American Journal of Information Science and Technology
2025, Vol. 9, No. 3, pp. 211-224
https://doi.org/10.11648/j.ajist.20250903.15

Research Article

The Accuracy-Interpretability Dilemma: A Strategic Framework for Navigating the Trade-off in Modern Machine Learning

Partha Majumdar*

Department of Computer Science, Swiss School of Business and Management, Geneva, Switzerland

Abstract

This paper explores the enduring accuracy-interpretability trade-off in machine learning, highlighting its profound implications for model selection, regulatory compliance, and practical deployment across diverse industries. It begins by defining accuracy as a model's ability to generalise effectively on unseen data, measured through context-specific metrics. It contrasts it with interpretability, which ensures that model predictions are understandable and justifiable to human stakeholders. The paper maps models across the white-box to black-box spectrum, from inherently transparent techniques such as linear regression and decision trees to opaque but highly accurate methods like ensemble models and deep neural networks. It critiques the conventional view that increasing accuracy necessarily diminishes interpretability, presenting alternative perspectives such as the Rashomon effect, which suggests that equally accurate yet interpretable models often exist within the solution space. The paper emphasises two pathways: interpretability-by-design approaches, such as Generalised Additive Models and sparse decision trees, and post-hoc explainability tools like LIME and SHAP that enhance transparency in black-box models. Industry case studies in finance, healthcare, algorithmic trading, and business strategy illustrate the context-dependent balance between performance and explainability, shaped by legal mandates, trust requirements, and operational priorities. The framework proposed equips practitioners with strategic questions to guide model selection, incorporating considerations of compliance, end-user needs, and the relative costs of errors versus missed insights. The paper also anticipates future advancements in Explainable AI, inherently interpretable architectures, and causal machine learning that could dissolve the trade-off altogether by achieving high accuracy without sacrificing transparency. By reframing the dilemma as a strategic decision rather than a rigid constraint, it provides a structured roadmap for aligning model development with business objectives, ethical imperatives, and stakeholder trust, advocating a shift towards accuracy and interpretability as complementary rather than competing goals.

Keywords

Accuracy-interpretability Trade-off, Explainable Artificial Intelligence (XAI), Regulatory Compliance, Inherently Interpretable Models, Causal Machine Learning, Rashomon Set

*Corresponding author: partha.majumdar@hotmail.com (Partha Majumdar)

Received: 6 August 2025; **Accepted:** 15 August 2025; **Published:** 13 September 2025

1. Deconstructing the Trade-off: Foundations and Nuances

1.1. Defining the Poles: Accuracy and Interpretability

In the landscape of predictive modelling, practitioners continually navigate a foundational tension between two competing virtues: a model's predictive power and its transparency. This dynamic, often termed the accuracy-interpretability trade-off, shapes the selection of algorithms, the design of systems, and the ultimate utility of machine learning in real-world applications. Understanding the precise nature of these two poles is the first step toward developing a strategic framework for managing their inherent conflict.

Accuracy, in its most critical sense, refers to a model's ability to generalise and make correct predictions on new, unseen data. It is a measure of a model's performance in the wild, not merely on the data it was trained on. [17] While the term is often used colloquially, its technical evaluation involves a suite of metrics, the choice of which is highly dependent on the specific context of the problem. For classification tasks, metrics such as precision (the proportion of identifications that were actually correct), recall (the proportion of actual positives that were correctly identified), F1-score (the harmonic mean of precision and recall), and the Area Under the Receiver Operating Characteristic Curve (AUC) provide nuanced views of performance, especially in cases of class imbalance. For regression tasks, metrics like Mean Absolute Error (MAE) or Root Mean Squared Error (RMSE) quantify the magnitude of prediction errors. The ultimate goal of optimising for accuracy is to create a model that reliably produces the correct output, as the value of the model is often directly tied to its predictive performance. [17].

Interpretability, conversely, is the degree to which a human can understand the cause and effect behind a model's predictions [User Query]. It addresses the fundamental question: *Why* did the model make this specific decision? This concept is often used alongside, and sometimes interchangeably with, related terms like *explainability* and *transparency*, though important distinctions exist. Transparency refers to understanding the mechanics of the model itself-how the algorithm works and how it was trained. [14] A simple linear regression model is transparent because its mathematical form is straightforward. Explainability is the capacity to provide reasons or justifications for a specific output, often in human-understandable terms. [19] Interpretability is the broader quality that encompasses both; it is the link between the model's internal logic and a human's cognitive grasp of its behaviour. Unlike accuracy, which can be measured with objective scores, interpretability is inherently subjective and context-dependent. [17] A set of linear model coefficients may be perfectly interpretable to a data scientist but entirely opaque to a business executive or a government regulator, who may require a more narrative or visual explanation. The demand for interpretability arises from the need for trust, accountability, debugging, and the extraction of actionable insights from the model's logic. [2].

1.2. The Model Complexity Spectrum: From White Box to Black Box

The tension between accuracy and interpretability is most clearly observed when mapping machine learning algorithms along a spectrum of complexity. At one end lie simple, transparent models, often called "white-box" or "glass-box" models, whose decision-making processes are inherently understandable. At the other end are highly complex and opaque "black-box" models, which often achieve superior accuracy at the cost of transparency. [21]

Inherently Interpretable ("White Box") Models

These models are designed in a way that their internal logic is accessible and can be directly inspected by a human analyst. [26] Their simplicity is their primary virtue, making them ideal for applications where understanding the "why" is paramount.

1) Linear and Logistic Regression: These are among the most fundamental and interpretable models. In linear regression, the relationship between features and the outcome is modelled as a weighted sum. Each feature's coefficient represents its contribution to the prediction; a positive coefficient indicates that an increase in the feature's value leads to a corresponding increase in the predicted outcome, holding other features constant. [37] Logistic regression extends this concept to classification problems, where coefficients represent the change in the log-odds of the outcome. These log odds can be transformed into odds ratios, providing a powerful and intuitive explanation. [12] For instance, in a healthcare model predicting heart disease, a logistic regression coefficient for the feature "chest pain type" could be translated to mean that a one-unit increase in this feature increases the odds of having heart disease by 120.1%. [12] This direct, quantifiable link between a feature and the outcome makes the model's reasoning explicit.
2) Decision Trees: These models function as a set of hierarchical if-then-else rules, which can be easily visualised as a flowchart. [35] To understand a prediction for a specific instance, one can trace its path from the root node down to a terminal leaf node, following the decision rules at each split. [12] Each path represents a clear and logical explanation for the resulting classification. For example, a decision tree for a heart disease prediction might have a root node that splits on "chest pain type." If the condition is met, the path goes left; if not, it goes right, continuing through subsequent nodes based

on other features like "number of major vessels" or "sex" until a final prediction is reached. [12] However, this inherent interpretability is fragile. As a decision tree grows in depth and complexity to capture more intricate patterns in the data, it can quickly become a dense, sprawling web of rules that is no longer easily comprehensible to a human, thus becoming a "white box" that has effectively turned grey. [17]

Opaque ("Black Box") Models

Black-box models are systems that produce useful outputs without revealing their internal workings; the logic connecting inputs to outputs is unknowable or too complex for human analysis. [22] These models typically achieve state-of-the-art accuracy by modelling highly complex, non-linear relationships in the data, a capability that inherently requires a level of complexity that defies simple explanation. [21]

1) Ensemble Methods (Random Forests, Gradient Boosting): These powerful algorithms operate by combining the predictions of many individual, often simple, models (typically decision trees). A Random Forest, for example, builds hundreds or thousands of decision trees on different subsets of the data and features, and then averages their predictions (for regression) or takes a majority vote (for classification). [37] While each tree might be interpretable, the aggregation of their outputs into a single prediction makes the overall decision process opaque. [21] Similarly, Gradient Boosting Machines (GBMs) like XGBoost build trees sequentially, with each new tree correcting the errors of the previous ones. This sequential, additive process results in highly accurate models that are extremely difficult to interpret directly. [31]
2) Support Vector Machines (SVMs) with Non-Linear Kernels: SVMs work by finding an optimal hyperplane that separates data points into different classes. For data that is not linearly separable, SVMs employ the "kernel trick," which maps the data into a higher-dimensional space where a linear separation becomes possible. [21] While powerful, this transformation means the separating hyperplane is defined in a space that is not the original feature space, obscuring the direct relationship between the original input features and the final decision boundary. [35]
3) Deep Neural Networks (DNNs): DNNs are the archetypal black-box models and are the engine behind many of the most significant advances in AI, from image recognition to natural language processing. [21] They consist of multiple layers of interconnected nodes (or "neurons"), each performing a simple computation. When stacked in deep architectures with non-linear activation functions, these networks can learn incredibly complex and hierarchical patterns from data. However, with potentially millions or even billions of weighted parameters, tracing a single prediction back through the network to understand the contribution of each input feature is practically impossible. [27] This opacity presents significant challenges, as it becomes difficult to trust the model's outputs, debug its failures, or ensure it is not operating on biased or irrelevant factors. [22]

To provide a clear reference for practitioners, the following table summarises the key characteristics of these models along the accuracy-interpretability spectrum.

Table 1. *A Comparative Analysis of Machine Learning Models on the Accuracy-Interpretability Spectrum.*

Model	Inherent Interpretability	Typical Accuracy Potential	Computational Complexity	Common Use Cases
Linear/Logistic Regression	High	Low to Moderate	Low	Baseline models, inference, regulated industries (e.g., credit scoring)
Decision Trees	High (if shallow)	Moderate	Low	Rule-based systems, explainable classification, feature exploration
Generalised Additive Models (GAMs)	High	Moderate to High	Moderate	Interpretable modelling with non-linear effects (e.g., churn analysis)
Random Forest	Low	High	Moderate to High	Complex classification/ regression, robust prediction
Gradient Boosting Machines (e.g., XGBoost)	Very Low	Very High	High	Winning competitions, state-of-the-art performance on tabular data
Support Vector Machines (non-linear)	Very Low	High	High	High-dimensional classification, image recognition
Deep Neural Networks	None	Very High	Very High	Image/speech recognition, NLP, autonomous systems

This structured overview provides a foundational map of the modelling landscape. A practitioner facing a problem that demands high interpretability for regulatory compliance, such as credit lending, can immediately see that models like Logistic Regression or GAMs are strong initial candidates, while DNNs would present significant challenges. Conversely, for a task where raw predictive performance is the sole objective, such as certain types of algorithmic trading, the table points toward high-complexity models like GBMs and DNNs. This codification of the field's common wisdom serves as a crucial first-pass filter in the model selection process.

1.3. Beyond the Simple Trade-off: A Nuanced Reality

The narrative that greater accuracy requires sacrificing interpretability is a powerful and persistent one in machine learning. The conventional wisdom holds that model complexity is the central mechanism: as a model becomes more complex (e.g., by adding more features, parameters, or non-linear transformations), its capacity to capture intricate patterns in the data increases, leading to higher accuracy. However, this same complexity inevitably reduces its interpretability. [27] While this inverse relationship is a useful heuristic and frequently observed in practice, it is not a strict, inviolable law. [2].

Emerging research and a more critical examination of the trade-off reveal a more nuanced reality. First, the relationship is not always monotonic; that is, increasing complexity does not always guarantee an increase in accuracy. [2] In some applications, simpler, more interpretable models have been shown to outperform their black-box counterparts, particularly in domains like fault diagnosis or certain environmental predictions. [2] Furthermore, the axes of accuracy and interpretability are not always in direct opposition; they can be orthogonal. It is possible to have a model that is both inaccurate and difficult to interpret, demonstrating that complexity and opacity are not sufficient conditions for high performance. [23]

A more profound challenge to the conventional trade-off comes from the concept of the "Rashomon Set," a term popularised by Professor Cynthia Rudin. [4] The Rashomon effect, named after the 1950 Kurosawa film where multiple characters provide contradictory but equally plausible accounts of the same event, describes a scenario in machine learning where there exists a large set of different models that all achieve the same state-of-the-art level of predictive accuracy for a given problem. [4] The existence of such a set has a powerful implication: if many different models are all optimally accurate, probably, at least one of them is also simple and interpretable.

This perspective reframes the entire accuracy-interpretability dilemma. The trade-off we so often observe may not be a fundamental property of the problem itself, but rather an artefact of our modelling tools and practices. The most common and powerful algorithms, such as deep learning and gradient boosting, are designed and optimised with a single-minded focus: to find *any* model within the Rashomon Set that minimises prediction error. They are not designed to search for the *simplest* or *most interpretable* model within that set. Consequently, they often converge on a highly complex solution simply because it is the one their optimisation path happens to find first.

This shifts the challenge for the practitioner from one of a necessary sacrifice to one of search and discovery. The question is no longer, "How much accuracy must I give up to gain interpretability?" but rather, "An accurate and interpretable model likely exists; what methods and tools can I use to find it?" This suggests that the perceived trade-off is often a result of an incomplete search of the model space. By investing in algorithms that explicitly optimise for simplicity alongside accuracy (such as the optimised sparse decision trees discussed later) or by conducting a more thorough and creative model selection process, a practitioner may be able to achieve an "interpretable free lunch"-state-of-the-art performance without sacrificing transparency. [4] This understanding moves the conversation beyond a simple dichotomy and toward a more sophisticated, strategic approach to model building.

2. A Practitioner's Guide to Interpretability

Navigating the accuracy-interpretability landscape requires a practical toolkit. There are two primary strategic pathways for a practitioner to achieve interpretability. The first is to build models that are transparent from the ground up, an approach known as interpretability by design. The second is to take a high-performing but opaque black-box model and apply post-hoc analysis techniques to explain its behaviour, a field broadly known as Explainable AI (XAI). The choice between these strategies depends heavily on the context of the problem, particularly the stakes involved in the decision-making process.

2.1. Interpretability by Design: The Power of Intrinsic Models

The most robust and trustworthy path to interpretability is to use models that are inherently transparent by design. [26] This approach avoids the potential pitfalls and approximations of post-hoc methods by ensuring that the model's logic is never obscured in the first place. For high-stakes decisions, such as in finance or healthcare, where unreliable outcomes can have severe consequences, there is a strong argument that inherently interpretable models should be the default choice over black-box alternatives. [2]

1)Generalised Additive Models (GAMs) and Explainable Boosting Machines (EBMs): These models represent a powerful middle ground between simple linear models and complex black boxes. A GAM works by modelling the outcome as a sum of smooth, non-linear functions of individual features. [26] Instead of a single coefficient for each feature, a GAM fits a flexible function (often a spline) that can capture non-linear relationships. The key to their interpretability is that the effect of each feature is additive and can be isolated and visualised independently. An Explainable Boosting Machine (EBM) is a modern, tree-based implementation of GAMs that also automatically detects and includes pairwise interaction terms. [31] For example, in a customer churn prediction model, a GAM or EBM can generate a plot for the feature "customer tenure." This plot might reveal a highly non-linear effect: churn risk is very high in the first few months, drops sharply, and then plateaus. This granular insight is directly interpretable and actionable, unlike a single coefficient from a linear model or an opaque prediction from a random forest. [31]

2)Globally Optimised Sparse Decision Trees (GOSDT): Traditional decision trees, like CART, are built using a greedy, recursive splitting process, which often results in suboptimal and overly complex trees. In contrast, algorithms like Globally Optimised Sparse Decision Trees (GOSDT) use integer programming and other optimisation techniques to find the provably optimal sparse decision tree for a given dataset. [31] This approach directly balances accuracy with a penalty for complexity (e.g., the number of leaves), resulting in a simple, highly interpretable tree that does not sacrifice performance unnecessarily. It directly addresses the search problem highlighted by the Rashomon Set concept, aiming to find the simplest model at the frontier of accuracy.

3)Sparse Linear Models (e.g., with L0/L1 Regularisation): Another way to enforce simplicity is through regularisation. Techniques like LASSO (L1 regularisation) penalise the absolute size of the coefficients in a linear or logistic regression model, forcing the coefficients of less important features to become exactly zero. This effectively performs automatic feature selection, resulting in a "sparse" model that only includes the most impactful predictors. [31] More advanced tools like L0Learn use L0 regularisation, which directly penalises the number of non-zero coefficients, to find the best model for a given level of sparsity. The result is a simple, parsimonious model that is easy to interpret and less prone to overfitting.

By prioritising these intrinsically interpretable models, practitioners can build systems where the explanation is the model itself. This eliminates the need for post-hoc approximations and provides a more solid foundation for trust, debugging, and regulatory compliance.

2.2. Opening the Black Box: The Rise of Explainable AI (XAI)

While interpretable-by-design models are ideal, there are many scenarios where the superior predictive performance of a black-box model is either necessary or has already been deployed. In these cases, the goal shifts to explaining the model's behaviour after it has been trained. This is the domain of Explainable AI (XAI), a collection of post-hoc techniques designed to peer inside the black box. [26] This approach can be thought of as "adding interpretability" to an existing model, allowing data scientists to leverage the full arsenal of machine learning algorithms and explain their scoring after the fact. [17] XAI methods can be broadly categorised by their scope: global techniques that explain the model's overall behaviour, and local techniques that explain a single, individual prediction.

2.2.1. Global Interpretation Techniques (Explaining the "What")

Global interpretation methods aim to provide a high-level understanding of what a model has learned from the data as a whole. They reveal general trends and the average importance of different features.

1) Permutation Feature Importance: This is a widely used, model-agnostic technique for assessing the global importance of a feature. [37] The process is intuitive: first, the model's prediction error is calculated on a dataset. Then, the values of a single feature column are randomly shuffled, breaking the relationship between that feature and the target variable. The model's prediction error is calculated again on this permuted data. The increase in error after shuffling serves as the measure of that feature's importance-the more the model's performance degrades, the more it must have been relying on that feature. [26] This process is repeated for all features to produce a ranked list of their importance.

2) Partial Dependence Plots (PDPs) and Accumulated Local Effects (ALEs): These are visualisation tools that illustrate the marginal effect of one or two features on the predicted outcome of a model, averaged over all other features. [37] A PDP for a single feature shows how the average prediction changes as the value of that feature varies across its range. For example, a PDP could show how the predicted price of a house changes as the "square footage" feature increases. While simple and intuitive, PDPs have a significant drawback: they assume the feature being analysed is independent of all other features. If features are correlated, PDPs can be highly misleading because they may average over unrealistic data points that would never occur in the real world. [37] Accumulated Local Effect (ALE) plots were developed as a more robust alternative that is not biased by correlated features, making them a better choice in most practical scenarios. [28].

2.2.2. Local Interpretation Techniques (Explaining the "Why" for a Single Case)

While global methods are useful for understanding the model in general, many applications require an explanation for a specific decision. Local interpretation methods are designed to answer the question, "Why did the model make *this particular* prediction for *this particular* instance?"

1) LIME (Local Interpretable Model-agnostic Explanations): LIME is a popular model-agnostic technique that explains individual predictions by learning a simple, interpretable model in the "local" vicinity of the prediction. [5] The process involves several steps:

(a) Perturbation: Take the single instance you want to explain and generate a new dataset of "perturbed" or slightly modified versions of it. For tabular data, this might involve randomly changing some feature values. For an image, it might mean turning off certain super-pixels (groupings of pixels). [37]

(b) Prediction: Use the original black-box model to make predictions on this new dataset of perturbed instances.

(c) Weighting: Assign a weight to each perturbed instance based on its proximity or similarity to the original instance.

(d) Fitting: Train a simple, interpretable model (e.g., a sparse linear regression or a small decision tree) on the weighted, perturbed dataset.

(e) Explanation: The trained interpretable model, which is only "locally faithful" to the black box, serves as the explanation. For example, the coefficients of the local linear model indicate which features were most influential for that specific prediction. [5] A classic example is explaining an image classifier's prediction of "tree frog." LIME would highlight the specific green and orange super-pixels of the frog in the image that had the highest positive weights in its local model, providing a visual explanation for the decision. [37]

2) SHAP (SHapley Additive exPlanations): SHAP is a powerful method for local explanation that is grounded in cooperative game theory and the concept of Shapley values. [5] It explains a prediction by treating each feature as a "player" in a game where the "payout" is the model's prediction. The Shapley value for a feature is its average marginal contribution to the prediction across all possible combinations (or coalitions) of features. [37] SHAP values have several desirable properties, most notably additivity: the sum of the SHAP values for all features for a given prediction equals the difference between that prediction and the baseline (average) prediction for the entire dataset. [5] This ensures a complete and accurate accounting of the prediction.

SHAP explanations are often visualised using force plots, which show how each feature's SHAP value acts as a "force" pushing the prediction away from the baseline. For instance, in a model predicting house prices, a force plot for a specific house might show that a high value for "LSTAT" (% lower status of the population) pushes the predicted price down. In contrast, a high value for "RM" (average number of rooms) pushes the price up. [37] This provides a detailed, quantitative breakdown of the factors driving a single prediction.

The following table serves as a practitioner's guide to these common XAI techniques, outlining their scope, advantages, and limitations.

Table 2. *A Guide to Explainable AI (XAI) Techniques.*

Technique	Explanation Scope	Model-Agnostic?	Output Type	Key Advantage	Key Limitation/ Caveat
Permutation Feature Importance	Global	Yes	Ranked list of feature scores	Simple, intuitive, and widely applicable.	Can be misleading for correlated features; computationally intensive.
PDP / ALE Plots	Global	Yes	2D or 3D plots	Provides clear visualization of marginal feature effects.	PDP is unreliable with correlated features; ALE is better but less known.
LIME	Local	Yes	Feature weights/rules for one prediction	Intuitive local explanations; works on text, image, tabular data.	Explanations can be unstable; sensitive to perturbation settings. [15]
SHAP	Local & Global	Yes	SHAP values, various plots (force, summary)	Strong theoretical foundation; guarantees accurate attribution.	Computationally very expensive; explanation depends on choice of baseline. [37]
Global Surrogate Models	Global	Yes	An entirely new, simple model	Can use any interpretable model; easy to measure fidelity (R2).	Interprets the black-box model, not the data itself; may not be faithful globally. [37]

While this toolkit of XAI methods is incredibly powerful, it does not eliminate the trade-off entirely. Instead, it introduces a new layer of complexity and a "meta-interpretability" challenge. These post-hoc explanations are, by their nature, approximations of the true black-box model. Their fidelity is not guaranteed. For example, LIME's explanations are known to be potentially unstable; two very similar data points can sometimes produce very different local explanations, which undermines trust in the method. [15] SHAP, while theoretically more robust, is computationally demanding, and its explanations are critically dependent on the choice of the "background" or "baseline" dataset used for comparison. A different baseline can lead to different SHAP values and, therefore, a different interpretation of the same prediction. [37]

This reality means that the practitioner's job is not simply to run an XAI tool and present the output. They must critically evaluate the explanation itself. In a regulated environment like finance, a bank using SHAP to generate an adverse action notice for a denied loan must be prepared to defend not only the black-box model but also its choice of baseline dataset for the SHAP calculation, as a different choice could have yielded a different "principal reason" for the denial. [6] The trade-off is not gone; it has been shifted to a new, more subtle level. The practitioner now faces a trade-off between the fidelity, stability, and computational cost *of the explanation method itself*, adding another dimension to the strategic navigation of the accuracy-interpretability dilemma.

3. The Trade-off in Practice: Industry Deep Dives

The theoretical tension between accuracy and interpretability manifests with vastly different implications across various industries. The optimal balance is not a universal constant but is dictated by the specific context of the application, including its regulatory environment, the stakes of its decisions, and its ultimate business objective. Examining how this trade-off plays out in high-stakes finance, mission-critical healthcare, performance-centric applications, and strategic business analysis provides a concrete understanding of the practical challenges and solutions.

3.1. High-stakes Finance: Navigating Regulation and Risk

In the financial services industry, particularly in consumer credit, the accuracy-interpretability trade-off is not merely a technical consideration; it is a matter of legal and regulatory compliance. The need for explainable decisions is mandated by law, forcing institutions to prioritise transparency even when it might seem to conflict with maximising predictive performance.

3.1.1. The Regulatory Mandate: ECOA and the CFPB's Stance on AI

The central piece of legislation governing this domain in the United States is the Equal Credit Opportunity Act (ECOA). First enacted in the 1970s to prevent discrimination in lending, ECOA and its implementing Regulation B require creditors to provide applicants with a "statement of specific reasons" for any adverse action taken, such as denying a loan application or unfavourably changing the terms of an account. [11] This "right to explanation" is a cornerstone of consumer protection in finance. [13]

In recent years, the Consumer Financial Protection Bureau (CFPB), the primary federal agency enforcing ECOA, has issued a series of advisory opinions and circulars that directly address the use of artificial intelligence and complex algorithms in credit decisions. [20] The CFPB's position is unequivocal: the legal requirements of ECOA apply fully to all credit decisions, regardless of the technology used to make them. [10] A creditor cannot use the complexity or opacity of its model as an excuse for failing to provide a specific, accurate, and understandable explanation for an adverse action. The argument that a "black-box" model is too difficult to interpret is not a valid legal defence. [30].

The CFPB has specifically targeted the common practice of using generic, check-box reasons from sample adverse action forms. The agency has clarified that these sample forms are not a safe harbour; if the checklist items do not accurately capture the principal reason(s) for the adverse action, the creditor violates the law. [34] This is particularly relevant for models that use non-traditional or "behavioural" data. For example, the CFPB has stated that if a consumer's credit limit is reduced because their purchasing patterns are deemed risky (e.g., based on the "type of establishment at which a consumer shops"), a generic reason like "purchasing history" would be insufficient. The notice must be specific enough to inform the consumer about the actual behaviour that influenced the decision. [25] This regulatory pressure forces lenders to invest in models and systems that can produce these granular explanations.

3.1.2. Case Study: Fair Lending, Bias Auditing, and Adverse Action Notices

Consider a practical scenario that synthesises these challenges. A large bank develops and deploys a highly accurate AI model for mortgage underwriting, aiming to improve efficiency and reduce defaults. The model is trained on decades of historical loan data. However, this historical data may reflect past societal biases, and the model inadvertently learns to associate certain features, which act as proxies for protected characteristics like race or national origin, with higher risk. [32] For instance, a case brought by the Massachusetts Attorney General alleged that a student loan company's AI

model used a "cohort default rate" (CDR) variable from the Department of Education, which disproportionately and negatively impacted Black and Hispanic applicants. [7].

In this situation, interpretability serves two critical functions. First, it is essential for bias auditing and fair lending analysis. Before deploying the model, the bank must use interpretable models or XAI tools to audit its behaviour. By examining which features are most influential, the bank can identify and mitigate potential sources of discriminatory impact. If the model is found to be weighing a proxy for a protected class too heavily, it must be retrained or redesigned. This is a crucial step in risk management to avoid disparate impact liability. [3].

Second, once the model is deployed, interpretability is essential for generating compliant adverse action notices. When the model denies an applicant, the bank cannot simply state "low score on the credit model." It must provide the specific principal reasons. As suggested by legal and technical experts, post-hoc XAI tools like SHAP or LIME can be used to analyse the individual denial and identify the key features that drove the decision. [6] The bank could use SHAP to determine that for a specific applicant, the three principal reasons for denial were "income insufficient for the amount of credit requested," "limited credit experience," and a high debt-to-income ratio calculated from non-traditional data sources. [13] This allows the bank to generate a notice that is both specific and accurate, satisfying its obligations under ECOA and the CFPB's stringent guidance. [18]

The intense regulatory environment in finance fundamentally alters the nature of the trade-off. For a data scientist, a 2% gain in accuracy from an XGBoost model over a logistic regression model might seem like a clear win. However, for the institution's Chief Risk Officer and General Counsel, that 2% gain is meaningless if the XGBoost model cannot produce legally defensible explanations for its decisions. The potential cost of a CFPB enforcement action or a class-action lawsuit for systemic ECOA violations far outweighs the marginal increase in predictive performance. [3] This external pressure elevates the need for interpretability from a desirable technical feature to a non-negotiable component of corporate governance and legal risk management, making it a central pillar of the entire model development lifecycle.

3.2. Mission-critical Healthcare: Building Trust and Ensuring Safety

In healthcare, the stakes of predictive modelling are arguably the highest, as decisions can directly impact patient well-being and survival. Here, the trade-off is not simply between accuracy and interpretability, but between a model's statistical performance and its ability to be safely and effectively integrated into clinical practice. Trust and safety are paramount, making interpretability an essential prerequisite for the adoption of AI in medicine.

3.2.1. The Clinician-in-the-loop: Overcoming the Trust Deficit

Medical professionals are trained to operate within a framework of evidence-based practice, where clinical decisions must be justified by scientific evidence, clinical expertise, and patient values. [1] They are legally and ethically accountable for the outcomes of their patients. Consequently, clinicians are often deeply sceptical of "black-box" AI systems that provide recommendations without transparent reasoning. [14] A doctor cannot responsibly act on a recommendation they do not understand, as they would be unable to justify their decision if something were to go wrong.

This trust deficit is compounded by the risk of the "Clever Hans" effect, where a model achieves high accuracy on a training dataset for entirely the wrong reasons. [22] The name comes from a horse in the early 20th century that appeared to be able to perform arithmetic but was actually responding to subtle, unintentional cues from its handler. In machine learning, this occurs when a model learns to exploit spurious correlations or artefacts in the training data rather than the true underlying patterns. A widely cited example in healthcare is an AI model developed to diagnose COVID-19 from chest X-rays. The model achieved impressive accuracy but was later found to be making its predictions based on irrelevant factors, such as the text annotations or markings on the X-ray images, which happened to be more common on scans from sicker patients who were more likely to have COVID-19. [22] Without interpretability, such a dangerous flaw could go undetected, leading to misdiagnoses and improper treatment when the model is deployed in a real-world setting with different imaging protocols. Interpretability is the essential diagnostic tool for the model itself, allowing humans to verify that it has learned valid medical patterns.

3.2.2. Case Study: From AI-powered Diagnosis to Clinical Confidence

The value of interpretability becomes clear when examining the clinical workflow for AI-powered diagnosis. Consider applications in detecting complex conditions like Alzheimer's disease from brain scans, identifying cancerous lesions in pathology slides, or predicting cardiac events from patient records. [14]

An AI model might analyse a patient's MRI scan and output a high probability of early-stage Alzheimer's disease. A black-box system provides this number, leaving the clinician in a difficult position. They must either accept the opaque recommendation blindly or discard it, negating the potential benefit of the AI's analytical power.

An explainable AI system, in contrast, would augment this prediction with an explanation. Using techniques like saliency maps or Grad-CAM for images, the system could produce a "heat map" that highlights the specific regions of the brain scan that were most influential in its decision. [14] For a model using tabular data, a SHAP analysis could identify the

key biomarkers or clinical history data points that contributed most to the risk score. This explanation transforms the AI from an inscrutable oracle into a sophisticated diagnostic assistant. The clinician can now use their medical expertise to evaluate the AI's reasoning. They can examine the highlighted regions of the scan and determine if they correspond to known pathological indicators of Alzheimer's. [33] This process of validation builds trust and allows the clinician to confidently integrate the AI's insight into their comprehensive assessment of the patient. The AI becomes a powerful "second opinion," augmenting rather than replacing human expertise.

This dynamic reveals that in healthcare, raw statistical accuracy is insufficient. The critical trade-off is not between accuracy and interpretability, but between a model's standalone accuracy and its potential for effective and safe clinical action. An uninterpretable model, no matter how accurate in a lab setting, can be clinically useless or even dangerous if it cannot be trusted and validated by the human expert who bears the ultimate responsibility for the patient's care. [14] Interpretability is the essential bridge that allows a human expert to responsibly translate a model's probabilistic output into a safe and effective real-world decision. Without this bridge, high accuracy can become a liability rather than an asset.

3.3. Performance-centric Applications: When Accuracy Is the Primary Mandate

While regulated and high-stakes domains demand interpretability, there are other areas of machine learning where the pendulum swings heavily toward raw predictive performance. In these applications, the value is derived directly from the accuracy, speed, and quality of the model's output, and the need for human-readable explanations is secondary or non-existent for the end-user.

3.3.1. Algorithmic Trading: The Need for Speed

High-Frequency Trading (HFT) is a prime example of a domain where accuracy and speed are the paramount virtues. [8] HFT firms use powerful computer systems to analyse market data and execute a large number of orders at extremely high speeds, often in microseconds. [8] The goal is to capitalise on small, fleeting price discrepancies and market inefficiencies that are imperceptible to human traders.

In this context, the decision-making loop is entirely automated. Complex algorithms for strategies like statistical arbitrage, market making, and momentum ignition are designed to make and execute trades without any human intervention for individual transactions. [8] The "why" behind a single trade that lasts for a fraction of a second is irrelevant. What matters is that, over millions of trades, the model's predictions are accurate enough to generate a profit. A marginal improvement in the model's predictive accuracy or a reduction in its latency (the time it takes to make a decision) can translate directly and immediately into significant financial gains. [8] Here, the black-box nature of a sophisticated model is not a drawback but a feature, as it enables the complexity required to find subtle patterns in noisy financial data. Interpretability is sacrificed for the competitive edge that comes from superior performance.

3.3.2. Recommendation Engines and Spam Filtering: The User Experience

Another class of applications where accuracy often takes precedence is in systems that directly shape a user's digital experience, such as e-commerce recommendation engines and email spam filters. [9].

In spam filtering, the user's primary concern is a clean inbox. They care that the filter correctly identifies and removes unwanted emails while not mistakenly flagging important messages (i.e., high precision and recall). [35] Whether the filter uses a simple Naive Bayes classifier or a complex deep learning model is of little consequence to the end-user, as long as it performs its function effectively. [35] The value is in the outcome - a less cluttered digital life - not in understanding the probabilistic calculations behind each classification.

Similarly, for recommendation engines on platforms like Amazon or Netflix, the goal is to enhance user engagement and drive sales by suggesting relevant products or content. [9] A shopper on Amazon cares that the "Customers who bought this also bought" feature suggests a product they are genuinely interested in, which improves their shopping experience and increases the likelihood of a purchase. [9] Research indicates that these recommendations drive a significant portion of revenue for major e-commerce and media companies. [9] The success of the system is measured by metrics like click-through rates, conversion rates, and increased average order value, all of which are direct functions of the recommendation's accuracy and relevance. [24] The complex collaborative filtering or deep learning algorithms that power these engines remain a black box to the consumer, and this is perfectly acceptable because the quality of the recommendation itself is the sole measure of success from their perspective.

However, it is a mistake to assume that interpretability has no value at all in these domains. While not critical for the end-user, it is highly valuable for the developers, data scientists, and business analysts who build and maintain these systems. Interpretability is crucial for:

1)Debugging and Model Improvement: If a recommendation engine starts suggesting bizarre or irrelevant items, developers need to understand why.

2)Identifying and Mitigating Bias: A model might learn only to recommend blockbuster movies or best-selling products, reducing the diversity of recommendations and creating a "filter bubble." Interpretability tools can help identify and counteract this tendency. [36]

3)Defending Against Adversarial Attacks: Malicious actors, sometimes called "spammers" or an "internet water army," may attempt to manipulate a system by creating

fake user profiles and ratings to boost the visibility of a certain product or movie artificially. [36] Understanding how the model weighs different factors is essential for building robust systems that can detect and resist such attacks.

In these performance-centric applications, the trade-off is managed by context. Accuracy is prioritised for the user-facing output, while interpretability remains a vital tool for the internal teams responsible for the system's health, fairness, and security.

3.4. Business Strategy: From Prediction to Actionable Insight

In many business contexts, the ultimate goal of a modelling project is not simply to make a prediction, but to gain a deeper understanding of a business problem in order to drive strategic action. In these scenarios, a model's ability to provide clear, actionable insights can be far more valuable than a marginal gain in predictive accuracy. This is the classic distinction between prediction and inference.

Customer churn prediction is a quintessential example of a problem where inference is often more important than pure prediction. [31] Every business, from telecommunications providers to subscription services, wants to identify and retain customers who are at risk of leaving.

Imagine a telecommunications company that builds two models to predict which of its customers will churn (cancel their service) in the next month.

1)Model A is a highly-tuned XGBoost model, a black box. It analyses hundreds of features and predicts with 95% accuracy which 1,000 customers are most likely to churn. The output of this model is a list of 1,000 names. This prediction is highly accurate, but it is not strategically useful on its own. What should the business do with this list? Offer everyone a discount? This would be expensive and inefficient, as the reasons for churn likely vary.

2)Model B is an interpretable model, such as an Explainable Boosting Machine (EBM) or a Generalised Additive Model (GAM). [31] This model may be less accurate, predicting churn with 92% accuracy. However, because it is interpretable, it provides more than just a list of names. By examining the model's components, the business can extract clear, actionable insights.

The interpretable model might reveal several key drivers of churn. For example, an analysis of the model's feature importance plots and individual spline functions (as demonstrated in a churn project using GAMs) could uncover the following insights [31]:

1)Contract Type is Key: Customers on a flexible "month-to-month" contract are overwhelmingly more likely to churn than those on one- or two-year contracts.

2)Early Tenure is a Danger Zone: The model might show a non-linear relationship with tenure, where the probability of churn is extremely high for customers in their first four months, after which it drops significantly and stabilises.

3)Payment Method Matters: Customers who pay by mailed check are more likely to churn than those who use automatic bank withdrawal or credit card payments.

4)Age is a Surprising Factor: Contrary to intuition, the model might show that older customers are slightly more likely to churn, perhaps due to difficulty with new technology or services.

These insights are strategically invaluable. The black-box model's list of 1,000 names is sterile; it is a prediction without an explanation. The interpretable model's insights, despite its slightly lower accuracy, provide a clear roadmap for action. The business can now design targeted, data-driven retention strategies:

1)Launch a marketing campaign specifically aimed at month-to-month customers, offering a significant discount to switch to a one-year contract.

2)Focus on customer success and onboarding efforts intensely on new customers during their first four months to navigate the critical danger zone.

3)Create an incentive program, such as a small monthly credit, for customers to switch from paper checks to electronic payment methods.

In this context, the small loss in predictive accuracy is more than compensated for by the immense gain in strategic understanding. The interpretable model empowers the business to move beyond simply predicting the future to actively changing it. This demonstrates that when the goal is to understand the "why" behind a phenomenon to inform decision-making, interpretability is not a secondary concern but the primary source of value.

4. Synthesis and Strategic Recommendations

The exploration of the accuracy-interpretability trade-off reveals a complex, context-dependent dilemma rather than a simple, universal rule. Navigating this landscape requires a strategic mindset that moves beyond a binary choice and toward a nuanced decision-making process tailored to the specific demands of each project. By synthesising the lessons from various industries and looking toward the future of the field, we can construct a practical framework for practitioners and anticipate the evolution of this fundamental challenge.

4.1. A Decision Framework for Model Selection

To move from theory to practice, a practitioner needs a structured way to decide where on the accuracy-interpretability spectrum their project should lie. The optimal choice is not determined by the data alone, but by a holistic assessment of the project's objectives, constraints, and stakeholders. The following series of questions can guide this decision-making process at the outset of any modelling endeavour:

1)What is the primary business objective? Is the main goal to achieve the highest possible predictive accuracy for an automated task (e.g., high-frequency trading)? Is it to gain strategic insights and understand the drivers of a business outcome (e.g., customer churn analysis)? Or is it to make a high-stakes, accountable decision about an individual that is subject to external scrutiny (e.g., credit lending)? The answer to this question sets the initial priority.

2)Are there legal, ethical, or regulatory constraints? Does the application fall under a legal framework that mandates a "right to explanation," such as the Equal Credit Opportunity Act (ECOA) in the U.S. or the General Data Protection Regulation (GDPR) in Europe? [13] The presence of such regulations immediately elevates interpretability from a "nice-to-have" to a "must-have."

3)Who is the end-user of the model's output? Is the prediction being fed directly into another automated system, where only the output matters? Is it being presented to a human expert, like a clinician, who needs to validate the reasoning before taking action? Is it being delivered to a customer, who may require a simple, non-technical explanation for a decision that affects them? Or is it for a business strategist who needs to understand the underlying drivers to formulate a plan? The nature of the audience dictates the required level of transparency and the form the explanation must take.

4)What is the cost of a wrong prediction versus the cost of a missed insight or a compliance failure? This risk assessment is critical. In algorithmic trading, the cost of a wrong prediction is a direct financial loss. In customer churn analysis, the cost of a missed insight is a lost opportunity to improve the retention strategy. In credit lending, the cost of a compliance failure could be a multi-million dollar lawsuit and significant reputational damage. Weighing these disparate costs helps to quantify the relative value of accuracy and interpretability for the specific problem.

The following table synthesises these considerations into a high-level, contextual framework, providing a strategic guide for mapping common application domains to a recommended modelling approach.

Table 3. *Contextual Decision Framework for Model Prioritisation.*

Application Domain	Primary Goal	Key Constraints	Dominant Factor	Recommended Approach
Credit Scoring / Lending	Accountable, fair decision-making	High; Legal (ECOA, GDPR), ethical (fairness), regulatory (CFPB)	Interpretability	Prioritize inherently interpretable models (GAMs, sparse linear) or use black-box models with rigorously validated XAI for adverse action notices.
Medical Diagnosis	Accurate prediction with expert validation	High; Patient safety, clinician trust, ethical accountability	Interpretability (as a prerequisite for trust)	Use high-performance models (e.g., DNNs) but pair them with robust XAI (LIME, SHAP, saliency maps) to make them transparent to clinicians. [16]
Algorithmic Trading (HFT)	Automated, high-speed profit generation	Low (in terms of interpretability); speed and performance are key	Accuracy	Deploy high-complexity, black-box models optimized for speed and predictive precision. Interpretability is for offline analysis and debugging only.
Customer Churn Analysis	Strategic insight and business understanding	Moderate; business value depends on actionable insights	Interpretability	Favor inherently interpretable models (EBMs, GAMs, decision trees) that reveal the "why" behind churn, even at the cost of a small drop in accuracy.
Spam Filtering / Recommendation	Enhance user experience	Low (for end-user); moderate for developers (bias, robustness)	Accuracy	Use high-accuracy models. Interpretability is a secondary tool for internal teams to debug, audit for fairness, and defend against attacks.

This framework demonstrates that there is no single best practice. Instead, the "right" approach is a strategic choice informed by a clear-eyed assessment of the problem's unique context. It moves the practitioner from a purely technical evaluation of algorithms to a more comprehensive, business- and risk-aware mode of thinking.

4.2. The Future of Interpretability: Dissolving the Trade-off

The accuracy-interpretability trade-off has been a defining feature of the current era of machine learning, but it is not necessarily a permanent or fundamental law of artificial intelligence. The active frontiers of research are focused on developing new methods and models that aim to reconcile these two competing goals, potentially dissolving the trade-off altogether.

One promising direction is the maturation of the "add interpretability" paradigm. [17] This reframes the problem not as a sacrifice of accuracy for interpretability, but as a two-step process: first, build the most accurate model possible using any available technique; second, apply a suite of increasingly sophisticated XAI tools to add the necessary layer of explanation [26]. As XAI methods become more robust, computationally efficient, and better at handling challenges like feature correlation and explanation stability, this approach will become more viable even in high-stakes environments.

An even more fundamental approach is the development of inherently interpretable-by-design architectures. [26] This research goes beyond using existing simple models. Instead, it focuses on creating entirely new classes of models that are designed from the ground up to be both powerful and transparent. This includes novel neural network architectures that incorporate prototype-based reasoning, or advanced tree ensembles that are constructed in a way that their feature contributions remain separable and understandable. [26] The goal of this research is to create models that reside in the "top-right quadrant" of the accuracy-interpretability plane, achieving high performance without opacity.

Perhaps the most profound long-term development is the rise of Causal AI. The majority of current machine learning models, including the most complex deep learning systems, are fundamentally correlation-based. They are masters at identifying statistical associations in data (e.g., feature X is associated with outcome Y), but they have no understanding of cause and effect. The emerging field of Causal Machine Learning aims to build models that can learn and reason about the underlying causal mechanisms of a system-to understand that *A causes B*. A model that understands causality would represent the pinnacle of interpretability; it could provide the ultimate explanation for a prediction by outlining the causal chain of events that led to it. Furthermore, a model that understands the true data-generating process should, in theory, be more robust and achieve a higher level of generalisation accuracy than one that merely learns superficial correlations.

These future directions suggest that the current tension between accuracy and interpretability is a symptom of the limitations of our present-day tools, which are largely optimised for correlation-based pattern recognition. [29] As the field evolves toward models that can incorporate structural knowledge, reason about causality, and are designed with transparency as a core principle, the dilemma may not simply be navigated or managed, but ultimately resolved. The future of AI may not require a trade-off, but instead deliver both superior performance and perfect explanation as two sides of the same coin.

Abbreviations

AI	Artificial Intelligence
ALE	Accumulated Local Effects
AUC	Area Under Receiver Operating Characteristic Curve
CART	Classification and Regression Trees
CFPB	Consumer Financial Protection Bureau
COVID	Coronavirus Disease
DNN	Deep Neural Network
EBM	Explainable Boosting Machine
ECOA	Equal Credit Opportunity Act
GAM	Generalised Additive Models
GDPR	General Data Protection
GOSDT	Globally Optimised Sparse Decision Trees
Grad-CAM	Gradient-weighted Class Activation Mapping
HFT	High-frequency Trading
LIME	Local Interpretable Model-agnostic Explanations
MAE	Mean Absolute Error
MRI	Magnetic Resonance Imaging
NLP	Natural Language Processing
PDP	Partial Dependence Plot
RMSE	Root Mean Square Error
SHAP	SHapley Additive exPlanations
SVM	Support Vector Machine
XAI	Explainable Artificial Intelligence

Author Contributions

Partha Majumdar is the sole author. The author read and approved the final manuscript.

Conflicts of Interest

The author declares no conflicts of interest.

References

[1] Abdullah, T. A. A., Soperi, M., Zahid, M., & Ali, W. (2021, December 17). *A Review of Interpretable ML in Healthcare: Taxonomy, Applications, Challenges, and Future Directions*. Retrieved August 30, 2025, from https://www.mdpi.com/2073-8994/13/12/2439

[2] Atrey, P., Brundage, M. P., Wu, M., & Dutta, S. (2025, March 10). *Demystifying the Accuracy-Interpretability Trade-Off: A Case Study of Inferring Ratings from Reviews*. Retrieved August 30, 2025, from https://arxiv.org/html/2503.07914v1

[3] auxiliobits (n.d.). *Explainable AI in Credit Risk Assessment: Balancing Performance and Transparency*. Retrieved August 30, 2025, from https://www.auxiliobits.com/blog/explainable-ai-in-credit-risk-assessment-balancing-performance-and-transparency/

[4] Babic, B., & Hanyin, Z. (2024). *Reframing the Accuracy/Interpretability Trade-Off in Machine Learning*. Retrieved August 30, 2025, from https://borisbabic.com/research/IAT_August2024.pdf

[5] Bak, U. (2024, March 12). *Techniques for Explainable AI: LIME and SHAP*. Retrieved August 30, 2025, from https://www.unnatbak.com/blog/techniques-for-explainable-ai-lime-and-shap

[6] Bhatnagar, V. (2024, March 28). *Interpretable Algorithms as a Potential Solution to CFPB's Guidance on AI-driven Credit Denials*. Retrieved August 30, 2025, from https://www.goodwinlaw.com/en/insights/blogs/2024/03/interpretable-algorithms-as-a-potential-solution-to-cfpbs-guidance-on-ai-driven-credit-denials

[7] Charlton, M., & Oak, L. S. (2025, August 4). *Disparate Impact Discrimination in AI Underwriting Alleged Against Student Loan Company*. Retrieved August 30, 2025, from https://frostbrowntodd.com/disparate-impact-discrimination-in-ai-underwriting-alleged-against-student-loan-company/

[8] Chen, J. (2025, August 10). *Understanding High-Frequency Trading (HFT): Basics, Mechanics, and Example*. Retrieved August 30, 2025, from https://www.investopedia.com/terms/h/high-frequency-trading.asp

[9] Choudhary, I. (2023, July 21). *The power of data: How recommendation engines are driving business growth*. Retrieved August 30, 2025, from https://www.mindtheproduct.com/the-power-of-data-how-recommendation-engines-are-driving-business-growth/

[10] Consumer Financial Protection Bureau (2024, August 28). *Consumer Financial Protection Circular 2022-03*. Retrieved August 30, 2025, from https://www.consumerfinance.gov/compliance/circulars/circular-2022-03-adverse-action-notification-requirements-in-connection-with-credit-decisions-based-on-complex-algorithms/

[11] Dankworth, C., Gesser, A. S., Patterson, J., Litt, L., & Mogul, A. (2023, October 30). *Adverse Action Notice Compliance Considerations for Creditors That Use AI*. Retrieved August 30, 2025, from https://www.americanbar.org/groups/business_law/resources/business-law-today/2023-november/adverse-action-notice-compliance-considerations-for-creditors-that-use-ai/

[12] De Harder, H. (2020, January 4). *Interpretable Machine Learning Models*. Retrieved August 30, 2025, from https://towardsdatascience.com/interpretable-machine-learning-models-aef0c7be3fd9/

[13] Demajo, L. M., Vella, V., & Dingli, A. (n.d.). *EXPLAINABLE AI FOR INTERPRETABLE CREDIT SCORING*. Retrieved August 30, 2025, from https://arxiv.org/pdf/2012.03749

[14] Ennab, M., & Mcheick, H. (2024, November 28). *Enhancing interpretability and accuracy of AI models in healthcare: A comprehensive review on challenges and future directions*. Retrieved August 30, 2025, from https://pmc.ncbi.nlm.nih.gov/articles/PMC11638409/

[15] Fathima, S. (2024, February 26). *LIME vs SHAP: A Comparative Analysis of Interpretability Tools*. Retrieved August 30, 2025, from https://www.markovml.com/blog/lime-vs-shap

[16] GeeksforGeeks (2025, July 23). *Explainable AI(XAI) Using LIME*. Retrieved August 30, 2025, from https://www.geeksforgeeks.org/artificial-intelligence/introduction-to-explainable-aixai-using-lime

[17] Goodrum, W. (2016, November 4). *Balance: Accuracy vs. Interpretability in Regulated Environments*. Retrieved August 30, 2025, from https://www.elderresearch.com/blog/balance-accuracy-vs-interpretability-in-regulated-environments/

[18] Griffith, M. A. (2023, January 3). *AI Lending and ECOA: Avoiding Accidental Discrimination*. Retrieved August 30, 2025, from https://scholarship.law.unc.edu/cgi/viewcontent.cgi?article=1570&context=ncbi

[19] Hammann, D., & Wouters, M. (2025, January 24). *Explainability Versus Accuracy of Machine Learning Models: The Role of Task Uncertainty and Need for Interaction with the Machine Learning Model*. Retrieved August 30, 2025, from https://www.tandfonline.com/doi/full/10.1080/09638180.2025.2463961

[20] Kawski, M. V., McFall, A., & Lomax, S. (2023, October 11). *CFPB Issues New Guidance for AI-Driven Credit Decisions*. Retrieved August 30, 2025, from https://www.huschblackwell.com/newsandinsights/cfpb-issues-new-guidance-for-ai-driven-credit-decisions

[21] Kenton, W. (2025, August 21). *What Is a Black Box Model? Definition, Uses, and Examples*. Retrieved August 30, 2025, from https://www.investopedia.com/terms/b/blackbox.asp

[22] Kosinski, M. (2024, October 29). *What is black box AI?* Retrieved August 30, 2025, from https://www.ibm.com/think/topics/black-box-ai

[23] Kumar, R. (2024). *Balancing Model Accuracy and Interpretability: How do you navigate this trade-off?* Retrieved August 30, 2025, from https://www.kaggle.com/discussions/questions-and-answers/519788

[24] Lee, S. (2025, March 27). *10 Stats on How Recommendation Systems Elevate E-commerce Retail*. Retrieved August 30, 2025, from https://www.numberanalytics.com/blog/10-stats-recommendation-systems-ecommerce-retail

[25] McElroy, J. (2023, October 26). *CFPB Issues Guidance on Use of AI in Credit Decisioning*. Retrieved August 30, 2025, from https://www.saul.com/insights/alert/cfpb-issues-guidance-use-ai-credit-decisioning

[26] Molnar, C. (n.d.). *Interpretable Machine Learning*. Retrieved August 30, 2025, from https://christophm.github.io/interpretable-ml-book/

[27] Ndungula, S. (2022, November 25). *Model Accuracy and Interpretability*. Retrieved August 30, 2025, from https://samuelndungula.medium.com/model-accuracy-and-interpretability-3d875439942c

[28] Nesvijevskaia, A., Ouillade, S., Guilmin, P., & Zucker, J. D. (2021, July 5). *The accuracy versus interpretability trade-off in fraud detection model*. Retrieved August 30, 2025, from https://www.cambridge.org/core/journals/data-and-policy/article/accuracy-versus-interpretability-tradeoff-in-fraud-detection-model/3BA1586A6B635E08BE556FAB89AD8770

[29] Nussberger, A. M., Luo, L., Celis, L. E., & Crockett, M. J. (2022, October 3). *Public attitudes value interpretability but prioritize accuracy in Artificial Intelligence*. Retrieved August 30, 2025, from https://pmc.ncbi.nlm.nih.gov/articles/PMC9528860/

[30] Pace, R. (2024, May 6). *Using Explainable AI to Produce ECOA Adverse Action Reasons: What Are The Risks?* Retrieved August 30, 2025, from https://www.paceanalyticsllc.com/post/ecoa-adverse-actions-and-explainable-ai

[31] Potdar, D. (2023). *Churn Prediction using Interpretable Machine Learning*. Retrieved August 30, 2025, from https://github.com/dhavalpotdar/interpretable-churn-prediction

[32] Thought Leadership (2024, July 2). *The Importance of Interpretable AI in the Financial Services Industry*. Retrieved August 30, 2025, from https://www.ncino.com/news/importance-interpretable-ai-financial-services-industry

[33] Vimbi, V., Shaffi, N., & Mahmud, M. (2024, April 5). *Interpreting artificial intelligence models: A systematic review on the application of LIME and SHAP in Alzheimer's disease detection*. Retrieved August 30, 2025, from https://pmc.ncbi.nlm.nih.gov/articles/PMC10997568/

[34] Welch, D. M., & Burcat, B. A. (2024, January 24). *CFPB Applies Adverse Action Notification Requirement to Artificial Intelligence Models*. Retrieved August 30, 2025, from https://www.skadden.com/insights/publications/2024/01/cfpb-applies-adverse-action-notification-requirement

[35] Zhang, C. (2024). *Enhancing Spam Filtering: A Comparative Study of Modern Advanced Machine Learning Techniques*. Retrieved August 30, 2025, from https://www.itm-conferences.org/articles/itmconf/pdf/2025/01/itmconf_dai2024_04013.pdf

[36] Zhang, C., Liu, J., Qu, Y., Han, T., Ge, X., & Zeng, A. (2018, November 1). *Enhancing the robustness of recommender systems against spammers*. Retrieved August 30, 2025, from https://pmc.ncbi.nlm.nih.gov/articles/PMC6211683/

[37] Zhou, X. (2019, June 25). *Interpretability Methods in Machine Learning: A Brief Survey*. Retrieved August 30, 2025, from https://www.twosigma.com/articles/interpretability-methods-in-machine-learning-a-brief-survey/