ISSN: 3033-4136
DOI: https://doi.org/10.17868/strath.00089720

Promoting Fairness and Exploring Algorithmic Discrimination in Financial Decision Making Through Explainable Artificial Intelligence

FRIL
Financial Regulation Innovation Lab

FinTech Scotland®

University of Glasgow

University of Strathclyde Glasgow

Financial Regulation Innovation Lab

Who are we?

The Financial Regulation Innovation Lab (FRIL) is an industry-led collaborative research and innovation programme focused on leveraging new technologies to respond to, shape, and help evolve the future regulatory landscape in the UK and globally, helping to create new employment and business opportunities, and enabling the future talent.

FRIL provides an environment for participants to engage and collaborate on the dynamic demands of financial regulation, explore, test and experiment with new technologies, build confidence in solutions and demonstrate their ability to meet regulatory standards worldwide.

What is Actionable Research?

FRIL will integrate academic research with an industry relevant agenda, focused on enabling knowledge on cutting-edge topics such as generative and explainable AI, advanced analytics, advanced computing, and earth-intelligent data as applied to financial regulation. The approach fosters cross sector learning to produce a series of papers, actionable recommendations and strategic plans that can be tested in the innovation environment, in collaboration across industry and regulators.

Locally-led Innovation Accelerators delivered in partnership with DSIT, Innovate UK and City Regions

Department for Science, Innovation & Technology

UKRI Innovate UK

GLASGOW CITY REGION

FRIL White Paper Series

Promoting fairness and exploring algorithmic discrimination in financial decision making through explainable artificial intelligence

Kushagra Jain* James Bowden* Mark Cummins*

* University of Strathclyde

June 2024

Abstract: In this white paper a comprehensive toolbox is developed, grounded in an ethical "rights to explanation" framework, deploying state-of-the-art machine learning/artificial intelligence models, through the lens of explainability. Harnessing these explainable artificial intelligence algorithms within the toolbox, we propose implementing an ensemble of model-agnostic techniques, to improve fairness in financial decision making, with a particular focus on US home mortgage loan applications with a granular public dataset. We also highlight variability in these techniques, imposing various pragmatic scenarios that explore real-world decision making, alongside equality of opportunity and equality of outcome conditions. We highlight potential pitfalls, nuances, and possible innovations in applying these techniques, while providing the ability to simultaneously assess the impact of any specific variable in decision making, and a model's performance in such decision making, with established machine learning criteria. Furthermore, we showcase the trade-off between fairness and model performance optimization with a protected characteristic (age) that might form the basis of plausibly discriminatory practices in such a context. Our study aims to be in the spirit of Agarwal, Muckley, & Neelakantan (2023), Kelley, Ovchinnikov, Hardoon, & Heinrich, (2022), Kozodoi, Jacob, & Lessmann (2022), and Kim & Routledge (2022), among others. We lastly identify areas for future research.

Table of Contents

1. Problem Statement

The discipline of business ethics is overlooking novel harms and marginalized stakeholders in emerging and impactful technology industries (Martin, 2023). Large industries exist whose business models use marginalized stakeholders' data against them, acting in ways that are demeaning and objectionably exploitative (Martin, 2023). Businesses increasingly utilize proprietary algorithms[1] that are data-trained sets of decision rules (i.e., the output of processes that are often "machine learning" based) and implement decisions with little or no human intervention that have significant impacts on humans (Kim & Routledge, 2022). Algorithms that make and support such decisions are a prevalent and growing component of economic life (Kim & Routledge, 2022). Uncritical assumptions are made concerning their efficiency and accuracy, and important critical examination of the type of "progress" being made is lacking (Martin, 2023). In algorithmic contexts, morally objectionable errors can occur non-negligently in an unpredictable manner. For instance, "algorithms could exhibit [discriminatory] tendencies even if they have not been manually programmed to do so, whether on purpose or by accident." (Kim & Routledge, 2022).

The problem that emerges is that stakeholders have a right to explanation around such decision making. This is the problem that we tackle in this study. Using the above motivation, we outline the rationale behind these rights (expounded in further detail later in the paper) and we illustrate how to fulfil such rights. An application exploring US home mortgage loan decisions is showcased. It demonstrates how explainable artificial intelligence algorithms can deliver these rights to explanation to consumers. Our application sheds light on real-world decision making. Though there is no way of knowing how data is used by borrowers, our toolbox allows assessment of the underlying drivers of such decision making. This is regardless of whether decisions are made by humans or algorithms.[2]

We further explore scenarios imposing equality of opportunity and equality of outcome within our framework. The approaches we employ test various scenarios for adherence to norms of fairness codified in ethical rights. We illustrate the bias-performance trade-off inherent in this setting, and associated caveats.[3] The explanations elicited from the toolbox can promote greater transparency and fairness in such decision making, but beneficially are agnostic to the model, data or setting employed.[4] Finally, we provide the ability to facilitate a comparison between models and the approaches/techniques used for investigating the fairness of decision making from interrogation of the data. We lastly identify future research avenues that we will explore in subsequent white papers.

2. Literature Review

The problem is best illustrated with the use of a salient example from Kim & Routledge (2022) in the context of the use of algorithms, artificial intelligence, and machine learning models for decision making by financial technology firms, and traditional financial institutions, such as banks.

David H. Hansson and his wife, Jamie H. Hansson, applied for Apple Cards — developed in partnership with Goldman Sachs — when

1 An algorithm is a set of rules and procedures that leads to a decision. Businesses have been using algorithms for a long time. However, algorithms that have their roots in data-driven artificial intelligence/ machine learning that results in decisions being implemented with no (or little) human intermediation, are relatively new (Kim & Routledge, 2022).

2 In other words, real-world empirical data is scrutinized within an explainable artificial intelligence framework. It allows one to assess if lending practices are plausibly discriminatory. Even if all the decision making is manual, the algorithm still sheds light on the underlying drivers of the decisions being made.

3 By performance, we mean the accuracy of the models in making predictions in the test data, and by bias we mean the plausible discrimination spoken of hitherto.

4 In other words, our open-source framework is malleable and readily adopted to other decision-making contexts. Binary or continuous variables can be used in an out-of-sample setting. Any models compatible with the Shapley framework can be deployed. Indeed, many of other tools developed for such a purpose (edified upon in the literature review) may also be incorporated into the fold of the toolbox as well.

launched in August 2019. The husband received a credit limit that was twenty times higher than his wife's, even though they file joint tax returns, and her credit score was higher than his. When the applicant contacted Apple's customer service department, a representative blamed the result on its black box algorithm, which automatically decides such issues as credit limits. Subsequently, Goldman Sachs shared a statement stating, "We have not and will not make decisions based on factors like gender." But it is well known that machine learning–derived algorithms can discriminate based on gender or race without using such data as classifiers (Kim & Routledge, 2022).

Do Apple and Goldman Sachs have an obligation to provide a meaningful explanation to Jamie H. Hansson? The Equal Credit Opportunity Act — with philosophical foundations based on equal treatment and fairness — already demands that financial firms provide decision rationales to customers in the United States. Equal treatment is an important moral value upholding a right to explanation for cases like credit card limits or approval of loan applications (Kim & Routledge, 2022).[5]

This right to explanation is thereafter reconceptualized by Kim & Routledge (2022) as seen in Table 1.[6]

Table 1. Contours of a Right to Explanation

	Ethicists' view (philosophical)	Data scientists' view (operational)
Ex ante	The traditional notion of informed consent (or a right to ex ante explanation)	Offering an ex ante generic explanation
Ex post	A right to remidial explanation (or ex post explanation for redress)	Offering both an ex post generic explanation and an ex post specific explanation
	A right to updating explanation (or an ex post explanation for opt-out)	Offering both an ex post generic explanation and an ex post specific explanation

From an ethical point of view, and related to ex-post rights, Kim & Routledge (2022) state generally that when a company harms (or wrongs) a person by its use of an automated algorithmic system, the harmed (or wronged) party with a right to an ex-post explanation (both generic and specific) is entitled to demand the company explain what happened — and why — in an intelligible manner. To be specific, those who are not harmed or wronged do not have this right. Kim & Routledge (2022) call this a "right to remedial explanation" (see Table 1).

The second kind of a right to an ex-post explanation (both generic and specific) is a right that data subjects can make legitimate claims without harm (or wrong) being done to them as a result. Consumers need to know whether they can continue trusting companies, and thus a right to updated explanations should exist without suffering harms or wrongs. This is the "right to an updating explanation" (see Table 1). This right exists even after a decision is made using the consumer's data whether harm or wrong has been done or not. This is as the consumer's data might continue to be used in a way that can harm or wrong them. When consumers give companies data for decision-making, their consent may not extend to a continuing (possibly never-ending) process. Thus, the quality of informed consent provided by consumers matters in algorithmic contexts, giving rise to the right to updating explanation (Kim & Routledge, 2022).

Table 1 maps the above ethical view to the data science view. In this context, there can be three different kinds of possible explanations with respect to algorithmic decisions: 1) an ex-ante explanation about system functionality (or an ex-ante generic explanation), 2) an ex-post explanation about system functionality (or an

5 This example can also be used to illustrate that discrimination can occur not only at an individual level, but also at a group or subgroup level. This controversy triggered a now pending government investigation where several other viral anecdotal accounts from Twitter also indicated that Goldman Sachs' credit limit policy discriminated against women. Reviewing the details of the claims and Goldman Sachs' response, one can speculate that the policy may have penalized homemakers relative to their working partners by virtue of prioritizing personal income and employment status while ignoring household-level financials. Most homemakers in the United States today are female (Weber, Yurochkin, Botros, & Markov, 2020). Thus, a significant subgroup of women may indeed have received disparate treatment, even while all women as a protected group may have been treated with statistical parity to men, which one may safely assume since the policy passed Goldman Sachs' model risk assessments before release. This is illustrative of the argument that it is not only the whole group that is exposed to discriminatory risk, but sufficiently large subgroups (e.g., homemakers) within a protected meta-group (women) (Weber, Yurochkin, Botros, & Markov, 2020).

6 It is reasonable for data subjects to expect companies to assure them up front that, if harms or wrongs occur, the company will respond in a fair and responsible manner. It can be argued that if the right is to remediation, then there need not be an explanation.

ex-post generic explanation), or 3) an ex-post explanation about a specific decision (or an ex-post specific explanation).[7] Ex-ante generic explanation is a technical name for the traditional understanding of a right to be informed (Kim & Routledge, 2022).

An ex-post generic explanation differs from an ex-ante generic explanation - even if both are about system functionality - because during training or processing, logic can change. Thus, ex-post generic explanations add value in such cases. However, this alone may not be enough to meaningfully satisfy the right to explanation. For instance, in the Apple Card application case, the company used a black box system to decide credit limit and, in response to an applicant seemingly disfavoured because of her gender, offered a generic ex-post explanation about how its decision process generally worked for all applicants (i.e., "The black box algorithm made a decision, and gender was not used as a factor"). If the applicant has a right to an ex-post explanation, the company should offer a meaningful and intelligible explanation (*both generic and specific*) about why and how the algorithmic system created a disparate impact upon the applicant, including specific features used in the data processing (Kim & Routledge, 2022).[8]

2.1 Explainable AI

A growing number of researchers are attempting to develop explainable AI (XAI) systems in response to the concerns raised earlier, which may help implement the outlined rights to explanation. But there are still problems to overcome. Different researchers have different ideas about the term explanation, so it is not yet clear how to objectively know which form of XAI is good or better/worse than others for a specific domain. To answer this, some form of "goodness" criteria are needed. But there is a lack of literature about which form of explanation (e.g., global, local, counterfactual) is best and how much information is suitable for human data subjects. Thus, researchers need to attempt to theoretically and empirically develop goodness criteria for the practical use of AI. A core research problem is to understand the features that make for a beneficial explanation of an AI system. This can refer to the output features of a machine learning algorithm, textual explanation, or a written explanation of certain algorithmic outputs. The answer should come up with a philosophical, theoretical definition and a framework of good explanations (e.g., objective understanding). Researchers should work to operationalize the notion of a "good explanation" in various contexts. For example, in the context of textual explanation, researchers may be given several different descriptions of a certain concept to share with data subjects (Kim & Routledge, 2022).

The generic criteria discussed by Kim & Routledge (2022) are not sufficient to address the difficulties associated with the complexities of explaining algorithms. The objective is to show that companies should not only develop explainable AI, but also seriously study what types of explanation are useful for users. Different stakeholders may need different types of explanation. Users may need simple or complex ones (depending on context). It is difficult to answer all these questions without further studying the criteria - theoretically and empirically.

Do companies have an incentive to develop XAI? Interestingly, the drive to explain AI models is not inconsistent with performance. Better understanding guards against overfitting and facilitates fine-tuning. It is also worth noting that advances in techniques to explain nonlinear models have followed their empirical success. Presumably, had the models not been useful, there would have been little effort to understand them. Choosing an algorithm that has poorer predictive performance — but is more easily explained – may also be a rational choice (Kim & Routledge, 2022).

Considering the above concerns, several advancements have been made in XAI, and so-

7 Computer scientists often use the term global instead of general and the term local instead of specific.

8 We provide greater detail on Kim & Routledge (2022)'s trust-based framework outlining a decision algorithm's process and the different types of stakeholders and their rights in the appendices.

called fairness techniques which attempt to solve these problems and improve fairness. A myriad of such techniques exist that employ several different approaches to strike a balance between optimizing fairness, accuracy and explainability, depending on the definition of fairness in question. For instance Weber, Yurochkin, Botros, & Markov (2020) highlight techniques that address subgroup discrimination.[9] Similarly, Castelnovo, et al. (2020) implement a toolkit called BeFair that employs a combination of existing approaches including AIF360 (Bellamy, et al., 2018), Fairlearn (Bird, et al., 2020), Causal Discovery Toolbox and CausalNex (Beaumont, et al., 2021). Fairlearn is also deployed by Dudik, et al. (2020) to reduce credit/loan outcome disparity based on gender from 8 to 1 percentage point without any (statistically significant) impact on the cost to the financial services organization.

Blattner, Stark, & Spiess, (2022) list a combination of proprietary and open-source tools they employ in their study, including SHAP and LIME. Kusner, Loftus, Russell, & Silva (2017) propose a counterfactual based definition of fairness which they implement here. Karimi, Khan, Liu, Derr, & Liu, (2022) propose to enhance individual fairness through propensity score matching. Kozodoi, Jacob, & Lessmann (2022) empirically implement multiple fairness techniques and evaluate them using the Fair Credit Scoring toolbox. Wan, Zha, Liu, & Zou (2023) review the progress of in-process fairness techniques. Finally, Chen, Giudici, Liu, & Raffinetti (2022) propose a general methodology framework for explainable credit scoring to provide interpretability of each individual variable and measure fairness. It can detect important variables and quantifies their individual impact on a firm's credit classification via the Shapley-Lorenz metric; and it quantifies the degree of discrimination, conditional on the endogenous effects generated by the variables, via the Kolmogorov-Smirnov test.

3. Solution Framework

The solution framework we propose for the issues outlined thus far is rooted in the ethical framework that motivated us to study this research question (Kim & Routledge, 2022), capable of fulfilling each of the rights to explanation defined therein. Furthermore, we build upon this foundation by employing a framework that is model, data and setting-agnostic in every aspect. This includes the preprocessing/hybrid over-under sampling and hyperparameter tuning we utilize for achieving fairness in the spirit, and augmenting the approaches, of both Agarwal, Muckley, & Neelakantan (2023) and Kelley, Ovchinnikov, Hardoon, & Heinrich (2022). This can be envisioned as creating a training dataset which creates *equality of opportunity* for loan decisions, *equality of outcome* between protected and unprotected classes of a protected characteristic (e.g. gender), or both. Further, it enables the use of any state-of-the-art, open-source and well-known artificial intelligence/machine learning algorithms, with explainability delivered via Shapley values (at both a global and local level, although we focus on the global level of explainability herein, as we are interested in overall rather than individual outcomes). Finally, it allows for assessment in terms of model performance/operational optimization using well known machine learning criteria, derived from a confusion matrix, or the area under the curve.

Although our demonstration is centred on binary predictions of loan outcomes with binary predictors, it is easily extended with minor modifications to a continuous/multivariate predictive setting using continuous/multivariate predictors. Our approach enables one to elicit Shapley values that indicate the contribution of each predictor to the decision variable (loan decision) for a given US state in a year in the test dataset, based on a model trained on optimized hyperparameters for the training data using

9 See the SenSR and EXPLORE Fair metric learning toolboxes.

Stratified K-fold Cross Validation and Random/Grid Search.[10]

From an industry practitioner's perspective, we also provide a way to compare the resulting distributions of Shapley values (in our case for states in a year but say for a bank this could be for branches in a year, or any other context). We utilise two distinct test statistics that allow such a comparison of the Shapley value distributions for single predictor variables and multiple variables at the same time (albeit with modifications for one of the test statistics in the multivariate case). Shapley values allow industry professionals to assess the magnitude of impact a particular variable (or indeed variables) is having in their decision making and relative to other predictor variables.

Our focus in this context is on protected characteristics that can promote fairer practices in decision making and the role protected characteristics play in such decision making. Further introducing these test statistics adds another comparative dimension to our toolbox. Specifically, they enable practitioners to compare (pairwise) the Shapley distributions of a variable or variables that are generated by different cases or models. This adds practical utility because by interpreting these test statistics, one may explicitly assess if one model has Shapley values that are lower than the other model. Again, our emphasis is on protected characteristics, this allows a professional to see if one model is fairer than the other, and to what degree. Applying this iteratively, one can rank cases or models in order of how fair they are, adding a layer of abstraction to compare not only the importance of such a variable relative to others, but across distinct cases or models across multiple panel instances. The reason for using two different test statistics is to impose different thresholds of differences in distributions, as in some cases distributions may differ, but this may not be identified with the first test statistic, as the condition it tests for is very strong. Such differences in turn will be identified with the second one, which although imposing a strong condition for distinct distributions has a lower threshold than the first. We provide the technical detail for these in a footnote for interested readers.[11]

Finally, we touch upon potential pitfalls in deploying the framework, such as preprocessing on just the protected characteristic or applying preprocessing before the training and test split, both of which are problematic in this setting, creating issues such as possibly classifying all outcomes into one category or the other and biasing the test data by introducing a look-ahead bias from the in-sample training data respectively. We also note the preprocessing that we utilize is applicable iteratively to further balance other characteristics as well, albeit at the cost of creating imbalances in the characteristics balanced earlier that grows with each iteration and can have consequences like making the protected minority class the majority in the sample, with higher imbalance and lesser equality of outcome with each iteration.

The preprocessing technique we apply has a simple intuition. To be fairer in decision-making, Shapley values can only highlight a protected characteristic's role in decision making if it has enough instances to compare of 1) loan acceptances and rejections and 2) minority and majority category applications. These are the conditions we term equality of outcome (1) and equality of opportunity (2). Given the imbalance between categories otherwise present in real-world data, it would not be possible to promote fairness because the minority cases are so few, the algorithm will be unable to promote fairer outcomes.

10 Indeed, the bias-performance trade-off can be studied explicitly in a different setting by tuning these hyperparameters across the range iterated over in the Random/Grid Search, storing the relevant bias and performance metrics for each of those models for comparison. However, as this becomes case-specific rather than model/data/setting-agnostic, i.e., each case, model and data would yield different optimal hyperparameters, we do not explore this separately.

11 We also propose and implement a robust first- and restricted second-order stochastic dominance approach to identify distinctions between Shapley values and model performance criteria, readily applicable within other contexts, building on the contribution of Chen, Giudici, Liu, & Raffinetti (2022). However, we use stochastic dominance for facilitating comparisons between Shapley value distributions of differently specified models/cases, rather than against a standard uniform distribution. This framework is readily scaled to compare distributions of Shapley values and model performance criteria in a multivariate context (multiple sets of Shapley values distributions).

Likewise, a similar logic exists for loan decisions, and without correcting for the applicant categorical or class imbalance, the algorithm cannot discern with relatively better precision or accuracy what drives the decisions for the class with sparse observations. By deploying hybrid over-under sampling, this can be accounted for while preserving the underlying statistical structure and properties of the sample to remain representative of the actual data.

The above logic is also backed by empirical evidence: both Agarwal, Muckley, & Neelakantan (2023) and Kelley, Ovchinnikov, Hardoon, & Heinrich (2022) approach their studies in similar spirits, and we assimilate both their considerations accordingly into our toolbox. Furthermore, the validity of our approach is corroborated in the different cases we test in the use case and appendices: with or without rebalancing the test data, the inferences we draw are the same i.e. in order to elicit a fairer impact of a protected characteristic on loan decision making, accounting for its class imbalance is neccesary over and above the decision making class imbalance.

4. Use Case Demonstration

The data employed for the use case is from the publicly available US data disclosed on the Federal Financial Institutions Examination Council (FFIEC) and Consumer Financial Protection Bureau (CFPB)'s website for the Home Mortgage Disclosure Act (HMDA). As specified on their website, this is the most comprehensive source of publicly available information on the U.S. mortgage market. The HMDA requires many financial institutions to maintain, report, and publicly disclose loan-level information about mortgages. These data help show whether lenders are serving the housing needs of their communities; they give public officials information that helps them make decisions and policies; and they shed light on lending patterns that could be discriminatory.

The public data are modified to protect applicant and borrower privacy and available for the time period 2000-2022 at the time of writing. HMDA was originally enacted by Congress in 1975 and is implemented by Regulation C. It captures the bulk of residential mortgage lending activity in the United States (Cortés & Strahan, 2017), and has been used in several studies and contexts for mortgages (Dlugosz, Gam, Gopalan, & Skrastins, 2023).

The take-away message we hope to propound for the industry with this use case is a demonstration that this toolbox will enable the identification of any particular variable's contribution to loan decision making and achieve fairer outcomes in practice and provide a comprehensive framework where this can be redressed in a rights-based sense. This is done while retaining cognizance of business pragmatism by providing performance measures to allow practitioners to tune the bias-performance trade-off to the desired level in a bespoke and tailored way (reducing bias as much as possible without miscategorising excessively to ensure operational risk is not exacerbated by approving future defaulters and desired loan performance levels are retained).

We apply a simplified version of the solution framework for the loan decision making problem defined above (i.e. the loan decision – we use the variable *Loan Decision*, where we assume a value of 1 if the loan is rejected and 0 if the loan is accepted). Thus, for brevity and for the sake of a "human-friendly" explanation of the framework, we focus on a single US state (Mississippi) in a single year (2018), with a single type of XAI model augmented with a single plausible source of discrimination (old age – we use the variable *Old*, which takes a value of 1 if the applicant's age is above 62 and 0 otherwise) alongside variables established in the literature (Agarwal, Muckley, & Neelakantan, 2023). We apply preprocessing in the training data to balance:

1. the outcome variable (i.e. *Loan Decision*)

2. both the protected characteristic (i.e. *Old*) and the outcome variable and use it on the unbalanced test data (i.e. the test data subsample from the original data split into test

and training subsets, without any rebalancing to correct for the class imbalance illustrated earlier) to reflect the performance and fairness of the model in a real-world pragmatic scenario.

We refrain from additionally applying hyperparameter tuning using Random/Grid Search complemented with Stratified K-Fold Cross Validation on the training data and opt for default parameters for the same reason, alongside computation time considerations. Using Grid Search in this fashion across the systematically selected broadest range of parameters can be very computationally expensive. This also simultaneously helps us avoid overfitting concerns on the training data accidentally, although we are cognizant of overfitting risks and consider them when selecting our hyperparameter ranges. Finally, we demonstrate two different cases and compare the Shapley values of the protected characteristics, and the performance metrics in terms of classification accuracy across the cases: as this is a single state-year, the comparison can be facilitated directly without any need for the stochastic dominance framework outlined. The cases are:

i. Preprocessing is applied only on the outcome variable.

ii. Preprocessing is applied on both the outcome variable and protected characteristic.

As can be seen from the Shapley value plots below, in both cases, the variable *Old* plays a role in explaining the decision making of the model. Moreover, in the second case (balancing both the protected characteristic with the outcome variable) relative to the first, its Shapley value is higher in the model's decision-making process. This is in line with Kelley, Ovchinnikov, Hardoon, & Heinrich, (2022), in that feature selection that is blind to the protected characteristic leads to discrimination. The second case has a worse area under the curve (AUC) in terms of its performance: specifically, the AUC in Case i is 0.6721 while the AUC in Case ii is 0.6486.

To see where this performance dip comes from, we separately assess the proportion of rejected loans misclassified by the model and the proportion of loans correctly classified by the model. For the proportion of rejected loans misclassified by the model, we observe that Case ii's performance (0.5159) is worse than that of Case i's (0.4471). If one is more interested in the proportion of loans correctly classified by the model, Case i (0.6805) underperforms Case ii (0.6830), albeit by a much slighter margin.

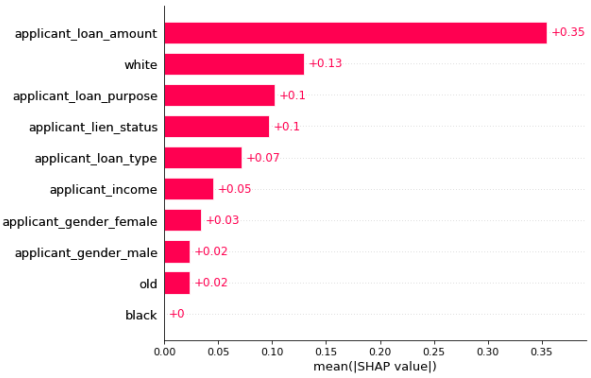
Notably, given Case ii misclassifies more rejections as acceptances, "fairer" outcomes can be said to have been achieved. This is due to the context of the rejections in this setting. Specifically, rejections are those loan applications initially approved by the loan guarantor (usually a Government Sponsored Enterprise), but subsequently failed to meet the lender's requirements.[12] From this context, coupled with marginally less misclassification for acceptances, one may conclude case ii provides overall "fairer" outcomes than case i.

Overall, this suggests using case ii vs case i, one can promote fairer outcomes based on a protected characteristic in the data, but the model's (fairer) classification performance suffers overall and for rejections (but not acceptances). A bias-performance trade-off is thus made evident by comparing these cases.

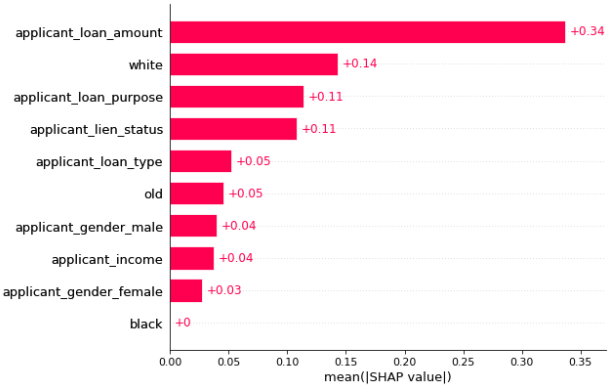
The rationale behind the increase in *Old*'s Shap value ranking ties back to the intuition edified behind the rebalancing undertaken through preprocessing (i.e., hybrid over-under sampling). To be able to ascertain more clearly the impact of an imbalanced protected characteristic in real world decision making, this needs to be corrected for while preserving the underlying statistical properties of the

12 More specifically, loan rejections occur if a loan application initially satisfies the approval requirements of guarantors of loans (i.e., a Government Sponsored Enterprise (GSE) – Fannie Mae and Freddie Mac – or the Federal Housing Administration (FHA)), though it subsequently fails in meeting the lender's requirements.

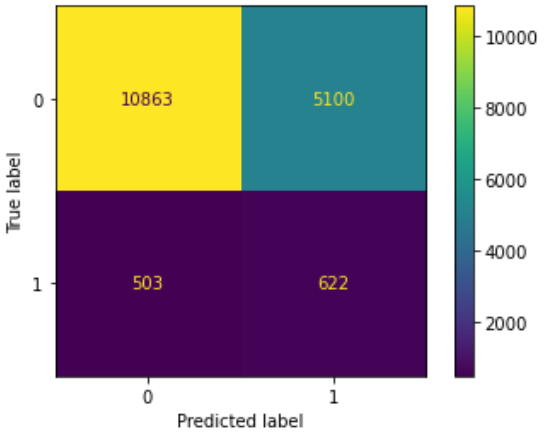
data, to be able to predict fairer outcomes in the unaltered test subset.



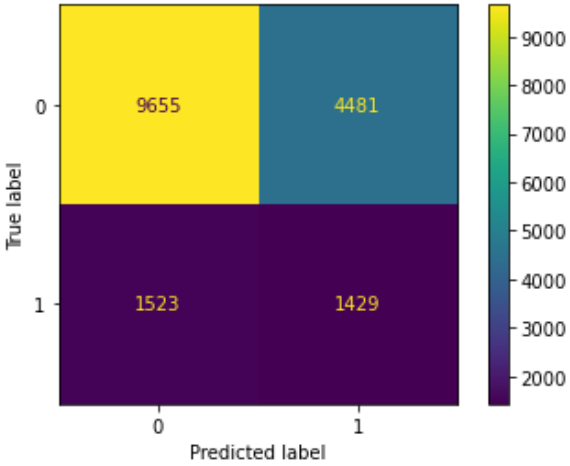
Case i: Bar plot of mean absolute Shapley values where we rebalance only training data for outcome variable (loan decision, which take a value of 1 if loan is rejected and 0 if loan is accepted)



Case ii: Bar plot of mean absolute Shapley values where we rebalance only training data for outcome variable (loan decision, which take a value of 1 if loan is rejected and 0 if loan is accepted) and protected characteristic (old, which takes a value of 1 if the applicant's age is above 62 and 0 otherwise)



Case i: Confusion matrix where we rebalance only training data for outcome variable (loan decision, which take a value of 1 if loan is rejected and 0 if loan is accepted)



Case ii: Confusion matrix where we rebalance only training data for outcome variable (loan decision, which take a value of 1 if loan is rejected and 0 if loan is accepted) and protected characteristic (old, which takes a value of 1 if the applicant's age is above 62 and 0 otherwise)

5. Conclusion and Recommendations

More and more businesses have begun to use bespoke black box algorithms for crucial decision making with negligible human involvement, but significant human impact. Regulatory authorities and computer scientists have consequently called for transparency through algorithmic accountability. Such algorithmic decision-making creates ex-ante and ex-post rights to explanation for all relevant stakeholders.

Further, algorithmic discrimination has been highlighted in recent years for individuals, groups, and subgroups. Such discrimination occurs based on protected characteristics that should have no bearing on outcomes. We propose a framework, using an ensemble of explainable Artificial Intelligence (XAI), and other so called fairness techniques to provide the information, at different stages of its execution, to satisfy these rights to explanation

for all relevant stakeholders. We further demonstrate how they may be used by banks and financial institutions using algorithms for decision making in one context (US individual home mortgage loan applications) but which may be readily extended to more contexts (including credit ratings, other loan applications, loan terms, to name a few). Based on their preferred definition of fairness, this allows stakeholders to assess algorithmic fairness in decision making, while studying the trade-off in the algorithm's performance in predicting outcomes

References

Agarwal, S., Muckley, C., & Neelakantan, P. (2023). Countering racial discrimination in algorithmic lending: A case for model-agnostic interpretation methods. *Economics Letters*, 111117.

Bartlett, R., Morse, A., Stanton, R., & Wallace, N. (2022). Consumer-lending discrimination in the FinTech era. *Journal of Financial Economics*, 30-56.

Beaumont, P., Horsburgh, B., Pilgerstorfer, P., Droth, A., Oentaryo, R., Ler, S., . . . Leong, W. (2021). CausalNex. *https://github.com/quantumblacklabs/causalnex*.

Bellamy, R. K., Dey, K., Hind, M., Hoffman, S. C., Houde, S., Kannan, K., . . . others. (2018). AI Fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias. *arXiv preprint arXiv:1810.01943*.

Bird, S., Dudik, M., Edgar, R., Horn, B., Lutz, R., Milan, V., . . . Walker, K. (2020). Fairlearn: A toolkit for assessing and improving fairness in AI. *Microsoft, Tech. Rep. MSR-TR-2020-32*.

Blattner, L., Stark, P.-R., & Spiess, J. (2022). Machine Learning Explainability \& Fairness: Insights from Consumer Lending. *FinRegLab Whitepaper*.

Castelnovo, A., Crupi, R., Del Gamba, G., Greco, G., Naseer, A., Regoli, D., & Gonzalez, B. S. (2020). Befair: Addressing fairness in the banking sector. *2020 IEEE International Conference on Big Data (Big Data)*, 3652-3661.

Chen, Y., Giudici, P., Liu, K., & Raffinetti, E. (2022). Measuring Fairness in Credit Scoring. *Available at SSRN 4123413*.

Cortés, K., & Strahan, P. (2017). Tracing out capital flows: How financially integrated banks respond to natural disasters. *Journal of Financial Economics*, 182–199.

Dlugosz, J., Gam, Y. K., Gopalan, R., & Skrastins, J. (2023). Decision-Making Delegation in Banks. *Management Science*, 1-21.

Dudik, M., Chen, W., Barrocas, S., Inchiosa, M., Lewins, N., Oprescu, M., . . . Wallach, H. (2020). Assessing and mitigating unfairness in credit models with the Fairlearn toolkit. *Microsoft and EY White Paper*.

Fuster, A., Goldsmith-Pinkham, P., Ramadorai, T., & Walther, A. (2022). Predictably unequal? The effects of machine learning on credit markets. *The Journal of Finance*, 5-47.

Karimi, H., Khan, M. F., Liu, H., Derr, T., & Liu, H. (2022). Enhancing individual fairness through propensity score matching. *2022 IEEE 9th International Conference on Data Science and Advanced Analytics (DSAA)*, 1-10.

Kelley, S., Ovchinnikov, A., Hardoon, D., & Heinrich, A. (2022). Antidiscrimination Laws, Artificial Intelligence, and Gender Bias: A Case Study in Nonmortgage Fintech Lending. *Manufacturing & Service Operations Management*, 3039–3059.

Kim, T., & Routledge, B. (2022). Why a Right to an Explanation of Algorithmic Decision-Making Should Exist: A Trust-Based Approach. *Business Ethics Quarterly*, 75–102.

Kozodoi, N., Jacob, J., & Lessmann, S. (2022). Fairness in credit scoring: Assessment, implementation and profit implications. *European Journal of Operational Research*, 1083–1094.

Kusner, M. J., Loftus, J., Russell, C., & Silva, R. (2017). Counterfactual fairness. *Advances in neural information processing systems*.

Martin, K. (2023). Who Counts in Business Ethics. *Business Ethics Quarterly*, 216-243.

Wan, M., Zha, D., Liu, N., & Zou, N. (2023). In-processing modeling techniques for machine learning fairness: A survey. *ACM Transactions on Knowledge Discovery from Data*, 1-27.

Weber, M., Yurochkin, M., Botros, S., & Markov, V. (2020). Black loans matter: Distributionally robust fairness for fighting subgroup discrimination. *arXiv preprint arXiv:2012.01193*.

About the Authors

Kushagra Jain is a Research Associate at the Financial Regulation Innovation Lab (FRIL), University of Strathclyde. His research interests include artificial intelligence, machine learning, financial/regulatory technology, textual analysis, international finance and risk management, among others. He is a recipient of doctoral scholarships from the Financial Mathematics and Computation Cluster (FMCC), Science Foundation Ireland (SFI), Higher Education Authority (HEA) and Michael Smurfit Graduate Business School, University College Dublin (UCD). Previously, he worked within wealth management and as a statutory auditor. He completed his doctoral studies in Finance from UCD in 2024, and obtained his MSc in Finance from UCD, his Accounting Technician accreditation from the Institute of Chartered Accountants of India and his undergraduate degree from Bangalore University. He was a FMCC Database Management Group Data Manager, Research Assistant, PhD Representative and Teaching Assistant for undergraduate, graduate and MBA programmes.

Email: kushagra.jain@strath.ac.uk

Dr James Bowden is Lecturer in Financial Technology at the Strathclyde Business School, University of Strathclyde, where he is the programme director of the MSc Financial Technology. Prior to this, he gained experience as a Knowledge Transfer Partnership (KTP) Associate at Bangor Business School, and he has previous industry experience within the global financial index team at FTSE Russell. Dr Bowden's research focusses on different areas of financial technology (FinTech), and his published work involves the application of text analysis algorithms to financial disclosures, news reporting, and social media. More recently he has been working on projects incorporating audio analysis into existing financial text analysis models and investigating the use cases of satellite imagery for the purpose of corporate environmental monitoring. Dr Bowden has published in respected international journals, such as the European Journal of Finance, the Journal of Comparative Economics, and the Journal of International Financial Markets, Institutions and Money. He has also contributed chapters to books including "Disruptive Technology in Banking and Finance", published by Palgrave Macmillan. His commentary on financial events has previously been published in The Conversation UK, the World Economic Forum, MarketWatch and Business Insider, and he has appeared on international TV stations to discuss financial innovations such as non-fungible tokens (NFTs).

Email: james.bowden@strath.ac.uk

Professor Mark Cummins is Professor of Financial Technology at the Strathclyde Business School, University of Strathclyde, where he leads the FinTech Cluster as part of the university's Technology and Innovation Zone leadership and connection into the Glasgow City Innovation District. As part of this role, he is driving collaboration between the FinTech Cluster and the other strategic clusters identified by the University of Strathclyde, in particular the Space, Quantum and Industrial Informatics Clusters. Professor Cummins is the lead investigator at the University of Strathclyde on the newly funded (via UK Government and Glasgow City Council) Financial Regulation Innovation Lab initiative, a novel industry project under the leadership of FinTech Scotland and in collaboration with the University of Glasgow. He previously held the posts of Professor of Finance at the Dublin City University (DCU) Business School and Director of the Irish Institute of Digital Business. Professor Cummins has research interests in the following areas: financial technology (FinTech), with particular interest in Explainable AI and Generative AI; quantitative finance; energy and commodity finance; sustainable finance; model risk management. Professor Cummins has over 50 publication outputs. He has published in leading international discipline journals such as: European Journal of Operational Research; Journal of Money, Credit and Banking; Journal of Banking and Finance; Journal of Financial Markets; Journal of Empirical Finance; and International Review of Financial Analysis. Professor Cummins is co-editor of the open access Palgrave title *Disrupting Finance: Fintech and Strategy in the 21st Century*. He is also co-author of the Wiley Finance title *Handbook of Multi-Commodity Markets and Products: Structuring, Trading and Risk Management*.

Email: mark.cummins@strath.ac.uk

Get in touch

FRIL@FinTechscotland.com

FinTech Scotland®

University of Glasgow

University of Strathclyde Glasgow