WHITE PAPER

Model Explainability—Re-explained

TABLE OF CONTENTS

ZEST AI

Lenders have been increasingly adopting and including AI-driven technology. However, the adoption of newer technology over legacy methods can seem daunting. When adopting new technology, such as Zest AI machine learning (ML) models, creditors must be able to answer questions as to how new technology meets long-standing requirements of transparency and accuracy to both end-consumers and internal stakeholders.

One of the most basic questions a creditor must be able to answer when reviewing an application for credit is why an applicant may be denied. This is true whether the creditor is using a paper scorecard or a machine learning model.

Zest AI has always held itself to the long-standing requirements of Regulation B, that all credit underwriting models, irrespective of technology, must be able to provide the actual and accurate reasons for reaching its decision. Zest AI pioneered the use of rigorous, game-theoretic methods of explaining machine learning credit underwriting models and has led the discussion of how inferior methods can result in inaccurate notices and undermine a lender's ability to comply with consumer lending law.

In this whitepaper, after a brief introduction, we (i) describe the technology Zest AI uses to enable users to understand their ML models and how to provide explanations of denied credit to consumers, (ii) address arguments that legacy industry participants have promulgated in an attempt to sway the industry towards so-called "interpretable" models, (iii) distinguish Zest AI's methods from the problematic "post hoc" methods and (iv) discuss the results of a recent experiment Zest AI conducted, empirically validating the accuracy of its explainability technology. You will find that both the science and law support the methods Zest AI employs to provide accurate adverse action reasons from machine learning models used to underwrite consumer loans.

What's a model and what's a model explanation?

A model is a system of rules or mathematical expressions that takes a collection of inputs (called attributes or variables) and produces a score. One famous example is a legacy credit score, which is frequently used to assess the likelihood a consumer will repay a loan or be a responsible tenant or employee. When we talk about "explaining" a model or allowing someone to "interpret" the model, we mean helping a human understand how the inputs into the model (bankruptcies, on-time payments, income, etc.) contribute to the output (a credit score).

In financial services, explainability is required to comply with long-standing law. This is true if you are still using a paper credit score card or "advanced decision" technologies like AI/ML models. If a lender cannot accurately tell consumers which features in the model triggered a denial of credit for the consumer, then the lender can't use the model. Simple models composed of a series of if/then rules are seemingly easy to explain: you can look at the model and see what the rules are and trace the model's logic to the outcome, approved or denied.

This is one of the reasons that Zest AI has always been at the forefront of advanced explainability, and its proprietary and patent-pending technologies ensure that it can accurately explain and return model reasons. While there has been a rash of proposed state legislation directed to AI governance, in the world of financial services, technology-agnostic laws already exist to direct lenders in their use of AI/ML decisioning technology in a safe and compliant manner.

Models like logistic regression models, neural networks, or decision trees ***seem*** less straightforward to explain because the rules they utilize are more complex. (Note the qualifier "seem" in the preceding sentence. More below.)

It has historically been a challenge for model developers to create accurate explanations for more complex models. Many practitioners did not know the methods required to accurately analyze the model to understand which inputs drove the outcome, so some modeling methods were called "black boxes" and deemed impossible to explain. Machine learning models were often lumped into that category and thus deemed unsuitable for making lending decisions.

Shapley's mathematical tools and proofs are the only defensible way to explain how ML models make decisions. The "players" in the games he studied correspond to the variables in ML models. Similarly, the "games" are like the models, and the final "scores" are like the model outputs (in credit, usually the probability of defaulting on a loan). Since then, academics and researchers have consistently applied Shapley's method to explain and interpret ML models.

In 2017, two published papers proposed applying Shapley's methods to explaining machine learning models. These papers also validated the methods as being accurate. The first,[2] by Sundararajan and Yan, found that Shapley's methods "can be applied to a variety of deep networks, and [have] a strong theoretical justification." (Id. at p. 8.) The second,[3] by Lundberg and Lee, demonstrated how certain computer-based implementations of Shapley's theories, when used to explain ML models, "show improved computational performance and/or better consistency with human intuition than previous approaches." (Id. at p. 1.) In other words, computer-based implementations of Shapley's methods provide accurate and intuitive explanations of ML models. Several more recent studies have further validated these findings, showing that Shapley values can be used to explain particular model decisions (referred to as "local" explanations) as well as model behavior overall (referred to as "global" explanations).[4] For example, in his 2020 paper, Lundberg explains that "we developed an algorithm that computes local explanations based on exact Shapley values in polynomial time. This provides local explanations with theoretical guarantees of local accuracy and consistency" (Id. at p. 1.). He also "show[s] that combining many local explanations lets us represent global structure while retaining local faithfulness to the original model, which produces detailed and accurate representations of model behavior." Zest AI has relied on the Shapley values to generate model explanations and model risk documentation for its clients long before Lundberg wrote his 2020 paper, published in the journal **Nature**,.

In March 2022, Stanford University partnered with FinRegLab[5] to conduct a study on the use of AI and ML in credit underwriting.

Their report noted that "[a]mong the model diagnostic tools we evaluated, some tools can reliably identify features in the model that are related to adverse credit decisions for individual loan applicants." (Id. at p. 8.) They found that "[t]he high-fidelity tools all use a version of Shapley Additive Explanation ('SHAP') feature importance measures, and identify drivers as those with the largest positive values (contributing most to a high default prediction for a particular applicant)." (Id. at p. 29.) Zest AI's technology (tested in the Stanford/FinRegLab study) uses proprietary extensions of the SHAP methods that do not change the fundamental properties that render explanations valid.

(1) Lloyd S. Shapley, Notes on the n-Person Game– II: The Value of an n-Person Game, Rand Corp. (1951).

(2) Mukund Sundararajan et al., Axiomatic Attribution for Deep Networks (2017).

(3) Scott Lundberg & Su-In Lee, A Unified Approach to Interpreting Model Predictions (2017).

(4) Scott Lundberg et al., From local explanations to global understanding with explainable AI for trees (2020).

(5) FinRegLab et al., Machine Learning Explainability & Fairness: Insights from Consumer Lending (2022).

Some have advocated using certain breeds of interpretable models, which are less accurate and still hard to understand without the help of computers

Despite the overwhelming academic and industry support for the accuracy of explanations produced by Shapley-based methods, some legacy players in the financial services industry and a handful of academics have argued that models should meet the standard of being "inherently or intrinsically interpretable." However, the term "interpretable" is important to understand in the context of machine learning models, as it can be used in misleading ways.

As applied to modeling, the term "interpretable" lacks a well-accepted meaning. Some refer to Shapley-based methods as forms of post-hoc interpretability and refer to tree-based machine learning models as being interpretable.[6]

Some use the term to mean that models must be simple enough that a practitioner can understand why the model generated a given output just by looking at the model's equation.[7] The latter view suffers from several significant defects.

First, and most fundamentally, this view does not accurately describe the real-world process of explaining even traditional models. Practitioners today rely on "post-hoc" methods executed on computers to produce reason codes, even for simple models. They also use computers and "post hoc" methods to test the validity of the reason codes empirically. No one sits down with paper and pencil to calculate the impact of the various features on a lending decision. Even if they did, it's hard to see how that would benefit consumers.

Second, so-called "interpretable" models cannot be fully understood by looking at the model's equation. Even so-called "simple" underwriting models still rely on a few dozen features, each with its own weight or importance.[8] Model features are also frequently "compound" features or ratios of

multiple data points, making them harder to interpret. What's more, the value produced by the algorithm can't even be used in raw form. Instead, the value has to be normalized across the entire population of credit applicants. (It's no use seeing a raw value without understanding the value in the context of other data points.) So, while a reasonably-skilled credit analyst might be able to use a paper and pencil to compute a score, they could not look at the algorithm and intuit how an input value will impact a credit decision without knowing more.

Third, even if one could gather some rudimentary intuition by looking at a credit model, that intuition is meaningless without also understanding how the model behaves on a range of inputs and how those inputs are distributed in the model development data. Just because a model says it weights a variable highly doesn't mean that that variable has a practical impact. The variable might not change enough from applicant to applicant to impact the model's calculation of the applicant's default risk. To properly gather the desired intuition, an analyst would have to have computer-generated read-outs showing the distribution of each feature in the development data and its impact on the model's score.

Fourth, choosing a more "interpretable" model almost always means giving up significant accuracy. In credit risk assessment, lower accuracy can only lead to two things: (1) loans given to consumers who cannot repay them, which leads to higher default rates, more collection activity, and in some instances, even bankruptcy and financial ruin for those consumers who were granted loans inappropriately, or (2) fewer credit-worthy consumers being granted loans, which locks deserving borrowers out of financial services and impairs the social mobility of those borrowers.

Inaccurate models disproportionately affect African American and Hispanic borrowers because African American and Hispanic borrowers are more likely to be denied than whites, especially when oversimplified, legacy modeling techniques make the decisions. According to a recent study of HMDA data, African American borrowers are denied 3x more often than whites for home loans (15% denial rate for African American borrowers vs. 5% for whites).[9] In our work with lenders, we

have seen that switching to more accurate machine learning models can significantly improve financial inclusion, in many cases increasing approval rates for Black and Hispanic borrowers by 40% without increasing the risk of default. These results not only improve financial outcomes, but they are also consistent with the goals eradicating discrimination and barriers to access in the consumer financial services marketplace.

Virtually all key factors for the purposes of generating adverse action notices are computed using “post hoc” methods with the help of computers, even for purportedly intrinsically interpretable models. They are audited and validated using computers. Relying on human intuition to validate algorithms—even simple ones—is dangerous and misleading. Perhaps more importantly, it undermines the goal of advancing financial inclusion by unnecessarily restricting the types of models used. What matters is not whether you can build an intuition about the model just by looking at it but whether the tools used to interpret models are valid theoretically and empirically.

(6) C. Molnar, A Guide for Making Black Box Models Explainable, Interpretable Machine Learning (Mar. 29, 2022), https://christophm.github.io/interpretable-ml-book/.

(7) Agus Sudjianto & Scott Zoldi, Breaking Down “Black Box” AI with Interpretable Models Linkedin Live, Youtube (Apr. 21, 2022), https://www.youtube.com/watch?v=F-8PNWimSHc; Agus Sudjianto & Aijun Zhang, Designing Inherently Interpretable Machine Learning Models (2021); Agus Sudjianto, Interpretable Machine Learning (2019); Steve Marlin, Wells touts new explainability technique for AI credit models, Risk.net (Aug. 16, 2021), https://www.risk.net/risk-management/7865541/wells-touts-new-explainability- technique-for-ai-cr-edit-models.

(8) This applies to simple models like logistic regression and more complex models designed to be more interpretable, such as those proposed in Agus Sudjianto et al., Linear Iterative Feature Embedding: An Ensemble Framework for Interpretable Model (2021). It is not obvious how these more complex models which borrow structure from neural networks make their decisions just by inspecting the model, they require post-hoc analysis. As the paper admits, “a new interpretation tool is introduced to detect main and interaction effects”.

(9) Black and Hispanic people have been found to be more likely to be denied mortgage loans. Zeninjor Enwemeka et al., Black and Hispanic people are more likely to be denied mortgage loans in Boston, WBUR (Mar. 30, 2022), https://www.wbur.org/news/2022/03/30/home-loans-mortgages-boston-denials.

Shapley-based methods are not based on model approximations

Some early attempts to explain and interpret complex models were not accurate. Some practitioners began using those methods in an effort to comply with the adverse action requirement, even though the methods they were using were not theoretically and empirically sound. In doing so, it hurt consumers in more ways than one: it likely meant that they received inaccurate denial reasons in their adverse action notices, and it slowed the acceptance of machine learning underwriting and its unique ability to expand access to credit to women, people of color, thin-file borrowers, and others.

Permutation feature importance, also referred to as drop one, is one early technique used (wrongly) to explain machine learning models. With drop one, lenders test which model variables contribute most to the model score by removing one variable from the model and measuring the change in the score as a means of quantifying the importance or influence of the removed variable.

Drop one essentially says, “Let’s see whether so and so would have been denied if they didn’t have X variable in their credit file or if their X variable were closer to everyone else’s.”

While that sounds reasonable and is a method commonly used to explain logistic regression models, it’s not accurate in the machine learning context because drop one can’t account for variable interactions, which frequently occur in ML models. For this reason, there is no support for using such methods to explain ML models. Zest AI does not and has never used this method and has been vocal about the dangers of using it for many years: such as in the CFPB’s October 2020 **Tech Sprint on Adverse Action Notices**, and in a December 18, 2020 **letter** responding to the CFPB’s request for information on AI/ML in financial services.

Similarly, local interpretable model-agnostic explanations (or "LIME") is another method that seemed to have promise early on but has also shown to be inaccurate. The LIME technique involves approximating the machine learning model using a series of linear models and then explaining the linear models. It's essentially like using a series of straight lines to approximate and explain a curvy one. LIME was shown to be inaccurate when used to produce explanations of machine learning underwriting models in the recent Stanford / FinRegLab study[10]. The CFPB rightly criticized methods where the "explanations approximate models." Zest AI doesn't use LIME or methods like it.

In sharp contrast, the Shapley-based methods that Zest AI uses are assuredly not based on model approximations. Lundberg's 2020 paper describes the Shapley-based algorithms he implemented as "directly measur[ing] local feature interaction effects,"[11] i.e., not measuring approximations. "By focusing specifically on trees [in the paper], we developed an algorithm that computes local explanations based on exact Shapley values in polynomial time," Lundberg states. "This provides local explanations with theoretical guarantees of local accuracy and consistency." (Id. p. 1., emphasis added)

Even proponents of so-called "interpretable" models recognize that Shapley-based explainability methods are not based on model approximations. Sudjianto, for example, classified explainability research as follows: "There are, broadly speaking, three inter-related model-based areas of research: (a) global diagnostics (Sobol & Kucherenko (2009) [16], Kucherenko (2010) [17]); (b) local diagnostics (Sundararajan et al. (2017) [24], Ancona et al. (2018) [2]); and c) development of approximate or surrogate models that may be easier to understand and explain."[12] Shapley-based methods fall into the "local diagnostics" category as evidenced by the reference and citation to Sundararajan's 2017 paper.

Sudarajan and Yan's Integrated Gradients method, which is the Shapley-based method Zest uses to explain neural networks, relies on the model's gradient itself, which can be directly retrieved from the model: "It can be implemented using a few calls to the gradients operator, can be applied to a variety

of deep networks, and has a strong theoretical justification."[13] (Page 8). Again, this method does not rely on an approximation of the model. It relies on the analysis of the model itself to determine how the model generates outcomes. Zest AI used these methodologies to generate accurate adverse action reasons long before any of these papers were published.

(10) FinRegLab et al., Machine Learning Explainability & Fairness: Insights from Consumer Lending 29 (2022).

(11) Scott M. Lundberg et al., From local explanations to global understanding with explainable AI for trees, Nature Machine Intelligence (2d ed. 2020).

(12) Agus Sudjianto et al., Linear Iterative Feature Embedding: An Ensemble Framework for Interpretable Model, Wells Fargo (2021).

(13) Mukund Sundararajan et al., Axiomatic Attribution for Deep Networks (2017).

The accuracy of the key factors identified using Zest AI's Shapley-based methods have been empirically validated

Not only have the methods that Zest AI uses to enable lenders to understand their models been theoretically validated, shown to be superior to other methods, and survived the rigorous peer-review process of academic journals, they have also been empirically validated.

The FinRegLab/Stanford study,[14] in which Zest AI was a participant, described and executed a method of empirically validating explanations associated with denials from various credit risk models using various explainability methods. The FinRegLab/Stanford study demonstrated that certain Shapley-based approaches generated high-fidelity explanations for any kind of model, while other methods, such as LIME and other usages of Shapley-based methods, do not.

Notably, the study showed that the best fidelity achieved by any explainability method was achieved by a Shapley-based approach, which performed the same on simple models (Table 4, page 38) as machine learning models (Table 3, page 36). No other approach for explaining models (simple or complex) performed better.

The FinRegLab/Stanford study had some limitations: it didn't use actual underwriting models nor group model variables into "reason codes" or "key factors," as is commonly done in the industry to make consumer adverse action notices more useful and understandable. Zest AI's method of validating adverse action methodologies is inspired by the FinRegLab/Stanford method but extends the method to accommodates the industry practice of grouping individual variables into adverse action reasons.

The Zest AI method replaces the applicant's attributes associated with a decline reason identified by the adverse action methodology with values corresponding to approved borrowers, rescores the model, and generates new adverse action reasons. If the risk score improved and the adverse action reason corresponding to the modified attributes was no longer present in the list of principal adverse action reasons, the adverse action reason is validated.

To understand how the validation method works, consider a simplified but illustrative example:

> Marie applied for a loan to pay off several overdue bills, but was declined. The model had assigned Marie a 58% likelihood of defaulting on a new loan, and the top reason for denial in her adverse action letter was Recent Delinquencies. The model used many input variables, and three variables were associated with the adverse action reason, Recent Delinquencies. Marie had the following values:

Table 1: Marie's Recent Delinquencies

Variable	Value
Delinquencies in the last 30 days	5
Delinquencies in the last 60 days	7
Delinquencies in the last 90 days	10

> To validate whether Recent Delinquencies was an appropriate adverse action decline reason, the validation method replaces the values of Marie's recent delinquency variables with values from approved applicants, creating "counterfactual" versions of Marie that have all the same attributes of Marie but with better recent delinquency values.[15] One of these counterfactual versions of Marie might look something like this:

Table 2: One "Counterfactual" Marie's Recent Delinquencies

Variable	Value
Delinquencies in the last 30 days	0
Delinquencies in the last 60 days	0
Delinquencies in the last 90 days	0

When scored by the model, this “counterfactual” Marie, with no delinquencies, now gets assigned a 27% likelihood of defaulting on a new loan. The top reason for her adverse action notice would no longer be Recent Delinquencies. This process is repeated for other values corresponding to approved borrowers, and the average is taken. If the score returned by the model gives each “counterfactual” Marie a consistently better score, and the Recent Delinquencies adverse action consistently moves out of the top 5, Marie’s first decline reason is validated, and the process repeats to validate the rest of Mary’s decline reasons.

Below, we show the results from this process on a real Zest AI-developed underwriting model whose adverse action reasons were generated using Zest AI’s Shapley value-based adverse action methodology.

Table 3: Adverse Action Validation Results

Stated Adverse Action Reason	Avg. Relative Score Improvement after Substituting Approved Borrower Values	Avg. Change in Rank after Substituting Approved Borrower Values
Credit Limit Amount	53%	-16
Recent Delinquencies	28%	-13
Credit Utilization	24%	-13
Recent Inquiries	20%	-9
Credit History	19%	-8
Inquiries	18%	-10
Amount Past Due	19%	-12
Collection Account Payment Past Due	15%	-5

As expected and confirmed in the above table, when borrowers are denied and given the “stated adverse action reason” the variables associated with the reason are replaced with values from good borrowers, the applicant’s score improves, and the adverse action reason decreases in ranking. To understand how to read the Table 3, it shows, for example, that when a borrower is denied for “Recent Delinquencies” and attributes associated with recent delinquencies are substituted for approved borrower values, on average, the score increases 28%, and the rank of the “Recent Delinquencies” adverse action reason decreases 13 positions. The validation shows that the explainability method Zest used actually reflects the principal reasons for the denial and that the denial reasons that would be provided in an adverse action notice based on these methods are accurate.

(14) FinRegLab et al., Machine Learning Explainability & Fairness: Insights from Consumer Lending (2022).

(15) A “counterfactual” is a “what-if” scenario that didn’t really happen.

CONCLUSION

Zest AI's methods of helping creditors provide adverse action reasons for machine learning underwriting models are theoretically and empirically sound

Regulators, including the CFPB, and state law makers have long acknowledged and reviewed the use of complex algorithms in making underwriting decisions. While some uses of AI/ML technology are novel or may require additional legislation, long-standing laws already govern the information that must be disclosed to consumers in connection with credit applications. Ensuring that your technology partner can clearly and empirically demonstrate that their models produce accurate reasons is a base requirement, one that Zest AI also provides in each client's model risk management document.

By using the best and most accurate explainability methods, we enable financial institutions to compliantly unlock the benefits that machine learning brings to consumers: higher approval rates holding risk constant; fewer defaults without a decrease in approvals; expanded access to credit for everyone, including thin-file and no-file borrowers. At Zest AI, we remain committed to using the best available technology to help financial institutions of all sizes make fair and transparent credit available to everyone.

AI for a thriving lending ecosystem

Since 2009, Zest AI has been innovating and perfecting AI lending technology. A pioneer in the field, serving more than $5.6T in assets with over 600 active models, Zest AI is helping financial institutions build thriving lending ecosystems with purpose-built, best-in-class AI.

To learn more visit us at zest.ai or reach out at hello@zest.ai.