

WHITE PAPER

# **CFPB Circular 2022-03 Reminds Law-Abiding Lenders That They Must Generate Accurate Adverse Action Reasons; Zest AI is Here to Help**



---

# TABLE OF CONTENTS

01	Introduction	What's a Model and What's a Model Explanation?
02	Zest AI's Shapley-Based Explainability Methods	How and Why They Work to Explain Machine Learning Models Accurately
05	The Problem with Interpretable Models	Some Have Advocated Using Certain Breeds of Interpretable Models, Which Are Less Accurate And Still Hard to Understand Without the Help of Computers
08	The Problem with Shapley-Based Methods	Shapley-Based Methods Are Not Based on Model Approximations, the Concern Raised in the CFPB Circular's First Footnote
11	Accuracy with Zest AI	The Accuracy of the Key Factors Identified Using Zest AI's Shapley-based Methods Have Been Empirically Validated
15	Conclusion	Zest AI's Methods of Helping Creditors Provide Adverse Action Reasons for Machine Learning Underwriting Models Are Theoretically and Empirically Sound and Consistent with the CFPB Circular

On May 26, 2022, the CFPB issued a [circular](#) confirming that the Equal Credit Opportunity Act (“ECOA”) and Regulation B (“Reg B”) require lenders to provide notices of adverse action to consumers who are denied credit even when those lenders rely on complex algorithms to make their credit decisions. Zest AI applauds the CFPB for holding all models used in consumer lending to a high bar and clarifying that accurate adverse action notices are required when using AI models in decisioning. Zest AI is among those that raised this issue with the CFPB, both during the CFPB’s October 2020 [Tech Sprint on Adverse Action Notices](#) and in Zest AI’s December 18, 2020 [letter](#) responding to the CFPB’s request for information on AI/ML in financial services.

In the press release announcing the circular, the CFPB acknowledged that “[l]aw-abiding financial companies have long used advanced computational methods as part of their credit decision-making processes, and they have been able to provide the rationales for their credit decisions.” Zest AI has always held itself to this standard. Indeed, 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 that the CFPB criticizes in footnote 1 of the circular, 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 the 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 the law. 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.

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 was a challenge for model developers to create accurate explanations for more complex models until recently. 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. However, breakthroughs in the last several years have changed all that.

# Shapley-Based Explainability Methods Accurately Explain ML Models

The explanations provided by Zest AI's software are based on the work of Nobel laureate Lloyd Shapley<sup>1</sup> and numerous academics and researchers that followed him. Using [Shapley values](#) is the only rigorous method that provides a mathematically defensible way to explain machine learning models. That is why Zest AI always applies Shapley-based approaches to explain and document credit underwriting models. Below we describe the methodology and review some academic literature supporting this approach.

In the 1960s and '70s, mathematicians, sociologists, and economists became interested in a field that has come to be called "game theory." Those mathematicians, Shapley included, were trying to quantify how much individual players in cooperative games like soccer contribute to the final score of the game, taking into account the number of

baskets, touchdowns, or goals each player scored on their own, as well as the player's assists, passes, and blocks (their interactions with the other players on the team).

Shapley's method simulates all possible atomic game scenarios and their outcomes when a given player is present or absent. Importantly, Shapley proved that his method was the only one capable of accurately allocating credit to each player in the game. In his seminal paper, he states that "[o]ur procedure will be to enunciate a set of axioms related to the values of games and combinations of games; and then to show that the axioms determine that value uniquely." (Shapley 1951, p. 2.)

While theoretically sound, his method is somewhat impractical for sports – you would have to play too many games to use it. But as a mathematical construct, it was beneficial because it rigorously defined what it meant to contribute to the outcome of a game. And now that we have powerful computers and efficient algorithms to reduce the number of games you need to simulate, it has become far more efficient and valuable.

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,<sup>2</sup> 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,<sup>3</sup> 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).<sup>4</sup> 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 behaviour." 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.

Most recently, in March 2022, Stanford University partnered with FinRegLab to conduct a study on<sup>5</sup> 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 software (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.” While the CFPB did not take this position in its circular, its references to “interpretability” bear further analysis because some have used that term 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.<sup>6</sup>

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.<sup>7</sup> 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, as we showed in a previous [whitepaper](#), 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.<sup>8</sup> 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).<sup>9</sup> 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. We are confident that these results are consistent with the CFPB's focus on 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 CFPB's mission 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, as the CFPB said in its circular.

(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. Zeninor 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, the Concern Raised in the CFPB Circular's First Footnote

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: first in a 2008 [whitepaper](#), then during the CFPB’s October 2020 [Tech Sprint on Adverse Action Notices](#), and, most recently, in Zest AI’s 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.<sup>10</sup> 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, published in the journal [Nature](#), describes the Shapley-based algorithms he implemented as “directly measur[ing] local feature interaction effects,”<sup>11</sup> 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.”<sup>12</sup>

Shapley-based methods fall into the “local diagnostics” method he mentions, 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.”<sup>13</sup> (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 can also be empirically validated.

The FinRegLab/Stanford study,<sup>14</sup> 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 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 accommodate 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.<sup>15</sup> 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 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 reason stated. 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 53%, 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.

# **Zest AI's Methods of Helping Creditors Provide Adverse Action Reasons for Machine Learning Underwriting Models Are Theoretically and Empirically Sound and Consistent with the CFPB Circular**

The CFPB has long acknowledged that law-abiding creditors can make use of complex algorithms in making underwriting decisions. The circular reminded the industry that failing to pay adequate attention to the accuracy of adverse action reasons can land them in hot water. Zest AI's methods have been validated both theoretically and empirically and are not the sorts of methods that the CFPB was warning about.

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; and fairer credit outcomes for women and people of color. 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.

