AI in Financial Services: *Explainability in Credit Underwriting*

Frequently Asked Questions

September 2020

FinRegLab's wide-ranging investigation into the use of artificial intelligence (AI) in financial services provides the basis for our AI FAQs. This resource is designed to provide financial services stakeholders with accessible information about the technological, market, and policy issues related to the use of these advanced analytical techniques in the service of a vibrant and inclusive financial marketplace. This edition of our AI FAQs focuses on model transparency and explainability as critical threshold issues for firms in highly regulated areas like financial services and on the use of AI for credit underwriting.

FinRegLab is investigating the state of AI in financial services to understand better several key issues:

» **How the use of advanced analytical processes can help drive our financial system toward a more rapid and inclusive recovery from the COVID-19 pandemic—one that improves the financial resiliency of families, businesses, and communities over the longer term**

» **How the use of AI in financial services is shaping the evolution of these technologies, especially with respect to improvements in the explainability, reliability, and fairness of AI and machine learning models**

FinRegLab is a nonprofit research organization founded on the premise that independent, rigorous research is a primary ingredient in developing market norms and policy solutions that will enable responsible innovation and a more inclusive financial system.

- **How financial institutions, vendors, and policy makers evaluate obstacles to the use of AI and machine learning in various applications**

- **How policy, law, and regulation may need to evolve to promote responsible development and use of AI in the financial system**

The use of machine learning in credit underwriting is of particular interest as an outgrowth of FinRegLab's research on one form of alternative data — the use of cash-flow data in consumer and small business lending. Research into advanced analytical methods like machine learning also complements our work to evaluate how data and technology can foster an inclusive recovery from the pandemic and improve the financial foundations of millions of individuals and businesses.

To create a resource for financial services stakeholders, we have designed these FAQs to share insights from our investigation of the use of AI and machine learning in financial services. Our AI FAQs provide foundational information and help explain areas of potential confusion such as nuances in how various stakeholder groups use terms like explainability and bias. This edition of FAQs focuses on the issues and debates about model transparency and explainability and the implications of using machine learning for credit underwriting. The following questions are answered in this edition:

- **What is model transparency? Why do we need it?**
- **What is model interpretability?**
- **What is model explainability?**
- **Why is model transparency especially important in the context of AI and machine learning models?**
- **What techniques can make machine learning models more transparent?**
- **How are *post hoc* explainability techniques being used in practice?**
- **What are global and local model explanations?**
- **How can machine learning be used in credit underwriting?**
- **Why is model transparency especially important in the context of machine learning models used for credit underwriting?**
- **Who needs information about how a credit underwriting model works?**
- **What potential risks are important to consider when lenders replace incumbent underwriting models with machine learning?**
- **What legal and regulatory frameworks apply to the use of machine learning credit underwriting models?**

- **What prudential frameworks apply to the use of machine learning credit underwriting models?**
- **What kind of statistical biases are most important with respect to underwriting models?**
- **What consumer protection frameworks apply to the use of machine learning credit underwriting models?**
- **What are the sources of discrimination or unfairness in underwriting models and other predictive models?**
- **What are the core concerns about adopting machine learning underwriting models from the perspective of fair lending and financial inclusion?**
- **How can we measure the fairness of a model?**

In time, our AI FAQs will be presented in digital form on our website to facilitate their use as a reference. Until then, this document refers to, rather than repeats, content previously covered in our AI FAQs. One question in the prior edition of AI FAQs provides particularly important context for what follows: What is the basis for believing that machine learning could improve credit underwriting?

Readers can find the full set of AI FAQs here or use the links in the cross-reference sections included with specific questions below.

What is model transparency? Why do we need it?

Model transparency refers to the ability of stakeholders in a particular model – such as its developers, risk managers, and regulators – to access the information about the model's design, use, and performance that they need. Those needs may vary based on the stage of the model lifecycle and the purpose that such information serves.

Machine learning models do not inherently need to be transparent to make predictions, and existing law or regulation do not generally require users of AI or machine learning model users to meet defined thresholds for model transparency or measure it at any particular points in the model lifecycle.

But this quality is critical to using models in practice, especially in highly regulated sectors. Model transparency is an important component of establishing the trustworthiness of a model and of developing a broad consensus about the public use and oversight of AI. Model transparency is also needed in many, but not all areas, to enable oversight of whether a model is being operated in compliance with laws and regulations applicable to its use case, including fairness and privacy expectations. Notable too is the potential for model transparency to advance policy aims broader than regulatory compliance – for example, in improving consumers' understanding of their credit score or ways in which they might improve their financial position and creditworthiness in the future.

The type and complexity of the machine learning model being used shapes how model developers can achieve transparency. Some models have a higher degree of transparency by virtue of their structure and design. These models are said to be more interpretable.[1] Others may lack architecture that is transparent by design and therefore require the use of additional models, visualizations, or other techniques designed to explain the model – that is, to improve stakeholders' ability to access information about the model's behavior and the bases of its decisions. These interventions add an "observable component" to such complex models in order to enhance stakeholders' ability to understand the model's behavior and accept or challenge its decisions.[2]

Although many stakeholders use the terms interpretability and explainability interchangeably and both contribute to a model's transparency, the distinction between the two is important to understanding the choices that model developers make when designing and operating specific models and to considering the evolution of law, policy, and regulation to support the trustworthiness of AI models.

Further Reading

Leilani Gilpin, David Bau, Ben Yuan, Ayesha Bajwa, Michael Specter, and Lalana Kagal, Explaining Explanations: An Overview of Interpretability of Machine Learning, Massachusetts Institute of Technology (February 3, 2019), available at https://arxiv.org/pdf/1806.00069.pdf

Henrike Mueller and Florian Ostmann, AI Transparency in Financial Services, The Alan Turing Institute (February 18, 2020), available at https://www.turing.ac.uk/news/ai-transparency-financial-services

P. Jonathon Phillips, Carina Hahn, Peter Fontana, David Broniatowski, and Mark Przybocki, Four Principles of Explainable Artificial Intelligence, National Institute of Standards and Technology (NIST) (August 2020), available at https://doi.org/10.6028/NIST.IR.8312-draft

The Royal Society, Explainable AI: The Basics (November 2019), available at https://royalsociety.org/-/media/policy/projects/explainable-ai/AI-and-interpretability-policy-briefing.pdf

Andrew D. Selbst and Solon Barocas, The Intuitive Appeal of Explainable Machines, Fordham Law Review (December 2018), available at https://ir.lawnet.fordham.edu/flr/vol87/iss3/11/

Related FAQs

What is model interpretability?

What is model explainability?

Why is model transparency especially important in the context of AI and machine learning models?

Why is model transparency especially important in the context of machine learning models used for credit underwriting?

What is model interpretability?

Model interpretability refers to the ability to explain or to present models and results in terms that are understandable to a human and in so doing convey a basic sense of how the technology works.[3]

Interpretable models are ones where model stakeholders can relatively easily identify correlations or relationships used by the model to predict an outcome because of the model's design or structure.[4] Interpretable models include models with comparatively simple structures, such as short decision trees models,[5] that can be "inspected"[6] or may have a limited number of features or parameters, which make them easier to parse directly and without the use of additional models or tools. Developers can also apply constraints in the course of initial model building to make models interpretable.[7] Examples of constraints to increase model transparency include limiting input data or calculative processes in order to constrain the nonlinearity of the model so that its feature interactions are, for example, three-way instead of involving dozens or hundreds of interactions.[8]

Interpretable models may also be paired with additional modelling techniques to enhance their explainability[9] in order to meet the needs of certain requirements that require enhanced insight into model behavior – like adverse action notices which call on lenders to provide a statement of the primary bases of certain kinds of credit decisions.

Although the structure of these models may easily permit review and oversight, they may also involve tradeoffs in performance compared to more complex AI models because the same structure that facilitates interpretability may limit the model's capacity to identify predictive relationships among data points.[10]

Further Reading

Diogo Carvalho, Eduardo Pereira, Jaime Cardoso, Machine Learning Interpretability: A Survey on Methods and Metrics, MDPI (2019), available at https://www.mdpi.com/2079-9292/8/8/832

Patrick Hall and Navdeep Gill, An Introduction to Machine Learning Interpretability: An Applied Perspective on Fairness, Accountability, Transparency, and Explainable AI, Dataiku (2018), available at https://pages.dataiku.com/hubfs/ML-interperatability.pdf

The Royal Society (2019).

Related FAQs

What is model transparency? Why do we need it?

What is model explainability?

What techniques can make machine learning models more transparent?

What are global and local model explanations?

What is model explainability?

Model explainability refers to the ability of model stakeholders to understand model behavior – that is, how or why a particular prediction was made or result was reached.[11] This can include, how the model's predictions distribute across the population or subpopulations in the data set, or how the model's outputs may vary depending on different inputs and features.[12]

Like interpretability, explainability is a critical and specific component of establishing model transparency and trustworthiness. A primary purpose of explainability is to render the operation of a complex AI or machine learning model sufficiently transparent that its processes can be reviewed, although the need for this will vary based on the specific requirement in question. For instance, a particular type of review may focus on establishing the accuracy or quality of the model's predictions or on whether the model operates in compliance with applicable law, regulation, and firm policies. Neural networks and other forms of deep learning typically require use of explainability techniques since outputs are produced from numerous layers of nonlinear mathematics that identify and evaluate feature interactions.[13]

The explainability of any predictive model can be evaluated, but this issue is particularly important for AI models that may not be particularly interpretable or explainable without the use of additional models, visualizations or other techniques after the model has been trained. Such *post-hoc* techniques may be able to satisfy transparency needs without significantly affecting the predictiveness of the underlying complex model, but may impose other costs on the model user, such as slowing down the model or requiring access to additional data. They also raise independent trustworthiness questions, because they have the effect of reducing high levels of complexity in the underlying model into approximations of the model that are more readily understood.

A model developer's choice between building an interpretable model and pairing complex models with *post hoc* explainability techniques reflects an important, ongoing academic debate.[14] More research is needed to understand how specific explainability techniques work in applied and theoretical contexts, including developing a consistent framework for evaluating the transparency of AI models and identifying the strengths and weaknesses of various *post hoc* explainability techniques.

Further Reading

Christoph Molnar, Interpretable Machine Learning: A Guide for Making Black Boxes Explainable (2019), available at https://christophm.github.io/interpretable-ml-book/

The Royal Society (2019).

Related FAQs

What is model transparency? Why do we need it?

What is model interpretability?

What techniques can make machine learning models more transparent?

How are post hoc explainability techniques being used in practice?

What are global and local model explanations?

Why is model transparency especially important in the context of AI and machine learning models?

The technical demands of enabling the transparency of AI models, especially for more complex models, has heightened the importance of transparency in debates about the use of AI. These challenges include:

- The complexity of deciphering variable and feature interactions in non-linear models that may use tens or hundreds of thousands of data points to make predictions
- The difficulty of pinpointing specific reasons for a decision from a much larger set of variables, features, and interactions than traditional models use
- The absence of operating history to understand performance of models in all stages of use, through economic cycles, and in a range of applications

Controversies related to early applications of AI[15] have further increased the attention given to model transparency as an important threshold issue for adoption of AI and machine learning models. This is especially true in "high stakes" use cases in fields like medicine, criminal justice, and financial services, where models can deeply affect human lives. The specific regulatory frameworks that have been designed to promote practices that are both prudent and fair in these sensitive areas may implicitly or explicitly require model transparency. In these contexts, the processes through which a model reaches predictions, how various data points or interactions affected the decision, or the fidelity of its performance over time must be reviewable and in some case revisable to permit mitigation of problematic findings.[16] Absence of transparency in these area limits understanding or oversight of the model by developers and users, by regulators, and by people affected by its predictions whether they are investors in a company or users of its products. Given this, concerns about the transparency of AI models may ultimately delay or prevent adoption of AI models in specific use cases.

Related FAQs

Why is model transparency especially important in the context of machine learning models used for credit underwriting?

Who needs information about how a credit underwriting model works?

What techniques can make machine learning models more transparent?

A model developer can choose from an array of techniques to improve the transparency of a complex machine learning model. These techniques are relatively new and rapidly developing. Three types of *post hoc* explainability techniques, each of which do not fundamentally alter the model's internal operations, are particularly important:[17]

- **Surrogate Models:** Surrogate models are more interpretable models that are designed to approximate how the original model's processes work to make predictions. LIME is an example of a *post hoc* explainability technique that uses a surrogate model.

- **Visualizations:** A variety of visualization techniques can improve the interpretability of complex models. Visualizations can be simplified reproductions of a neural network's layers that can facilitate review of the processes that a model uses to make a prediction. They can also involve the generation of graphs that, for example, show how certain features correlate to a prediction.[18] Examples of visualizations for model behavior include partial dependence plots.[19]

- **Post-Perturbation Methods:** Post-perturbation[1] explainability methods, like SHAP, assess the impact of changing individual features to attribute their significance to the model's outcome. The resulting values can be averaged to form global feature importance metrics and visualized in numerous ways to describe model behaviors or global or local feature importance.[20]

Related FAQs

What is model explainability?

How are post hoc explainability techniques being used in practice?

What are global and local model explanations?

How are *post hoc* explainability techniques being used in practice?

Given the relative novelty of *post hoc* explainability techniques and their rapid evolution, emerging practice points in the direction of using multiple *post hoc* explainability techniques at once with complex models. Further, in some cases, model developers might choose to use *post hoc* explainability techniques with interpretable models to mitigate risk of errors, to get additional comfort on the performance and transparency of the underlying model, and to meet specific requirements like adverse action reporting that call for an explanation of a particular credit decision.[21] Additional information provided by the *post hoc* explainability techniques can refine the developer's understanding of the interpretable models and produce insight into how the explainability technique works and can be improved.

More research is needed to understand whether and in what circumstances the quality of explanations from *post hoc* techniques changes when those techniques are paired with interpretable models. Since those techniques provide explanations in the form of summaries, there is the potential that applying *post hoc* explainability techniques to interpretable models will produce higher quality summaries or approximations of the underlying model.[22]

Further Reading

BLDS, LLC., Discover Financial Services Inc., and H2O.ai, Machine Learning: Considerations for Expanding Access to Credit Fairly and Transparently, H2O.ai (July 2020), available at https://www.h2o.ai/resources/white-paper/machine-learning-considerations-for-fairly-and-transparently-expanding-access-to-credit/

Ilknur Kaynar Kabul, Interpret Model Predictions with Partial Dependence Plots and Individual Conditional Expectation Plots, SAS (June 12, 2018), available at https://blogs.sas.com/content/subconsciousmusings/2018/06/12/interpret-model-predictions-with-partial-dependence-and-individual-conditional-expectation-plots/

Molnar (2019).

Marco Tulio Ribeiro, Sameer Singh, Carlos Guestrin, "Why Should I Trust You?": Explaining the Predictions of Any Classifier, The University of Washington (August 9, 2016), available at https://arxiv.org/abs/1602.04938

Ram Sagar, 8 Explainable AI Frameworks Driving A New Paradigm for Transparency in AI, Analytics India Magazine (October 18, 2019), available at https://analyticsindiamag.com/8-explainable-ai-frameworks-driving-a-new-paradigm-for-transparency-in-ai/

Related FAQs

What is model explainability?

What techniques can make machine learning models more transparent?

What are global and local model explanations?

Model explanations can operate at different levels of specificity and scale by producing either global or local results.

Global explanations describe a model's outcomes and processes across an entire set of data.[23] They relate a group of inputs, class of objects, or data sets with the predictions of the dependent or target variable generated by the model. A global explanation speaks holistically to the analytical processes that describe how it functions when making predictions, although the information provided may be "highly approximate" under certain conditions.[24]

Local explanations describe model outcomes over a subset of the data or through a single feature, input, or variable.[25] They point to the reason or reasons for a specific decision, such as why a specific loan application was rejected.[26] The accuracy of local explanations tends to be higher than that of global explanations, because defined sections of machine learning models are more likely to be linear and interpretable.[27]

A weather app that is designed to predict the weather in the user's location might use a "poll of polls" model to derive its predictions. The app collates and analyzes a variety of individual weather forecasts. A global explanation of that model will describe aspects of the model's operation such as how the app's algorithm weighs individual forecasts based on the accuracy of recent forecasts in particular locations and how it determines weather patterns or conditions are most predictive in various seasons or locations. A local explanation of the model will state which particular forecasts and weather patterns or conditions reported therein were most important to its prediction of the weather for a given location at a given time.

Further Reading

Hall & Gill (2018).

How can machine learning be used in credit underwriting?

Many lenders are currently investigating how to responsibly develop and implement machine learning models that evaluate the credit risk posed by individuals or small businesses seeking a credit card, consumer loan, or some other form of credit as part of the application decisioning process. But AI and machine learning can be and are already being used to affect underwriting in a variety of additional ways, including:

- **Screening for fraud**: Fraud screening is a well-established use for both varied types of digital data and complex AI models like neural nets given the data-intensive, iterative processes needed to identify individual illicit acts based on rapidly changing patterns within massive volumes of streaming activity.[28] These models can be used both to determine which credit applications are evaluated in full underwriting processes and to evaluate individual transactions involving open-end credit, such as credit cards.

- **Developing marketing strategies:** AI and machine learning can help lenders sift through vast volumes of digital data to identify potential customers for their products and services and support the creation of pre-screened offers of credit.

- **Identifying predictive relationships:** An underwriting model using traditional modelling techniques may nevertheless apply rules or use interactions that the model builder identified through analysis of large data sets using machine learning. This can give lenders the benefit of machine learning's insight and ability to analyze large volumes of diverse data without incurring the costs of changing their lending platform or incurring certain regulatory risks.

- **Servicing loans:** Lenders can use machine learning to monitor the performance of their portfolio, to help identify borrowers who are most likely to falter in repayment, and to determine appropriate loan terms in a modification or workout.

Further Reading

Joseph Breeden, A Survey of Machine Learning in Credit Risk (May 30, 2020), available at https://www.researchgate.net/publication/341804274_A_Survey_of_Machine_Learning_in_Credit_Risk

Related FAQs

How are AI and machine learning being used in financial services?

What forms of AI and machine learning are most commonly used in financial services? How do they work?

What is the basis for believing that machine learning could improve credit underwriting?

Why is model transparency especially important in the context of machine learning models used for credit underwriting?

The importance of model transparency is further heightened in the context of evaluating applications for consumer and small business credit because of their high stakes compared to other uses of machine learning in financial services. These determinations may deeply affect the applicant's financial trajectory and expose the firm's investors and taxpayers to significant financial losses. Some of the legal and regulatory requirements that have long been used to manage these risks depend on model transparency and explainability as a means to satisfying specific compliance obligations. The sensitivity of this context requires oversight and information that will let internal and external stakeholders determine whether individual models deploy sound logic, can be used in compliance with all applicable laws and regulation, and are ethically defensible.[29]

Given that, lenders, regulators, and advocates alike have moved more slowly to fully accept machine learning in the context of underwriting than in other financial services contexts, like trading or fraud screening. Three particular risk management areas make interpretability a critical threshold issue for lenders that want to shift from traditional underwriting models to machine learning models: model risk management, fair lending, and adverse action reporting. For example, firms and agencies can detect analytical findings that warrant review as a potential disparate impact using traditional regression-based techniques, but in order to mitigate the sources of those adverse findings, users of machine learning models need to be able to identify which variables, features, or interactions cause them.

In each of these areas, solving technical challenges to deliver accurate and meaningful information about model behavior is an essential input to demonstrating appropriate regulatory compliance and developing the basis for safe and responsible use of machine learning underwriting models. For their part, regulators may also need to adapt existing guidance and expectations to reduce interpretative uncertainty around applicable frameworks that pre-date the advent of machine learning in this context.

Further Reading

Breeden, (2020).

Related FAQs

Who needs information about how a credit underwriting model works?

What legal and regulatory frameworks apply to the use of machine learning credit underwriting models?

Who needs information about how a credit underwriting model works?

A variety of stakeholders have a general need to understand how a credit underwriting model works and, in some cases, a particular need to understand individual predictions made by a model:

- A firm's business executives, who commit capital based on the model's predictions and need to establish the model's fitness-for-use
- A firm's legal and risk management teams, which review a model's compliance with laws, regulations, and firm policies relevant to a model's specific use case

- A firm's regulators, who review the firm's decisions about model development and use from the perspective of compliance with individual consumer financial protection requirements and monitoring prudential risks where applicable

- A firm's customers and potential customers, who want to understand the basis for the firm's decisions on applications and who are best able to detect promptly the use of erroneous data

- A firm's investors, who supply capital based on confidence in management's business judgment and performance of the loans or asset-backed securities

The needs of each of these stakeholders to understand how a model works are not fundamentally different for machine learning models as compared to incumbent underwriting models, but firms are still working to develop and test forms of machine learning that consistently meet these needs in credit underwriting.

Related FAQs

What are key policy debates about using AI in financial services?

What legal and regulatory frameworks apply to the use of machine learning credit underwriting models?

What prudential frameworks apply to the use of machine learning credit underwriting models?

What consumer protection frameworks apply to the use of machine learning credit underwriting models?

What potential risks are important to consider when lenders replace incumbent credit underwriting models with machine learning?

The shift to machine learning underwriting models may accentuate a number of risks that also occur with models built using more traditional techniques. Foremost among those are:

- **Overfitting:** Overfitting refers to the risk that the machine learning algorithm fits the predictive model too narrowly to the specific characteristics of training data, which may result in unnecessary complexity and increase the fragility of the model's performance. The effects of overfitting may be more severe to the extent that the training data is under-representative, inaccurate or otherwise flawed.

- **Data drift:** Data drift can occur when the data that a model uses to make a prediction differs in important ways from the data on which it was trained. Seasonal changes in behavior – sunscreen purchases are more common in the summer than winter – provide a simple example of the kind of data change that might affect model performance. The economic conditions brought by Covid-19 also pose a potential data drift risk because most underwriting models currently in use were developed and updated based on an unusually prolonged period of economic expansion and have shown signs of fragility in response to changes in conditions from their training data and prior operating data.[30]

» **Discrimination:** The expansion of the scale and types of data processed by machine learning models and the complexity of the resulting models heightens risks related to discrimination that affects groups specifically protected by law and regulation.[31] For example, in lending, machine learning models have the potential to replicate or amplify historical discrimination in whether and how credit has been provided due to reliance on lending data and the way in which such models are developed, used, and managed.

Further Reading

Breeden (2020).

Related FAQs

How different are AI and machine learning from other common forms of predictive modelling?

How can we evaluate a specific use of AI or machine learning to understand relevant differences when compared to incumbent models?

What kind of statistical biases are most important with respect to underwriting models?

What are the sources of discrimination or unfairness in underwriting models and other predictive models?

What legal and regulatory frameworks apply to the use of machine learning credit underwriting models?

The use of machine learning for credit underwriting will require firms to meet legal and regulatory expectations that apply to other kinds of underwriting models.[32] Both prudential and consumer protection requirements apply to various aspects of these models when used by a bank, whereas some consumer protections apply to lenders regardless of their legal form.

Here, we will highlight briefly where adapting machine learning underwriting models to existing requirements may be particularly challenging for firms. These include:

» Prudential requirements about the performance and governance of the models throughout the model lifecycle at the portfolio or lender level[33]

» Fair lending requirements, particularly with regard to facially neutral practices that have an impermissible disparate impact on a prohibited basis[34]

- Reporting requirements to provide applicants with individualized "adverse action" notices explaining why they were denied credit or offered less favorable terms[35]

Related FAQs

What are the key policy debates about using AI in financial services?

What potential risks are important to consider when lenders replace incumbent underwriting models with machine learning?

What prudential frameworks apply to the use of machine learning credit underwriting models?

What consumer protection frameworks apply to the use of machine learning credit underwriting models?

What are the sources of discrimination or unfairness in underwriting models and other predictive models?

What are the core concerns about adopting machine learning underwriting models from the perspective of fair lending and financial inclusion?

How can we measure the fairness of a model?

What prudential frameworks apply to the use of machine learning credit underwriting models?

A variety of prudential expectations apply to credit underwriting. Model risk management is one area where adapting machine learning underwriting models to existing requirements may be particularly challenging for firms.

Prudential requirements for banks include adhering to the modern model risk management framework articulated by the Board of Governors of the Federal Reserve Board and the Office of the Comptroller of the Currency in 2011 and Federal Deposit Insurance Corporation in 2017.[36] This guidance drives firms to document the design and operation of models, as well as their performance and reliability in various contexts. Demonstrating a model's fitness for use typically has a number of aspects that touch on how well firms can interpret and explain machine learning models:

- **Performance:** A core inquiry for model governance focuses on the quality, accuracy, and stability of model predictions in different circumstances. In the context of machine learning, this will likely necessitate assessment of risks that are more acute than for traditional regression-based models such as overfitting and data drift. Once in use, model performance must be monitored closely for signs that changes in data, economic conditions, or other factors do not undermine the robustness of the model's performance.

- **Statistical debiasing**: Part of assessing and refining the quality of a model's predictions involves considering several types of statistical bias, such as sample and measurement bias and feature interactions, that might undermine the quality of the model's predictions or create other problems. Model design documentation will typically address data selection and debiasing, including in certain cases the performance and reliability of particular debiasing techniques used to make the data or model appropriate for use and steps taken to adhere the firm's specific data governance and privacy policies, if applicable.

- **Risk management:** Model users will need to demonstrate that they understand the specific risks related to using a model in its proposed application and have an appropriate plan for monitoring and mitigating those risks. In the context of underwriting, this will normally include, for example, testing for fair lending risks and plans for monitoring those risks while the model is in use.

Further Reading

Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, Aram Galstyan, A Survey on Bias and Fairness in Machine Learning (September 2019), available at https://arxiv.org/pdf/1908.09635.pdf.

Related FAQs

What are the key policy debates about using AI in financial services?

What potential risks are important to consider when lenders replace incumbent underwriting models with machine learning?

What legal and regulatory frameworks apply to the use of machine learning credit underwriting models?

What kind of statistical biases are most important with respect to underwriting models?

What consumer protection frameworks apply to the use of machine learning credit underwriting models?

What are the sources of discrimination or unfairness in underwriting models and other predictive models?

What are the core concerns about adopting machine learning underwriting models from the perspective of fair lending and financial inclusion?

What kind of statistical biases are most important with respect to underwriting models?

Developers of underwriting models spend considerable energy trying to understand and correct for biases in their models in order to increase the accuracy of the model's predictions and limit lending losses once the model is in use. In this context, bias takes on a broader meaning than it has in anti-discrimination law and regulation and refers to systematic deviations between the model's predictions and observed results. Many forms of statistical biases can affect models and there is no standardized taxonomy of statistical biases, but the following are particularly relevant to the development and use of underwriting models:[37]

- **Representation Bias:** Occurs when defining and sampling a population to support development of a model. It reflects divergence in characteristics, behaviors, and outcomes for individuals in the data set used to develop the model and the data that the model will encounter when in use. Under-representation in the training data can mean that the model's predictions do not generalize well once the model is in use. Its causes include sampling methods that only reach a narrow population and changes between the population of interest and sample that are not captured in data used for model development.

- **Measurement Bias**: Arises when a variable in the model is mis-measured. A simple example is using test scores as a measurement of ability. This may mean that the model leaves out important factors or that the selection or creation of features or labels introduces group- or input-dependent noise that affects model performance. It can be caused by measurement processes that vary among groups, or by an oversimplified approach to defining the model's task.

- **Historical Bias:** Describes the effect that occurs when the data available from current or past practice is accurate and correctly sampled, but skewed in ways that means the model may produce outcomes that are not desirable from broader perspectives. For example, an algorithm designed to select which applicants for an engineering job or academic program merit interviews may successfully replicate the historical results from a period during which this decision resulted from human review of applications, but be nonetheless undesirable for institutions that want to include more women and minorities.[38]

- **Aggregation Bias:** Reflects the use of a generalized model for subpopulations with different conditional distributions. This may result in a model that is a poor descriptor of any one subpopulation or that describes only a dominant population among the sample. For example, in the development of medications, testing results for women of child-bearing age may be unduly affected by other populations, since clinical trials tend to include fewer participants in that subpopulation.

- **Omitted Variable Bias:** Occurs when a model's target variable is affected by an explanatory variable that is not included in the model.[39] For example, a model that was designed to predict hourly wages would ideally consider the education levels and innate ability of people in the data set. The latter is likely to be an omitted variable, because it difficult data to measure or obtain. The degree to which the omitted variable affects the overall accuracy of the model depends on the likely relationship of the omitted variable to those considered in the model. If people with more innate ability tend to be more productive for reasons not captured in the data on educational attainment, then this variable's omission is more likely to bias the model's predictions. These problems do not go away by expanding the observations in training data sets and are difficult to overcome entirely given the costs of and other practical constraints related to data acquisition. Unfortunately, omitted-variable bias is common, though the strength of correlation between omitted variables and protected attributes or other uncollected data varies.[40]

These biases can all result in faulty predictions and distortions of data. Many of these can also give rise to fair lending and discrimination problems, depending on the circumstances.

The shift to machine learning from incumbent underwriting models may amplify the importance of some of these risks, but it also presents an opportunity for practitioners and policymakers to rethink how credit is provided and to consider how AI can be adapted to help truly overcome the tendency in lending decisions to reflect past practices.

Further Reading

Jongbin Jung, Sam Corbett-Davies, Ravi Shroff, and Sharad Goel, Omitted and Included Variable Bias in Tests of Disparate Impact (August 29, 2019), available at https://arxiv.org/abs/1809.05651

Mehrabi, Morstatter, Saxena, Lerman, & Galstyan (2019).

Harini Suresh & John V. Guttag, A Framework for Understanding Unintended Consequences of Machine Learning (Feb. 17, 2020), available at https://arxiv.org/pdf/1901.10002.pdf

Related FAQs

What are the key policy debates about using AI in financial services?

What potential risks are important to consider when lenders replace incumbent underwriting models with machine learning?

What prudential frameworks apply to the use of machine learning credit underwriting models?

What are the sources of discrimination or unfairness in underwriting models and other predictive models?

What are the core concerns about adopting machine learning underwriting models from the perspective of fair lending and financial inclusion?

What consumer protection frameworks apply to the use of machine learning credit underwriting models?

A variety of consumer protection laws and regulations apply to credit underwriting. Fair lending and adverse action reporting are two areas where adapting machine learning underwriting models to existing requirements may be particularly challenging for firms.

Fair Lending

Bank and nonbank lenders have a general obligation to provide non-discriminatory access to credit under the Equal Credit Opportunity Act (ECOA) and the Fair Housing Act (FHA). ECOA makes three forms of discrimination against protected classes (including, but not limited to, race, ethnicity, sex, or age) unlawful:[41]

- **Overt Discrimination:** Overt discrimination involves blatant use of a protected class status for an impermissible purpose. A firm that offers loan with a limit of up to $500 for applicants of one race and $1,500 for applicants for all other applicants would engage in impermissible overt discrimination.

- **Disparate Treatment:** Disparate treatment occurs when a firm treats similarly situated applicants differently based a prohibited characteristic, like race or gender, with no credible, nondiscriminatory explanation even if there is no evidence of prejudice or a conscious intent to discriminate. Redlining based on neighborhood demographics is also considered an example of impermissible disparate treatment.

- **Disparate Impact:** Disparate impact occurs when a facially neutral policy or practice has a disproportionate impact on a protected class, unless that policy or practice meets a legitimate business need that cannot reasonably be achieved as well by alternatives that create less disparate impact. For example, a rule that prohibited applications for mortgages from people less than six feet tall would disproportionately exclude women and would not be justified under the other prongs of the analysis.

Managing fair lending risk can get more challenging in the context of machine learning underwriting models and less traditional data. With respect to both disparate treatment and disparate impact, the deeper insight machine learning models are able to derive, the scale of data they can use, and the evolution of rules they use to make decisions enhance the need for monitoring underwriting models closely while in use and using models where weight given to specific rules and variables can be reliably evaluated. Specific open questions include:

- Whether the evaluation takes into account protected class information

- Whether similarities and patterns inferred from the data create interactions that function as impermissible proxies for protected class information

- How to identify which variables or combinations of variables drive disparities in the frequency or pricing of offers made to different groups.

The more ambitious the expansion of data and the choice of specific modelling techniques, the more complicated the challenges of documenting and managing fair lending risks in credit models. At the same time, the arrival of machine learning underwriting models may usher in tools and oversight processes that provide lenders with more insight about the disparate impact in their models, which may in turn improve oversight of disparate impact for all models, not just those that use machine learning. Similarly, machine learning models used for underwriting or to debias underwriting models may also produce an enhanced array of options for mitigating disparate impact – essentially providing options to reduce discrimination that require fewer tradeoffs in the model accuracy because instead of removing variables that correlate with adverse impacts on protected classes, machine learning models can alter the influence of a variable's correlation with protected class status.[42]

Adverse Action Reporting

Requirements to provide an explanation to applicants who receive an adverse credit decision[43] pursuant to ECOA and the Fair Credit Reporting Act (FCRA) may represent the most technically difficult explainability challenge for lenders using machine learning underwriting models, even if the regulatory requirements provide lenders with some latitude as to how they determine and articulate the principal reasons for a particular decision.

Here, firms must provide the primary bases for a denial of credit or other adverse action. When adverse action is based in whole or in part on a credit score obtained from a consumer reporting agency, firms must disclose that score and key factors that adversely affected tit, the name and contact information of the score provider, and additional content.[44] These requirements "serve important anti-discrimination, educational, and accuracy purposes."[45]

Using machine learning heightens the challenges of mapping explanatory variables to reasons that generally apply to the provision of adverse action notices. There are potentially thousands of explanatory variables in a model and controlling for their interplay in a specific decision is needed to produce a statement of up to four primary bases for the lender's decision. Adding to this challenge is that the audience for this explanation is not a data scientist, a credit expert, or a regulatory lawyer – it is a person of ordinary experience and understanding.

The specific choices that a lender makes about the number of data inputs and features in their underwriting model are particularly relevant. The more ambitious the number of potential determinants of a prediction and the more complex their potential effect on each other, the more compression there will be in articulating the specific bases for a decision.[46]

Further Reading

Solon Barocas and Andrew D. Selbst, Big Data's Disparate Impact, California Law Review (September 30, 2016), available at https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2477899

Patrice Ficklin, Tom Pahl, and Paul Watkins, Innovation Spotlight: Providing Adverse Action Notices when using AI/ML Models, Consumer Financial Protection Bureau (July 7, 2020), available at https://www.consumerfinance.gov/about-us/blog/innovation-spotlight-providing-adverse-action-notices-when-using-ai-ml-models/

Related FAQs

What are the key policy debates about using AI in financial services?

What potential risks are important to consider when lenders replace incumbent underwriting models with machine learning?

What legal and regulatory frameworks apply to the use of machine learning credit underwriting models?

What prudential frameworks apply to the use of machine learning credit underwriting models?

What are the sources of discrimination or unfairness in underwriting models and other predictive models?

What are the core concerns about adopting machine learning underwriting models from the perspective of fair lending and financial inclusion?

How can we measure the fairness of a model?

What are the sources of discrimination or unfairness in underwriting models and other predictive models?

Fairness concerns – including but not limited to legally defined forms of discrimination such as disparate treatment and disparate impact – may be introduced into a model in a variety of ways. The main sources of unfairness include:

- **Data:** Models use historical data of one kind or another to make predictions about a future event or future behavior, though data may vary in how directly related it is to a prediction. If that data is unrepresentative or inaccurate or contains mistakes,[47] the model's predictions will be less reliable. In underwriting and credit scoring specifically, data used to model which people are more likely to default is primarily derived from prior lending activity. That means the data used to evaluate current applicants may not be able to assess well the credit risk posed by people who have not been able to obtain credit or have had to rely on products whose structure and terms increased their likelihood of default.[48]

- **Model Design and Governance:** Automated models, including AI models, can replace subjective decision-making processes that can be subject to unfairness and bias, but still require human decisions in design and governance than can introduce unfairness and bias.[49] Models may be designed in ways that reflect assumptions about economic structures or business models with embedded inequalities especially with respect to the design of the target variable the algorithm will optimize. For example, an AI algorithm that a British medical school used to determine which applicants to interview was found to be biased against women and those with non-European names.[50] The algorithm was designed to match human admissions decisions, with at least 90 percent accuracy. Similar issues may result where algorithms are designed to serve larger groups rather than distinct or differentiated subpopulations.

- **Personnel:** Lack of representativeness among personnel who design, operate, and govern models can increase the chances of biases being built into model architecture and design. This may also weaken organization's ability to recognize and respond to evidence of unfairness in all phases a model's development and use.[51]

Further Reading

Mehrabi, Morstatter, Saxena, Lerman, & Galstyan (2019).

Related FAQs

How are machine learning models developed?

What are the core concerns about adopting machine learning underwriting models from the perspective of fair lending and financial inclusion?

How can we measure the fairness of a model?

What are the core concerns about adopting machine learning underwriting models from the perspective of fair lending and financial inclusion?

Concern about bias and discrimination are among the most important obstacles to the widespread use of machine learning for credit underwriting. Three concerns about the models themselves are paramount:

- The ability of machine learning models to triangulate data points to determine protected class information, especially when using expansive data sets
- The challenge of detecting proxies for protected class information among much larger volume of data and the complexity of variable and feature interactions[52]
- The difficulty of identifying alternatives to proxies that create less disparate impact

Notwithstanding these concerns, there is also significant interest in the ability of machine learning to usher in more inclusive lending. Machine learning models may have the capacity to incorporate more forms of data than traditional forms of statistical prediction and can detect positive credit attributes with greater precision and faster than incumbent underwriting models.

There is a debate about whether the shift to AI and machine learning – with or without the use of expanded data – can be used in ways that will help overcome historical lending patterns or will simply replicate or enhance them. In response to concerns about data limitations and the risk of coding human bias into new technologies,[53] some argue that machines have a greater likelihood than humans of operating without bias.[54] Viewed in this light, machine learning approaches that reflect higher levels of human intervention – supervised and reinforcement learning – may also be prone to other sources of discrimination problems, since the decisions made to define the training data and impose constraints on the model may introduce further bias.

Further Reading

BLDS, LLC., Discover Financial Services Inc., and H2O.ai (2020).

Mehrabi, Morstatter, Saxena, Lerman, & Galstyan (2019).

Related FAQs

What is the basis for believing that machine learning could improve credit underwriting?

What legal and regulatory frameworks apply to the use of machine learning credit underwriting?

What kind of statistical biases are most important with respect to underwriting models?

What consumer protection frameworks apply to the use of machine learning credit underwriting models?

What are the sources of discrimination or unfairness in underwriting models and other predictive models?

How can we measure the fairness of a model?

How can we measure the fairness of a model?

The shift to machine learning has opened a significant debate among data scientists about how to measure fairness. That debate has spawned over twenty different metrics that can be used to measure fairness – some may be best used to identify an adverse analytical finding that constitutes a disparate impact and others may capture aspects of unfairness that neither disparate treatment nor disparate impact address. A further debate among law professors and practitioners assesses how well these metrics serve the purposes of current law and regulation[55] and meet existing requirements.[56]

These metrics can be thought of as falling into the following categories:[57]

- **Decision-Based Statistical Measures:** These encompass statistical parity and conditional statistical parity and answer the question: do outcomes systematically differ between particular population groups? A version of this approach is used at the first stage of disparate impact analyses under fair lending laws, which typically focus on whether there are substantial variations in approval rates among protected classes.

- **Binary Error Measures:** These all target the ratio of false positive rates (accepting a negative case) and false negative rates (rejecting a positive case).

- **Calibration Measures:** These account for probability statistics, such as an individual's probability of defaulting on a loan.

- **Input and Distance Measures:** These dictate either the inclusion or omission of protected classes, such as gender, in the model.

- **Counterfactual and Structural Measures:** These determine fairness through causal graphs and mapping relationships between variables. Counterfactual fairness is an approach that evaluates the effect of sensitive attributes by replacing those in the model.

- **Welfare-Based Measures:** These metrics focus on impact or the perceived benefit that groups receive, based on how model developers define the benefit in question.

Recent academic research has focused on trade-offs involved in using these fairness metrics, finding that optimizing models for one measure of fairness generally has the effect of making others deteriorate and that conditions for avoiding those trade-offs are exceedingly rare.[58]

Data availability may limit the utility of specific metrics in particular contexts, and further research is needed to understand the benefits and tradeoffs related to the use of individual alternative fairness metrics.

Further Reading

Solon Barocas and Moritz Hardt, Fairness in Machine Learning, NeurIPS (December 4, 2017), available at https://nips.cc/Conferences/2017/Schedule?showEvent=8734

Solon Barocas, Moritz Hardt, and Arvind Narayanan, Chapter 2: Classification, Fairness, and Machine Learning (December 6, 2019), available at https://fairmlbook.org/classification.html

Deborah Hellman, Measuring Algorithmic Fairness, Virginia Law Review, Forthcoming (July 16, 2019), available at https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3418528

Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan, Inherent Trade-Offs in the Fair Determination of Risk Scores, Proceedings of Innovations in Theoretical Computer Science (November 17, 2016), available at https://arxiv.org/pdf/1609.05807.pdf

Arvind Narayanan, Translation Tutorial: 21 Fairness Definitions and Their Politics, Conference on Fairness, Accountability, and Transparency (Feb. 23, 2018), available at https://www.youtube.com/watch?v=jIXIuYdnyyk

Nicholas Schmidt and Bryce Stephens, An Introduction to Artificial Intelligence and Solutions to the Problems of Algorithmic Discrimination (November 8, 2019), available at https://arxiv.org/pdf/1911.05755.pdf

Sahil Verma and Julia Rubin, Fairness Definitions Explained, ACM/IEEE International Workshop on Software Fairness (May 29, 2018), available at http://fairware.cs.umass.edu/papers/Verma.pdf

Related FAQs

What are the key policy debates about using AI in financial services?

What kind of statistical biases are most important with respect to underwriting models?

What consumer protection frameworks apply to the use of machine learning credit underwriting models?

What are the sources of discrimination or unfairness in underwriting models and other predictive models?

What are the core concerns about adopting machine learning underwriting models from the perspective of fair lending and inclusion?

Endnotes

What is model transparency? Why do we need it?

[1] Christoph Molnar, Interpretable Machine Learning: A Guide for Making Black Boxes Explainable (2019), available at https://christophm.github.io/interpretable-ml-book/.

[2] Jonathan Johnson, Interpretability vs. Explainability: The Black Box of Machine Learning, BMC (July 16, 2020), available at https://www.bmc.com/blogs/machine-learning-interpretability-vs-explainability/; Leilani Gilpin, David Bau, Ben Yuan, Ayesha Bajwa, Michael Specter, and Lalana Kagal, Explaining Explanations: An Overview of Interpretability of Machine Learning, Massachusetts Institute of Technology (February 3, 2019), available at https://arxiv.org/pdf/1806.00069.pdf.

What is model interpretability?

[3] Finale Doshi-Velez and Been Kim, Towards a Rigorous Science of Interpretable Machine Learning (March 2, 2017), available at https://arxiv.org/pdf/1702.08608.pdf; The Royal Society, Explainable AI: The Basics (November 2019), available at https://royalsociety.org/-/media/policy/projects/explainable-ai/AI-and-interpretability-policy-briefing.pdf.

[4] Johnson (2020).

[5] Patrick Hall and Navdeep Gill, An Introduction to Machine Learning Interpretability: An Applied Perspective on Fairness, Accountability, Transparency, and Explainable AI, Dataiku (2018), available at https://pages.dataiku.com/hubfs/ML-interperatability.pdf.

[6] Arun Rai, Explainable AI: From Black Box to Glass Box, Journal of the Academy of Marketing Science (December 17, 209), available at https://link.springer.com/article/10.1007/s11747-019-00710-5.

[7] Cynthia Rudin, Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead, Nature Machine Intelligence (May 13, 2019), available at https://www.nature.com/articles/s42256-019-0048-x; Maya Gupta, Andrew Cotter, Jan Pfeifer, Konstantin Voevodski, Kevin Canini, Alexander Mangylov, Wojciech Moczydlowski, Alexander van Esbroeck, Monotonic Calibrated Interpolated Look-Up Tables, Journal of Machine Learning Research (2016), available at https://jmlr.org/papers/v17/15-243.html.

[8] Diogo Carvalho, Eduardo Pereira, Jaime Cardoso, Machine Learning Interpretability: A Survey on Methods and Metrics, Electronics (July 26,2019), available at https://www.mdpi.com/2079-9292/8/8/832.

[9] Gilpin, Bau, Yuan, Bajwa, Specter, and Kagal (2019).

[10] Hall & Gill (2018).

What is model explainability?

[11] The Royal Society, (2019).

[12] *Ibid.*

[13] Rai (2019); Amanda Thomas, Model Transparency and Explainability, Ople (March 24, 2020), available at https://ople.ai/ai-blog/model-transparency-and-explainability/.

[14] Molnar (2019); Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador García, Sergio Gil-López, Daniel Molina, Richard Benjamins, Raja Chatila, Francisco Herrera, Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges toward Responsible AI (December 26, 2019), available at https://arxiv.org/pdf/1910.10045.pdf; Rudin (2019); Cynthia Rudin and Joanna Radin, Why Are We Using Black Box Models in AI When We Don't Need To? A Lesson from an Explainable AI Competition, Harvard Data Science Review (November 22, 2019), available at https://hdsr.mitpress.mit.edu/pub/f9kuryi8/release/5

Why is model transparency especially important in the context of AI and machine learning models?

[15] James Manyika, Jake Silberg, and Brittany Presten, What Do We Do About the Biases in AI?, Harvard Business Review (October 25, 2019), available at https://hbr.org/2019/10/what-do-we-do-about-the-biases-in-ai; Jeffrey Dastin, Amazon Scraps Secret AI

Recruiting Tool that Showed Bias Against Women, Reuters (Oct. 8, 2018), https://reut.rs/2Po4ZJi; Nicole Martin, The Main Concerns About Facial Recognition Software, Forbes (September 25, 2019), available at https://www.forbes.com/sites/nicolemartin1/2019/09/25/the-major-concerns-around-facial-recognition-technology/#67e57a34fe3e; Julie Angwin & Jeff Larson, Bias in Criminal Risk Scores is Mathematically Inevitable, Researchers Say, ProPublica (Dec. 30, 2016), available at https://www.propublica.org/article/bias-in-criminal-risk-scores-is-mathematically-inevitable-researchers-say.

[16] XBRL, Data Amplified 2019: Explainable AI in Finance (October 24, 2019), available at https://www.xbrl.org/news/data-amplified-2019-explainable-ai-in-finance/; Ron Schmelzer, Understanding Explainable AI, Forbes (July 23, 2019), available at https://www.forbes.com/sites/cognitiveworld/2019/07/23/understanding-explainable-ai/#70e65da97c9e.

What techniques can make machine learning models more transparent?

[17] Subsequent FAQs will explore the data science behind *post hoc* explainability techniques.

[18] Molnar (2019).

[19] Subsequent FAQs will explore the data science behind *post hoc* explainability techniques

[20] Scott Lundberg, Gabriel Erion, and Su-In Lee, Consistent Individualized Feature Attribution for Tree Ensembles, University of Washington (March 7, 2019), available at https://arxiv.org/pdf/1802.03888.pdf.

How are *post hoc* explainability techniques being used in practice?

[21] BLDS, LLC., Discover Financial Services Inc., and H2O.ai, Machine Learning: Considerations for Expanding Access to Credit Fairly and Transparently, H2O.ai (July 2020), available at https://www.h2o.ai/resources/white-paper/machine-learning-considerations-for-fairly-and-transparently-expanding-access-to-credit/.

[22] Patrick Hall, SriSatish Ambati, Wen Phan, Ideas on Interpreting Machine Learning, O'Reilly (March 15, 2017), available at https://www.oreilly.com/radar/ideas-on-interpreting-machine-learning/.

What are global and local model explanations?

[23] Molnar (2019); Philippe Bracke, Anupam Datta, Carsten Jung, and Shayak Sen, Machine Learning Explainability in Finance: An Application to Default Risk Analysis, Bank of England (August 2019), available at https://www.bankofengland.co.uk/-/media/boe/files/working-paper/2019/machine-learning-explainability-in-finance-an-application-to-default-risk-analysis.pdf?la=en&hash=692E8FD8550DFBF5394A35394C00B1152DAFCC9E.

[24] Hall & Gill (2018).

[25] Leon Kopitar, Leona Cilar, Primoz Kocbek, Gregor Stiglic, Local vs. Global Interpretability of Machine Learning Models in Type 2 Diabetes Mellitus Screening, International Workshop on Knowledge Representation for Health Care (January 3, 2020), available at https://link.springer.com/chapter/10.1007/978-3-030-37446-4_9.

[26] Doshi-Velez and Kim (2017).

[27] Hall & Gill (2018).

How can machine learning be used in credit underwriting?

[28] Sushmito Ghosh and Douglas L. Reilly, "Credit card fraud detection with a neural-network," The Twenty-Seventh Hawaii International Conference on System Sciences (January 1994), available at https://ieeexplore.ieee.org/document/323314.

Why is model transparency especially important in the context of machine learning models used for credit underwriting?

[29] Ning Xie, Gabrielle Ras, Marcel van Gerven, and Derek Doran, Explainable Deep Learning: A Field Guide for the Uninitiated (April 30, 2020), available at https://arxiv.org/abs/2004.14545.

What potential risks are important to consider when lenders replace incumbent underwriting models with machine learning?

[30] For example, online retailers' algorithms governing inventory, fraud, and marketing went haywire when demand shifted dramatically from phone accessories and toys to toilet paper, disinfectant wipes, and kettle bells. Similarly, machine learning underwriting models that lenders rely on to detect heterogeneity within credit bands are unproven in a downturn and may reflect "brittle correlations calculated during good economic times." Both examples show the need for informed oversight of machine learning models as used with traditional predictive methods for the same purposes. (See: Will Douglas Heaven, Our Weird Behavior During the Pandemic is Messing with AI Models, MIT Technology Review (May 11, 2020), available at https://www.technologyreview.com/2020/05/11/1001563/covid-pandemic-broken-ai-machine-learning-amazon-retail-fraud-humans-in-the-loop/; Jacob Kosoff, BankThink AI Models Could Struggle to Handle the Market Downturn, American Banker (March 19, 2020), available at https://www.americanbanker.com/opinion/ai-models-could-struggle-to-handle-the-market-downturn)

[31] Jennifer Miller, Is an Algorithm Less Racist than a Loan Officer?, The New York Times (September 18, 2020), available at https://www.nytimes.com/2020/09/18/business/digital-mortgages.html; Kyle Wiggers, Facebook's Discriminatory Ad Targeting Illustrates the Dangers of Biased Algorithms, Venture Beat (August 28, 2020), available at https://venturebeat.com/2020/08/28/ai-weekly-facebooks-discriminatory-ad-targeting-illustrates-the-dangers-of-biased-algorithms/; Llana James, Race-Based COVID-19 Data May Be Used to Discriminate Against Racialized Communities, The Conversation (September 14, 2020), available at https://theconversation.com/race-based-covid-19-data-may-be-used-to-discriminate-against-racialized-communities-138372; Karen Hao, The UK Exam Debacle Reminds Us that Algorithms Can't Fix Broken Systems, MIT Technology Review (August 20, 2020), available at https://www.technologyreview.com/2020/08/20/1007502/uk-exam-algorithm-cant-fix-broken-system/.

What legal and regulatory frameworks apply to the use of machine learning credit underwriting models?

[32] Federal Reserve Board, Federal Fair Lending Regulations and Statutes, available at https://www.federalreserve.gov/boarddocs/supmanual/cch/fair_lend_over.pdf.

[33] Board of Governors of the Federal Reserve System, Office of the Comptroller of the Currency, Supervisory Guidance on Model Risk Management (April 4, 2011), available at https://www.federalreserve.gov/supervisionreg/srletters/sr1107a1.pdf ; Federal Deposit Insurance Corporation, Adoption of Supervisory Guidance on Model Risk Management (June 7, 2017), available at https://www.fdic.gov/news/financial-institution-letters/2017/fil17022.html.

[34] 12 CFR Part 1002 Supp. I Sec. 1002.4(a)-1, 1002.6(a)-2.

[35] 12 U.S.C. § 1691(d); 12 C.F.R. § 1002.9; 15 U.S.C. § 1681;12 C.F.R. 1022.70-75, 1022.130, 1022.136-38; see also Patrice Alexander Ficklin, Tom Pahl, Paul Watkins, Innovation Spotlight: Providing Adverse Action Notices when using AI/ML Models, Consumer Financial Protection Bureau (July 7, 2020), available at https://www.consumerfinance.gov/about-us/blog/innovation-spotlight-providing-adverse-action-notices-when-using-ai-ml-models/.

What prudential frameworks apply to the use of machine learning credit underwriting models?

[36] Board of Governors of the Federal Reserve System & Office of the Comptroller of the Currency (2011); Federal Deposit Insurance Corporation (2017); Office of the Comptroller of the Currency, Sound Practices for Model Risk Management: Supervisory Guidance on Model Risk Management (April 4, 2011), available at https://www.occ.gov/news-issuances/bulletins/2011/bulletin-2011-12a.pdf.

What kind of statistical biases are most important with respect to underwriting models?

[37] Harini Suresh & John V. Guttag, A Framework for Understanding Unintended Consequences of Machine Learning, MIT (Feb. 17, 2020), available at https://arxiv.org/pdf/1901.10002.pdf

[38] Manyika, Silberg, and Presten (2019).

[39] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, Aram Galstyan, A Survey on Bias and Fairness in Machine Learning (September 2019), available at https://arxiv.org/pdf/1908.09635.pdf; Jongbin Jung, Sam Corbett-Davies, Ravi Shroff, and Sharad Goel, Omitted and Included Variable Bias in Tests for Disparate Impact (August 30, 2019), available at https://arxiv.org/pdf/1809.05651.pdf

[40] *Ibid.*

What consumer protection frameworks apply to the use of machine learning credit underwriting models?

[41] Consumer Financial Protection Bureau, Equal Credit Opportunity Act (ECOA) examination procedures (October 30, 2015), available at https://www.consumerfinance.gov/policy-compliance/guidance/supervision-examinations/equal-credit-opportunity-act-ecoa-examination-procedures/.

[42] Nicholas Schmidt and Bryce Stephens, An Introduction to Artificial Intelligence and Solutions to the Problems of Algorithmic Discrimination (November 8, 2019), available at https://arxiv.org/abs/1911.05755; BLDS, LLC., Discover Financial Services Inc., and H2O.ai, (2020).

[43] An adverse action is a credit decision in which a lender declines to provide credit in the amount or terms requested, provides credit on terms that are materially different to a substantial proportion of consumers, or makes a negative change to an existing account. Federal law requires lenders to provide disclosures to consumers and small businesses after taking an adverse action to explain the principal reason(s) for the decision. A statement of the principal reasons for a decision need not describe how or why the disclosed factor(s) adversely affected the application or how the factor relates to creditworthiness in the derivation of a score. (see 12 CFR pt. 1002, comment 9(b)(2)-3, 4).

[44] 12 U.S.C. § 1691(d); 12 C.F.R. § 1002.9. 15 U.S.C. § 1681m(a), (b); Interagency Alternative Data Statement at 2. See also p. 90, FinRegLab, the Use of Cash-Flow Data in Underwriting Credit (February 2020), available at https://finreglab.org/wp-content/uploads/2020/03/FinRegLab_Cash-Flow-Data-in-Underwriting-Credit_Market-Context-Policy-Analysis.pdf; Patrice Alexander Ficklin, Tom Pahl, Paul Watkins, Innovation Spotlight: Providing Adverse Action Notices when using AI/ML Models, Consumer Financial Protection Bureau (July 7, 2020), available at https://www.consumerfinance.gov/about-us/blog/innovation-spotlight-providing-adverse-action-notices-when-using-ai-ml-models/.

[45] Patrice Alexander Ficklin, Tom Pahl, Paul Watkins, Innovation Spotlight: Providing Adverse Action Notices when using AI/ML Models, Consumer Financial Protection Bureau (July 7, 2020), available at https://www.consumerfinance.gov/about-us/blog/innovation-spotlight-providing-adverse-action-notices-when-using-ai-ml-models/.

[46] Patrick Hall, SriSatish Ambati, Wen Phan, Ideas on Interpreting Machine Learning, O'Reilly (March 15, 2017), available at https://www.oreilly.com/radar/ideas-on-interpreting-machine-learning/.

What are the sources of discrimination or unfairness in underwriting models and other predictive models?

[47] Solon Barocas, Moritz Hardt, and Arvind Narayanan, Legal Background and Normative Questions, Fairness and Machine Learning (December 6, 2019), available at https://fairmlbook.org/.

[48] Mark MacCarthy, Fairness in Algorithmic Decision Making, Brookings Institution (December 6, 2019), available at https://www.brookings.edu/research/fairness-in-algorithmic-decision-making/; Schmidt and Stephens (2019).

[49] Schmidt and Stephens (2019).; Nicol Turner Lee, Paul Resnick, Genie Barton, Algorithmic Bias Detection and Mitigation: Best Practices and Policies to Reduce Consumer Harms, Brookings Institution (May 22, 2019), available at https://www.brookings.edu/research/algorithmic-bias-detection-and-mitigation-best-practices-and-policies-to-reduce-consumer-harms/; BLDS, LLC., Discover Financial Services Inc., and H2O.ai, (2020).

[50] Manyika, Silberg, and Presten (2019).

[51] National Fair Housing Alliance, Fair Housing Groups Settle Lawsuit with Facebook: Transforms Facebook's Ad Platform Impacting Millions of Users (March 18, 2019), available at https://nationalfairhousing.org/2019/03/18/national-fair-housing-alliance-settles-lawsuit-with-facebook-transforms-facebooks-ad-platform-impacting-millions-of-users/; Kenneth Holstein, Jennifer Wortman Vaughan, Hal Daumé III, Miro Dudík, Hanna Wallach, Improving Fairness in Machine Learning Systems: What do Industry Practitioners Need?, 2019 ACM CHI Conference on Human Factors in Computing Systems (January 7, 2019), available at https://arxiv.org/abs/1812.05239.

What are the core concerns about adopting machine learning underwriting models from the perspective of fair lending and financial inclusion?

[52] MacCarthy (2019).

[53] Chris DeBrusk, the Risk of Machine Learning Bias (And How to Prevent it), MIT Sloan Management Review (March 26, 2018), available at https://sloanreview.mit.edu/article/the-risk-of-machine-learning-bias-and-how-to-prevent-it/; Kathryn Hume and Alex LaPlante, Managing Bias and Risk at Every Step of the AI-Building Process, Harvard Business Review (October 30, 2019), available at https://hbr.org/2019/10/managing-bias-and-risk-at-every-step-of-the-ai-building-process; Tobias Baer and Vishnu Kamalnath, Controlling Machine Learning Algorithms and their Biases, McKinsey & Company (November 10, 2017), available at https://www.mckinsey.com/business-functions/risk/our-insights/controlling-machine-learning-algorithms-and-their-biases#.

[54] Sendhil Mullainathan, Biased Algorithms Are Easier to Fix Than Biased People, The New York Times (December 6, 2019), available at https://www.nytimes.com/2019/12/06/business/algorithm-bias-fix.html.

How can we measure the fairness of a model?

[55] Deborah Hellman, Measuring Algorithmic Fairness, Virginia Law Review, Forthcoming (July 16, 2019), available at https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3418528; Sam Corbett-Davies and Sharad Goel, The Measure and Mismeasure of Fairness: A Critical Review of Fair Machine Learning, Stanford University (September 11, 2018), available at https://arxiv.org/abs/1808.00023; MacCarthy (2019); Nina Grgic-Hlaca, Muhammad Bilal Zafar, Krishna Gummadi, and Adrian Weller, The Case for Process Fairness in Learning: Feature Selection for Fair Decision Making, NIPS 2016 (2016), available at http://www.mlandthelaw.org/papers/grgic.pdf.

[56] Rules promulgated under Title VII of the Civil Rights Act of 1964 have defined "adverse impact" in the context of disparate impact as a substantially different rate of selection. 29 CFR 1607.16(B). Those rules also define "unfairness." 29 CFR 1607.14(8). Guidance from the Equal Employment Opportunity Commission also addresses the differences between unfairness, differential validity, and differential prediction. See: https://www.eeoc.gov/policy/docs/qanda_clarify_procedures.html.

[57] Sahil Verma and Julia Rubin, Fairness Definitions Explained, ACM/IEEE International Workshop on Software Fairness (May 29, 2018), available at http://fairware.cs.umass.edu/papers/Verma.pdf; Solon Barocas and Moritz Hardt, Fairness in Machine Learning, NeurIPS (December 4, 2017), available at https://nips.cc/Conferences/2017/Schedule?showEvent=8734

[58] Arvind Narayanan, Translation Tutorial: 21 Fairness Definitions and Their Politics, Conference on Fairness, Accountability, and Transparency (Feb. 23, 2018), available at https://www.youtube.com/watch?v=jIXIuYdnyyk ; Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan, Inherent Trade-Offs in the Fair Determination of Risk Scores, Proceedings of Innovations in Theoretical Computer Science (November 17, 2016), available at https://arxiv.org/pdf/1609.05807.pdf.