2022

Integrated Gradients is a Nonlinear Generalization of the Industry Standard Approach to Variable Attribution for Credit Risk Models

Jonathan Boardman
Kennesaw State University, jboardma@students.kennesaw.edu

Md Shafiul Alam

Xiao Huang
Kennesaw State University, xhuang3@kennesaw.edu

Ying Xie
Kennesaw State University, yxie2@kennesaw.edu

Recommended Citation

J. Boardman, M. S. Alam, X. Huang and Y. Xie, "Integrated Gradients is a Nonlinear Generalization of the Industry Standard Approach to Variable Attribution for Credit Risk Models," 2022 IEEE International Conference on Big Data (Big Data), Osaka, Japan, 2022, pp. 5012-5023, doi: 10.1109/BigData55660.2022.10020687.

Integrated Gradients is a Nonlinear Generalization of the Industry Standard Approach to Variable Attribution for Credit Risk Models

Jonathan Boardman
Data Science and Analytics
Kennesaw State University
Kennesaw, USA
jboardma@students.kennesaw.edu

Md Shafiul Alam
Data Science and Analytics
Kennesaw State University
Kennesaw, USA
malam6@students.kennesaw.edu

Xiao Huang
Economics, Finances, &
Quantitative Analysis
Kennesaw State University
Kennesaw, USA
xhuang3@kennesaw.edu

Ying Xie
Information Technology
Kennesaw State University
Kennesaw, USA
yxie2@kennesaw.edu

***Abstract*— In modern society, epistemic uncertainty limits trust in financial relationships, necessitating transparency and accountability mechanisms for both consumers and lenders. One upshot is that credit risk assessments must be explainable to the consumer. In the United States regulatory milieu, this entails both the identification of key factors in a decision and the provision of consistent actions that would improve standing. The traditionally accepted approach to explainable credit risk modeling involves generating scores with Generalized Linear Models (GLMs) - usually logistic regression, calculating the contribution of each predictor to the total points lost from the theoretical maximum, and generating reason codes based on the 4 or 5 most impactful predictors. The industry standard approach is not directly applicable to a more expressive and flexible class of nonlinear models known as neural networks. This paper demonstrates that an eXplainable AI (XAI) variable attribution technique known as Integrated Gradients (IG) is a natural generalization of the industry standard to neural networks. We also discuss the unique semantics surrounding implementation details in this nonlinear context. While the primary purpose of this paper is to introduce IG to the credit industry and argue for its establishment as an industry standard, a secondary goal is to familiarize academia with the legislative constraints – including their historical and philosophical roots – and sketch the standard approach in the credit industry since there is a dearth of literature on the topic.**

Keywords*—Credit Scoring, Explainable Artificial Intelligence, Machine Learning Deep Learning, Integrated Gradients, Trust, Transparency, Fairness

I. Introduction: Credit, Trust, and Explainability

Assessments of creditworthiness – the probability that a borrower will meet their financial obligations – are a critical component of the lending process. These assessments are derived from historical financial records (i.e. – credit reports) that, for consumers in the United States, are maintained primarily by 3 Consumer Reporting Agencies (CRAs), known colloquially as 'credit bureaus': Equifax, TransUnion, and Experian [1]. Risk models are often built on the raw information in these reports to generate credit scores, which quantify the probability of repayment for a potential borrower. Some scores are specific to particular market segments, while others are more general. Of the general scores, the most common are the 3 digit FICO8/9 developed by Fair Isaac Corp. [2] and the VantageScore 3.0/4.0 developed jointly through the big 3 CRAs [3].

Lenders are not the only ones that have come to rely on credit reports. Employers, landlords, insurance companies, and even prospective dates may wish to leverage this sensitive information on the grounds that good credit is associated with good moral character – trust, in particular [4]. While making broad ethical generalizations about individuals based on their creditworthiness may be specious, credit scores do seem to be quite effective at discriminating between borrowers who can safely be trusted to pay as agreed from those who cannot [5]–[7].

This is no accident. Credit reporting was developed as a solution to practical ethical concerns involving trust in financial relationships and many of the ethical concerns around this solution have their roots in a still largely unappreciated fundamental tension between trust and transparency [8], [9].

A. Trust and Transparency

The philosophy of trust (as well as other ethical concepts we touch on) is not a settled matter. Here, we consider **trust** to be a *justified* belief in the **trustworthiness** of another party, where we regard trustworthiness to be some function of the *willingness* and *capability* of that party to behave as desired. If the party is, in fact, trustworthy, then – as a true, justified, belief – trust arguably constitutes a kind of knowledge about that party's character, and this knowledge allows us to make ourselves vulnerable to them. Luck may still play a prominent role in the outcome, but we can at least count on that party to do their best. Of course, we can never know for certain that the other party will not betray our trust. The degree to which our reasons and evidence support our belief in the other party's trustworthiness is our **epistemic certainty**. Epistemic certainty is crucial to trust, distinguishing trust from blind faith.

2022 IEEE International Conference on Big Data (Big Data), Osaka, Japan, 2022, pp. 5012-5023, doi: 10.1109/BigData55660.2022.10020687.

Consider the following situation involving 2 parties – Alice and Bob: Alice and Bob enter into a relationship where Alice expects Bob to behave in a certain way and Bob expects Alice to behave in a certain way. For Alice and Bob to *rationally* and *voluntarily* enter into such a relationship and if one or both are made vulnerable, then the vulnerable party or parties must extend trust (Note that without vulnerability – either because there is no risk or there exists some guarantee of the other party's behavior – trust is unnecessary [10]). Under conditions of sub-optimal epistemic certainty, trust cannot be extended. One way to overcome this is by making up the difference with **transparency** requirements. As a form of surveillance, transparency requirements come at the cost of the other party's **privacy**, but they offer some degree of **security** to the vulnerable party. Power and power differentials factor in here, as well, though; if a vulnerable party has no ability to hold the other party **accountable** for their actions, then transparency may give little security. Finally, note that if transparency and accountability requirements are too onerous and invasive, they may infringe privacy to such a degree that **autonomy** and **dignity** of the surveilled party are infringed, perhaps even hamstringing their ability to behave in the manner expected by the other party [9].

Contrary to popular belief, then, transparency is not *required* for trust. It may, however, be required to *establish* trust. Transparency allows a trustor to make up for a lack of epistemic certainty, but, as evidence is accumulated over time, trust may be extended and transparency requirements may be withdrawn. Many of us have found ourselves at one end of this ethical quagmire at some point in our lives. While these concepts are usually considered in an intimate interpersonal context (e.g. – friends, family), they are also relevant in other social contexts, such as financial and political relationships.

In the United States in the early 1800s, this contextual distinction was rare. Most financial and political relationships involved people that were well-acquainted. By the middle of the 19th century, however, this milieu was changing. As settlers pushed westward and immigration surged, trade networks grew more complex and increasingly involved connections between relative strangers in ever growing chains of credit [11], [12]. For example, in the spring and fall, scores of merchants from across the country would descend on the port cities to procure goods from wholesalers [12]. According to Olegario, "…almost all applied to buy goods on credit from wholesalers who were obliged to decide, in the lingo of nineteenth century commerce, whether or not to 'trust' them" [11, p. 5]. In the highly competitive environment, the most imminent risk of losing customers often forced wholesalers to adopt overly liberal policies [11]. In 1837, the problem with all of this unwarranted trust became painfully apparent as the country plunged into a deep recession.

B. *The Need for Transparency and the Evolution of Credit Reporting in the United States*

In 1841, after being nearly ruined in the wake of the Panic of 1837, Lewis Tappan and his brother established the Mercantile Agency, the world's first credit reporting firm. The firm compiled information on "…all known businesses in the United States…[,] with detailed reports on the personal character, financial means, and local reputations of their proprietors" [12, p. 303]. The Tappans were not the only ones whose business had been shuttered in the recession, and the Bankrupty Act of 1841, which "…provided the insolvent with generous legal and financial protection" [12, p. 307], left many merchants feeling vulnerable and embittered. The Mercantile Agency offered much-needed transparency, allowing business owners to make more calibrated assessments and justify their decisions to extend trust.

The Mercantile Agency – later R.G. Dun and Company – and its primary competitor, Bradstreet Company, were concerned chiefly with *commercial* credit reporting. In fact, to this day Dun & Bradstreet – the result of a 1933 merger between the two rivals – continues in this vein. Consumer credit reporting took nearly half a century longer evolve [13]. Even in the largest cities in the latter half of the nineteenth century, most shopkeepers still knew their customers well enough that credit reports were deemed unnecessary. Moreover, participating in the emerging retail credit reporting infrastructure involved a quid pro quo that many found to be either too loathsome from the standpoint of competition or conscience or too administratively burdensome [13]. However, consumer credit found a toehold in department stores, installment houses, and other large retail stores. During the early 1900s, the evolution of filing and communication technology and the professionalization of retail credit management drove more widespread adoption, and credit reporting by local CRAs became an indispensable tool for assessing the financial trustworthiness of individuals and businesses alike.

While consumer credit reports offered transparency to *businesses*, consumers in the mid-twentieth century were often left in the dark, not only regarding the contents of these reports, but also regarding how these reports and other information collected by lenders were used to arrive at potentially life-altering decisions. In a mere hundred years, the pendulum had swung, and the consumer became a vulnerable party in the credit relationship. As we will see in Section 2, in the second half of the twentieth century, the United States enacted legislation around fair lending to make the process more transparent, more accurate, and less discriminatory. Importantly, lenders were required to provide consumers with explanations regarding adverse action (i.e. – denial of credit), allowing consumers to understand not only the *who* and the *what*, but the *why*.

C. *Explainable Credit Scoring in the Age of AI*

Exactly what constitutes an acceptable answer to a 'why' question, or, more precisely, what constitutes an *explanation* is complex. There is a rich body of work on this topic in philosophy [14]–[17], cognitive science [18], [19], social psychology [20], [21], and, increasingly, in machine learning [22], and, while we draw on this literature at points, our scope is narrower and grounded in traditional notions of explainability for credit scoring models in the context of the United States regulatory milieu. We outline the history and structure of relevant regulation in Section 2, and, in Section 3, detail the current industry standard approach to explainable credit modeling and extracting explanations. This approach relies on linear models, which are naturally interpretable and, usually, yield explanations that are both faithful to the mechanism and that are comprehensible by humans – fulfilling transparency requirements not only to the consumers, but to the lenders who rely on them. Modern machine learning techniques (deep neural networks, specifically) hold significant promise for the credit industry, but they are not *obviously* amenable to explanation in the way that linear models are. In this paper, we show that, in the context of the industry standard approach to credit scoring and with appropriately implemented monotonicity constraints, they are.

We argue that to accomplish this requires a principled approach to variable attribution, and, in Sections 4 and 5, we review the problem of variable attribution for differentiable, nonlinear models (e.g. – neural networks), survey the landscape of approaches, and conclude that a technique known as Integrated Gradients is not only desirable, but that it is actually *equivalent* to the industry standard approach for linear models and represents a generalization of the approach to differentiable, nonlinear models. In Section 6, we discuss interesting semantic considerations that appear in the context of baseline variable attribution methods (e.g. – Integrated Gradients) applied to nonlinear models that are absent in the linear case. In Section 7, we conclude with remarks on interesting future directions.

II. Creditworthiness: From Reputation to Regulation

We defined creditworthiness in the introduction as the probability that a borrower will meet their financial obligations, but how does one assess such a thing? Clearly, it cannot be measured directly, and the outcome can only be known with the benefit of hindsight. One important factor is certainly "…whether one [is] the sort of person who [feels] sufficiently constrained, by conscience or social obligation, to [repay one's debts]" [12, p. 307] (i.e. - trustworthiness). However, external factors play a role in whether even the most trustworthy and well-intentioned individual succeeds [23], and, in the case of failure, the ability to liquidate assets or lean on one's social network can attenuate losses. Thus, historical, economic, and environmental information cannot be ignored.

In the early days of credit reporting, these factors were captured by "the 'three Cs' of credit reporting: character, capacity, and capital" [12, p. 309]:

> Each category had its own implicit indicators. For character: the individual's work habits (hard working? conscientious?), local reputation (well liked? trusted?), and personal life (married? alcoholic? gambler? philanderer?). For capacity: age, experience in business, past employment, and known history of successes or failures. For capital: assets, liabilities, and property owned by the individual, as well as assets potentially available through well-to-do family or business connections who might rescue an individual in default [12, p. 309-310].

Despite its highly subjective nature, 'character' has historically held the preeminent role in credit evaluations; honest and hardworking debtors were preferential to the wealthy and capable, but unscrupulous, as the former would pay what they could while the latter were unpredictable. Indeed, early credit reports available through Tappan's Mercantile Agency were often little more than a brief encapsulation of one's public reputation, and any financial information was usually hearsay, as well. The rise of reference books with coded ratings in the latter half of the nineteenth century started the move towards quantification and, by the turn of the century, financial information was usually split out from general reputation. Consumer credit reporting inherited many of the practices of commercial credit reporting, and, in addition to ledger information, subjective character assessments remained a core component of both credit reports and credit evaluations into the mid twentieth century. As demand for consumer credit exploded during the 1910s and 20s, many credit managers relied on generalizations and blacklists to quickly sort applicants based on occupation, location of residence, and other perceived correlates of financial deviance [13]. Initial screenings were frequently followed by face-to-face interviews, which were potentially subject to any manner of idiosyncratic and socio-normative biases.

Despite the potential problems with such high-degrees of subjectivity in both data and decisioning, the issue that catalyzed an avalanche of congressional hearings between 1966 and 1969 was a more general issue with quality control in credit reporting [13]. In the late 1960s, with the exceptions of the Credit Data Corporation and the Retail Credit Company, credit reporting was largely distributed across thousands of local bureaus, and a lack of standardization and regulation meant that report quality varied significantly and data provenance was uncertain. Even if a consumer was able to ascertain the origins of their potential multiplicity of credit reports, they were usually not permitted to view the contents of those reports. In the case of adverse action (e.g. – denial of credit), they often received little to no explanation.

Congress responded to this need with the Consumer Credit Protection Act of 1968 (CCPA) and the Fair Credit Reporting Act of 1970 (FCRA), which, among other provisions, limited who could obtain credit reports, restricted the kinds of information contained in the reports, put time limits on derogatory information, outlined specific procedures for notifying consumers in the case of adverse action, including giving consumers the CRA's contact information and a right to know the contents of their credit report, and requiring CRAs to respond to disputes and correct inaccuracies in a timely manner. Later amendments gave consumers the right to a free annual copy of their credit report from each CRA and gave enforcement responsibility to the Consumer Financial Protection Bureau (CFPB) in addition to the Federal Trade Commission (FTC) (originally charged with enforcement in 1971) among several other provisions [24]. While the FCRA introduced much needed transparency and accountability into the reporting aspect of credit, it did little for transparency on the decisioning side where the subjectivity in credit managers' decisioning processes intersected growing public concerns about unfair discrimination.

Since the dawn of commerce, decisions to grant credit had been based on professional judgment and, therefore, were prone to influence from irrelevant information and undesirable biases. Even worse, human intuition is a black box. Explanations for such ‘system 1’ thinking are necessarily retroactive and ad-hoc [25]. Unconscious recall and framing biases may further pollute explanations in unpredictable and misleading ways [25], [26]. This opacity makes it difficult to determine whether and to what degree discrimination [27] or disparate treatment [28] has occurred.

The advent of statistical risk modeling in the 1930s provided a tool to reduce subjectivity in decisioning by forcing creditors to explicitly encode the factors they considered relevant to decision-making. These tools were also more interpretable and transparent than human decision-making. Linear models, in particular, lend themselves naturally to explanation (as we will see in next section). However, as these tools gained wider acceptance in the late 1960s, it became clear that the inherent transparency and explainability afforded by these tools did not eliminate issues of fairness and discrimination; it only made them explicit, accessible, and (to some degree) quantifiable in ways that were impossible when dealing with credit managers. According to Lauer:

> Consumer lenders and retailers were in business to make money. Generally speaking, they did not reject female credit applicants because of their gender, but because women typically earned less than men and often left the workforce (and their own incomes) to have and raise children. Likewise, lenders did not uniformly refuse to lend to African Americans as a class, but avoided dealing with residents of unstable, low-income inner-city neighborhoods where many African Americans lived…. Credit decisions that privileged men over women and whites over African Americans were a reflection of real structural inequalities in American society [13, p. 236].

A first step in addressing these issues was taken with the Equal Credit Opportunity Act of 1974 (ECOA) and subsequent amendments in 1976 – implemented in Regulation B, which forbade discrimination “(1) on the basis of race, color, religion, national origin, sex or marital status, or age (provided the applicant has the capacity to contract); (2) because all or part of the applicant’s income derives from any public assistance program; or (3) because the applicant has in good faith exercised any right under this chapter” [29]. It did allow the use of age under the caveat that its use could not harm judgments of creditworthiness in older applicants.

Most relevant for the purposes of this paper, however, the ECOA as implemented in Regulation B introduced new adverse action disclosure requirements. Both the FCRA and the ECOA / Regulation B have requirements surrounding adverse action notifications. Those in the FCRA [30] are more about the who and the what (i.e. – who is responsible and what information). Those in the ECOA / Regulation B [31] also have (different) who/what requirements, but they also contain requirements on why. The creditor must not only explain what action was taken, but they must give the consumer specific reasons about why:

> **(2) Statement of specific reasons.** The statement of reasons for adverse action required by paragraph (a)(2)(i) of this section must be specific and indicate the principal reason(s) for the adverse action. Statements that the adverse action was based on the creditor's internal standards or policies or that the applicant, joint applicant, or similar party failed to achieve a qualifying score on the creditor's credit scoring system are insufficient [31].

Together, the FCRA and ECOA / Regulation B are the legal bedrock of transparency and accountability for consumers in the U.S. credit industry. In the rest of this paper, we will focus specifically on the portion of this legislation related to explaining adverse credit decisions to the consumer – how it is done, how it should be done, and how it can be extended in a way that is compatible with modern deep learning systems and that is mutually beneficial for both lenders and consumers. We begin this journey with the relevant section of the CFPB’s current interpretation of the fair lending legislation:

> ***Credit score and key factors disclosed***
>
> In addition to the notice, a creditor, such as a financial institution, must also disclose the credit score, the range of possible scores, the date that the score was created, and the “key factors” used in the score calculation. “Key factors” are all relevant elements or reasons adversely affecting the credit score for the particular individual, listed in the order of their importance, and based on their effect on the credit score. The total number of factors to be disclosed must not exceed four. However, if one of the key factors is the number of inquiries into a consumer’s credit information, then the total number of factors must not exceed five. These key factors come from information the consumer reporting agencies supplied with any consumer report that was furnished containing a credit score (Section 605(d)(2)) [24, p. 28].

This interpretation still leaves room for debate. In the next section, we will look at the current industry standard approach to regulatory-compliant explainable credit risk modeling.

III. The Industry Standard Approach

Based on much of the academic literature, the landscape of credit scoring models appears as diverse as any other field within machine learning [32]–[35]. An extensive survey conducted by Louzada et al. covering the period from January 1992 through December 2015 indicated that neural networks and Support Vector Machines (SVMs) were the dominant classes of models used in credit scoring, not counting hybrid models and model ensembles [35]. Tree-based models and Bayesian networks made strong showings, as well. Use of logistic regression was tied with neural networks for first place only in recent years, but, generally, according to the study, appeared to be just one of many methods in common use.

The industry viewpoint contrasts starkly. Logistic regression has been the industry standard approach to binary risk classification for decades [6], [7], and it is only due to recent innovations in monotonically constrained neural nets and tree-based models that the landscape has begun to shift [36]. Compliance with regulation (see Section 2) is paramount, and Generalized Linear Models (GLMs) – logistic regression, specifically – are viewed by lenders and regulators alike as the gold standard. Not only are they well understood [37] and decently predictive in the credit scoring context [5]–[7], but they naturally yield explanations that satisfy the ECOA as implemented by Regulation B.

To illustrate, consider the following logistic regression model with m predictor variables:

$$\begin{aligned}\ln\left(\frac{p}{(1-p)}\right) = f(\boldsymbol{x}) &= \beta_0 + \beta_1 \cdot x_1 + \cdots + \beta_m \cdot x_m \\ &= \beta_0 + \boldsymbol{\beta} \cdot \boldsymbol{x}\end{aligned}$$

This logistic regression models the natural logarithm of the odds (logodds) of defaulting (p being the probability of defaulting) as a linear combination of predictor variables $\boldsymbol{x}$, where $\boldsymbol{\beta}$ is an associated vector of coefficients and β_0 is the intercept. In practice, the predictors are engineered in a way that allows the model output to be converted into a fixed range score where higher is better. For now, however, we will simply work with the logodds of default $f(\boldsymbol{x})$.

Suppose a customer with a given vector of attributes $\boldsymbol{x}$ is seeking credit from a lender, but the customer's predicted logodds of default $f(\boldsymbol{x})$ exceeds the lender's pre-defined threshold. When the lender denies credit to that customer, the fair lending laws discussed in Section 2 may be triggered. Among the requirements is an obligation to report the specific and principal reasons for the adverse action corresponding to the top 4 or 5 key factors listed in order of importance based on size of effect on the score.

The standard approach to generating such reasons begins by identifying the attribute vector $\boldsymbol{x}'$ corresponding to the hypothetical ideal/perfect customer. This process is straightforward if the following conditions are met:

1. All predictors have a known maximum, and their minimum values are 0 (As binning is standard practice, this is usually not an issue).
2. The relationship between each predictor and the default probability is monotonic.
3. For each predictor:
    a. If the predictor is positively correlated with default probability, then leave it as is.
    b. Otherwise, transform the predictor by subtracting it from its maximum possible value

If these conditions are met, then the ideal customer corresponds to the zero vector. The predicted logodds of default for this ideal customer $f(\boldsymbol{x}')$ is just the intercept term β_0.

The difference in the predicted logodds of default between the prospective customer and the theoretical ideal is as follows:

$$\begin{aligned} f(\boldsymbol{x}) - f(\boldsymbol{x}') &= \beta_0 + \boldsymbol{\beta} \cdot \boldsymbol{x} - (\beta_0 + \boldsymbol{\beta} \cdot \boldsymbol{x}') \\ &= \boldsymbol{\beta} \cdot (\boldsymbol{x} - \boldsymbol{x}') \\ &= \boldsymbol{\beta} \cdot \boldsymbol{\Delta x} \\ &= \beta_1 \cdot \Delta x_1 + \cdots + \beta_m \cdot \Delta x_m \end{aligned}$$

Thus, the difference in predicted values can be expressed as the sum of the contributions from the marginal differences in each predictor variable, where the contribution is the coefficient of that variable times the difference in the values of that variable for the prospective customer and the theoretical ideal. This linearity makes it simple to list the predictors in descending order of their contribution to the difference; the top 4 predictors (5 if number of enquiries is amongst them) are regarded as the key factors in the adverse decision. The corresponding reasons are presented to the customer [38].

Since logistic regression assumes the relationships between independent (predictor) variables and the logodds are linear and monotonic, if the customer takes rational action to improve their score, it will always improve. For example, say adverse action is taken because credit utilization is too high. The monotonic relationship between credit utilization and logodds of default means that if the customer decreases their credit utilization, then they can be confident that their predicted logodds of default will decrease. It will never be the case that decreasing credit utilization will *increase* the logodds of default. Moreover, monotonicity ensures that it will never be the case that 2 different customers will receive reasons with conflicting implications regarding to the same attribute. For example, it will never be the case that customer A will be penalized for having a credit utilization that is too high, while customer B will be penalized for having a credit utilization that is too low. Monotonicity ensures that explanations are consistently actionable.

While consistent actionability is not an explicit requirement of the ECOA or Regulation B, not satisfying this requirement would clearly contradict at least some aspect of the teleological intent underpinning this legislation. As we discussed in Section 2, transparency for consumers in the lending approval process is one part of this intent, and the legitimacy of the explanation as a transparent window into the decisioning process requires the possibility of counterfactual reasoning. For example, a claim that the value of some factor was too high implies that a lower value would be better. If this were not so, then the reason is underspecified; it is unclear how the factor was used to reach a decision.

So, there are at least 3 requirements that adverse action explanations must meet:

1. The explanations must be consistently actionable.
2. The explanations must correspond to the 4 (or 5) key factors in the decision [24].
3. The explanations must be rank-ordered by the size of the contributions of their corresponding key factors to the decision [24].

Logistic regression readily yields explanations that meet these requirements due to the linearity in the relationship between the predictors and the logodds of default. Not only does linearity ensure monotonicity, it ensures that the model coefficients are the same for all possible combinations of attributes. Because these coefficients are constant, it is simple to determine how much a change in a predictor will change the score (see Figure 1), and, conversely, linear separability makes it easy to attribute any change in score to the predictors. For nonlinear models, these properties may not hold. The relationship with the output may be nonmonotonic, and, in the case of differentiable nonlinear models, the gradient of the output with respect to the predictors is not constant (i.e. – the coefficients of a local linear approximation will vary – see Figure 1). This latter fact makes it difficult to determine the degree to which each predictor contributes to a change in score. This is known as the problem of variable attribution.

Ensuring that the model output is monotonic in its predictors satisfies the first of the three requirements. To satisfy the second and third requirements, we must have a way to do variable attribution for nonlinear models.

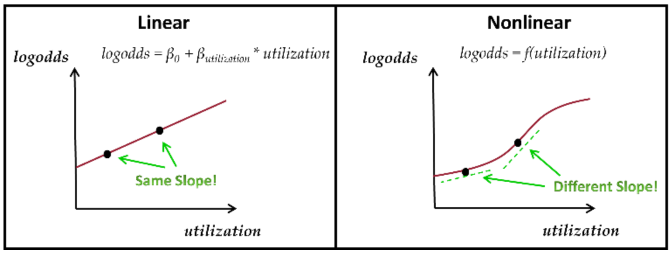

Fig. 1. Linear and Nonlinear univariate models of logodds of default as a function of credit utilization. For the linear model, $\beta_{utilization}$ is constant making it easy to convert a change in utilization into a change in logodds or vice versa (left). For the nonlinear model, the conversion rate covaries with the value of utilization, making variable attribution more difficult (right).

IV. Variable Attribution

Research into variable attribution for differentiable nonlinear machine learning models has exploded in recent years due to the rise of deep learning. While linear models are not necessarily more interpretable or explainable than simple neural networks [22], deep learning models are generally quite complex, often regarded as black boxes whose inner workings cannot be explained or understood in simple terms [39]. The quest to understand complex machine learning models has led to rapid growth in the field of eXplainable AI (XAI) in recent years [40]–[43], though some believe this post hoc approach to be misguided [44]. Some have succeeded in building flashlights to peer directly into the box [45]. However, many fruitful approaches have instead focused on analyzing the input-output behavior of neural networks using variable attribution methods.

The problem of variable attribution concerns how to quantify or distribute responsibility amongst a model's predictors for a given prediction made using a particular input. As such, it is concerned with local, instance-level explanations rather than global explanations of the model itself [46], [47]. Some variable attribution approaches rely on local gradients (or a modified analogous quantity) to identify important features (e.g. – Explanation Vectors [48], Saliency Maps [49], Deconvolutional Networks [45], Guided Backpropagation [50], Layer-wise Relevance Propagation (LRP) [51], Class Activation Mapping (CAM) [52], Gradient-weighted Class Activation Mapping (Grad-CAM) [53]). The features identified by these methods are those that (locally) would have the greatest impact on the score if changed, but they are not necessarily those that were most important in calculating the score. As Kindermans et al. point out, the gradient does not necessarily align with the signal in the data [54]. Additionally, the contributions from the most important features may be saturated at the input value [55]. In other words, it is possible that a feature may contribute greatly to the signal, but lightly perturbing it will have minimal or no impact on the prediction. For example, credit utilization – the ratio of outstanding balances to total credit – is extremely important in credit risk modeling, but for sufficiently low utilization, a small change is unlikely to have a significant impact. Gradient methods are not the only ones susceptible to this problem. Other approaches relying on occlusions, ablations, or perturbations [45], [56]–[58] may also fail to properly account for saturated features [55], [59].

Overcoming the saturation problem requires the use of baselines [59] (alternatively, reference points [55] or root points [54], [60]), and several approaches have been put forward (e.g. – DeepLIFT [55], Deep Taylor Decomposition (DTD) [60], Integrated Gradients (IG) [59]). The purpose of a baseline is to act as a contrast for the input. It could be an impartial benchmark, an instance-specific counterfactual foil, or even a random noninformative or neutral input. Choosing an appropriate baseline is not trivial, and we discuss this in more detail in Section 6, but, ideally, the baseline should differ from the input in all and only the task-relevant respects. For the industry standard approach to credit scoring, the baseline is the ideal customer.

For baseline variable attribution methods, the goal is to attribute the differences in how the model scored the input and the baseline to the differences between the input and baseline themselves. Sundararajan et al. explicitly define the problem of variable attribution for neural networks as follows:

> Formally, suppose we have a function $F: R^n \rightarrow [0,1]$ that represents a deep network, and an input $x = (x_1, \ldots, x_n) \in R^n$. An attribution of the prediction at input x relative to a baseline input x' is a vector $A_F(x, x') = (a_1, \ldots, a_n) \in R^n$ where a_i is the contribution of x_i to the prediction $F(x)$ [59, pg. 1].

There are many possible approaches to assigning attributions, but not all of them are desirable. There are certain properties that we may like to see in an attribution method. For example, a lack of what Sundararajan et al. call *sensitivity(a)* is why the aforementioned gradient methods have issues with saturated features [59]. Roughly, this axiom says that if an input and baseline receive different scores and only differ in one feature, then that feature should receive some attribution. A stronger version of this – variously called *completeness* [59], *efficiency* [61], *summation-to-delta* [55], and *conservative* [60] – requires that the sum of the attributions across all features be equal to the difference in scores for the input and the baseline. We noted in Section 3 that this ability to comprehensively decompose the difference onto the features was a useful property of linear models, and it would be equally useful if we could do this with nonlinear credit models.

Sundararajan et al. enumerate several other properties that, as we show in the Appendix, are also satisfied by the industry standard approach outlined in Section 3. These additional properties include *sensitivity(b)*, *linearity*, and *implementation invariance* [59]. Roughly, *sensitivity(b)* requires 0 attribution be given to unused features, *linearity* preserves attribution under linear combination, and *implementation invariance* requires attributions to be invariant across functionally equivalent networks, where "[t]wo networks are *functionally equivalent* if their outputs are equal for all inputs, despite having different implementations" [59, p. 2]. Only attribution methods based on path integrated gradients satisfy completeness, sensitivity(b), linearity, and implementation invariance (see [59] for details). If we add one additional requirement, that the attribution method be *symmetry-preserving* in the sense that "for all inputs that have identical values for symmetric variables and baselines that have identical values for symmetric variables, the symmetric variables receive identical attributions" [59, p. 5], then there is only one attribution method that can satisfy these requirements: Integrated gradients [59].

Sundararajan et al. define Integrated Gradients as follows [59]:

> The integrated gradient along the i^{th} dimension for an input x and baseline x' is defined as follows. Here, $\frac{\partial F(x)}{\partial x_i}$ is the gradient of $F(x)$ along the i^{th} dimension.
>
> $$IntegratedGrads_i(x) = (x_i - x_i') \times \int_{\alpha=0}^{1} \frac{\partial F\left(x' + \alpha \times (x - x\prime)\right)}{\partial x_i} \partial\alpha$$

Essentially, integrated gradients calculates the average gradient along the straightline path obtained by interpolating between the baseline and the input, and the resulting attributions are the elementwise product of this path averaged gradient vector and the input-baseline difference vector. The entries in the path averaged gradient vector are analogous to the coefficients in the logistic regression model from Section 3. From this perspective, integrated gradients bears a strong resemblance to the industry standard approach. Furthermore, since integrated gradients is provably the only variable attribution method satisfying the 5 desirable properties we touched on, and since the industry standard approach also shares these properties (see Appendix), the approaches must be the same at some level. In the next section, we will show that, indeed, integrated gradients is a generalization of the industry standard approach to variable attribution in nonlinear spaces.

V. From Industry Standard to Integrated Gradients

Let us now return to the industry standard approach to variable attribution. Consider the following univariate model:

$$y = f(x)$$

where y is the output (e.g. – logodds of default), x is some predictor (e.g. – credit utilization), and $f(\cdot)$ is a function relating the two. Let us begin by assuming that $f(\cdot)$ is a linear function (see Figure 2):

$$y = \beta_0 + \beta_1 x \quad x' \leq x \leq x^*$$

where y is the output (e.g. – logodds of default), x' is the baseline (e.g. – hypothetical ideal), and x^* is an input (e.g. – prospective customer). While credit scoring models are usually constructed in such a way that they are bounded (see Section 3), it would be unusual for x^* to correspond to the worst possible customer as would seem to be the case with the above construction. This need not be the case. For our purposes, we are only interested in the behavior of the model between the baseline and input, so the model has been defined in this way for pedagogical simplicity.

As we saw in Section 3, the industry standard approach yields:

$$\begin{aligned} \varDelta y &= f(x^*) - f(x') \\ &= \beta_1(x^* - x') \\ &= \beta_1 \varDelta x \end{aligned}$$

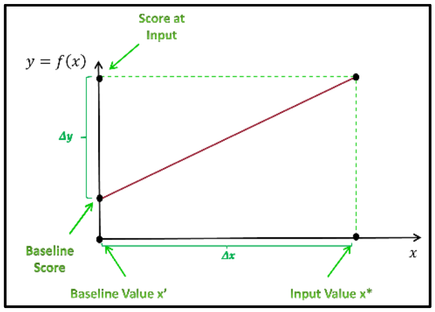

Fig. 2. y is a linear function of x, and the difference in y evaluated at input and at baseline is simply the difference in input and baseline multiplied by the slope.

Now, let us assume that $f(\cdot)$ is a piecewise linear function with 2 segments (see Figure 3):

$$f(x) = \begin{cases} \beta_0^1 + \beta_1^1 x & x' \leq x \leq c_1 \\ \beta_0^2 + \beta_1^2 x & c_1 < x \leq x^* \end{cases}$$

The industry standard approach can be applied to both segments separately, and their contributions added together yield the total:

$$\begin{aligned} \varDelta y &= \varDelta y_1 + \varDelta y_2 \\ &= \beta_1^1(c_1 - x') + \beta_1^2(x^* - c_1) \end{aligned}$$

Assuming that

$$(c_1 - x') = (x^* - c_1) = \varDelta x$$

Then,

$$\varDelta y = (\beta_1^1 + \beta_1^2)\varDelta x$$

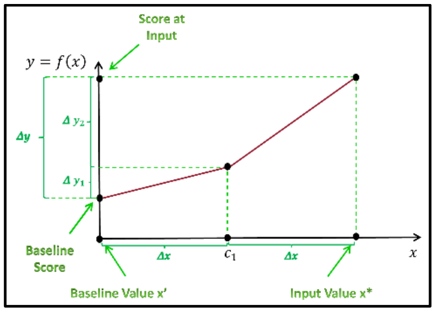

Fig. 3. y is a piecewise linear function of x with 2 segments, and the difference in y evaluated at input and at baseline is simply the sum of the products of the difference in x along each segment and the slope of each segment.

Now, let us assume that $f(\cdot)$ is a piecewise linear function with m segments (see Figure 4):

$$f(x) = \begin{cases} \beta_0^1 + \beta_1^1 x & x' \leq x \leq c_1 \\ \beta_0^2 + \beta_1^2 x & c_1 < x \leq c_2 \\ \quad \ldots & \quad \ldots \\ \beta_0^m + \beta_1^m x & c_{m-1} < x \leq x^* \end{cases}$$

Once again, assuming that the change in x along each segment is the same

$$(c_1 - x') = (c_2 - c_1) = \cdots = (x^* - c_{m-1}) = \varDelta x$$

And following same logic as before, we get the following:

$$\varDelta y = \varDelta x \sum_{k=1}^{m} \beta_1^k$$

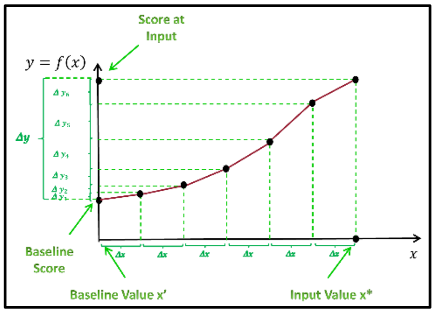

Fig. 4. y is a piecewise linear function of x with m segments, and the difference in y evaluated at input and at baseline is simply the sum of the products of the difference in x along each segment and the slope of each segment.

Note that

$$\Delta x = \frac{1}{m}(x^* - x')$$

And

$$\beta_1^k = \frac{\partial f\left(x' + \frac{k}{m}(x^* - x')\right)}{\partial x}$$

Plugging in and rearranging

$$\Delta y = (x^* - x') \times \sum_{k=1}^{m} \frac{\partial f\left(x' + \frac{k}{m}(x^* - x')\right)}{\partial x} \times \frac{1}{m}$$

This is exactly the Riemann approximation of the integrated gradients integral we saw in Section 4 in the case where the model is univariate (See equation 3 in [59]). The extension of our argument to the multivariate case is straightforward mathematically, though it is less amenable to visualization (see Appendix). In short, as long as the change along each marginal of each segment remains constant (i.e. – the interpolation between baseline and input occurs along a straightline path), then the analogous (full) result holds. The industry standard method applied to an ever-finer-grained piecewise linear approximation of a nonlinear function yields the Riemann approximation of the IG integral.

We have shown that the industry standard approach to variable attribution yields integrated gradients as we move from linear to nonlinear models. It is also easy to go the other direction, showing that Integrated Gradients reduces to the industry standard approach for linear models:

$$IntegratedGrads_i(x) = (x_i - x_i') \times \int_{\alpha=0}^{1} \frac{\partial F\big(x' + \alpha \times (x - x')\big)}{\partial x_i} \partial\alpha$$

When $F(x)$ is linear $\frac{\partial F(x)}{\partial x_i} = \beta_i$ is a constant value, and IG is equivalent to the industry standard approach:

$$IntegratedGrads_i(x) = (x_i - x_i') \times \beta_i$$

We have shown both that:

1. Integrated gradients is a natural extension to nonlinear models of the industry standard approach to variable attribution for linear models.
2. The industry standard approach is a special case of integrated gradients; specifically, when the model is a linear function of the inputs.

These results have major implications for the use of differentiable nonlinear models (e.g. – neural networks) in credit scoring. If the industry standard approach to variable attribution is an acceptable approach to identifying and ranking 'key factors', then integrated gradients must also be an acceptable approach to identifying and ranking 'key factors'. Recall that the ability to identify and rank 'key factors' is all that is needed to satisfy the 2nd and 3rd requirements mentioned in Section 2, and a strict monotonic relationship between the output and the inputs is all that is needed to satisfy the 1st requirement. **So, *any* model architecture – linear or nonlinear – that is compatible with use of integrated gradients and which satisfies appropriate input-output monotonicity constraints arguably meets the requirements for generating regulatory compliant explanations.** This opens up a wide swath of machine learning models that previously, have been off the table. Furthermore, as we will touch on in our conclusions and future work, integrated gradients is a part of a larger, unified framework that extends in a straightforward way to *nondifferentiable* nonlinear models, potentially opening the door to even more (appropriately monotonically-constrained) machine learning models.

VI. Semantics Underlying Choice of Baseline and Path Dependence in Nonlinear Models

A key feature of path methods for variable attribution (e.g. – integrated gradients) is that they require choosing a baseline. In many applications (e.g. – computer vision tasks), the choice of baseline can be somewhat arbitrary. In credit modeling, however, as we saw in Section 3, there is a very natural baseline: the ideal customer. Another key feature of path methods is choice of path. For integrated gradients, this is the straightline path between baseline and input. The justification for this choice was an appeal to symmetry preservation as a naturally desirable property. It has been said that Occam's razor makes straight cuts; so, in the absence of countervailing arguments, this seems reasonable. Additionally, the choice of a straightline path (where movement along each dimension from baseline to input occurs with constant stepsize) was crucial to connecting integrated gradients to the industry standard, but this only implies that movement along any segment must occur along a straightline path; the trajectory from baseline to input could be piecewise linear if there were some semantically important reason for visiting particular points along the way [62]. For example, suppose that the input and baseline are temporal snapshots of the same entity at different points in a sequence. There may be a host of intermediate snapshots carving out a jagged path from baseline to input. Integrating the gradients over this path may yield very different attributions.

The choice of path and the choice of baseline have semantic implications. The attributions generated by baseline variable attribution methods are a form of contrastive explanation [14], [18], [21]. They answer questions of the form 'Why P rather than Q?' or, more precisely, 'Why [F(input)] rather than [F(baseline)]?' (in terms of differences in the input and baseline and, in the case of path methods, as assessed along some path). For example, suppose one asks why some customer received the score they did. The answer depends. Why they received the score they did as opposed to a higher score requires selecting a baseline that has a higher score. There are generally many such possible baselines, and each will differ from our customer in different ways resulting in different attributions [62].

Note that in the case of linear models, the attributions will be the same regardless of path choice. This is why the industry standard approach only needed to concern itself with choice of baseline. For nonlinear models, the attributions will generally vary depending on the chosen path.

We will look at 2 informative baselines and path choices for credit scoring models based on different explanatory goals assuming we have access to historical customer data. For a much more thorough treatment of the implications of various baseline and path choices when applying integrated gradients to explain neural credit risk models, see Alam et al. [62].

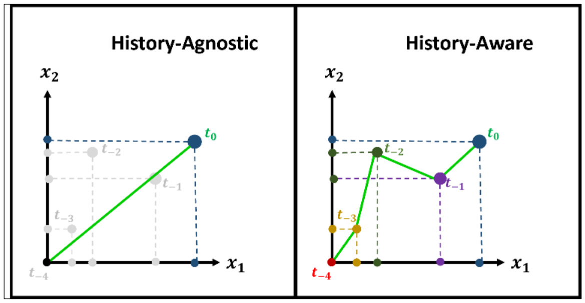

Fig. 5. Illustration of 2 baseline / path choices. The history-agnostic explanation uses a fixed benchmark as baseline (origin in this case) and attributions are accumulated along a straightline path (left). The history-aware explanation takes a sequence of inputs and accumulates attributions along the piecewise linear path defined by this sequence (right).

A. *Goal 1: History-Agnostic Explanation*

The goal of the history agnostic explanation is to give the customer a measure of their (history-agnostic) performance relative to some benchmark (see Figure 5). Here, we use the ideal customer – the attribute vector corresponding to the borrower with the lowest possible default probability – as baseline and a straight-line path between baseline and input. The choice of the ideal customer as baseline means we are explicitly interested in how differences in a customer and the theoretical ideal account for differences in their resultant scores. The ideal customer acts as a fixed and impartial standard, and, since every input uses the same baseline, it gives a logical way to compare the attributions of all the inputs. Here, the key factors will correspond to the most influential factors differentiating the customer from the ideal along the straightline path, but they do not necessarily reflect the relative influence of the factors in a historical causal sense.

B. *Goal 2: History-Aware Explanation*

The goal of the history-aware explanation is to give the customer insight into the relative importance of each factor in their unique history. Consider a situation where a customer asks why their credit score has decreased in a particular month. The history-agnostic approach cannot answer that question because it was implemented to take a straight-line path from some impartial benchmark to the input. However, the history-aware explanation uses a piecewise linear path defined by the unique history of the customer allowing it to keep track of changes in attributions from point to point (see Figure 5). It can reveal exactly which features contributed the most at any point in time.

Consider the following equation:

$$\sum_{i=0}^{p} A(x^{(i)}) = F(x) - F(x_0)$$

Here, $x^{(i)}$ is the i^{th} feature, $A(x^{(i)})$ is the attribution of i^{th} feature, x is the input, $F(x)$ is the model output at the input, x_0 is the baseline, and $F(x_0)$ is the model output at the baseline.

The baseline (x_0) can be replaced by a point in the history of the borrower (x_h). The interpretation will depend on which point is chosen. If we choose the previous month's data as baseline, we will get attribution for one month. The integrated gradients can be calculated along a straight-line path from the new baseline (x_h) to the input (x). The attribution of each individual feature is their contribution in one month. Similarly, it is possible to calculate contribution of the features in each month by using the previous month as baseline and current month as input. Feature-wise addition of all the attributions for each month will give the total attribution with respect to the earliest month as baseline.

Let's consider three months of borrower data x_0, x_1, and x_2. x_0 is the earliest month while x_2 is the current month. Let $A_{j,k}$ be the attribution of input data at j^{th} month with respect to the baseline data at the k^{th} month along a straight-line path from the baseline x_k to the input x_j. Now,

$$\begin{aligned}\textstyle\sum_{i=0}^{p} A_{2,1}({x_2}^{(i)}) &= F(x_2) - F(x_1)\\ \textstyle\sum_{i=0}^{p} A_{1,0}({x_1}^{(i)}) &= F(x_1) - F(x_0)\\ \textstyle\sum_{i=0}^{p} A_{2,0}({x_2}^{(i)}) &= F(x_2) - F(x_0)\end{aligned}$$

So,

$$\begin{aligned}\textstyle\sum_{i=0}^{p} A_{2,1}({x_2}^{(i)}) + \sum_{i=0}^{p} A_{1,0}({x_1}^{(i)}) &= F(x_2) - F(x_1) + F(x_1) - F(x_0)\\ &= F(x_2) - F(x_0)\\ &= \textstyle\sum_{i=0}^{p} A_{2,0}({x_2}^{(i)})\end{aligned}$$

This equation guarantees that the sum of the attributions of all features will be the same regardless of the path chosen. However, the distribution of this total attribution amongst the individual features will generally be different along the direct straightline path vs. the piecewise path through the borrower's credit history. Attribution along the borrower's credit history allows the customer to receive reason codes specific to their unique trajectory.

VII. Conclusions and Future Work

Credit is a critical component of modern commerce, but credit requires trust, which does not scale due to epistemological limitations. These limitations can be overcome with increased transparency in the form of credit reporting, but increased reliance on the credit reporting infrastructure by merchants makes consumers vulnerable to CRAs. Additionally, the displacement of subjective in-person interviews by automated credit risk models has the potential to decrease unwanted bias in the decisioning process; however, **credit risk modeling only delivers increased transparency and accountability for both sides if consumers can understand what is in their credit report, who compiled and used it, and why it resulted in a particular decision**. The FCRA and ECOA as implemented in

Regulation B codify these requirements into U.S. fair lending law. Therefore, explainability is not only ethically desirable in credit risk models, but it is a legal requirement.

In this paper, we sketched the industry standard approach to generating regulatory-compliant explanations from linear models. We also enumerated 3 requirements that explanations must meet, one of which is satisfied by monotonically constraining the relationships between the inputs and the output of the model. The other requirements necessitate a principled approach to variable attribution. We then showed that the industry standard approach to variable attribution is a special case of a more general technique known as integrated gradients, which applies to differentiable nonlinear models like neural networks. On these grounds, we argued that if the industry standard is an acceptable approach to variable attribution for linear models, then integrated gradients must be an acceptable variable attribution approach for differentiable nonlinear models.

It turns out that integrated gradients is itself a special case of a more general framework (i.e. – Shapley values) that encompasses *nondifferentiable* nonlinear models, as well [61]. This could potentially open the door to using any appropriately monotonically constrained machine learning approach in credit risk modeling. Additionally, recent work that builds on integrated gradients for interaction attributions [63]–[65] and group attributions [66] may be promising avenues for future research. Finally, there is much work to be done on explainable risk modeling that incorporates time-series information, that operates in dynamic 'online' contexts, or that relies on reinforcement learning agents. These latter paradigms move beyond the standards developed for static, cross-sectional data and will require significant and careful consideration.

Acknowledgment

We would like to thank Equifax for providing invaluable insight into the standard practices in the credit industry. We would especially like to thank Dr. Michael McBurnett, Dr. Jeff Dugger, and Dr. Matthew Turner for their feedback on this project.

References

[1] "Credit Reports and Scores," Federal Government of the United States, Apr. 27, 2022. https://www.usa.gov/credit-reports (accessed Jun. 19, 2022).

[2] "FICO® Scores Versions," Fair Isaac Corporation, 2022. https://www.myfico.com/credit-education/credit-scores/fico-score-versions (accessed Jun. 19, 2022).

[3] "The VantageScore Model," VantageScore Solutions, LLC, 2022. https://vantagescore.com/lenders/why-vantagescore/our-models/ (accessed Jun. 19, 2022).

[4] S. Arya, C. Eckel, and C. Wichman, "Anatomy of the credit score," J Econ Behav Organ, vol. 95, pp. 175–185, Nov. 2013, doi: 10.1016/j.jebo.2011.05.005.

[5] B. Baesens, T. van Gestel, S. Viaene, M. Stepanova, J. Suykens, and J. Vanthienen, "Benchmarking state-of-the-art classification algorithms for credit scoring," Journal of the Operational Research Society, vol. 54, no. 6, pp. 627–635, Jun. 2003, doi: 10.1057/palgrave.jors.2601545.

[6] S. Lessmann, B. Baesens, H.-V. Seow, and L. C. Thomas, "Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research," Eur J Oper Res, vol. 247, no. 1, pp. 124–136, Nov. 2015, doi: 10.1016/j.ejor.2015.05.030.

[7] B. R. Gunnarsson, S. vanden Broucke, B. Baesens, M. Óskarsdóttir, and W. Lemahieu, "Deep learning for credit scoring: Do or don't?," Eur J Oper Res, vol. 295, no. 1, pp. 292–305, Nov. 2021, doi: 10.1016/j.ejor.2021.03.006.

[8] J. Whittlestone, R. Nyrup, A. Alexandrova, and S. Cave, "The Role and Limits of Principles in AI Ethics," in Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society, Jan. 2019, pp. 195–200. doi: 10.1145/3306618.3314289.

[9] C. T. Nguyen, "Transparency is Surveillance," Philos Phenomenol Res, p. phpr.12823, Aug. 2021, doi: 10.1111/phpr.12823.

[10] C. McLeod, "Trust," The Stanford Encyclopedia of Philosophy (Fall 2021 Edition), 2021. https://plato.stanford.edu/archives/fall2021/entries/trust/ (accessed Jun. 21, 2022).

[11] R. OLEGARIO, A Culture of Credit. Harvard University Press, 2006. doi: 10.4159/9780674041639.

[12] J. Lauer, "From Rumor to Written Record: Credit Reporting and the Invention of Financial Identity in Nineteenth-Century America," Technol Cult, vol. 49, no. 2, pp. 301–324, 2008, Accessed: Jun. 21, 2022. [Online]. Available: http://www.jstor.org/stable/40061517

[13] J. Lauer, Creditworthy. Columbia University Press, 2017. doi: 10.7312/laue16808.

[14] P. Lipton, "Contrastive explanation," Royal Institute of Philosophy Supplements, vol. 27, pp. 247–266, 1990.

[15] D. Lewis, "Causal Explanation," in Philosophical Papers Volume II, Oxford University Press, 1987, pp. 214–240. doi: 10.1093/0195036468.003.0007.

[16] J. Woodward and L. Ross, "Scientific Explanation," The Stanford Encyclopedia of Philosophy (Summer 2021 Edition), 2021. https://plato.stanford.edu/archives/sum2021/entries/scientific-explanation/ (accessed Jun. 21, 2022).

[17] A. Brenner, A.-S. Maurin, A. Skiles, R. Stenwall, and N. Thompson, "Metaphysical Explanation," The Stanford Encyclopedia of Philosophy (Winter 2021 Edition), 2021, Accessed: Jun. 21, 2022. [Online]. Available: https://plato.stanford.edu/archives/win2021/entries/metaphysical-explanation/

[18] T. Lombrozo, "The structure and function of explanations," Trends Cogn Sci, vol. 10, no. 10, pp. 464–470, Oct. 2006, doi: 10.1016/j.tics.2006.08.004.

[19] T. Lombrozo, "Explanation and Abductive Inference," in The Oxford Handbook of Thinking and Reasoning, K. J. Holyoak and R. G. Morrison, Eds. Oxford University Press, 2012. doi: 10.1093/oxfordhb/9780199734689.013.0014.

[20] T. Miller, P. Howe, and L. Sonenberg, "Explainable AI: Beware of inmates running the asylum or: How I learnt to stop worrying and love the social and behavioural sciences," 2017.

[21] T. Miller, "Explanation in artificial intelligence: Insights from the social sciences," Artif Intell, vol. 267, pp. 1–38, Feb. 2019, doi: 10.1016/j.artint.2018.07.007.

[22] Z. C. Lipton, "The Mythos of Model Interpretability: In machine learning, the concept of interpretability is both important and slippery," Queue, vol. 16, no. 3, pp. 31–57, 2018.

[23] T. Nagel, "Moral Luck," in Mortal Questions, Cambridge University Press, 2012, pp. 24–38. doi: 10.1017/CBO9781107341050.005.

[24] "Fair Credit Reporting Act (FCRA) examination procedures," Oct. 2012. Accessed: Jun. 22, 2022. [Online]. Available: https://files.consumerfinance.gov/f/documents/102012_cfpb_fair-credit-reporting-act-fcra_procedures.pdf

[25] D. Kahneman, Thinking, fast and slow. Macmillan, 2011.

[26] S. Shiffman, A. A. Stone, and M. R. Hufford, "Ecological Momentary Assessment," Annu Rev Clin Psychol, vol. 4, no. 1, pp. 1–32, Apr. 2008, doi: 10.1146/annurev.clinpsy.3.022806.091415.

[27] R. Binns, "Fairness in Machine Learning: Lessons from Political Philosophy," in Conference on Fairness, Accountability and Transparency, 2018, vol. 81, pp. 149–159.

[28] "ECOA examination procedures," Oct. 2015. Accessed: Jun. 22, 2022. [Online]. Available:

https://files.consumerfinance.gov/f/documents/201510_cfpb_ecoa-narrative-and-procedures.pdf

[29] “15 U.S.C. § 1691 - Scope of prohibition.” Jan. 07, 2011.

[30] “15 U.S.C. § 1681m - Requirements on users of consumer reports.” Dec. 18, 2010.

[31] “12 C.F.R. § 202.9 - Notifications.” Nov. 09, 2007.

[32] D. J. Hand and W. E. Henley, “Statistical Classification Methods in Consumer Credit Scoring: a Review,” J R Stat Soc Ser A Stat Soc, vol. 160, no. 3, pp. 523–541, Sep. 1997, doi: 10.1111/j.1467-985X.1997.00078.x.

[33] M. Vojtek and E. Kocenda, “Credit Scoring Methods,” Czech Journal of Economics and Finance (Finance a uver), vol. 56, no. 3–4, pp. 152–167, 2006.

[34] H. A. Abdou and J. Pointon, “CREDIT SCORING, STATISTICAL TECHNIQUES AND EVALUATION CRITERIA: A REVIEW OF THE LITERATURE,” Intelligent Systems in Accounting, Finance and Management, vol. 18, no. 2–3, pp. 59–88, Apr. 2011, doi: 10.1002/isaf.325.

[35] F. Louzada, A. Ara, and G. B. Fernandes, “Classification methods applied to credit scoring: Systematic review and overall comparison,” Surveys in Operations Research and Management Science, vol. 21, no. 2, pp. 117–134, Dec. 2016, doi: 10.1016/j.sorms.2016.10.001.

[36] M. Turner and M. McBurnett, “Optimizing neural networks for risk assessment,” U.S. Patent 10,133,980, Nov. 20, 2018

[37] D. W. Hosmer, S. Lemeshow, and R. X. Sturdivant, Applied Logistic Regression, 3rd ed. Hoboken, NJ, USA: John Wiley & Sons, Inc., 2013.

[38] “What Are Credit Score Reason Codes?,” Fair Isaac Corporation, 2022. https://www.myfico.com/credit-education/blog/reason-codes.

[39] R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, and D. Pedreschi, “A Survey of Methods for Explaining Black Box Models,” ACM Comput Surv, vol. 51, no. 5, pp. 1–42, Sep. 2019, doi: 10.1145/3236009.

[40] A. Adadi and M. Berrada, “Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI),” IEEE Access, vol. 6, pp. 52138–52160, 2018, doi: 10.1109/ACCESS.2018.2870052.

[41] A. Barredo Arrieta et al., “Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI,” Information Fusion, vol. 58, pp. 82–115, Jun. 2020, doi: 10.1016/j.inffus.2019.12.012.

[42] S. Mohseni, N. Zarei, and E. D. Ragan, “A Multidisciplinary Survey and Framework for Design and Evaluation of Explainable AI Systems,” ACM Trans Interact Intell Syst, vol. 11, no. 3–4, pp. 1–45, Dec. 2021, doi: 10.1145/3387166.

[43] F. Doshi-Velez and B. Kim, “Towards A Rigorous Science of Interpretable Machine Learning,” Feb. 2017.

[44] C. Rudin, “Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead,” Nat Mach Intell, vol. 1, no. 5, pp. 206–215, May 2019, doi: 10.1038/s42256-019-0048-x.

[45] M. D. Zeiler and R. Fergus, “Visualizing and Understanding Convolutional Networks,” 2014, pp. 818–833. doi: 10.1007/978-3-319-10590-1_53.

[46] A. Rai, “Explainable AI: from black box to glass box,” J Acad Mark Sci, vol. 48, no. 1, pp. 137–141, Jan. 2020, doi: 10.1007/s11747-019-00710-5.

[47] C. Molnar, G. Casalicchio, and B. Bischl, “Interpretable Machine Learning – A Brief History, State-of-the-Art and Challenges,” 2020, pp. 417–431. doi: 10.1007/978-3-030-65965-3_28.

[48] D. Baehrens, T. Schroeter, S. Harmeling, M. Kawanabe, K. Hansen, and K.-R. Muller, “How to Explain Individual Classification Decisions,” Journal of Machine Learning Research, vol. 11, pp. 1803–1831, 2010.

[49] K. Simonyan, A. Vedaldi, and A. Zisserman, “Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps,” Dec. 2013.

[50] J. T. Springenberg, A. Dosovitskiy, T. Brox, and M. Riedmiller, “Striving for Simplicity: The All Convolutional Net,” Dec. 2014.

[51] S. Bach, A. Binder, G. Montavon, F. Klauschen, K.-R. Müller, and W. Samek, “On Pixel-Wise Explanations for Non-Linear Classifier Decisions by Layer-Wise Relevance Propagation,” PLoS One, vol. 10, no. 7, p. e0130140, Jul. 2015, doi: 10.1371/journal.pone.0130140.

[52] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, “Learning Deep Features for Discriminative Localization,” Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2921–2929, 2016.

[53] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, “Grad-CAM: Visual Explanations From Deep Networks via Gradient-Based Localization,” Proceedings of the IEEE International Conference on Computer Vision (ICCV), pp. 618–626, 2017.

[54] P.-J. Kindermans et al., “Learning how to explain neural networks: PatternNet and PatternAttribution,” May 2017.

[55] A. Shrikumar, P. Greenside, and A. Kundaje, “Learning Important Features Through Propagating Activation Differences,” Proceedings of the 34th International Conference on Machine Learning, PMLR, vol. 70, pp. 3145–3153, 2017.

[56] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, “Object Detectors Emerge in Deep Scene CNNs,” ICLR, 2015.

[57] L. M. Zintgraf, T. S. Cohen, T. Adel, and M. Welling, “Visualizing Deep Neural Network Decisions: Prediction Difference Analysis,” ICLR, 2017.

[58] M. T. Ribeiro, S. Singh, and C. Guestrin, “‘Why Should I Trust You?,’” in Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Aug. 2016, pp. 1135–1144. doi: 10.1145/2939672.2939778.

[59] M. Sundararajan, A. Taly, and Q. Yan, “Axiomatic attribution for deep networks,” in International conference on machine learning, Jul. 2017, pp. 3319–3328.

[60] G. Montavon, S. Lapuschkin, A. Binder, W. Samek, and K.-R. Müller, “Explaining nonlinear classification decisions with deep Taylor decomposition,” Pattern Recognit, vol. 65, pp. 211–222, May 2017, doi: 10.1016/j.patcog.2016.11.008.

[61] M. Sundararajan and A. Najmi, “The Many Shapley Values for Model Explanation,” Proceedings of the 37th International Conference on Machine Learning, PMLR, vol. 119, pp. 9269–9278, 2020.

[62] M. S. Alam, J. Boardman, X. Huang, and M. Turner, “Applications of Integrated Gradients in Credit Risk Modeling,” unpublished.

[63] J. D. Janizek, P. Sturmfels, and S.-I. Lee, “Explaining Explanations: Axiomatic Feature Interactions for Deep Networks,” Journal of Machine Learning Research, vol. 22, no. 104, pp. 1–54, 2021.

[64] K. Dhamdhere, A. Agarwal, and M. Sundararajan, “The Shapley Taylor Interaction Index,” Proceedings of the 37th International Conference on Machine Learning, PMLR, vol. 119, pp. 9259–9268, 2020.

[65] M. Tsang, S. Rambhatla, and Y. Liu, “How does this interaction affect me? interpretable attribution for feature interactions,” Adv Neural Inf Process Syst, vol. 33, pp. 6147–6159, 2020.

[66] S. Sikdar, P. Bhattacharya, and K. Heese, “Integrated Directional Gradients: Feature Interaction Attribution for Neural NLP Models,” in Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), 2021, pp. 865–878. doi: 10.18653/v1/2021.acl-long.71.

APPENDIX

A. Derivation of IG in multivariate case

Let’s consider a multivariate logistic regression model with p predictors:

$$y = f(\boldsymbol{x}) = \beta_0 + \sum_{i=1}^{p} \beta_i \boldsymbol{x_i}$$

For an input x^* and a baseline x', the output at these two points are

$$y^* = f(\boldsymbol{x}^*) = \beta_0 + \sum_{i=1}^{p} \beta_i x_i^*$$
$$y' = f(\boldsymbol{x}') = \beta_0 + \sum_{i=1}^{p} \beta_i x_i'$$

Now the change in y from x' to x^* is given by

$$\Delta y = \; y^* - y' = f(\boldsymbol{x}^*) \; - f(\boldsymbol{x}') \; = \sum_{i=1}^{p} \beta_i (x_i^* - x_i')$$

Each variable contributes independently to the change in loggodds (Δy), and the contribution of a variable only depends on the corresponding β and the change in the value of that particular variable. The total contribution is the sum of the contributions of all variables. So, the attribution for i^{th} variable is given by

$$\Delta y_i = \beta_i(x_i^* - x_i')$$

Following the construct in the univariate case and considering m piecewise steps from the baseline (x_i') to the input (x_i^*) to approximate the attribution of the i^{th} variable, we can rewrite the above equation,

$$\Delta y_i = (x_i^* - x_i') \times \sum_{k=1}^{m} \frac{\partial f\left(x' + \frac{k}{m}(x^* - x')\right)}{\partial x_i} \times \frac{1}{m}$$

B. Industry Standard Attribution

Consider a linear model in p variables given by the function

$$y = f(\boldsymbol{x}) = \beta_0 + \boldsymbol{\beta} \cdot \boldsymbol{x} = \beta_0 + \sum_{i=0}^{p} \beta_i x_i$$

Here, y is the logodds, f is a function of the input $\boldsymbol{x}$, β_0 is the intercept, and β_i is the coefficient of the i^{th} variable.

For a baseline $\boldsymbol{x}'$ and an input $\boldsymbol{x}^*$, the attribution for the i^{th} variable is defined as

$$\Delta y_i = \beta_i * (x_i^* - x_i')$$

C. Completeness

The completeness axiom states that the attributions add up to the difference between the output at the input ($\boldsymbol{x}^*$) and at the baseline $(\boldsymbol{x}')$ [59]. Mathematically,

$$f(\boldsymbol{x}^*) - f(\boldsymbol{x}') = \sum_{i=0}^{p} A_i$$

where A_i is the attribution for the i^{th} variable. For Industry Standard Attribution above

$$A_i = \beta_i * (x_i^* - x_i')$$

So, completeness requires

$$f(\boldsymbol{x}^*) - f(\boldsymbol{x}') = \sum_{i=0}^{p} \beta_i * (x_i^* - x_i')$$

From the definition of the linear model above we get,

$$f(\boldsymbol{x}^*) = \beta_0 + \sum_{i=0}^{p} \beta_i x_i^*$$
$$f(\boldsymbol{x}') = \beta_0 + \sum_{i=0}^{p} \beta_i x_i'$$

So,

$$\begin{aligned} f(\boldsymbol{x}^*) - f(\boldsymbol{x}') &= \beta_0 + \sum_{i=0}^{p} \beta_i x_i^* - \beta_0 - \sum_{i=0}^{p} \beta_i x_i' \\ &= \sum_{i=0}^{p} \beta_i (x_i^* - x_i') = \sum_{i=0}^{p} A_i \end{aligned}$$

D. Sensitivity(b)

The axiom states that if the model output does not depend on a variable, then the attribution for that variable should always be zero [59]. In other words, if for any change in variable x_i, the output $f(\boldsymbol{x})$ does not change, the attribution $\Delta y_i = \beta_i * (x_i^* - x_i')$ should be equal to zero.

We can rewrite the linear model definition as follows,

$$f(\boldsymbol{x}) = \beta_0 + \sum_{i=0, i\neq r}^{p} \beta_i \boldsymbol{x}_i + \beta_r \cdot \boldsymbol{x}_r$$

Let's say the r^{th} variable does not have any effect on the output y. So, changing the value of the r^{th} variable from x_r^* to x_r' will not change the value of the function f.

$$\beta_0 + \sum_{i=0, i\neq r}^{p} \beta_i x_i + \beta_r \cdot x_r^* = \beta_0 + \sum_{i=0, i\neq r}^{p} \beta_i x_i + \beta_r \cdot x_r'$$
$$\beta_r x_r^* - \beta_r x_r' = 0$$
$$\beta_r (x_r^* - x_r') = 0$$

Since $(x_r^* - x_r') \neq 0$, β_r must be zero. Since the attribution for the r^{th} variable is defined as $\beta_r * (x_r^* - x_r')$, the attribution will always be zero.

E. Linearity

The linearity axiom states that if a model (derived model) is a linear combination of two or more models (source models), then the attributions of the derived model should also be a linear combination of the attributions of the source models [59]. Let's consider the following two models,

$$f(\boldsymbol{x}) = \beta_{f_0} + \sum_{i=0}^{p} \beta_{f_i} x_i$$
$$g(\boldsymbol{x}) = \beta_{g_0} + \sum_{i=0}^{p} \beta_{g_i} x_i$$

For a baseline x' and an input x^* the attributions of the above two models can be given by

$$f(\boldsymbol{x}^*) - f(\boldsymbol{x}') = \sum_{i=0}^{p} \beta_{f_i} (x_i^* - x_i')$$
$$g(\boldsymbol{x}^*) - g(\boldsymbol{x}') = \sum_{i=0}^{p} \beta_{g_i} (x_i^* - x_i')$$

Let's consider a model $h(x)$, which is a linear combination of f and g such that

$$\begin{aligned} h(\boldsymbol{x}) &= a * f(\boldsymbol{x}) + b * g(\boldsymbol{x}) \\ &= a * \left(\beta_{f_0} + \sum_{i=0}^{p} \beta_{f_i} x_i\right) + b * \left(\beta_{g_0} + \sum_{i=0}^{p} \beta_{g_i} x_i\right) \end{aligned}$$

Now the attribution of the variables in the model h can be given by,

$$\begin{aligned} & a * \left(\beta_{f_0} + \sum_{i=0}^{p} \beta_{f_i} x_i^*\right) + b * \left(\beta_{g_0} + \sum_{i=0}^{p} \beta_{g_i} x_i^*\right) - a * \Big(\beta_{f_0} + \\ & \sum_{i=0}^{p} \beta_{f_i} x_i'\Big) - b * \left(\beta_{g_0} + \sum_{i=0}^{p} \beta_{g_i} x_i'\right) \\ &= a * \left(\sum_{i=0}^{p} \beta_{f_i} x_i^* - \sum_{i=0}^{p} \beta_{f_i} x_i'\right) + b * \left(\sum_{i=0}^{p} \beta_{g_i} x_i^* - \sum_{i=0}^{p} \beta_{g_i} x_i'\right) \\ &= a * \sum_{i=0}^{p} \beta_{f_i} * (x_i^* - x_i') + b * \sum_{i=0}^{p} \beta_{g_i} (x_i^* - x_i') \end{aligned}$$

Clearly the attribution of each variable of h is a linear combination of the attribution of the corresponding variables of f and g.

F. Implementation Invariance

Two models are functionally equivalent if their outputs are equal for all inputs, despite having very different implementations. Implementation invariance axiom states that the attribution should be identical for functionally equivalent model [59]. Let's consider the following two models,

$$f(\boldsymbol{x}) = \beta_{f_0} + \boldsymbol{\beta}_f \cdot \boldsymbol{x}$$
$$g(\boldsymbol{x}) = \beta_{g_0} + \boldsymbol{\beta}_f \cdot \boldsymbol{x}$$

If $f(\boldsymbol{x}) = g(\boldsymbol{x}), \forall \boldsymbol{x}$, then then the two networks are functionally equivalent and

$$\beta_{f_0} + \boldsymbol{\beta}_f \cdot \boldsymbol{x} = \beta_{g_0} + \boldsymbol{\beta}_g \cdot \boldsymbol{x}$$
$$\boldsymbol{\beta}_f = \boldsymbol{\beta}_g$$

Thus, if two linear models are functionally equivalent, they must be the same model, and, therefore, their attributions must be identical.

G. Symmetry-preserving

If swapping two variables does not have any impact on the output of a function then these variables are called symmetric with respect to the function. For example, if for all values of x_1 and x_2, $f(x_1, x_2) = f(x_2, x_1)$, then x_1 and x_2 are symmetric with respect to f. The symmetry-preserving axiom states that "if for all inputs that have identical values for symmetric variables and baselines that have identical values for symmetric variables, the symmetric variables receive identical attributions" [59].

Let's consider two symmetric variables x_1 and x_2 with respect to the log odds $f(\boldsymbol{x})$

$$f(x_1, x_2) = \beta_0 + \beta_1 \cdot x_1 + \beta_2 \cdot x_2$$
$$f(x_2, x_1) = \beta_0 + \beta_1 \cdot x_2 + \beta_2 \cdot x_1$$

Now,

$$\beta_0 + \beta_1 \cdot x_1 + \beta_2 \cdot x_2 = \beta_0 + \beta_1 \cdot x_2 + \beta_2 \cdot x_1$$
$$\beta_1(x_1 - x_2) - \beta_2(x_1 - x_2) = 0$$
$$(\beta_1 - \beta_2)(x_1 - x_2) = 0$$

Since x_1 and x_2 are symmetric with respect to f for all values of x_1 and x_2, β_1 must also be equal to β_2.

Now the attributions for the variable x_1 is

$$\beta_1 * (x_1^* - x_1')$$

and the attributions for the variable x_2 is

$$\beta_2 * (x_2^* - x_2')$$

Since the coefficients are equal, input is identical, and the baseline is identical. So, the attribution for each symmetry preserving variable is identical.