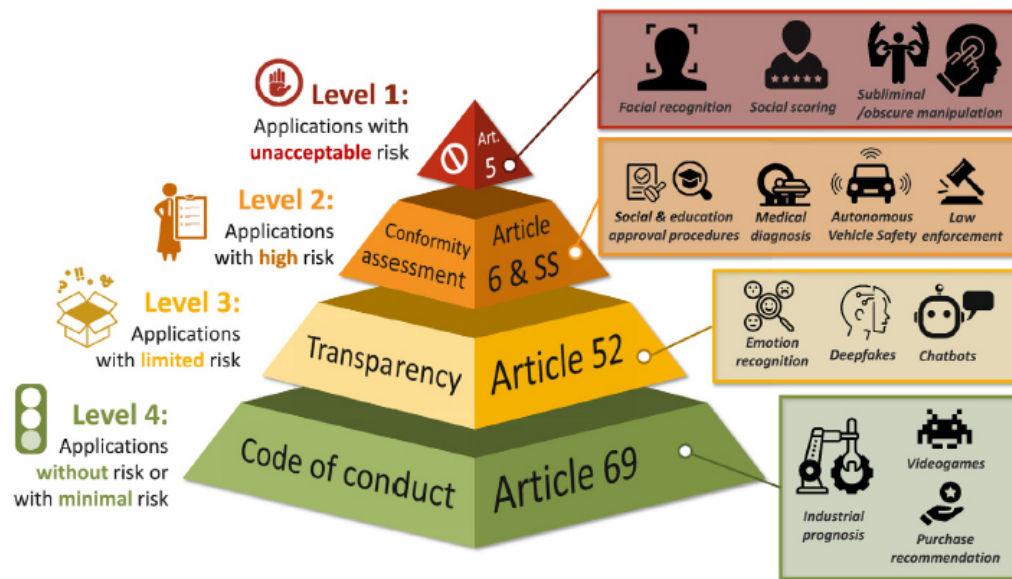


Figure 3: EU AI Act Risk Pyramid (EU AI Act and N. Díaz-Rodríguez et al. 2023)

For the European Union, EU AI Act entered into force on 1 August 2024 and machine learning algorithms accessing creditworthiness has been identified as high risk. These AI models are under Annex III use case and AI Act will apply by 24-month deadline. (EU AI Act Summary, 2024)

High-risk AI systems are subject to strict obligations before they can be put on the market (EU AI Act, 2024):

- adequate risk assessment and mitigation systems
- high quality of the datasets feeding the system to minimise risks and discriminatory outcomes
- logging of activity to ensure traceability of results
- detailed documentation providing all information necessary on the system and its purpose for authorities to assess its compliance
- clear and adequate information to the deployer
- appropriate human oversight measures to minimise risk
- high level of robustness, security and accuracy

EU AI Act allows free use for algorithms that are identified as low-risk (e.g. AI-enabled video games or spam filters), however for high-risk AI systems, like credit scoring algorithms it is essential to implement trustworthy AI solutions.

3.2.2. Fairness – Data and Model Bias

Fairness plays a crucial role in credit scoring machine learning models. In the US, Equal Credit Opportunity Act considers activities unlawful if “creditor discriminate against any applicant, with respect to any aspect of a credit transaction on the basis of race, colour, religion, national origin, sex or marital status, or age.” (Equal Credit Opportunity Act, 2024)

In 2019, some Apple Card customers reported in the US that their credit card’s issuer, Goldman Sachs gave women far lower credit limit, even if they share all assets and accounts with their spouse. Some married couples found the husband received a credit limit that was 10 times higher than the wife’s. (Apple credit bias, 2024) That was an example when the machine learning model produced bias however, there are cases where not the algorithm itself, but the underlying data is the cause of bias.

Data bias in case of credit scoring happens when minority and low-income groups produce limited credit history. Blattner and Nelson showed in their research (Blattner et al, 2021) that because of lack of data these groups are more likely get rejected in credit application process. These effects could cause misallocation of credit that also could prevent further wealth accumulation for those groups. (Inequality in Credit, 2024). To resolve this data issue, one solution might be that institutions start to credit those minority groups. Affirm, a company that wants to fill this gap, tries to provide an alternative to credit cards so they provide mini loans that are not just based on credit history and do not impose high cost because of delayed payments. That could help minority groups to get access to credit. (Affirm, 2024)

FICO, the leading Fintech company providing credit scoring services, plans to integrate alternative data to their AI models. With that additional information, machine learning models could use more information to access the applicant’s creditworthiness so more groups could get access to credit.

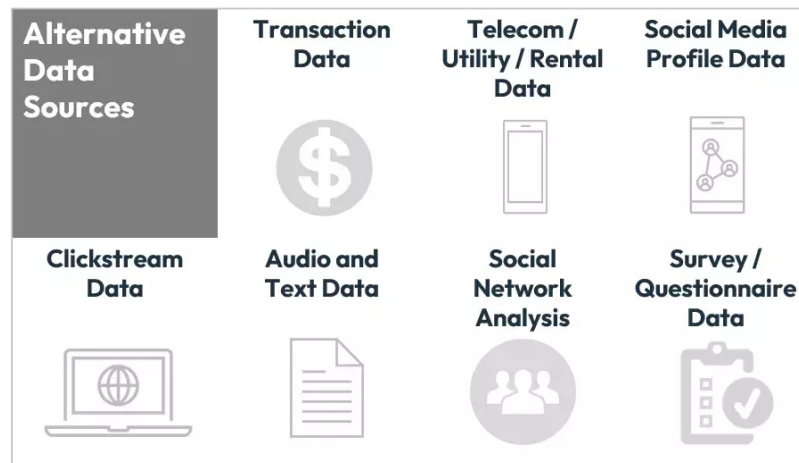


Figure 4: Alternative Data for Credit Scoring (FICO Alternative Data, 2024)

There are tools available to assess model bias in the machine learning algorithms: these produce a plenty of metrics to identify unfairness and provide solutions to mitigate those. However, “fairness in AI is a sociotechnical challenge that so no software toolkit will “solve” fairness in all AI systems.” (Microsoft, 2020) These are the first steps in the way to achieve bias-free algorithms.

3.2.3. Explainability and Transparency

If we want to trust in machine learning models, one of the key requirements is that we would need transparency in a way to understand how algorithms make decisions. Finance is a heavily regulated industry which implies that only those products and services could be implemented to production that comply with rules and regulations, have transparent procedures, documented and properly risk managed. We could see based on the example of the subprime mortgage crises³ back in 2008 what could happen if participants lose trust in the financial sector which was due to irresponsible crediting practices at that time. After this incident regulators worldwide demand even more rigorous requirements from banks in financial services industry. We could see that there is a trade-off between model complexity and interpretability. Once we move to the direction of complex models, interpretability decreases. For credit scoring machine learning models, the industry standard is to use gradient boosting algorithms. For these models, tree-based XAI models could be the basis for the analysis.

³ https://en.wikipedia.org/wiki/Subprime_mortgage_crisis

3.2.4. Environmental Impact

The paper on AlexNet published in 2012 became the catalyst on the current assumptions on AI, where “bigger-is-better” paradigm started to shape the culture of AI development: “The AlexNet authors viewed scale as a key source of their model’s benchmark beating performance: they stated that “all of our experiments suggest that our results can be improved simply by waiting for faster GPUs and bigger datasets to become available” (Krizhevsky et al, 2012). It was further reinforced by Sutton in his “bitter lesson” article (Sutton, 2019), which states that approaches based on more computation win in the long term, as the cost of compute decreases with time” (Varoquaux et al, 2024).

Recently, a new perspective on AI has emerged: an increasing number of research started to focus on the environmental impact of machine learning models that could have a huge impact. These papers mainly investigated regarding large language and image processing models; however, it is also important to look at all type of machine learning models because all impacts could add up.

In some cases, we could get better performance with smaller models: in case of tabular data that is currently the basis for credit scoring machine learning algorithms, it has been shown that tree-based models produce better prediction and are much faster compared to neural networks. (Varoquaux et al, 2024)

To move forward to a more sustainable AI, it is essential to start to be transparent on the energy consumption of the machine learning algorithms. Prominent Sustainable AI Researchers, like Alexandra Luccioni proposed an “AI Energy Star project” that could categorize the AI models where the algorithms that consumes the least energy could get the highest rating. (Luccioni, 2024)

Current credit scoring AI models use tabular data, but soon alternative data could be introduced to creditworthiness analysis which would be based on large language models. Given large language models have significant environmental impact, it will be even more important to disclose environmental impact of credit scoring model in the future and consider how these impacts could be decreased with efficient models.

3.2.5. Privacy and Security

Protecting privacy and ensuring security in AI systems also play an important role in sustainable AI. However, this part of the framework is not accessed in this research. This could be an interesting layer to be investigated in future analysis.

3.3. Operationalize Sustainable AI in Credit Scoring – Case Study

To demonstrate the above Sustainable AI framework, we use a case study where the fairness, explainability and environmental impacts have been assessed.

3.3.1. Dataset and Preprocessing

For credit scoring analysis, we chose an open-source dataset available on UCI Machine Learning⁴ Repository (UCI Dataset, 2024). It contains data of 30.000 credit card customers in Taiwan where the target variable indicates if that person will declare default next month.

The dataset has 23 features, as we can see in the table below:

Features	Description
LIMIT BAL	Amount of the given credit (NT dollar)
SEX	Male/Female
EDUCATION LEVEL	Graduate School/University/High School/Others
MARRIAGE	Married/Single/Others
AGE	Age
PAY_1 – PAY 6	Repayment status for the last six month (pay duly/number of months of payment delay)
BILL_AMT 1- BILL_AMIT 6	Amount of bill statement (in NT dollar) in the last six months
PAY_AMT_1-PAY_AMT_6	Amount of previous payments in the last six months

Table 1: Model features

The dataset is imbalanced in terms of the target variable: 22% of the customers labelled as declaring default next month and 78% non-default. To correct this imbalance, random under-sampling from Imbalanced Learn library⁵ has been used to avoid potential bias towards

⁴ <https://archive.ics.uci.edu/dataset/350/default+of+credit+card+clients>

⁵ https://imbalanced-learn.org/stable/references/generated/imblearn.under_sampling.RandomUnderSampler.html

favouring majority class at the expense of the minority class. Although down sampling could cause loss of information, given the size of the dataset, after rebalancing we still have enough data for training and data would be less prone to overfitting as it would have been with oversampling.

From the 23 features there are 3 categorical variables that have been label encoded. Also, before training the features have been normalized with the Standard Scaler⁶ from the Sklearn library.

3.3.2. Machine Learning Models Used in Training

For machine learning-based creditworthiness analysis gradient boosting algorithms are increasingly becoming the preferred method due to their accuracy (FICO White Paper, 2024). Gradient boosting algorithms, invented by Friedman in 2001, are part of the ensemble machine learning models. “Ensemble learning refers to the generation and combination of multiple inducers to solve a particular machine learning task. The intuitive explanation for the ensemble methodology stems from human nature and the tendency to gather different opinions and weigh and combine them to make a complex decision. The main idea is that weighing and aggregating several individual opinions will be better than choosing the opinion of one individual”. (Sagi et al, 2018).

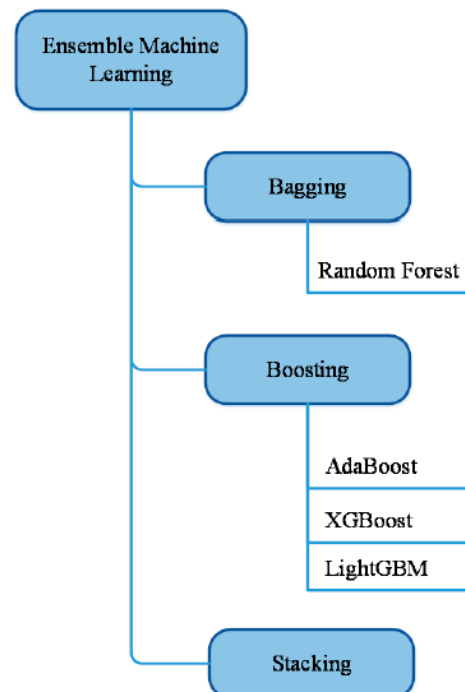


Figure 5: Ensemble machine learning framework (Li et al, 2020)

⁶ <https://scikit-learn.org/dev/modules/generated/sklearn.preprocessing.StandardScaler.html>

Gradient boosting is part of the boosting family: in boosting a sequence of weak classification trees is computed, and “the model parameters of each weak classifier are obtained by minimizing the loss function of the previous model on the training set”. (Li et al, 2020). Gradient boosting is different compared to the other boosting techniques, like AdaBoost in a way that “updates the sample data in the direction of the negative gradient to make the algorithm converge globally. Gradient boosting decision trees has the advantages of strongly generalizing ability and not being easy to overfitting.” From gradient boosting, two of these methods, XGBoost and LightGBM are the most popular choices. (Li et al, 2020)

EXtreme Gradient Boosting (XGBoost), proposed by Chen and Guestrin in 2016, shows high performance: it introduces a regularization term in the loss function to prevent overfitting (Li et al, 2020):

$$\mathcal{L}_K(F(x_i)) = \sum_{i=1}^n \Psi(y_i, F_K(x_i)) + \sum_{k=1}^K \Omega(f_k),$$

$$\Omega(f) = \gamma T + \frac{1}{2} \lambda \|\omega\|^2.$$

$F_K(x_i)$: prediction on the i-th sample at the K-th boost

Ψ : loss function that measures the differences between the prediction and the actual label

T: number of leaves

$\Omega(f_k)$: regularization term

γ : the complexity parameter and represents the complexity of leaves

λ : penalty parameter

ω : output of each leaf node

“The main cost in Gradient Boosting Decision Trees lies in learning the decision trees, and the most time-consuming part in learning a decision tree is to find the best split points. One of the most popular algorithms to find split points is the pre-sorted algorithm, which enumerates all possible split points on the pre-sorted feature values.” (Ke et al, 2017)

Light Gradient Boosting Machine (LightGBM, 2024) was invented by Microsoft Research that is a further enhancement of XGBoost. Compared to XGBoost, it can reduce the memory usage and improve the training speed by using the decision tree algorithm based on histogram as opposed to pre-sorted features. “Histogram-based algorithm buckets continuous feature values into discrete bins and uses these bins to construct feature histograms during training.” (Ke et al, 2017)

Also, in XGBoost trees grow level-wise, while in LightGBM leaf-wise with depth limitation. “Leaf-wise may cause over-fitting when data is small, so LightGBM includes the “max_depth” parameter to limit tree depth. However, trees still grow leaf-wise even when “max_depth” is specified”. (LightGBM documentation, 2024)

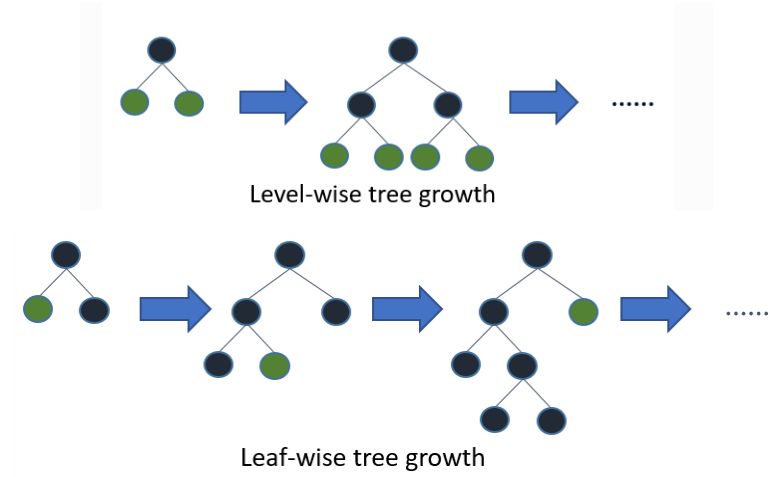


Figure 6: Leaf vs level-wise tree growth (LightGBM documentation, 2024)

In this analysis, we use LightGBM model along with the traditional logistic regression model to access sustainability AI framework.

3.3.3. Fairness Analysis

In this case study we used the open-source IBM AI Fairness 360 library⁷ because it includes a variety of methods to assess and mitigate bias based on user’s preferences. (Bellamy et al, 2019)

In fairness analysis the below terminology is used:

Favourable label	The label whose value corresponds to an outcome that provides an advantage to the recipient.
Protected attribute	An attribute that partitions a population into groups that have parity in terms of benefit received: e.g. race, gender... These are not universal but are application specific.
(Un)Privileged protected attribute	A value of a protected attribute indicating a group that has historically been at systematic (dis)advantage.
Group fairness	The goal of groups defined by protected attributes receiving similar treatments or outcomes.

⁷ <https://github.com/Trusted-AI/AIF360/blob/main/README.md>

Individual fairness	The goal of similar individuals receiving similar treatments or outcomes.
Fairness metric	Quantification of unwanted bias in training data or models.
Bias	In the context of fairness, we are concerned with unwanted bias that places privileged groups at a systematic advantage and unprivileged groups at a systematic disadvantage.
Bias mitigation algorithm	Procedure for reducing unwanted bias in training data or models.

Table 2: Terminology in Fairness Analysis (Bellamy et al, 2019)

For fairness metrics I used the following⁸:

Disparate impact: ratio of the rate of favourable outcomes for unprivileged group to the favourable outcomes for privileged group. The ideal value is 1 and if it is < 1 , it implies that privileged group receives higher benefit.

Statistical Parity Difference: the difference of the rate of favourable outcomes for unprivileged group to the favourable outcomes for privileged group. The ideal value is 0 and if it is < 0 , it implies that privileged group receives higher benefit.

Equal Opportunity Difference: this metric is computed as the difference of true positive rates between the unprivileged and the privileged groups. The ideal value is 0 and if it is < 0 , it implies that privileged group receives higher benefit.

In terms of bias mitigation algorithms, we could use pre-processing, in-processing, and post-processing. There are no one-size fits all solution so “the choice among algorithm categories can partially be made based on the user persona's ability to intervene at different parts of a machine learning pipeline. If the user is allowed to modify the training data, then pre-processing can be used. If the user is allowed to change the learning algorithm, then in-processing can be used. If the user can only treat the learned model as a black box without any ability to modify the training data or learning algorithm, then only post-processing can be used. AIF360 recommends the earliest mediation category in the pipeline that the user has permission

⁸ <https://medium.com/ibm-data-ai/fairness-explained-definitions-and-metrics-9690f8e0a4ea>

to apply because it gives the most flexibility and opportunity to correct bias as much as possible.” (Bellamy et al, 2019)

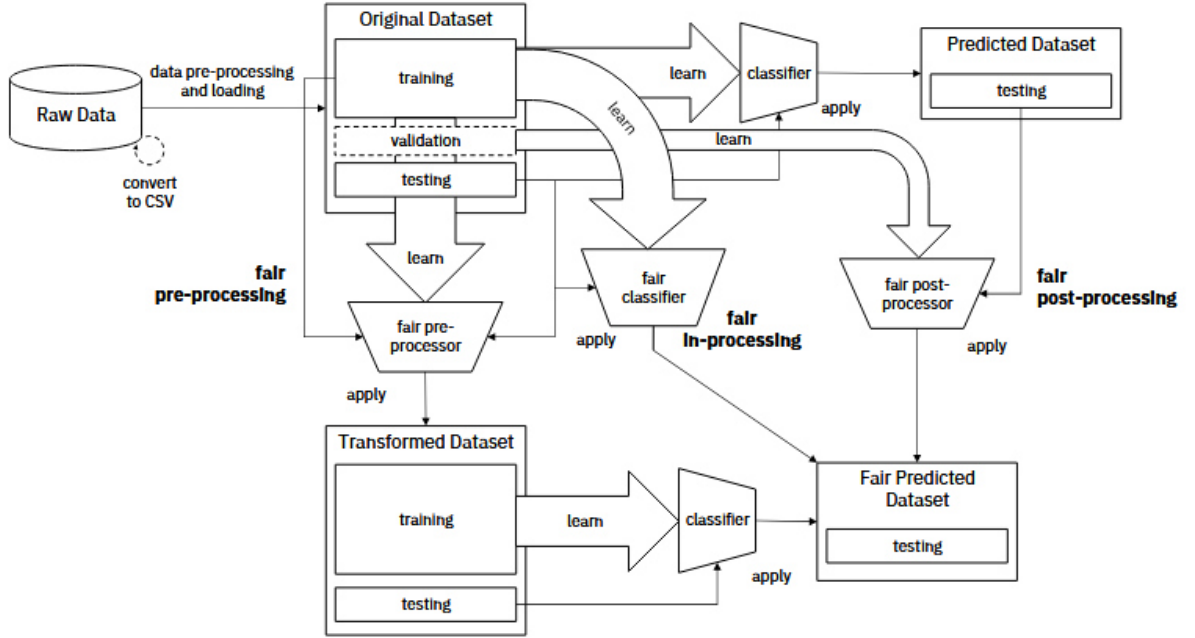


Figure 7 : Bias mitigation algorithms (Bellamy et al, 2019)

Based on the above recommendations, I chose a pre-processing algorithm, Reweighting so we could intervene in the training process as early as possible. This algorithm was implemented based on the research paper from Kamiran and Calders (Kamiran et al, 2011).

Let discrimination of classifier C be defined as the following:

$$disc_{S=b} := P(C(X) = + | X(S) = w) - P(C(X) = + | X(S) = b)$$

S : binary sensitive attribute where $S = b$ and $S = w$ are deprived and favoured communities

X : random unlabelled data object

C : class with domain $\{-, +\}$ where “+” is the desirable class

Reweighting algorithm does not change the label or features in the underlying data. Instead, it would assign weights to the training samples. These weights are calibrated so the training dataset could be discriminate free with respect to S . (Kamiran et al, 2011)

After preprocessing the data in our Taiwan Credit Card dataset using random undersampling, I investigated regarding potential bias in the dataset. Features “SEX” and “AGE” could be a good choice: Mean difference from the BinaryLabelDatasetMetric class has been calculated that “compare the percentage of favourable results for the privileged and unprivileged groups,

subtracting the former percentage from the latter. A negative value indicates less favourable outcomes for the unprivileged groups.” (IBM Credit Scoring Tutorial, 2024).

Given Microsoft Research already made a study on this dataset based on the feature “SEX” (Fairlearn Case Study, 2024), I investigated regarding the bias on feature “AGE”, so it was chosen as protected attribute. I wanted to check if there is any bias between applicants that are younger than 25 and older than 25 years old. I defined privileged group as applicants 25 years old or older and unprivileged group as younger than 25 years old.

I calculated the mean difference on original rebalanced dataset between these groups and found that it was -0.079157 which indicates bias in favour of the privileged group. I applied the above Reweighing algorithm on the dataset that decreased this number close to zero: -0.000000. The first simple measure indicates that we could expect improvements during reweighing. To assess that I trained the dataset before and after reweighing. I checked the impact on the traditional logistic regression and on the more advanced LightGBM algorithm.

We could also observe that after applying bias mitigation techniques, in most cases the balanced accuracy could decrease that shows a trade-off between fairness and balanced accuracy metrics (IBM Reweighing Tutorial, 2024). In these mitigation exercises we want to find the optimal classification threshold from the validation set that tries to maximize balanced accuracy while minimizing bias. On the below graphs we could see this based on the logistic regression model: we would like to maximize the blue line, while we would like to see the red one close to zero. After Reweighing this has been achieved so we decreased bias while balanced accuracy remained roughly the same:

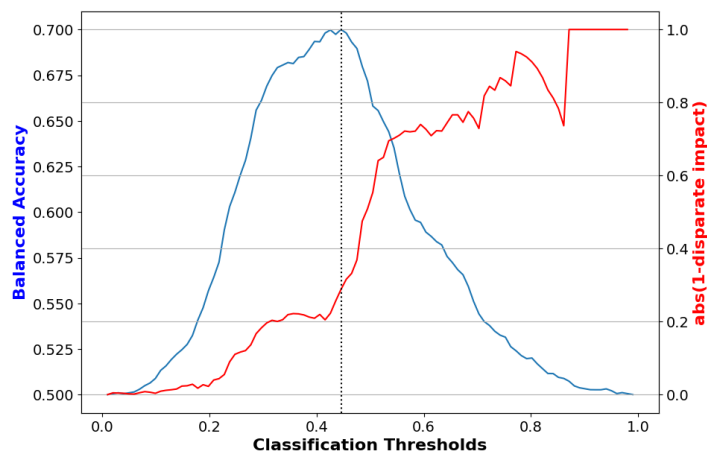


Figure 8: Logistic regression (before Reweighing)

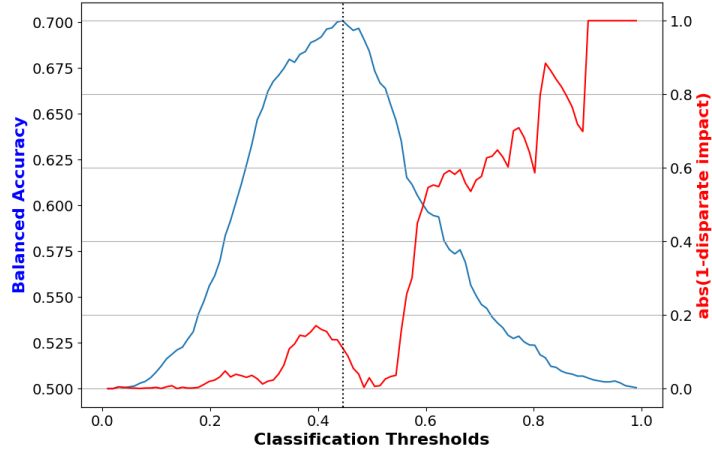


Figure 9: Logistic regression (after Reweighing)

Bias Mitigator Applied	Machine Learning Model	Balanced Accuracy	Statistical Parity Difference	Disparate Impact	Equal Opportunity Difference
No	Logistic Regression	0.7001	-0.1918	0.7119	-0.1613
Reweighting	Logistic Regression	0.7007	-0.073	0.8895	-0.0308
No	LightGBM	0.7068	-0.1258	0.7934	-0.0857
Reweighting	LightGBM	0.7067	-0.1296	0.7995	-0.0716

Table 3: Results on bias mitigation using Reweighing algorithm

Based on the results we could see that Reweighing decreased the bias on both algorithms. However, based on that dataset logistic regression shows a better performance in terms of bias mitigation. More investigations are needed to compare other bias mitigation techniques on the dataset to determine if we could provide better fairness results on the LightGBM model.

As discussed earlier in this paper these are just the first steps to understand AI-related bias. More research and analysis are required to fully understand the underlying characteristics of fairness since bias could also be a representation of mechanism in our society.

3.3.4. Explainability Analysis

To explain the results of the machine learning models used for the Taiwan credit card dataset, I chose a model-agnostic XAI model, SHAP (SHapley Additive exPlanations) to provide insight into model decision-making (Molnar, 2024).

Model-agnostic models have the advantage to be flexible since these are not model specific methods and could be applied to any machine learning models like tree-based or neural network models. From model-agnostic models, SHAP also has the main benefit that it could not only provide explanation locally on individual prediction but globally on the entire model as well. (Molnar, 2024)

SHAP is based on Shapley values, a method that comes from coalitional game theory where the “feature values of a data instance act as players in a coalition”. Shapley values estimate how much each feature contributes to the total prediction of the model: it is “the average marginal contribution of a feature value to the prediction in different coalitions and is not the difference in prediction when we would remove the feature from the model.” (Molnar, 2024)

In the context of our creditworthiness analysis, the features are the repayment history, limit balance, bill amount and other features we specified earlier.

Some SHAP methods, like KernelSHAP have the disadvantage of huge computing time, however this issue has been resolved for TreeSHAP models that compute in polynomial time instead of exponential.

However, TreeSHAP has the disadvantage that “it defines the value function using the conditional expectation instead of the marginal expectation. The problem with the conditional expectation is that features that have no influence on the prediction function can get a TreeSHAP estimate different from zero. The non-zero estimate can happen when the feature is correlated with another feature that has an influence on the prediction.” (Molnar, 2024)

For the analysis I used the SHAP open-source library⁹ that provides great visualization to highlight the results from the model.

⁹ <https://github.com/shap/shap>

SHAP feature importance:

SHAP feature importance plot shows the importance for each feature: “features with large Shapely values are important. For global importance we average the absolute Shapley values per feature across the data.” (Molnar, 2024)

For our analysis we calculated the feature importance plot for both LightGBM and logistic regression. For LightGBM TreeSHAP, and for logistic regression LinearExplainer has been used from SHAP library.

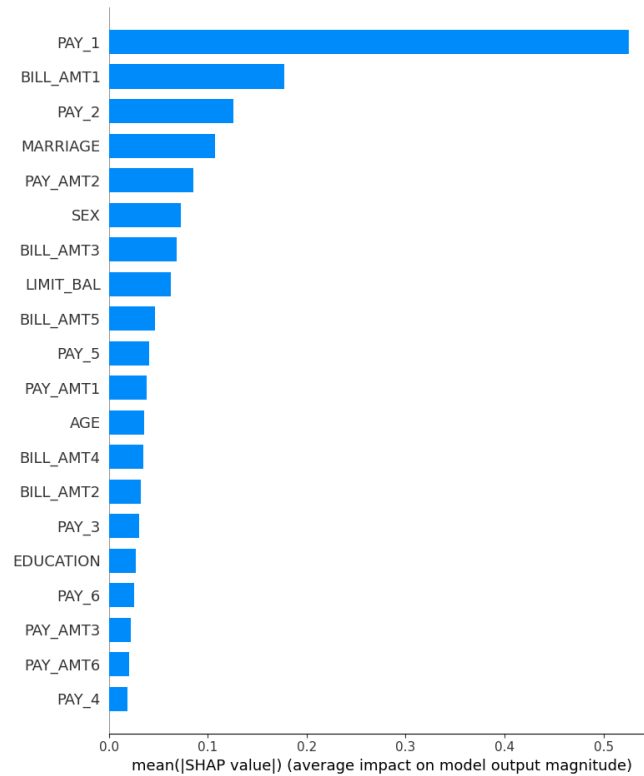


Figure 10: SHAP feature importance for logistic regression

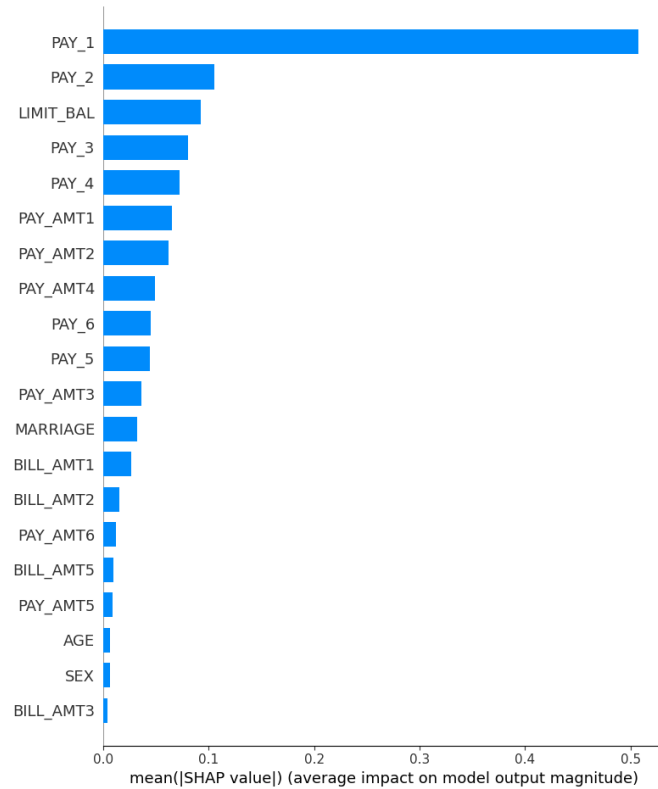


Figure 11: SHAP feature importance for LightGBM

We can see that on each model we have slight differences between the important features, however the current month's repayment status (PAY1) is the most significant feature based on both models.

SHAP Summary Plot:

In summary plot, features are also ordered based on their importance, however we could also see the distribution of Shapely values per feature. Each point is a Shapely value, and colours represent the feature value from low to high:

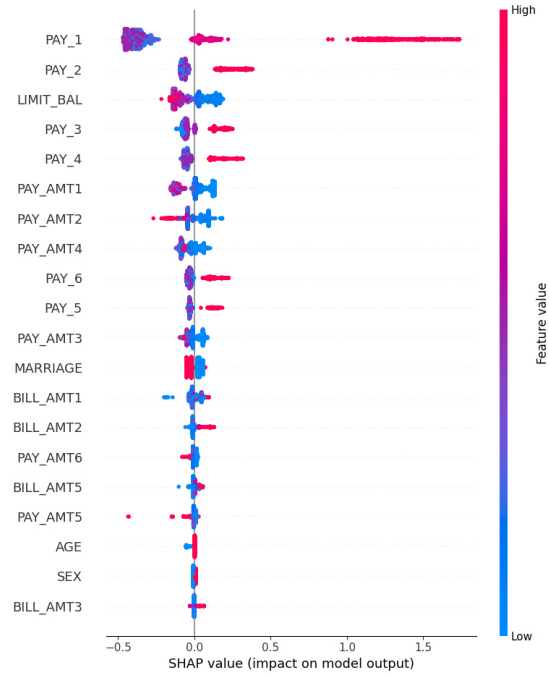


Figure 12: Summary plot for LightGBM

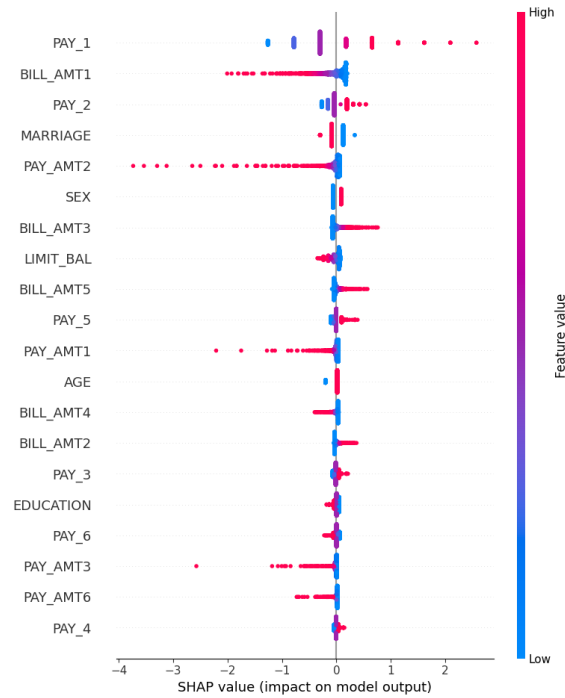


Figure 13: Summary plot for logistic regression

PAY1 shows higher value when a credit card customer has more delay to repay his/her loan (e.g. value of 2 means there is a 2 month-delay repaying the credit). This plot highlights that a customer with more repayment delay has a higher risk to declare default. It is also the most important feature of the model since it is on the top of the plot.

3.3.5. Environmental Impact

To access the environmental impact of our machine learning algorithms, I chose CodeCarbon¹⁰ tool that is an open-source library. In my analysis I calculated the emissions in CO₂-equivalents and the kWh energy used by these algorithms. I not only assessed the energy consumption of the model training phase but also computed the environmental impact of calculating SHAP values.

Model	Calculation Type	Duration of the Compute (seconds)	Emissions as CO ₂ -equivalents [CO ₂ eq] (kg)	Total Energy consumed (kWh)
Logistic Regression	Training	0.0894548	0.00000013936	0.00000068248
Logistic Regression	SHAP	0.1084471	0.00000017563	0.00000086014
LightGBM	Training	0.2117826	0.00000000405	0.00000170268
LightGBM	SHAP	2.550616	0.00000004954	0.00002083881

Table 4: Environmental Impact

The calculation has been performed on a Windows 10 operating system and on Intel(R) Core (TM) i5-4200H CPU @ 2.80GHz.

We can see that although the environmental impact is not that high, it is interesting to notice that calculating SHAP values for TreeSHAP consumes much more energy than training the underlying LightGBM model.

Further analysis must be taken to analyse the environmental impact of other layers (e.g. fairness) besides XAI in Sustainable AI to understand the possible interactions between these layers and the environment. That is also aligned with the concept of sustainable development which is based on the interactions of its pillars.

¹⁰ <https://mlco2.github.io/codecarbon/>

4. Conclusions

In my research we tried to demonstrate the gaps of the current responsible AI framework and proposed a sustainable AI approach that could highlight a new way to think about AI strategy and risk management.

To move forward in operationalizing sustainable AI, the proposed framework targeted a use case specific - a credit scoring specific - narrative since there is no one-size fits all solution when dealing with responsible AI strategies.

In the case study we measured the fairness, explainability and environmental impact of credit scoring algorithms in one research that had not been previously measured collectively in other research papers to the author's knowledge. We found that measuring fairness based on the current tools available depends on the actual use case and further studies required to assess the performance of each metric corresponding to the underlying machine learning model. Currently we could see that Reweighting algorithm works well with logistic regression. Explainability and environmental calculations have also been performed showing the importance of these layers. Further enhancement of this study could be to include privacy and security of AI systems in the analysis to get more insight on the proposed framework and additional investigation could be made to assess the interaction between the layers of Sustainable AI.

Hopefully with this research we are getting one step closer to a more sustainable AI that fosters trust in AI systems which is the foundation of the success in the long run.

5. References

Research Papers:

Bellamy, R. K. E., Dey K., Hind, M., Hoffman, S.C., Houde, S., Kannan, K., Lohia, P., Martino, J., Mehta, S., Mojsilovic, A., Nagar, S., Ramamurthy, K.N., Richards, J., Saha, D., Sattigeri, P., Singh, M., Varshney, K.R., Zhang, Y. AI Fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias *IBM Journal of Research and Development*, vol. 63, no. 4/5, pp. 4:1-4:15, 1 July-Sept. 2019, (2019) doi: 10.1147/JRD.2019.2942287.

Díaz-Rodríguez, N., Del Ser, J., Coeckelbergh, M., López de Prado, M., Herrera-Viedma, E., Herrera, F. Connecting the dots in trustworthy Artificial Intelligence: From AI principles ethics, and key requirements to responsible AI systems and regulation (2023); *Information fusion* (1872-6305), 99 , p. 101896. <https://doi.org/10.1016/j.inffus.2023.101896> (p 1-24)

Kamiran, F., Calders, T. Data preprocessing techniques for classification without discrimination *Knowledge and information systems* (0219-3116), 33 (1), p. 1. (2011) (p.1-33)

Krizhevsky, A., Sutskever, I., and Hinton, G. E. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25,(2012). (p 1-9)

Laura Blattner, Scott Nelson: Data and Disparities in Consumer Credit (2021) <https://arxiv.org/pdf/2105.07554> (p 1-86)

Li, Y., Chen, W. A Comparative Performance Assessment of Ensemble Learning for Credit Scoring (2020); *Mathematics (Basel)* (2227-7390), 8 (10), p. 1756. <https://doi.org/10.3390/math8101756> (p 1-19)

Luccioni, S. "Light bulbs have energy ratings — so why can't AI chatbots?". *Nature (London)* (2024) (1476-4687), 632 (8026), p. 736. <https://www.nature.com/articles/d41586-024-02680-3> (p. 736-738)

Sadek, M., Kallina, E., Bohné, T., Mougenot, C., Calvo, RA., Cave, S. Challenges of responsible AI in practice: scoping review and recommended actions *AI & Soc* (2024). <https://doi.org/10.1007/s00146-024-01880-9> (p 1-17)

Sagi O, Rokach L. Ensemble learning: A survey. *WIREs Data Mining Knowledge Discovery* (2018); 8:e1249. <https://doi.org/10.1002/widm.1249> (p 1-18)

van Wynsberghe, A.: Sustainable AI: AI for sustainability and the sustainability of AI. *AI Ethics* 1, 213–218 (2021). <https://doi.org/10.1007/s43681-021-00043-6> (p 1-6)

Varoquaux, G., Luccioni, A., Whittaker, M.: Hype, Sustainability, and the Price of the Bigger-is-Better Paradigm in AI, (2024) <https://doi.org/10.48550/arXiv.2409.14160> (p. 1-19)

Conference Paper:

Ke, G., Meng, Q., Finley, F., Wang, T., Chen, W., Ma, W., Ye, Q., Liu, T-Y. LightGBM: A Highly Efficient Gradient Boosting Decision Tree (2017)
31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA

Webpages:

Microsoft: *Assessing and mitigating unfairness in credit models with the Fairlearn toolkit* (2020) https://www.microsoft.com/en-us/research/uploads/prod/2020/09/Fairlearn-EY_WhitePaper-2020-09-22.pdf [Downloaded: October 2024]

UN sustainability: *Sustainability*: <https://www.un.org/en/academic-impact/sustainability> [Downloaded: October 2024]

Stakeholder: https://link.springer.com/referenceworkentry/10.1057/978-1-137-00772-8_270 [Downloaded: October 2024]

Sutton, R. The bitter lesson. Incomplete Ideas (blog), 13(1), (2019)
https://www.cs.utexas.edu/~eunsol/courses/data/bitter_lesson.pdf [Downloaded: October 2024]

Fairness Metrics: <https://medium.com/ibm-data-ai/fairness-explained-definitions-and-metrics-9690f8e0a4ea> [Downloaded: October 2024]

UCI Dataset: *Default of Credit Card Clients*:
<https://archive.ics.uci.edu/dataset/350/default+of+credit+card+clients> [Downloaded: October 2024]

FICO 1: *How-to-build-credit*: <https://www.myfico.com/credit-education/credit-scores/how-to-build-credit> [Downloaded: October 2024]

FICO 2: *What is in your credit score*: <https://www.myfico.com/credit-education/whats-in-your-credit-score> [Downloaded: October 2024]

XGBoost Documentation: <https://xgboost.readthedocs.io/en/stable/tutorials/model.html>
[Downloaded: October 2024]

LightGBM documentation:
<https://lightgbm.readthedocs.io/en/stable/Features.html#references> [Downloaded: October 2024]

EU AI Act: *Regulatory Framework*: <https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai> [Downloaded: October 2024]

EU AI Act Summary: *High-Level Summary* <https://artificialintelligenceact.eu/high-level-summary/> [Downloaded: October 2024]

IBM Responsible AI: <https://www.ibm.com/topics/responsible-ai> [Downloaded: October 2024]

Google RAI: <https://ai.google/responsibility/responsible-ai-practices/> [Downloaded: October 2024]

IBM HCAI: <https://research.ibm.com/blog/what-is-human-centered-ai> [Downloaded: October 2024]

FICO White Paper: <https://www.fico.com/en/latest-thinking/white-paper/machine-learning-and-fico-scores> [Downloaded: October 2024]

Fairlearn Case Study:

https://fairlearn.org/main/auto_examples/plot_credit_loan_decisions.html#sphx-glr-auto-examples-plot-credit-loan-decisions-py [Downloaded: October 2024]

Apple credit bias: <https://edition.cnn.com/2019/11/12/business/apple-card-gender-bias/index.html> [Downloaded: October 2024]

Equal Credit Opportunity Act: <https://www.govinfo.gov/content/pkg/USCODE-2011-title15/html/USCODE-2011-title15-chap41-subchapIV.htm> [Downloaded: October 2024]

Inequality in Credit: <https://hai.stanford.edu/news/how-flawed-data-aggravates-inequality-credit> [Downloaded: October 2024]

Podcast with Max Levchin: <https://open.spotify.com/episode/6n6crk7UZEMfcX2oPPPYKW> [Downloaded: October 2024]

Affirm: <https://www.affirm.com/how-it-works/why-affirm> [Downloaded: October 2024]

IBM Reweighting Tutorial: https://github.com/Trusted-AI/AIF360/blob/main/examples/demo_reweighting_preproc.ipynb [Downloaded: October 2024]

Molnar, C. Interpretable Machine Learning <https://christophm.github.io/interpretable-ml-book/> [Downloaded: October 2024]

FICO Alternative Data: <https://www.fico.com/blogs/how-use-alternative-data-credit-risk-analytics> [Downloaded: October 2024]

IBM Credit Scoring Tutorial: https://github.com/Trusted-AI/AIF360/blob/main/examples/tutorial_credit_scoring.ipynb [Downloaded: October 2024]

6. Tools and Libraries Used

- Python version 3.11
- Imbalanced Learn: Random UnderSampler https://imbalanced-learn.org/stable/references/generated/imblearn.under_sampling.RandomUnderSampler.html
- Sklearn: StandardScaler, LogisticRegression libraries
<https://scikit-learn.org/dev/modules/generated/sklearn.preprocessing.StandardScaler.html>
https://scikit-learn.org/1.5/modules/generated/sklearn.linear_model.LogisticRegression.html
- LightGBM library: <https://lightgbm.readthedocs.io/en/latest/Python-Intro.html>
- IBM AI Fairness 360:
<https://github.com/Trusted-AI/AIF360/blob/main/README.md>
<https://aif360.res.ibm.com/>
- IBM toolkit tutorials used to generate codes on AI Fairness analysis:
https://github.com/Trusted-AI/AIF360/blob/main/examples/tutorial_credit_scoring.ipynb
https://github.com/Trusted-AI/AIF360/blob/main/examples/demo_reweighting_preproc.ipynb

- SHAP library: <https://github.com/shap/shap>
- CodeCarbon: <https://mlco2.github.io/codecarbon/>
- ChatGPT: for educational purposes to understand and generate portions of codes that have been checked by official programming library documentations

7. Author's Own Contribution

In my research we introduced a framework that shifts from responsible AI to a sustainable AI approach: a new layer, environment will be added, and we will see sustainable AI as a dynamic system based on the interaction of its layers. We also used Edward Freeman's stakeholder theory to connect the dots between human-centred AI and sustainable AI approach that could highlight how we as industry practitioners could look at the problem from a different angle.

To demonstrate the above theory, we used credit scoring machine learning models as a case study and focused on the fairness, explainability and environmental impact of these algorithms in one analysis that was mainly investigated separately in other research papers.

To the author's knowledge, no study has been conducted to date to analyse these three impacts on credit scoring models in one research.

8. Acknowledgements

I would like to thank my supervisor, Laszlo Szathmary for all his guidance and advice on writing this research, whose feedback and insights were very valuable for me. I also would like to thank Balazs Harangi, Henrietta Toman and Robert Lakatos for their advice and contribution to this paper.