Machine Learning

Considerations for fairly and transparently expanding access to credit

Machine Learning:

Considerations for Fairly and Transparently Expanding Access to Credit

---

BLDS LLC, Discover Financial Services, and H2O.ai

July 2020: 1st Edition

by BLDS LLC, Discover Financial Services, and H2O.ai

Published by H2O.ai, Inc.
2307 Leghorn St.
Mountain View, CA 94043

July 2020: 1st Edition

Printed in the United States of America.

Contents

Abstract

There have long been cases of discrimination in access to capital and lending opportunities in financial services. With the introduction of machine learning, major concerns have arisen in regards to perpetuating historical human bias and in explaining credit decisions to consumers. In this whitepaper, lenders, vendors, and advisors create a common framework and set of definitions for key machine learning concepts, so that market participants, regulators, policymakers, and other stakeholders can all speak the same language when addressing novel opportunities and risks. We aim to define discrimination and interpretability in the context of fair lending and machine learning, and then to discuss important considerations for machine learning in the equitable and transparent use of algorithms to underwrite credit lending decisions.

Keywords: Machine Learning, Bias, Fairness, Explainable AI, Fair Lending, Machine Learning Interpretability, Responsible AI.

1 Introduction

The proliferation of machine learning (ML) has the potential to greatly improve outcomes for consumers, businesses, and other stakeholders across a wide range of applications and industries. While the promise of ML has already been realized, we anticipate the proliferation of ML will increase these benefits throughout various sectors of the economy. Within the financial services industry, lenders' use of ML tools to measure and identify risk in the provision of credit will likely benefit not only financial institutions (FIs), but also the consumers and businesses that obtain credit from the lenders. ML systems' increased accuracy of assessments of creditworthiness relative to traditional statistical modeling will not only allow lenders to manage risk and earnings more effectively, but will also enable lenders to expand access to credit to communities, individuals, and businesses that were previously unable to access the traditional mainstream financial system.

When FIs adopt ML systems, they are often better positioned than companies in other industries that lack expertise in mathematical modeling. In fact, many large FIs have considerable experience developing and deploying sophisticated mathematical and statistical models to support lending and decision-making functions. When skilled modelers and data scientists have access to high-quality data and tools, and if management is actively engaged in harnessing the value of statistical modeling, the adoption of ML systems is a natural next step in a lender's process of technological evolution. Further, lenders generally have strong guardrails in place for robust model testing and validation. Because of this wealth of experience and strong protections, the adoption of ML by FIs

is more likely to happen in a manner that helps to minimize the inherent risks associated with this technology.

On the other hand, challenges may arise since FIs are governed by FI-specific development, validation, and audit processes. Because the governance process is generally extensively prescribed, changing it to adopt new technology can be a slow-moving and arduous process. Further, a complex regime of state and federal regulations aim to promote transparency and stability for predictive modeling systems and prohibit illegal discriminatory outcomes.[1] Excessive and lagging layers of governance and regulatory requirements can detrimentally hinder lenders' ability to adopt ML systems, and may actually undermine the goals of transparency, stability, and fairness.

As such, one goal of this whitepaper is to continue the dialogue around the application of ML methodologies in the context of credit decisions. We propose to establish uniform definitions of key ML concepts, so that market participants, regulators, policymakers, and other stakeholders can all speak the same language. By sharing a common language and understanding the possibilities and risks presented by ML, we believe that the relevant stakeholders will be able to effectively identify considerations that need to be addressed so that this technology can be employed in the safest, fairest, and most timely way possible.

Another goal of the whitepaper is to describe how to mitigate adverse implications of this technology with the proper controls, and exhibit the various methodologies that have been proliferated in the ML arena. This whitepaper uses the current legal and regulatory environment as a foundation to add critical context to a broad discussion of relevant, substantial, and novel ML methodologies. We also provide a comparison between traditional methodologies and ML, show ways to potentially improve fairness, and discuss examples of explainable and interpretable models. The paper is organized as follows: Section 2 discusses the benefits of ML; Section 3 defines phrases related to discrimination, explanation, interpretable models, and scope; Section 4 provides details and concerns regarding existing and newer methodologies; and Section 5 concludes the whitepaper by stressing the importance of collaboration and governance to promote equitable, stable, and transparent lending models.

2 Benefits of Machine Learning

Like many past and present commercial technologies and models, ML presents both opportunities and risks. This section briefly addresses the principal op-

[1] Examples of legal and regulatory standards applicable to lenders include: The Equal Credit Opportunity Act (ECOA), The Fair Credit Reporting Act (FCRA), The Fair Housing Act (FHA), and regulatory guidance such as *Interagency Guidance on Model Risk Management* (Federal Reserve Board, SR Letter 11-7).

portunities for using ML in credit lending, while subsequent sections present analysis on various risks and tradeoffs. In this whitepaper, we consider why FIs would even consider ML for high-stakes use cases in credit lending.

ML has already become widespread in its applications in various sectors across the economy, including in many areas of consumer financial services, such as customer marketing, fraud detection, and, to a somewhat lesser extent, underwriting and pricing. Through advances in modern computing power, ML algorithms search for obscured patterns within and across features, enabling computers to learn to make decisions both faster and often with greater accuracy than can be achieved by humans. The ability to leverage this speed and efficiency has made ML methods effective and viable alternatives to traditional modeling methodologies. Additional advantages include ML's ability to impose fewer assumptions on the distribution of training data and its capacity to efficiently use a far greater number of input features, both of which often enable superior predictive accuracy.

ML models can be applied to a wide variety of data sources, some of which may help reduce the population of consumers and businesses that are currently "credit invisible" due to lenders' over-reliance on narrow datasets. Additionally, since ML can incorporate more information about a given person's creditworthiness in making its prediction, it streamlines the ability for lenders to incorporate so-called "alternative data", such as rent and utility payments that may allow lenders to assess the credit quality of thin file or no credit score applicants who previously would have been rejected for credit.[2] Since these applicants tend more frequently to be members of minority groups, ML's ability to incorporate this additional data can lead to more inclusiveness and a broader population receiving credit. Thus, ML models offer the potential to push the "efficiency frontier" by offering benefits to both customers (by potentially expanding credit access to new segments) and FIs (in the form of loan growth and decreased losses).

Perhaps counter intuitively, the complex nature of ML provides opportunities to make models fairer, as well. Traditional methods often presented limited options for making models fairer and frequently any changes to the model would cause its quality to deteriorate to the point where it was no longer useful from a business perspective. ML is generally the opposite. For a given dataset, data scientists are often able to identify large numbers of model specifications that provide virtually identical predictive quality. Because of this, it is not clear which of these model specifications will truly be the most predictive model once it is put into use (this is widely known as the "multiplicity of good models"). This multiplicity presents an opportunity for making lending fairer: since we

[2] See *Interagency Statement on the Use of Alternative Data in Credit Underwriting* for a working definition of alternative data and an outline of its opportunities and challenges.

have many models that are equal in predictive quality, we can also optimize fairness. Thus, among the approximately equally predictive models, lenders can choose one that is also the fairest. Finally, advances in the interpretability of ML now likely allow for such fair and accurate models to be explained to business partners, consumers, and regulators. Definitions and important considerations for preventing discrimination and increasing interpretability in ML are discussed in greater detail in the sections below.

3 Definitions

Establishing a common language is essential for stakeholders to accurately be able to discuss issues. Since ML is an evolving field of computational science, misconceptions and myths are bound to emerge in the public dialogue with terms and phrases that may, or may not, be relevant to the practice of ML in lending. Buzzwords and vague or inconsistent definitions are often counter-productive for nuanced discussions of complex topics, so Section 3 seeks to provide specific, uniform definitions for key concepts relating to ML in lending.

3.1 Discrimination: Protected Classes and Legal Standards

Since the 1960s, the US has had laws that prohibit illegal discrimination and establish a strong framework for safeguarding the rights of certain groups of consumers that have been historically disadvantaged and, thus, deemed "protected classes". For example, the Equal Credit Opportunity Act (ECOA) prohibits illegal discrimination in any aspect of a credit transaction based on an applicant's race, color, religion, national origin, sex, marital status, or age, as well as other "prohibited bases". Similarly, the Fair Housing Act (FHA) prohibits illegal discrimination in the mortgage lending or housing context.[3]

There are two principle theories of liability under ECOA and FHA for discrimination against members of protected classes: "disparate treatment" and "disparate impact".[4] Below we outline commonly accepted definitions of these terms.

Disparate Treatment: Disparate treatment occurs when a lender does not provide credit, or provides credit at less favorable terms, to an applicant or customer (respectively). In ML, disparate treatment may occur if a system explicitly considers protected class status or a very close substitute as an input to the system (commonly referred to as a "proxy"), such as including a feature

[3] Prohibited bases under FHA include race, color, religion, national origin, sex, familial status, and disability.

[4] *CFPB Supervision and Examination Manual*, Part II, Section C, Equal Credit Opportunity Act (October 2015).

or using segmentation. Disparate treatment discrimination is always illegal in lending.

Disparate Impact: Disparate impact occurs when a protected class experiences a larger share of less favorable outcomes as a result of an otherwise non-discriminatory and legitimate decision-making process. Disparate impact is not necessarily a violation of law and may be justified by a "business necessity", such as cost or profitability. However, there may still be a violation if the lender could have used an alternative policy or practice that had a less discriminatory effect.[5,6]

To illustrate, consider the effect of using a person's income when a lender decides whether to underwrite a loan. Of course, a lender would be remiss not to consider a borrower's income. Under disparate impact theory this would be considered a "facially neutral factor" – it is a business-justified factor because it is predictive of the borrower's ability to repay. However, among most cross sections of the US population, certain minority groups have lower incomes on average relative to Non-Hispanic Whites. Thus, when a lender considers income in its lending decision, the use of income could cause a disparate impact if the population being scored was one where the members of the minority group(s) had lower average incomes than Non-Hispanic Whites. Whether this disparate impact would rise to the level of a fair lending violation depends on whether other measures of the borrower's ability to repay could be used in place of, or in conjunction with, income, where that factor led to similarly predictive results, but lower levels of disparate impact.

Separate from the legally defined terms above, we also note that phrases and concepts such as "Ethical AI" or "ML Fairness", have become used widely in academic literature and in numerous media outlets, but are still somewhat amorphous or ill-defined. Because many similar concepts have already been discussed and evaluated in the context of employment, housing, and credit discrimination, techniques put forward under these new banners are often less attuned to regulations or legal precedent. However, many potentially exciting technological developments are revealing how ML can be made fairer. When these methods are aligned with accepted legal definitions of fairness, they not only benefit consumers, but are also more likely to lessen regulatory or legal risk.[7]

[5] *Id.*

[6] The US Supreme Court established the disparate impact theory in Griggs v. Duke Power Co. (1971), however there have been numerous subsequent court cases challenging whether disparate impact is cognizable under ECOA. These issues are outside the scope of this whitepaper.

[7] For additional information, interested readers may follow the evolution of the free textbook, **Fairness and Machine Learning** [1], in addition to other reputable publications. *An Introduction to Artificial Intelligence and Solutions to the Problems of Algorithmic*

3.2 Explanation, Interpretable Models, and Scope Definitions

One of the common myths of ML is that it creates "black-box" systems. However, this notion has been dispelled in recent years with the advent of a number of tools that increase transparency into ML-based decision-making. Transparency into the intricacies of ML systems is achieved today by two primary technical mechanisms: directly interpretable ML model architectures and the post-hoc explanation of ML model decisions. These mechanisms are particularly important in lending because, under ECOA's implementing regulation, Regulation B, and the Fair Credit Reporting Act (FCRA), many credit decisions that are adverse to the applicant must be summarized to consumers through a predefined set of short written explanations known as "adverse action notices".

Adverse Action Notices: Under Regulation B, lenders must notify an applicant in writing of the primary reasons for taking an adverse action on a loan application within a specific time period.[8] When using ML systems to make credit decisions, the principal reasons included on adverse action notices are explanations that are based on ML system input features that negatively affected the applicant's score or assessment. Until recently, generation of these customer-level explanations had been considered a serious impediment to the adoption of ML in lending, but beginning around 2016, open source and commercial implementations of more transparent ML approaches became more commonly available.

Regulation B provides standards for the factors lenders may use and how lenders must inform applicants of credit decisions when those decisions are based on credit scoring models.[9] For a rejected application, lenders must indicate the principal reasons for the adverse action and accurately describe the features actually considered. The notice must include a specific considered input feature, but is not required to state how or why a given feature contributed to an adverse outcome. An examination of the regulatory requirements for adverse action notices demonstrates that several of the methods described below could prove useful in ensuring that ML credit models are employed in a compliant manner.

Interpretability and Explainability: Although not strictly defined in the context of fair lending, relevant phrases from the academic literature and data

Discrimination is an additional resource that specifically addresses concerns about algorithmic fairness for FIs [2].

[8] See 12 CFR §1002.2(c). The term "adverse action" is defined generally to include a "refusal to grant credit in substantially the amount or on substantially the terms requested" by an applicant or a termination or other unfavorable change in terms on an existing account.

[9] See 12 CFR §1002.9(b) and the Official Commentary thereto included in Supplement I of Regulation B.

science community include "interpretable" and "explanation". Finale Doshi-Velez and Been Kim define interpretability as, "the ability to explain or to present in understandable terms to a human" [3]. Professor Sameer Singh of the University of California at Irvine defines an explanation in the context of an ML system as a "collection of visual and/or interactive artifacts that provide a user with sufficient description of a system's behavior to accurately perform tasks like evaluation, trusting, predicting, or improving a system" [4]. "Interpretable" is usually a descriptor for directly transparent or constrained ML model architectures, and "explanation" is often applied to a post-hoc process that occurs after model training to summarize main drivers of model decisions. Both concepts are important for adverse action notice reporting, because the more interpretable and explainable an ML system, the more accurate and consistent the associated adverse action notices.

Global v. Local Scope: A closely related concept to explanation is "scope". ML systems can be summarized "globally", meaning across an entire dataset or portfolio of customers, and "locally", meaning for only a subset of a dataset or a smaller group of customers, including a single customer. Both global and local explanations are important to FIs when deploying ML. Global explanation results are often documented as part of an FI's model governance processes to meet regulatory standards on model risk management (e.g., Federal Reserve Board SR Letter 11-7), while local customer-specific explanations are likely to be a primary technical process behind adverse action notice generation for FCRA and ECOA compliance.

Due to the proliferation of various local ML explanation techniques and their relevance to adverse action notice generation, this whitepaper proposes two additional definitions of locality: "data locality" and "value locality". "Data locality" refers to the subset of columns and rows of a dataset over which aggregate statistics are calculated to generate an explanation. An explanation method is said to have "value locality", if to determine the contribution of one input predictor for an individual ML-based decision, it is required to provide the values of all input predictors for that individual. For value-locality, an explanation depends on the rest of the predictor values and would change if those values also changed. See Figure 1 for an illustration of these concepts. In practical terms, value locality implies that adverse action reasons based on input features explaining why an application is rejected for credit, or any explanations for other types of ML predictions, are not independent and are intrinsically linked to the value of all other input features. For instance, decreases in income are often linked to increases in credit line utilization and changes in employment status. Value locality is likely a positive attribute for an explanation, since input features are rarely actually independent.

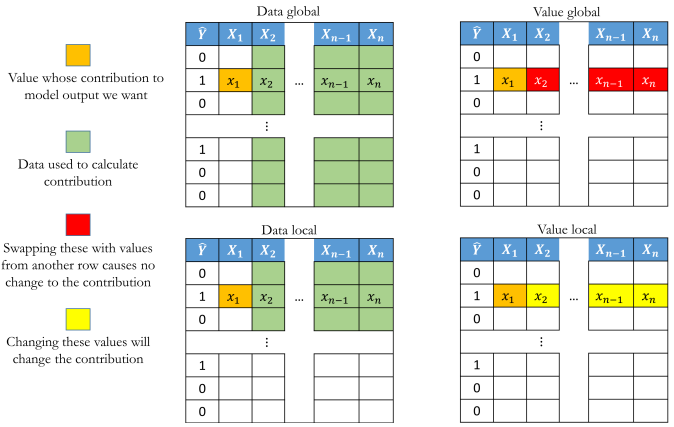


Figure 1: Illustration of scope concepts data global, data local, value global, and value local. Figure courtesy of Discover Financial Services.

4 Considerations

The boosted accuracy promised by ML lending models will require model developers and validators to carefully balance the high predictive capacity of ML with both well-established and novel considerations around discrimination, stability, and transparency. To better achieve this balance, Section 4 discusses some of the primary discrimination and interpretability concerns that arise when using ML in lending.

4.1 Discrimination

There are many ways that analysts and data scientists can define and mitigate discrimination in ML.[10] However, only a subset of the discrimination measurement and mitigation techniques available today are likely to be appropriate for fair lending purposes. This subsection puts forward a few established discrimination measurements before discussing some newer measures and mitigation techniques, and explains why, if some of these newer approaches are used, fair lending regulations must be carefully considered in order to properly mitigate compliance risk.

[10] See 21 Fairness Definitions and Their Politics.

4.1.1 Traditional Fair Lending Discrimination Definitions

Given the recent interest in fairness and ethics in ML, it may appear to some observers that algorithmic discrimination is a new issue. On the contrary, testing outcomes in education, lending, and employment for discrimination is a decades-old discipline [5]. Measures such as marginal effect (ME), adverse impact ratio (AIR), and standardized mean difference (SMD, which is also known as "Cohen's d") have been widely used in fair lending and are likely still some of the best ways to test for discrimination in ML lending models.

ME is used when reviewing a binary, yes-no, decision such as loan acceptance. ME is the difference between the acceptance rate for a control group (often Whites or males) and a protected demographic class of interest, reported as a percentage:

$$\text{ME} \equiv 100 \cdot (\text{Pr}(\hat{\mathbf{y}} = 1|X_c = 1) - \text{Pr}(\hat{\mathbf{y}} = 1|X_p = 1)) \tag{1}$$

where $\hat{\mathbf{y}}$ are the model decisions, X_c and X_p represent binary markers created from a demographic attribute, c denotes the control group, p indicates a protected group, and $\text{Pr}(\cdot)$ is the operator for conditional probability.[11]

Importantly, ME can only be interpreted within the context of a given lending scenario. As a hypothetical example, a male-female ME of 4% would often be a highly noteworthy difference in mortgage lending to a population of prime consumers, because class-control differences in credit quality among prime consumers is usually relatively small – particularly for this example because women often have higher average credit quality. On the other hand, a Minority-to-Non-Hispanic White ME of 4% might not be as unusual for something like a credit card offered to consumers across a wide spectrum of credit quality – particularly if the population includes those with thin credit files. This is because some minority groups are more frequently found at the lower end of the spectrum of perceived credit quality as measured using traditional scoring systems. Thus, while minorities with a given model score would be treated the same as a Non-Hispanic White who had the same score, the average offer rate by class would be lower, leading to the higher ME.[12]

[11] See Consumer Financial Protection Bureau, *Supervisory Highlights, Issue 9, Fall 2015*, p. 29.

[12] Regardless of the product, ME of 4% would warrant a fair lending review.

AIR is another well-known and widely accepted measure of discrimination used for evaluating yes-no decisions. AIR is the ratio of the acceptance rate for a protected class and the acceptance rate for the control group.[13]

$$\text{AIR} \equiv \frac{\text{Pr}(\hat{\mathbf{y}} = 1|X_p = 1)}{\text{Pr}(\hat{\mathbf{y}} = 1|X_c = 1)} \tag{2}$$

A lower AIR would indicate that a lower percentage of applicants within the protected class are accepted than applicants in the control group. Importantly, AIR is one measure of disparate impact that is associated with a published practical significance threshold. Specifically, in some employment law contexts, AIR values below 0.8 that are statistically significant may be considered *prima facie* evidence of illegal disparate impact discrimination.[14]

SMD is the difference in the mean protected class prediction, $\bar{\hat{\mathbf{y}}}_p$, minus the mean control class prediction, $\bar{\hat{\mathbf{y}}}_c$, divided by the standard deviation of the population prediction, $\sigma_{\hat{\mathbf{y}}}$:

$$\text{SMD} \equiv \frac{\bar{\hat{\mathbf{y}}}_p - \bar{\hat{\mathbf{y}}}_c}{\sigma_{\hat{\mathbf{y}}}} \tag{3}$$

A higher SMD indicates a greater disparity in outcomes between the protected class members and the control group. Like AIR, SMD also has frequently cited thresholds that indicate small (0.2), medium (0.5), and large (0.8) differences between predictions for two groups [6]. Unlike the two other measures defined in Eqs. 1 and 2, SMD is also appropriate for measuring continuous outcomes, such as credit limits and interest rates.

4.1.2 Recently Proposed Discrimination Definitions and Discrimination Mitigation Techniques

In recent years, ML and fair lending experts have explored ways to measure and mitigate discrimination in ML. These discrimination mitigation techniques

[13] See Uniform Guidelines on Employee Selection Procedures (1978) §1607.4.

[14] This threshold might be considered a "bright line" marker of concern, but regulators and courts are not bound to this threshold. Additionally, while this standard was put forward by the Equal Employment Opportunity Commission (EEOC) for matters of employment discrimination, it has not been given explicit approval by any FI regulators in the credit or housing context. Thus, simply because a model shows AIR values above 0.80 does not mean that it would not be found to be illegally discriminatory. For guidance in the employment context, see *Questions and Answers to Clarify and Provide a Common Interpretation of the Uniform Guidelines on Employee Selection Procedures*. Unfortunately, we know of no published guidance by FI regulators about these standards, but this knowledge comes from our experience working with lenders and regulators.

come in three forms: pre-processing, in-processing, and post-processing. Pre-processing techniques (e.g., reweighing [7]), diminish bias in the data used to train the ML models. In-processing methods (e.g., adversarial de-biasing [8]) are ML algorithms that themselves remove bias from their predictions as they learn. Post-processing techniques (e.g., reject option classification [9]) change the predictions that come out of an ML model in order to minimize discrimination.[15]

4.1.3 Considerations for Regulatory Compliance

It is imperative that FIs employ the appropriate use of discrimination testing and mitigation methods for regulated applications in fair lending because some methods may lead to counterproductive results or even result in non-compliance with anti-discrimination statutes.

Regarding measuring fairness, it is difficult to optimize on multiple metrics of fairness at one time – there is necessarily a trade-off where making a model fairer by one measure makes it appear less fair by another.[10] For example, choosing a model that provides more balanced error rates across protected classes will often correspond to choosing a model that gives loan offers to fewer protected class members. While academic literature on ML fairness has often focused on balanced error rates, regulators and courts have generally focused on lending rates. Certain open source and commercially available software have generally followed the academic practice and focused on measures of relative error rates.[16] While these are important and often useful measures of fairness, if a lender were to choose among models based on error rates alone, then they may cause traditional measures of disparate impact, such as ME or AIR (see Subsection 4.1.1), to become worse. Therefore, focusing on more established discrimination measures may be the safer route for today's practitioners in fair lending.

Similar scenarios can also arise for many newer discrimination mitigation techniques. Since ECOA proscribes the use of protected class status when making a lending decision – even if the lender intends to use it to make its lending decisions fairer – the discrimination mitigation methodologies that require explicit consideration of protected class status in training or production are not likely to be considered acceptable. In fact, because FIs are explicitly legally prohibited from collecting information such as race and ethnicity (apart from mortgage lending), these techniques are simply infeasible.

[15] We cite these specific references because they have influenced academic and popular debates about algorithmic discrimination, and are relevant for predictive modeling practitioners in general. However, these techniques sometimes fail to meet the requirements set forth by applicable regulations in fair lending as addressed in the following section.

[16] E.g., Aequitas or H2O Driverless AI.

Given such restrictions, mitigation approaches that perform feature selection and hyperparameter searches may be considered natural extensions of traditional approaches and are likely to be subject to less concern by regulators. Other methods that do not rebalance data or decisions and that do not explicitly consider protected class status may gain wider acceptance as they are used more frequently and are shown to be effective ways to decrease discrimination. Examples of these techniques may include methods that modify the objective function of an algorithm to minimize discrimination during the learning process.

4.2 Interpretability

Like discrimination testing and mitigation approaches, many new techniques for increasing transparency and understanding in ML models have been introduced in recent years. These new techniques create both transparent models and summaries of model decisions. They are already being used in the financial services industry today[17] and are likely to be deployed for lending purposes. This subsection introduces some of these techniques and important considerations for their use in the lending context.

4.2.1 Examples of Interpretable Machine Learning Models

In the past, ML researchers and practitioners operated under what appeared to be a natural trade-off: the more accurate a model, the more complex, and the harder to understand and explain. Due to adverse action notice disclosures, model documentation requirements, and discrimination concerns, black-box ML models were typically not considered a viable option for lending. Today, the landscape has changed for predictive modelers in credit lending with the advent of highly accurate and highly interpretable model architectures that appear to break the so-called "accuracy-interpretability trade-off". In fact, some leading scholars have posited that, for the kinds of structured tabular data used most commonly in lending models, black-boxes are likely not more accurate than interpretable ML models [10].[18]

Sophisticated and interpretable ML models are a particularly exciting breakthrough, as they enable practitioners to build accurate ML models that can also be documented, explained to consumers, and readily tested and remediated for discrimination. The diversity of interpretable ML models is also striking. Interpretable ML models include variations of linear models (e.g., explainable

[17] E.g., New Patent-Pending Technology from Equifax Enables Configurable AI Models; see also, *Deep Insights into Explainability and Interpretability of Machine Learning Algorithms and Applications to Risk Management*.

[18] See comparisons of EBM with other interpretable and black-box models by Microsoft Research: https://github.com/interpretml/interpret.

boosting machine (EBM, also known as GA2M) [11]), constrained tree-based models (e.g., optimal sparse decision trees (OSDTs) [12], monotonic gradient boosting machines (MGBMs)[19]), constrained neural networks (e.g., explainable neural networks (XNNs) [13]), novel or constrained rule-based models (e.g., scalable Bayesian rule lists (SBRLs) [14], CORELS [15]), and several others. Levels of interpretability vary from results that could only be understood by advanced technical practitioners (e.g., MGBMs or XNNs) to results that business and legal partners could likely consume directly, (e.g., OSDTs or SBRLs), to something in-between, (e.g., EBMs). Beyond their obvious advantages for adverse action notice requirements, interpretable ML models should also assist practitioners in model governance and documentation tasks, such as understanding which input features drive model predictions, how they do so, and which feature behavior under the model aligns with human domain knowledge. Moreover, interpretable models can help in discrimination testing and remediation by transparent weighting and treatment of input features. For reference, Figure 2 displays model architecture and diagnostic information from a MGBM model trained on data collected under the Home Mortgage Disclosure Act (HMDA).

4.2.2 Examples of Post-hoc Explanations

Post-hoc explanation techniques create summaries of varying types and accuracy about ML model behavior or predictions. These summaries can provide an additional, customizable layer of explanation for interpretable ML models, or they can be used to gain some amount of insight regarding the inner-workings of black-box ML models. Summary explanations can have global or local scopes, both of which are useful for adverse action notice generation, model documentation, and discrimination testing. Post-hoc explanations can be generated through numerous approaches including direct measures of feature importance (e.g., gradient-based feature attribution [16], Shapley values [17]), surrogate models (e.g., decision trees [18],[19], anchors [20], local interpretable model-agnostic explanations (LIME) [21]), and plots of trained model predictions (e.g., accumulated local effects (ALE) [22], partial dependence [23], and individual conditional expectation (ICE) [24]).

Figure 3a provides an example of global feature importance, calculated by the TreeSHAP method, which shows on average which input features are main drivers of model decisions for an entire dataset. Figure 3b gives an example of local TreeSHAP feature importance for three different individuals, showing which input features made each individual the most different from the overall average prediction of the model. Due to the ability to accurately summarize individual model predictions, variants of Shapley values, and TreeSHAP in

[19]Monotonic GBM, as implemented in XGBoost or h2o.

particular, seem to be gathering momentum for adverse action notice generation [25], [26]. Surrogate models are simple models derived from complex models. They can be global, such as a decision tree surrogate model, which creates a data-derived flow chart to describe the decision policies of a more complex model, or they can be local, as with LIME approaches that model small regions of a ML response function or decision boundary with an interpretable linear model. Plots of trained model predictions take many forms. Figure 2 displays examples of partial dependence and ICE plots acting as concise visual summaries of complex model behaviors. Of course, like all other ML techniques, post-hoc explanations approaches have pros and cons, and should never be regarded as a perfect view into complex model behaviors. (See Subsection 4.2.5 for a more detailed discussion of additional post-hoc explanation problems.)

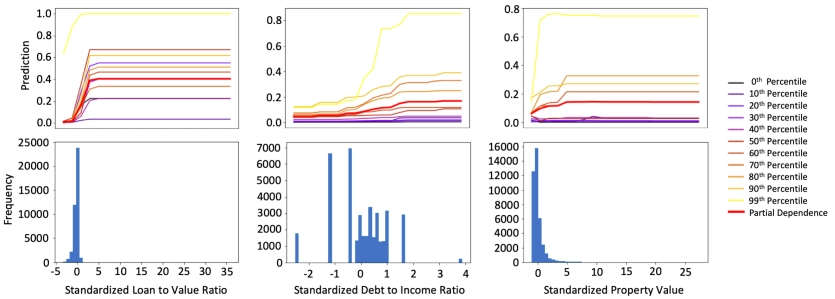


Figure 2: Partial dependence and individual conditional expectation (ICE) plots paired with histograms for a constrained MGBM model and HMDA data. Figure reproduced from *A Responsible Machine Learning Workflow* with permission of the authors [27].

Figures 2, 3a and 3b are displayed herein to give interested readers an illustration of a constrained ML model and post-hoc explanation workflow. Figure 2 shows partial dependence and ICE plots of a constrained MGBM that was trained to predict whether a consumer will receive a high priced mortgage, i.e., a loan with an annual percentage rate (APR) of at least 1.5% greater than similar loans awarded in the same year. Figure 2 uses partial dependence and ICE to verify that the MGBM learned a monotonically increasing relationship between loan to value ratio and debt to income ratio and the probability that a consumer receives a high-priced mortgage, independently. Because the data-global behavior of the model, portrayed by partial dependence (red), does not diverge from the data-local behavior of the model, represented by ICE curves at several deciles of predicted probability w.r.t loan to value ratio and debt to income ratio, partial dependence is likely to be an accurate global summary for the MGBM and these two input features. For property value, non-monotonic behavior is visible in the ICE curves at lower deciles of model predictions, indicating that the

model did not learn a monotonic relationship between property value and the probability that a customer receives a high priced loan. The non-monotonic ICE curves can also indicate the presence of interactions that are averaged out of the data-global partial dependence explanation or they can indicate that partial dependence is failing for other reasons, such as correlation between input features. Further model debugging, beyond the scope of this whitepaper, would be required to assess the root cause of the local non-monotonic behavior displayed for property value under the MGBM model.[20] Pairing the partial dependence and ICE plots of feature behavior in the MGBM with histograms (Figure 2) is a diagnostic technique that allows practitioners to spot areas of data sparsity, where predictions may be unstable or untrustworthy.

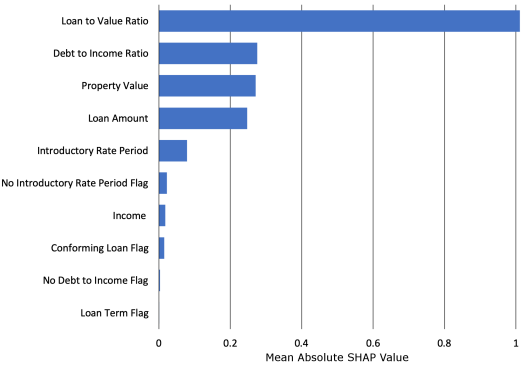
Figure 3a displays a quintessential global feature importance chart in which loan to value ratio, debt to income ratio, and property value are the three most important drivers of data-global MGBM behavior. Unlike inconsistent feature importance charts readers may have seen in the past, note that this feature importance chart was generated by aggregating accurate and consistent Shapley values across every row of a dataset.

Value-local Shapley attributions can be seen for three individual customers at various deciles of predicted probability in Figure 3b. These consumer-specific values can be interpreted as the contribution that an input feature makes to drive a prediction away from the average prediction. If this interpretation is not exactly aligned with a given FI's interpretation of the adverse action notice requirements of ECOA Regulation B, or other applicable statutes, practitioners can change background datasets or make causal modifications to arrive at Shapley values with more customized interpretations ([28], [29]), such as comparing a rejected customer to the average prediction for non-rejected customers. Furthermore, the information in Figures 2, 3a and 3b can also be added to model documentation or potentially be used to mitigate any discrimination problems that may arise from a model (see Subsection 4.2.6).

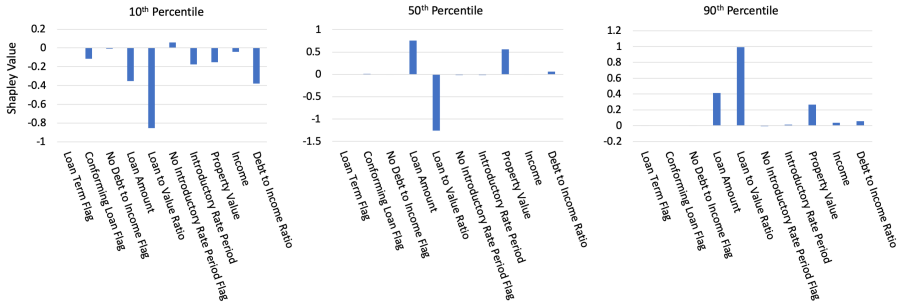
4.2.3 The Importance of Dual Scope Explanations

An important, and often discussed, aspect of ML interpretability is the scope of an explanation – whether an explanation is local or global. Many new research papers focus on local explanations for evaluating the impact of a feature at the individual customer level. However, seeing both a global and local view presents a more holistic picture of model outcomes. For example, it is important for a customer to understand what global factors resulted in an adverse action on their credit decision (e.g., length of credit history), while also understanding

[20] Model debugging is an emergent discipline focusing on advanced and exhaustive testing of sophisticated ML models, e.g., *Real-world Strategies for Model Debugging*.



(a) Global Shapley feature importance.



(b) Local Shapley feature importance for three individual customers at selected deciles of predicted probability.

Figure 3: Global and local feature importance for a constrained MGBM model and HMDA data. Figure reproduced from *A Responsible Machine Learning Workflow* with permission of the authors [27].

what are the local factors that are within their control to achieve a favorable outcome in the near future (e.g., lower utilization of credit limit). In addition to the use of interpretable models, explanation approaches that consider both global and local summaries are likely to be most informative to practitioners and to consumers (as in Figures 2, 3a and 3b). For convenience, Table 1 below buckets some common explanation techniques into different scopes as defined in Subsection 3.2.

4.2.4 Grouping Correlated Features for Explanation

Many explanatory techniques are less accurate in the presence of correlated input features [30], [31]. Grouping involves treating a group of correlated features,

Table 1: Scope of common post-hoc explanation techniques.

	Partial Dependence	Tree SHAP	LIME	ALE	Decision Tree Surrogate
Data Scope	Global (Subset for local)	Global (Subset for local)	Local	Local	Global (Subset for local)
Value Scope	Global	Local	Local	Global	Global

with strong correlations between features in the group and weak correlations with features outside of the group, as one from an explanation standpoint. Grouping presents benefits including the consideration of effects that features have on each other in the presence of correlation. Note that the success of grouping is likely contingent on the use of constrained models where measures of global Pearson correlation are more meaningful than in unconstrained ML models, which may focus almost entirely on complex local dependencies between features.

4.2.5 Some Concerns with Post-hoc Explanations

As mentioned in previous sections, post-hoc explanation is not yet a perfect science. Well-known pitfalls include partial dependence failing in the presence of correlated or interacting input features, inaccurate surrogate models that do not truly represent the underlying complex model they seek to summarize, and inconsistencies in feature importance values [30], [31], [32]. Moreover, as discussed in Subsection 4.2.4, few explanation methods are robust correlation between input features. Subsection 4.2.5 will explore issues beyond these direct mathematical concerns with specific explanation techniques, particularly the fundamental issue of explanation inconsistency and problems with human comprehension of explanations.

Inconsistent Explanations

Since many ML explanation techniques are inconsistent, different ML models, or even different configurations of the same ML model or refreshing the same ML model with new data, can result in different explanations for the same consumer if not controlled. Inconsistency bears special consideration in financial services, and especially for the generation of adverse action notices for credit lending decisions, where two models giving different explanations to the same applicant may raise questions, if not lead to regulatory non-compliance or reputational harm for the FI. **To mitigate risks associated with inconsistent explana-**

tions, FIs should pair post-hoc explanations with constrained and stable ML models, explicitly test for explanation stability, and consider using explanation techniques with consistency guarantees. It is the authors' view, as well as many experts, that TreeSHAP presents the strongest theoretical guarantees for accuracy and consistency, and Shapley values have a notable presence in economics and game theory literature dating back to the early 1950s [17], [33].

Human Comprehension of Explanations

Concerns have been raised that non-technical audiences (e.g., most credit applicants) cannot easily understand ML models and explanations [31], [34], [35]. In financial services, there are several less technical audiences to consider, including validation and audit personnel, business partners, legal, and consumers. The success of an explainable ML project often hinges on the comprehension of model behavior by less technical audiences. To ensure the project is successful and nontechnical experts are included, nontechnical stakeholders could be included in project planning for explainable ML endeavors. In doing so, technical practitioners can make best efforts toward enabling validation so nontechnical partners can understand their ML results. Additionally, some have suggested counterfactual explanations (i.e., explanations that tell a decision-subject what can be done to change their ML-based decision to receive a more positive outcome) are a preferred method for presenting ML decisions to a broader population [31], [36]. While counterfactual explanations could be more intuitive for some audiences, and perhaps suited for adverse action notice generation in some cases, without careful treatment, they may suffer from the inconsistency concerns raised in the subsection directly above.

4.2.6 Explanations for Discrimination Analysis

Explainable ML techniques can be used both to identify causes of discrimination and to enable lenders to find potentially less discriminatory, but similarly predictive alternative models. A difficulty with ML models is that they often use far more features than traditional methods, which means that each feature might only have a minimal effect on discrimination. This is problematic because traditional techniques for discrimination mitigation focus on dropping or adding just a few features – rarely more than two or three. In the case of an ML model with dozens or hundreds of features, disparate impact (if any) is generally caused by many features working in combination to cause the total negative effect on some protected class(es). This means that changing two or three features alone will often not provide a meaningful improvement in disparate impact. Further compounding this problem, when a model contains highly correlated features (which ML models often do), dropping one feature could only lead to another

feature taking its place in terms of its discriminatory effect. Until recently, these were merely ancillary problems because even identifying which features are drivers of discrimination appeared to be nearly impossible in black-box models. This is why it was challenging to use ML in credit decisioning. More recently, providers of discrimination mitigation software and academics working on these problems have shown that explanation techniques such as Shapley values can be used to guide an understanding of both the discriminatory and predictive impacts of each feature.[21] With this information, a model builder can structure a search for alternative models in a smarter way, by removing features that are of low importance and have large disparate impact and keeping features that are important and also beneficial to protected classes. This, in combination with high-powered computing, which makes testing a large number of possible alternative models feasible, allows model builders and compliance professionals to perform a more robust search for less discriminatory models that maintain their predictive ability [2].[22]

4.2.7 Considerations for Regulatory Compliance

ECOA and Regulation B do not prescribe a specific number of adverse action notices to share with consumers, nor do they prescribe specific mathematical techniques. However, regulatory commentary indicates that more than four reasons may not be meaningful to a consumer. FIs also have flexibility in selecting a method to identify principal reasons. Regulation B provides two example methods for selecting principal reasons from a credit scoring system, but allows creditors the flexibility to use any method that produces substantially similar results. One method is to identify the features for which the applicant's score fell furthest below the average score for each of those features achieved by applicants whose total score was at or slightly above the minimum passing score. Another method is to identify the features for which the applicant's score fell furthest below the average score for each of those features achieved by all applicants.[23] Both examples appear to be aligned with the general principles of Shapley values, counter-factual explanations, and other similar post-hoc explanation techniques described in Sections 3 and 4.

[21] See Explaining Measures of Fairness with SHAP, which provides examples of how Shapley values can be compared across protected class groups in order to understand the effects of a model's features on fairness metrics. This notebook was created by Scott Lundberg, the author of several authoritative papers on Shapley values.

[22] For example, through the use of Shapley values to weight the predictive and potentially discriminatory effect of features, BLDS, LLC's fairness software, ConsilienceML, has been shown to quickly converge to less discriminatory, but highly predictive alternative models that maintain business justification.

[23] See The Official Commentary of 12 CFR §1002.9(b)(2) for Regulation B.

5 Conclusion

The advent and increasingly widespread use of ML in lending has the promise of being more inclusive to historically underserved populations than traditional underwriting. We hope this whitepaper helps to dispel some of the myths and inaccuracies about ML and, in turn, provide a simplified, yet substantive, discussion of key definitions and considerations for using ML within the lending context. More importantly, we hope this whitepaper has demonstrated that the use of ML for lending is not only viable, but in many ways better suited to ensuring that FIs make credit decisions that are based on reliable predictive analytics and free of unlawful discrimination. As for advances in the interpretability of ML, additional options exist today to train sophisticated ML models and explain them to various audiences. While questions remain as to which methods will be viewed as most useful for ensuring compliance with regulatory requirements, such as adverse action notice regulations, variants of constrained models, Shapley values, and counterfactual explanations appear to be gaining some momentum in the US lending community.

In the fair lending context, there are well-established discrimination testing and mitigation methodologies that have been used for decades. In recent years, as researchers and engineers have developed ML technology with the goal of increasing the availability of credit while decreasing discrimination, some tension has arisen between traditionally accepted fair lending methodologies and the newly-developed ML fairness tools. Fair lending practitioners must work carefully with legal and compliance personnel to leverage recent ML advances without unintentionally neglecting compliance with existing regulatory and legal standards. Of course, discrimination and interpretability are only two of many concerns about ML for first-, second-, and third-line personnel at FIs. As models become more sophisticated and FIs become more dependent upon them, and as data privacy and artificial intelligence regulations grow in number and complexity, proper model governance and human review, and closer collaboration between legal, compliance, audit, risk, and data science functions will likely only increase in importance.

6 References

1. Solon Barocas, Moritz Hardt, and Arvind Narayanan. **Fairness and Machine Learning**. fairmlbook.org, 2019. URL: http://www.fairmlbook.org

2. Nicholas Schmidt and Bryce Stephens. An Introduction to Artificial Intelligence and Solutions to the Problems of Algorithmic Discrimination.

Conference on Consumer Finance Law Quarterly Report, 73(2):130–144, 2019. URL: https://arxiv.org/pdf/1911.05755.pdf

3. Finale Doshi-Velez and Been Kim. Towards a Rigorous Science of Interpretable Machine Learning. *arXiv preprint arXiv:1702.08608*, 2017. URL: https://arxiv.org/pdf/1702.08608.pdf

4. Patrick Hall, Navdeep Gill, and Nicholas Schmidt. Proposed Guidelines for the Responsible Use of Explainable Machine Learning. *NeurIPS Robust AI in Financial Services Workshop*, 2019. URL: https://arxiv.org/pdf/1906.03533.pdf

5. Ben Hutchinson and Margaret Mitchell. 50 Years of Test (Un) Fairness: Lessons for Machine Learning. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 49–58, 2019. URL: https://arxiv.org/pdf/1811.10104.pdf

6. Jacob Cohen. **Statistical Power Analysis for the Behavioral Sciences**. Lawrence Erlbaum Associates, 1988. URL: https://bit.ly/398IYLr

7. Faisal Kamiran and Toon Calders. Data Preprocessing Techniques for Classification Without Discrimination. *Knowledge and Information Systems*, 33(1):1–33, 2012. URL: https://bit.ly/2lH95lQ

8. Brian Hu Zhang, Blake Lemoine, and Margaret Mitchell. Mitigating Unwanted Biases with Adversarial Learning. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pages 335–340. ACM, 2018. URL: https://arxiv.org/pdf/1801.07593.pdf

9. Faisal Kamiran, Asim Karim, and Xiangliang Zhang. Decision Theory for Discrimination-aware Classification. In *2012 IEEE 12th International Conference on Data Mining*, pages 924–929. IEEE, 2012. URL: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.722.3030&rep=rep1&type=pdf

10. Cynthia Rudin. Please Stop Explaining Black Box Models for High Stakes Decisions and Use Interpretable Models Instead. *arXiv preprint arXiv:1811.10154*, 2018. URL: https://arxiv.org/pdf/1811.10154.pdf

11. Yin Lou, Rich Caruana, Johannes Gehrke, and Giles Hooker. Accurate Intelligible Models with Pairwise Interactions. In *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 623–631. ACM, 2013. URL: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.352.7682&rep=rep1&type=pdf

12. Xiyang Hu, Cynthia Rudin, and Margo Seltzer. Optimal Sparse Decision Trees. *arXiv preprint arXiv:1904.12847*, 2019. URL: https://arxiv.org/pdf/1904.12847.pdf

13. Joel Vaughan, Agus Sudjianto, Erind Brahimi, Jie Chen, and Vijayan N. Nair. Explainable Neural Networks Based on Additive Index Models. *arXiv preprint arXiv:1806.01933*, 2018. URL: https://arxiv.org/pdf/1806.01933.pdf

14. Hongyu Yang, Cynthia Rudin, and Margo Seltzer. Scalable Bayesian Rule Lists. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, 2017. URL: https://arxiv.org/pdf/1602.08610.pdf

15. Elaine Angelino, Nicholas Larus-Stone, Daniel Alabi, Margo Seltzer, and Cynthia Rudin. Learning Certifiably Optimal Rule Lists for Categorical Data. *The Journal of Machine Learning Research*, 18(1):8753–8830, 2017. URL: http://www.jmlr.org/papers/volume18/17-716/17-716.pdf

16. Marco Ancona, Enea Ceolini, Cengiz Oztireli, and Markus Gross. Towards Better Understanding of Gradient-based Attribution Methods for Deep Neural Networks. In *6th International Conference on Learning Representations (ICLR 2018)*, 2018. URL: https://www.research-collection.ethz.ch/bitstream/handle/20.500.11850/249929/Flow_ICLR_2018.pdf

17. Scott M. Lundberg and Su-In Lee. A Unified Approach to Interpreting Model Predictions. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems (NIPS)*, pages 4765–4774. Curran Associates, Inc., 2017. URL: http://papers.nips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions.pdf

18. Mark W. Craven and Jude W. Shavlik. Extracting Tree-Structured Representations of Trained Networks. *Advances in Neural Information Processing Systems*, 1996. URL: http://papers.nips.cc/paper/1152-extracting-tree-structured-representations-of-trained-networks.pdf

19. Osbert Bastani, Carolyn Kim, and Hamsa Bastani. Interpreting Blackbox Models via Model Extraction. *arXiv preprint arXiv:1705.08504*, 2017. URL: https://arxiv.org/pdf/1705.08504.pdf

20. Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. Anchors: High-Precision Model-agnostic Explanations. In *AAAI Conference on Artificial*

Intelligence, 2018. URL: https://homes.cs.washington.edu/~marcotcr/aaai18.pdf

21. Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. Why Should I Trust You?: Explaining the Predictions of Any Classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1135–1144. ACM, 2016. URL: http://www.kdd.org/kdd2016/papers/files/rfp0573-ribeiroA.pdf

22. Daniel W. Apley. Visualizing the Effects of Predictor Variables in Black Box Supervised Learning Models. *arXiv preprint arXiv:1612.08468*, 2016. URL: https://arxiv.org/pdf/1612.08468.pdf

23. Jerome Friedman, Trevor Hastie, and Robert Tibshirani. **The Elements of Statistical Learning**. Springer, New York, 2001. URL: https://web.stanford.edu/~hastie/ElemStatLearn/printings/ESLII_print12.pdf

24. Alex Goldstein, Adam Kapelner, Justin Bleich, and Emil Pitkin. Peeking Inside the Black Box: Visualizing Statistical Learning with Plots of Individual Conditional Expectation. *Journal of Computational and Graphical Statistics*, 24(1), 2015. URL: https://arxiv.org/pdf/1309.6392.pdf

25. Philippe Bracke, Anupam Datta, Carsten Jung, and Shayak Sen. Machine Learning Explainability in Finance: An Application to Default Risk Analysis. 2019. URL: https://bit.ly/2UXz4Uy

26. Niklas Bussman, Paolo Giudici, Dimitri Marinelli, and Jochen Papenbrock. Explainable AI in Credit Risk Management. *Credit Risk Management (December 18, 2019)*, 2019. URL: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3506274

27. Patrick Hall, Navdeep Gill, Kim Montgomery, and Nicholas Schmidt. A Responsible Machine Learning Workflow with Focus on Interpretable Models, Post-hoc Explanation, and Discrimination Testing. *Information*, 11, 2020. URL: https://www.mdpi.com/2078-2489/11/3/137

28. Christopher Frye, Colin Rowat, and Ilya Feige. Asymmetric Shapley Values: Incorporating Causal Knowledge into Model-agnostic Explainability. *Advances in Neural Information Processing Systems*, 33, 2020. URL: https://arxiv.org/pdf/1910.06358.pdf

29. Dominik Janzing, Lenon Minorics, and Patrick Blöbaum. Feature Relevance Quantification in Explainable AI: a Causal Problem. In *International Conference on Artificial Intelligence and Statistics*, pages 2907–2916. PMLR, 2020. URL: http://proceedings.mlr.press/v108/janzing20a/janzing20a.pdf

30. Christoph Molnar et al. *Limitations of Interpretable Machine Learning Methods*. LMU Munich, 2020. URL: https://compstat-lmu.github.io/iml_methods_limitations/

31. I. Elizabeth Kumar, Suresh Venkatasubramanian, Carlos Scheidegger, and Sorelle Friedler. Problems with Shapley-value-based Explanations as Feature Importance Measures. *arXiv preprint arXiv:2002.11097*, 2020. URL: https://arxiv.org/pdf/2002.11097.pdf

32. Patrick Hall. On the Art and Science of Machine Learning Explanations. In *KDD '19 XAI Workshop*, 2019. URL: https://arxiv.org/pdf/1810.02909.pdf

33. Lloyd S. Shapley, Alvin E. Roth, et al. *The Shapley Value: Essays in Honor of Lloyd S. Shapley*. Cambridge University Press, 1988. URL: http://www.library.fa.ru/files/Roth2.pdf

34. Forough Poursabzi-Sangdeh, Daniel G. Goldstein, Jake M. Hofman, Jennifer Wortman Vaughan, and Hanna Wallach. Manipulating and Measuring Model Interpretability. *arXiv preprint arXiv:1802.07810*, 2018. URL: https://arxiv.org/pdf/1802.07810.pdf

35. Richard Tomsett, Dave Braines, Dan Harborne, Alun Preece, and Supriyo Chakraborty. Interpretable to Whom? A Role-based Model for Analyzing Interpretable Machine Learning Systems. *2018 ICML Workshop on Human Interpretability in Machine Learning (WHI 2018)*, 2018. URL: https://arxiv.org/pdf/1806.07552.pdf

36. Sandra Wachter, Brent Mittelstadt, and Chris Russell. Counterfactual Explanations Without Opening the Black Box: Automated Decisions and the GPDR. *Harv. JL & Tech.*, 31:841, 2017. URL: https://arxiv.org/pdf/1711.00399.pdf

7 Contributors

Contributors listed in alphabetical order by institution.

BLDS, LLC
Nicholas Schmidt

Discover Financial Services
Steve Dickerson
Patrick Haggerty
Arjun Ravi Kannan
Kostas Kotsiopoulos
Raghu Kulkarni
Alexey Miroshnikov
Kate Prochaska
Melanie Wiwczaroski

H2O.ai
Benjamin Cox
Patrick Hall
Josephine Wang