# Explaining Denials: Adverse Action Codes and Machine Learning in Credit Decisioning

George Krivorotov Jeremiah Richey

June 10, 2022

**Abstract**

As ML/AI usage expands to a variety of high-stakes decisioning, there is an increasing need to provide transparent explanations. In credit decisioning, regulation has long stipulated the provision of "Adverse Action Codes" (AACs) explaining denial of credit. However, there is potentially a wide range of acceptable AAC methodologies that could be used. This paper compares AACs derived from four common methods - an axiomatically-backed Shapley-based approach, a Most Points Lost-based approach, a difference from Mean approach, and a Univariate binning approach - based on two XGBoost risk models predicting credit card and mortgage risk respectively. We find that, overall, the Univariate approach deviates the most from the Shapley-based approach, all derived differences are 'significant' within our novel placebo testing framework, differences are more pronounced for lower-risk customers on the border of the reject boundary, and the Most Points Lost and Mean approach are less robust to data perturbations than Shapley.

**JEL Classification:** D81, C4, G5
**Key Words:** Machine Learning Explainability, Credit Risk Modeling, Shapley, Adverse Action Codes

**Acknowledgements:** The views expressed in this paper are those of the authors alone and do not necessarily reflect those of the Office of the Comptroller of the Currency or the US Department of the Treasury. All errors are our own.

# 1 Introduction

Given the potential gains in accuracy of machine learning (ML) models in credit underwriting over traditional models, many lenders are increasingly relying on them for credit decisions. However, given the complex nature of many such models, often their decisioning logic is not readily explainable. This poses potential challenges not only in managing model risks for developers/validators, but also due to regulatory requirements that lenders provide borrowers the reasons for denials, should they occur - 'Adverse Action Codes' (AAC)[1]. These are stipulated under both the Equal Credit Opportunity Act (ECOA) implemented by Regulation B, as well as the Fair Credit Reporting Act (FCRA).

The ECOA states

**Equal Credit Opportunity Act (ECOA) - 12 CFR 1002.9(a)(2)**

*The creditor must ... provide the applicant with the specific principal reason for the action taken...*

Furthermore, Paragraph 9(b)(2) specifies that such provided reasons should "relate to and accurately describe the factors actually considered or scored" by the creditor. While in a slightly different context the FCRA stipulates:

**Fair Credit Reporting Act (FCRA) - 15 U.S. Code §1681g**

*A consumer reporting agency shall supply to the consumer ... a notice which shall include ... all of the key factors that adversely affected the credit score of the consumer in the model used, the total number of which shall not exceed 4 ...The term 'key factors' means all relevant elements or reasons adversely affecting the credit score for the particular individual, listed in the order of their importance based on their effect on the*

[1]This includes both assigning credit at time of booking, but also, in the case of revolving credit, for adjusting credit lines in response to borrower request, such as in reactive credit line increases for credit cards.

*credit score.*

However, the regulation is not prescriptive, and there are a wide range of methods that may be used to identify reasons for adverse actions. Indeed, paragraph 9(b)(2) makes clear that "The regulation does not require that any one method be used for selecting reasons for a credit denial or other adverse action that is based on a credit scoring system. Various methods may meet the requirements of the regulation. " This implies different lenders will likely utilize different methodologies to generate AACs for their ML models used in underwriting. Indeed, early adopters and vendors of ML/AI methods put forth or utilized several intuitively simple methods including (i) a "most points lost" (MPL) approach, (ii) a univariate binned predicted risk approach (Univariate), and (iii) a difference from means approach. These methods are simple and easy to compute, and can convey some essential, simplified information on denials. However, they may conceivably fall short if a model exhibits higher complexity, such as the presence of interactions or non-linearities.

In the face of these considerations, some lenders have started adopting alternative methodologies from the ML interpretability toolbox to generate AACs, most notably several implementations of Shapley values, an approach adapted from cooperative game theory. The motivation of Shapley is that the effect of a covariate $x_i$ on model output may differ depending on the values of the $\{x_{-i}\}$ other covariates if the variables interact. Shapley addresses this problem by taking the average of the effect of perturbations of $x_i$ on the output over all relevant perturbations of the $\{x_{-i}\}$ other covariates, thus incorporating information on these interactions. In addition, it is the only one that satisfies several desirable axiomatic properties, see Strumbelj and Kononenko (2014). Due to the need to take the average of many combinations of variable perturbations, Shapley can be computationally intensive, and there have been several implementations that simplify it at the cost of accuracy, such as TreeSHAP. In addition, there is a crucial choice of baseline/reference group that needs to be made, a topic we will only briefly touch on in this paper but address with more detail in

Krivorotov and Richey (forthcoming).

Unlike studies such as Strumbelj and Kononenko (2014) and Sundararajan et al. (2017) that take an axiomatic approach when justifying an interpretability method, the non-prescriptive nature of the regulatory requirement necessitates a more empirical approach in understand the differences between the methods. We do not make a stand on what the "correct" method would be, but simply compare their differences in key dimensions of interest. The "laboratory" where we will conduct our analysis will be 2 separate underwriting/account management style machine learning models estimated on data across two prominent credit portfolios: credit cards and mortgages. In this sense we are mimicking the setting where a lender has, for example, an account management cards model and is declining or approving requested line assignment increases thus necessitating the derivation of AACs. We set a reasonable reject bound for each model and calculate AACs for the entirety of the customers in our out-of-time reject sample across the 3 "legacy" methodologies - MPL, Univariate, and Difference from Mean, and one new methodology - Shapley. In this way, this paper is necessarily a set of case studies. However, we believe the results are generalizable qualitatively.

We perform four main analyses. First we simply document the differences in derived AACs between approaches. Given the theoretical justification for a Shapley approach, we set this as ground truth and then estimate several distance measures between derived AACs for individuals; we stress here that, though for the sake of our analysis we set it as a 'ground truth,' we base that off of its desirable axiomatic properties[2] and not the degree to which it is 'correct' in its relation to the requirements of the above cited regulation (see Section 2.5 where we discuss in more detail the conceptual goals of each method). Second, in order to give some meaning to these measured distances based on an objective measure of 'difference,' we take an approach that loosely follows the placebo testing done in the synthetic control literature (e.g.

[2]Specifically, efficiency, symmetry, dummy, and additivity (see Strumbelj and Kononenko (2010) for discussion)s. However, these axiomatic properties are not specified in ECOA and so we cannot say that having them would mean they are correct. ECOAs definition is broad enough so that multiple methods like Univariate, MPL, and Shapley can all be classified as correct.

Abadie et al. (2010) and Abadie et al. (2015)). This places between-AACs-methodologies distances within the distribution of between-model distances where this latter distribution is based on a simulation of multiple models with different seeds leading to between-model variations derived solely from the inherent randomness in the XGBoost algorithm. Third, we document the relationship between derived AAC distances and predicted risk. And lastly, we document AAC sensitivity to small perturbations in in a rejected applicant's data.

Our findings can be summarized as follows. First, the various AAC methods lead to clearly different AACs based on our chosen difference measures with MPL and the Mean approach being similar in their results and the Univariate approach leading to more pronounced differences. Second, based on our placebo testing approach, these differences are 'significant' as based on our derived pseudo p-value. Third, these seen differences are more pronounced for those with lower risk levels and closer to the accept boundary and thus more likely affected by AACs across all measures and methods. And lastly, we see differences in the deterioration of AACs with data perturbations with the Univariate approach being immune by its construction, the Shapley approach being least affected of the remaining and the MPL and Mean approaches deteriorating the fastest in somewhat similar fashion. These general results are robust across both credit portfolio models.

Our findings are closely related to recent research coming out of FinRegLab FinRegLab (2022). They investigate several methods for AACs (several Shapley approaches and several LIME approaches) and then analyze 'fidelity' and 'consistency'. Much of our analysis parallels, in part, their work on consistency which measures to what degree different approaches agree on derived AACs. Notabal differences are, however, that we compare a single Shapley approach to other approaches we have seen at lenders in the industry - MPL, Univarite, Difference from Mean - rather than several approaches of Shapley and LIME and we then go on to frame these differences in our placebo testing setting as well as document how these differences relate to predicted risk. Additionally, part of their fidelity analyis is centered

on a perturbation excercise which essentially mimics a MPL approach. So in a sense our Shapley-MPL comparison is related to their SHAP perturbation testing, though given the different aims the results are not directly comparable. Their main takeaway is that Shapley performs better in terms of fidelity than LIME and that there are considerable differences (inconsistency) even between different Shapley implementations; our main results find similar inconsistencies between a different set approaches and note the superior performance of Shapley in terms of stability rather than fidelity.

The rest of the paper proceeds as follows. Section 2 introduces the concept of AACs and discusses the derivation of the various AACs we utilize as well as conceptual differences between their goals. Section 3 discusses in greater details our analytical approach to investigating differences in AAC results. Section 4 discusses results and Section 5 concludes. Details of the two models employed and the data underlying them can be found in the appendix.

## 2 Deriving AACs

AACs, in this paper, are understood to be the model attributes determined to have the largest 'contributions'. What we mean by 'contributions' will vary according to AAC approach. To be clear, what we seek is some function $\phi$ that takes a model $f : \mathbb{R}^K \to [0, 1]$ and an input vector $x \in \mathbb{R}^K$ and yields a vector of 'contributions': $\phi(f, x) = (\phi_1, ...\phi_K) \in \mathbb{R}^K$. We then understand the subset of largest $\phi$ to be AACs (though at times in the paper we will refer to the entire set of $\phi$ as the AACs).

For standard underwriting models based on logistic regressions, we can outline a natural approach for deriving $\phi$. Such a model would estimate the log odds of default (i.e. the 'score') as a linear combination of attributes:

$$ln(\frac{Pr(y)}{1 - Pr(Y)}) = \beta_0 + X_1\beta_1 + ... + X_K\beta_K$$

From here, importantly, since the model is linear, the effect of $X_k$ is *constant* on the score *for everyone* and also *throughout the domain of* $X_k$. A natural approach in deriving contributions would be to set $\phi_j$ to $\beta_j(X_j - \bar{X}_j)$ given appropriately chosen baseline $\bar{X}_j$.

However, this simple approach can become infeasible in many ML models; particularly because most do not have an interpretable, closed-form expression. To that end, multiple solutions have been suggested to address this issue in practice. We will focus on four approaches that have been proposed as possible approaches or have been implemented at various financial institutions.

One approach, and one that is seen by many as the 'right' approach in some sense, is based on the concept of Shapley values and comes from work on game theory. In addition, Strumbelj and Kononenko (2014) and Sundararajan et al. (2017) show that it is the method

that satisfies several desirable axiomatic qualities, such as efficiency, additivity, and thus providing a theoretical justification for its usage, although as mentioned previously AAC guidance is nonprescriptive enough that methods do not necessarily need to follow these aixomatic properties. Another approach is to use a 'most points lost' approach where the 'contribution' is the difference in score between ones actual and the counterfactual if that variable were set at an individuals 'optimal' level.[3] A third approach is similar to an MPL approach, but instead considers the counterfactual if the variable is set to the average, and thus is a sort of measure of deviation from the average. Lastly, we will consider an approach that is based on univariate risk measure based on average predicted scores. Another, rather well-known approach known as LIME, will not be explored as we found it to be sensitive to too many seemingly subjective choices to make it unlikely to be used in a production environment;[4] furthermore we have not seen this used at any financial institutions.

## 2.1 Shapley Values

Shapley values are a decomposition of model score that assign points to each attribute such that they add up to the full model score. The concept is derived from game theory with the goal being to assign players' contributions to a game's score, explicitly taking account that one person's contribution depends on every other person's contribution.

The concept is simple. Assume there are three players (A,B,C). What is the contribution of player A? One answer is how the payoff changed when A entered the game: thus $\tilde{\phi}_1(f, A) = f(A, B, C) - f(B, C)$. But an equally correct answer is the change in the payoff after A entered when only B is present: $\tilde{\phi}_2(f, A) = f(A, B) - f(B)$, or after when only C is present: $\tilde{\phi}_3(f, A) = f(A, C) - f(C)$ or lastly the payoff with just A and no other player: $\tilde{\phi}_4(f, A) =$

[3]Note that in ML models, as opposed to logistic models, each individual would have their own optimal value.

[4]Even firms that do rely on some form of LIME to produce reason codes caution that their application should be done only after appropriate sensitivity checks and may not be valid for wide data (e.g. see Hall et. al. from the H20).

$f(A)$. The solution proposed is to take the average all of these possibilities, thus the Shapley value is: $\phi(f, A) = \frac{1}{4}\sum_i \tilde{\phi}_i(f, A)$.

This solution has several compelling properties related to the assigned contributions being 'fair' in some sense and is thus the basis for this approach being seen by many as the 'correct' answer to the question at hand (see Shapley (1953) for discussion and derivation or Strumbelj and Kononenko (2014) for more recent ML related discussion).

There is however a key issue to directly using this solution for the derivation of AACs for many ML models and that is "omitting" a certain variable from a function is not well-defined - we generally need values for all the attributes of the model to obtain a score. To discuss a practical approach to handling this issue we will introduce some notation.

Let $\mathbf{K} = \{1, 2, ..., K\}$ be a set of $K$ features, let $A$ be the feature space and $x = (x_i, x_2, ...x_K) \in A$ be an instance in the feature space. Furthermore let $f$ be a classifier of interest and $c$ the class of interest and so $f_c(x)$ the classifiers' prediction regarding class $c$. We begin by defining the *partial prediction*. This partial prediction $Pf_c(S)$ is the expected prediction if we only knew the values of a subset of the features $S \subset K$.

$$Pf_c(S) = \frac{1}{|A_{K\setminus S}|}\sum_{y \in A_{K\setminus S}} f_c(\tau(x, y, S)) \tag{1}$$

where $\tau(x, y, S) = (z_1, z_2, ..., z_K)$ such that $z_k = x_k$ if $k \in S$ else $z_k = y_k$.

What this represents is the average of all the predictions where the known attributes $(x)$ are set to their value and all other attributes $(y)$ are assigned to every possible combination of those attributes. So, for example, if $n = 6$ and we did not know $x_5$ and $x_6$ and they both took the value $\{0, 1\}$ then we would average over the four possible predictions where for each we set $(x_1, ..., x_4)$ to their known values and for the unknowns we had the possible combinations: $(0, 0); (0, 1); (1, 0); (1, 1)$. Of course with continuous attributes, or even those with multiple

values, this quickly becomes problematic in practice, a point we return to below regarding approximations.

Importantly, note how this partial prediction is done: it is an average of every possible mix of realizations of unknown attributes. In this sense it is 'data agnostic' and only relies on the model itself and does not leverage the development data (other than the domains for $x_k$); however, this differs from another practical alternative introduced in the next section.

To complete the Strumbelj and Kononenko (2014) formulation of Shapley, we will now let $\pi(K)$ be the set of all ordered permutations of $K$ and $Pre^k(O)$ be the set of attributes that proceed attribute $k$ in the order $O \in \pi(K)$. Now, we can write the Shapley value of attribute $k$ as:

$$\phi_k = \frac{1}{K!} \sum_{O \in \pi(K)} (Pf_c(Pre^k(O) \bigcup \{k\}) - Pf_c(Pre^k(O))) \tag{2}$$

This is the difference in the *partial prediction* with and without attribute $k$ averaged over every possible order that $k$ might enter into the prediction. Recall from the definition of $Pf_c(\cdot)$ that the partial prediction would compute the average prediction over all possible values of the omitted variables. However, it is unclear how to implement this in the case of continuous variables and even for categorical variables may be computationally difficult. In addition, one might want to weight certain classes or data ranges more or less when computing this average to have a more empirically reasonable baseline.

### 2.1.1 Sample Approximations

To circumvent this issue, Strumbelj and Kononenko (2014) suggest an approximation,which is the one employed in our application. To do this, first note we can adjust the partial prediction by double counting some instances and correcting the term with the proper average

adjustment; while this seems an added complication it simplifies the sampling strategy. More precisely, adjust the partial prediction to be:

$$Pf_c(S) = \frac{1}{|A|}\sum_{y\in A} f(\tau(x, y, S)) \tag{3}$$

What this does, compared to equation (2) is create many duplicates of each variation of the $y \in A_{K\setminus S}$, but this number of duplicates is equal for each $y \in A_{K\setminus S}$, so the average does not change. This implies another form for the Shapley value that lends to a sampling approximation:

$$\phi_k = \frac{1}{K!|A|}\sum_{O\in\pi(K)}\sum_{y\in A}(f(\tau(x, y, Pre^k(O)\bigcup\{k\})) - f(\tau(x, y, Pre^k(O)))) \tag{4}$$

From this we see that the Shapley value is an average where the sampling population is $\pi(K) \times A$, and in practice this may be infinite or, when not, still impractical to attempt to implement with any large model. However, since the value is equivalent to a population average, then it can be approximated with sample averages which will follow the central limit theorem. Thus an unbiased estimator of $\phi_k$ can be achieved by:

1. Select a random permutation $O \in \pi(K)$
2. Select a random mix of attributes $y \in A$
3. Compute $v_1 = Pf(\tau(x, y, Pre^k(O)\bigcup\{k\}))$
4. Compute $v_2 = Pf(\tau(x, y, Pre^k(O)))$
5. Compute $\tilde{\phi_k} = v_i - v_2$
6. Average values for Step 5 over a large number ($m$) of iterations of 1-5

Of course the size of $m$ is potentailly an issue. The estimator $\tilde{\phi}_k$ is governed by the central limit theorem and is approximately normal with mean $\phi_k$ and variance $\frac{\sigma_k^2}{m}$ where $\sigma_k^2$ is the population variance of the attribute's contribution. Given that contribution is bounded between -1 and 1 (since we are dealing with a classifier), the upper bound for this variance is 1. This leads to a simple power analysis to determine the appropriate $m$ to reach a given errror and confidence level. For any confidence level $(1-\alpha)$ and error $\epsilon$ we will wish the following to hold: $P(|\phi_k - \tilde{\phi}_k| < \epsilon) = 1 - \alpha$. And, assuming the approximate normality of the sampling distribution, we can derive the appropriate sample size: $m_i(1-\alpha, \epsilon) = \frac{Z_{1-\alpha}^2 \cdot \sigma_k^2}{\epsilon^2}$. Thus, for example, assuming the limit of 1 for $\sigma$, a 99% confidence, and an error of 0.01 we arrive at approximately 65,000 samples. In practice this will likely be much smaller and $\sigma^2$ could be estimated during the the sampling to determine an earlier stopping point.

### 2.1.2 Training Shapley

A final conceptual question to make the link from the game theoretic Shapley to a Shapley value that can be used for setting AACs is the choice of sampling distribution, in effect a baseline, which turns out to change the conceptual interpretation of the values and indeed give significantly different results. It is important to note that Shapleys with different baselines would still satisfy the same axioms, but the interpretation would be different, impacting the evaluation of derived AACs.

Let's say the model is trained on a wide range of customers, some of which may be in the regions of the feature space which would subsequently cause them to be rejected due to credit policy. If the Shapley sampling distribution is based on the empirical distribution of the training data, then calculating Shapley values for rejected customers would entail calculating marginal contributions relative to the *entire* set, including those customers that would subsequently have been rejected. This may muddle the interpretation of the reject reason since rejected customers may rather understand the main factors driving their difference

from the customers that would be accepted rather than from the full population. On the other hand, given how risk scores are used in practice within larger risk strategies that may continually change score cut-offs, Shapley values based on such subsets of the data could change overtime even if customers' data and scores remain constant.

Based on these considerations, as well as the observed approaches taken by lenders utilizing Shapley values for AACs, we will derive AACs from Shapley based on training them with respect to a baseline that consists of the full developmental population. However, it is worth noting that this would not be the only possible logical decision for a baseline. Other potential meaningful baselines would include 1) only the accepted customers from the underlying data, 2) only the accepted customers *closest* to the customer in question or 3) a point or area in the feature space *closest* to the customer in question.[5] For a more detailed study of the consequences of various baseline choices for Shapley explanations, see Krivorotov and Richey(forthcoming).

## 2.2 Most Points Lost

While Shapley provides a sophisticated method to explain models with interactions and nonlinearities, other, simpler techniques that were used commonly in the linear regression world are easier to compute and could conceivably provide similar explanations in some cases.

The first such approach we consider is known as "most points lost," (MPL). In this approach, we evaluate which *single* characteristic of a customer can be modified to create the maximal decrease in score (reflecting probability of default in this context), without changing any other attribute. In the simple logistic regression approach this would be, for each customer

[5]For either example, "closest" would depend on the feature space metric chosen and could vary significantly depending on objective for the explanation.

$x \in X$, $i \in A$,

$$\phi_i = \beta_i(x_i - \underline{x}_i)$$

where $\underline{x}_i$ is the minimal value in the domain if $\beta_k > 0$ and maximal value in the domain if $\beta_k < 0$.

This is simple to extend in the ML context by introducing the concept of an Individual Conditional Expectation (ICE) plot. These are used commonly to demonstrate how a prediction can change (potentially nonlinearly and nonmonotonically) after perturbing a *single* variable, keeping all other variables in the observation constant. [6]

To calculate the MPL with an ICE plot, one first finds the value $\underline{x}_i$ such that it minimizes the ICE curve for that particular observation. A key difference from the linear setting is that in an ML context, this may be unique for each observation due to interactions and non-monotonicities. We then derive $\phi_i$ by taking the difference between this minimal value and the observation itself. Thus

$$\phi_i = f(X_{-i}, \underline{x}_i) - f(X)$$

In our analysis we will, for each individual search the domain of $x_i$ with a quantile search using 100 grid points.

## 2.3 Divergence from Mean

A somewhat similar approach to MPL, though computationally easier, is to consider how a person's score would change had they had the average attribute value of the population; this is conceptually easy to grasp for many and may be deemed a natural approach. Thus

[6] It is also worth noting that this breaks cross-correlations in the data and can enter regions in the feature space that are not well represented in the data, which can either be a flaw or a benefit depending on the objective of the analysis.

for the 'Mean' approach, $\phi_i = f(X_{-i}, \bar{x}_i) - f(X)$ where $\bar{x}_i$ is the average from the model's development data.

## 2.4 Univariate Risk

The last approach we will explore divides each variable into decile bins and computes a mean model prediction for each bin. The $\phi_i$ for each customer and variable would be assigned based on the bin that each customer's level of variable $i$ falls in. Importantly, observations in each bin for variable $x_i$ are not fixed in any way for any other variable $x_k$ and vary within bin according to the natural joint distribution of those $(x_{-i})$ to $x_i$.

We can relate this methodology to Shapley and define it as a partial prediction (rather than a difference of partial predictions):

$$\phi_k = \tilde{P} f_c(x_k) = \frac{1}{|f(K_{N \setminus x_k}|x_k)|} \sum_{y \in f(A_{K \setminus x_k}|x_k)} f_c(x_k, y) \tag{5}$$

Note that this approach only considers one (conditional) partial prediction, that of introducing the attribute of interest first and then taking the *conditional* expectation of the classifier given that attribute as in Shapley, but only considering that one partial prediction. Thus it ignores all the other attribute values when assigning 'points' to the attribute of interest, in conflict with the idea behind the Shapley approach.

This methodology is a relatively common and computationally simple way to create AACs. While this provides an understanding of how risky on average a customer with variable $i$ is projected to be, it has the risk that it may ascribe high levels of importance to variables $x_i$ that are *correlated* to important variables $x_k$ but are not by themselves important. Say for example a credit risk model is trained on history of delinquencies but not FICO. This

method still ascribe high levels of importance to FICO even if it is not an input in the model since it is generally highly correlated to history of delinquencies.

## 2.5 Conceptual Comparison of AAC Approaches

Before moving on to analyze how different AAC methods are in the top attributes they select in application, we want to briefly discuss the different conceptual questions each method is answering. Since they are answering different questions, we can expect them to arrive at different answers, and we want to address these issues before moving on to measure how different they in fact are.

Shapley, as discussed above, is attempting, within a game-theoretic framework, to answer the question: "How much did each attribute contribute to the eventual score taking into account interactions from other attributes?" The MPL approach is asking a 'what if question'; specifically: "For each attribute, how much would your score change if we altered that attribute to the most beneficial level?" The Mean approach also asks a 'what if question'; specifically: "For each attribute, how much would your score change if we altered that attribute to the unconditional average level?" The univariate approach asks: "For each attribute, if the only information I knew about you was your level of that particular attribute, what risk level would my model predict you to have?"

The fact that the four approaches ask four fundamentally different questions notwithstanding, we still wish to answer, in practice, how much does this matter. This is the main contribution of the paper and our methodology is discussed in the next section.

# 3 Comparison Analyses

## 3.1 Measures for Comparing AACs

After running each of the four AAC methodologies on the full OOT reject sample, we base our analysis on four different measures of the cross-method differences in derived AACs (again noting we compare MPL, Mean, and Univariate to Shapley):

1. Kendall Tau distance on the full set of AAC rankings
2. Euclidean Distance on the full set of AAC rankings
3. A simple cardinal measure of how many of the top 4 AACs correspond between methods
4. a Euclidean distance between the rankings of the top 4 AACs for Shapley and the corresponding rankings for the other methods

In AACs, the information being provided to the customer is primarily ordinal in nature and so we focus on differences in rankings instead of magnitudes of attributions. To that end, the first metric, Kendall Tau, is focused on rank-ordering, providing a measure of pairwise agreement between the lists. The Euclidean distance on the other hand would provide more color on the difference in magnitude between some variable rankings - for example, if a variable went from the bottom to the top of the ranking or vice versa between the methodologies this would be highly weighted here.

The focus of the last two measures on top four attributes is due to two main concerns. The first is that the least important AACs are likely small in magnitude and highly likely to jump in order, so differences in those may cloud out differences in more meaningful/important AACs. Second, it is common practice in the industry to provide four AACs, thus this focus also has real world implications.

## 3.2 Placebo Testing

While the analysis above provides direct measures of differences between AACs derived by various methodologies, they do not give any measure as to whether these differences are meaningful in any way; for example does a Kendall Tau of 0.8 indicate meaningful differences in any objective way? To address this issue we loosely follow the placebo testing approach often done in the synthetic control literature. This approach compares a change in an outcome of interest for a unit which received a treatment to a distribution of outcome changes from units that did not receive the treatment; the idea being that some change in outcome is bound to be present simply due to underlying randomness in the data generating process and thus 'meaningful difference' should be in comparison to that distribution.

Our parallel to this approach uses a baseline measure of 'difference' as the one embedded inherently within the XGBoost model due to the random column and row sampling used in the algorithm. If we were to repeatedly fit our XGBoost model with the same hyperparameters, but a different seed, then recalculate our Shapley AACs, this would output different results. We then use the in-between-model differences in our four measures (Tau, Euclidean, Top 4, Top 4 Euclidean) as a quasi-objective measure of difference, and compare each individual's between-AAC-method distance measure to their unique in-between-model distance distribution.

Thus our approach is as follows. Optimally fit an XGBoost model (details in appendix) on our full development data set. Then, refit 50 XGBoost models using the same hyperparameters but different seeds (thus allowing randomness inherent in the XGBoost model to differ from the original model). Then, for each observation in an out-of-time (OOT) sample, extract Shapley values for each of the 50 models and create a distribution of 1,225 cross-model measurements of 'difference' between the Shapley values for all four measures discussed above.[7] We then use these observation-specific distributions as our comparison

[7] $50 * (50 - 1)/2 = 1,225$

when we investigate differences in AACs between the various approaches within our main model. We then compare our between-AAC-methodology distances to that observation's ECDF.

## 3.3 Distance vs. Risk

Next we investigate if there is a relationship between our derived distances in our first analysis and predicted risk. There is reason to believe it is likely that very risky individuals have clear red flags that lead to more or less the same reasons for such predictions, and thus different AAC methodologies may be more likely to agree as to the factors contributing to this risk. Conversely, those just on the threshold of being rejected may have more nuanced reasons for being rejected and thus the different methodologies may be less likely to agree as to main drivers of risk, in an "Anna Karenina"-like principle for credit worthiness.

If this principle holds, it is especially important to keep in mind as these are precisely the individuals most likely to have actionable insights on their AACs due to their proximity to being approved for credit. We perform these analysis with simple bivariate regressions (standard mean as well as quantile).

## 3.4 Sensitivity Analysis

Lastly we investigate the different AAC methodologies for their sensitivity to perturbations in the underlying data. Given that many credit risk drivers and measures that can change by small amounts over time, such as average balances, credit ratios or number of accounts, and that ML models may have many non-linearities and non-monotonicities in the relationships between attributes and predicted risk, we wish to see how much of this information is passed to the derived AACs.

We do this with a simulation exercise. First, we take each account from our OOT rejected pool and perturb each variable by 5% of the development datum's standard deviation in a random direction. Next, we loop through each variable and set it back to its real value and calculate that variable's $\phi$ for each account. Hence, each perturbed AAC has the variable of interest $x_i$ kept at the same value as the original, only the other variables $x_{-i}$ vary. Each of these derived $\phi_{i,\text{pert}}$ constitute the AACs. We then calculate the same distance measures as in the baseline analysis, however instead of calculating them across methodology, we calculate them *within* methodology, but between perturbed and non-perturbed. We then replicate this exercise in 5% intervals to a maximum of 50% of the SD of the data.[8] Then, for each AAC methodology we plot the mean distance of the perturbed AAC to its original one to see how the method deviates with such data changes.

[8]For the mortgage model which has several categorical variables we must take a slightly different approach from perturbing the data X% of the SD; rather we move X% of accounts to the next most likely category. While this is not exactly the same approach of course, we believe it a conceptually parallel move.

# 4 Results

## 4.1 Raw Differences

Tables 1 and 2 provide raw differences between Shapley and the other AAC methods for our four chosen measures for our Credit Card Model and Mortgage Model respectively; the tables provide means and standard deviations as well as select quantiles for our OOT rejected sample. For the credit card model, we can see that the MPL and Mean AAC approach lead to similar distributions of our difference measures with the Univariate approach diverging much more on all levels. Thus while MPL and the Mean approach lead to an average Kendall Tau of just under 0.5 and an average top 4-in-4 of slightly over 2, the Univariate approach leads to an average Tau of just under 0.25 with the 90th quantile only reaching 0.36 and an average 4-in-4 of under 1.5. Similarly for the Mortgage model, though not as drastic, the Univariate approach differs most from Shapley with an average 4-in-4 of 2.24 compared to 2.72 or 3.09 for the MPL or Mean approach; related, the 90th quantile for MPL and Mean are both 4 with Univariate only reaches 3 and the Euclidian distance for the Top 4 is 12 for the Univariate vs. around 5 for both Mean and MPL.

Of course, a keen observer will quickly be led to the question of whether these difference measures are meaningful in any sense. In standard statistical inquiries this is answered with statistical tests and resulting p-values. To parallel this we now proceed to our placebo testing approach.

Table 1: Raw Difference Between Shapley and Other AACs Methods (Credit Card Model)

| | Mean | SD | 10th Q | 25th Q | 50th Q | 75th Q | 90th Q |
|---|---|---|---|---|---|---|---|
| Shap-MPL Tau | 0.46 | 0.12 | 0.31 | 0.38 | 0.46 | 0.55 | 0.62 |
| Shap-MPL Eucl. Dist. | 36.77 | 7.11 | 27.02 | 31.63 | 37.10 | 41.83 | 45.78 |
| Shap-MPL 4in4 | 2.27 | 0.81 | 1.00 | 2.00 | 2.00 | 3.00 | 3.00 |
| Shap-MPL Eucl. Top 4 | 8.33 | 4.85 | 3.16 | 4.69 | 7.28 | 10.95 | 15.17 |
| | | | | | | | |
| Shap-Uni Tau | 0.23 | 0.10 | 0.10 | 0.16 | 0.23 | 0.30 | 0.36 |
| Shap-Uni Eucl. Dist. | 49.15 | 5.26 | 42.49 | 45.52 | 49.04 | 52.73 | 56.02 |
| Shap-Uni 4in4 | 1.46 | 0.81 | 1.00 | 1.00 | 1.00 | 2.00 | 2.00 |
| Shap-Uni Eucl. Top 4 | 18.08 | 7.17 | 8.54 | 13.19 | 17.97 | 23.02 | 27.48 |
| | | | | | | | |
| Shap-Mean Tau | 0.46 | 0.14 | 0.28 | 0.37 | 0.46 | 0.57 | 0.65 |
| Shap-Mean Eucl. Dist. | 36.62 | 8.16 | 25.57 | 30.66 | 36.88 | 42.14 | 47.12 |
| Shap-Mean 4in4 | 2.18 | 0.82 | 1.00 | 2.00 | 2.00 | 3.00 | 3.00 |
| Shap-Mean Eucl. Top 4 | 10.53 | 6.14 | 3.46 | 5.57 | 9.33 | 14.56 | 19.39 |

*Notes*: Numbers represent raw measured differences between Shapley AACs and other methods on the Out of Time sample.

Table 2: Raw Difference Between Shapley and Other AACs Methods (Mortgage Model)

| | Mean | SD | 10th Q | 25th Q | 50th Q | 75th Q | 90th Q |
|---|---|---|---|---|---|---|---|
| Shap-MPL Tau | 0.52 | 0.10 | 0.40 | 0.46 | 0.53 | 0.59 | 0.64 |
| Shap-MPL Eucl. Dist. | 38.35 | 6.53 | 30.17 | 33.80 | 38.08 | 42.57 | 46.97 |
| Shap-MPL 4in4 | 2.72 | 0.76 | 2.00 | 2.00 | 3.00 | 3.00 | 4.00 |
| Shap-MPL Eucl. Top 4 | 4.68 | 3.98 | 1.41 | 2.00 | 3.32 | 6.16 | 9.43 |
| | | | | | | | |
| Shap-Uni Tau | 0.49 | 0.12 | 0.34 | 0.41 | 0.50 | 0.58 | 0.64 |
| Shap-Uni Eucl. Dist. | 40.34 | 7.97 | 29.67 | 34.26 | 40.32 | 46.28 | 50.89 |
| Shap-Uni 4in4 | 2.24 | 0.85 | 1.00 | 2.00 | 2.00 | 3.00 | 3.00 |
| Shap-Uni Eucl. Top 4 | 12.10 | 9.40 | 3.32 | 5.20 | 7.04 | 22.15 | 25.51 |
| | | | | | | | |
| Shap-Mean Tau | 0.61 | 0.12 | 0.45 | 0.54 | 0.63 | 0.70 | 0.75 |
| Shap-Mean Eucl. Dist. | 32.29 | 7.95 | 22.27 | 26.40 | 31.71 | 37.79 | 43.05 |
| Shap-Mean 4in4 | 3.09 | 0.58 | 2.00 | 3.00 | 3.00 | 3.00 | 4.00 |
| Shap-Mean Eucl. Top 4 | 5.83 | 7.65 | 1.00 | 1.41 | 2.45 | 5.48 | 21.00 |

*Notes*: Numbers represent raw measured differences between Shapley AACs and other methods on the Out of Time sample.

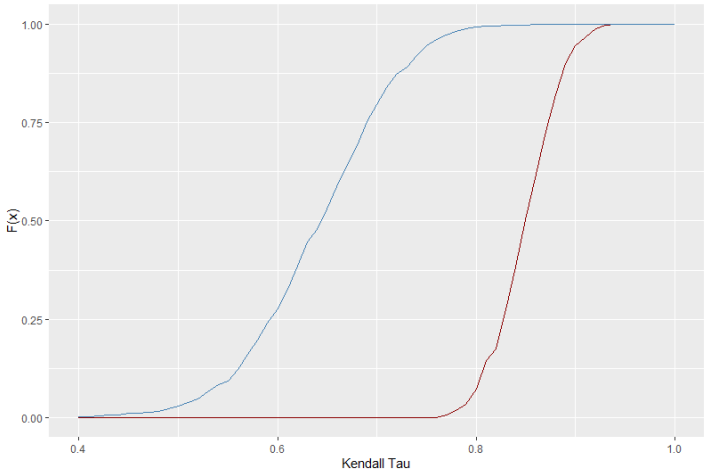
## 4.2 Placebo Testing

### 4.2.1 Within AAC Methodology Variation

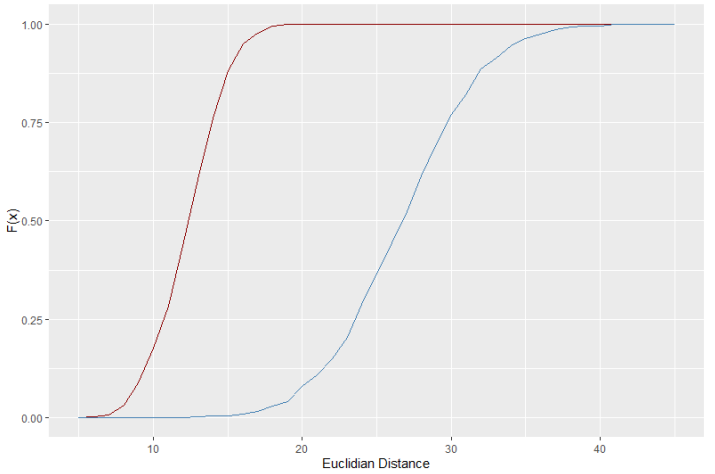
Before discussing the results of our placebo testing analysis, we first briefly note the difference in variation in cross-boostrap model ECDFs across observations. Recall that each observation has 50 different predictions and thus AACs within methodology from each bootstrap model. This means there are 1,225 cross-boostrap model AAC differences, and for *each observation*, we can construct an ECDF of these differences.

Let us take two sample observations from the OOT sample of the credit card model as examples. Figure (1) depicts the ECDFs of each these observations' 1,225 cross-boostrap model AAC differences. It is evident that these distributions are significantly different. Looking at Figure ( 1 ), we see that for one of these observations, nearly 100% of cross model Shapley's lead to Kendall Tau's of below 0.8, while for the other around only 5% lead to such measures. This continues to hold when looking at the Euclidean Distance measure, where one yields 100% under a measure of 20 while the other only about 5%. This large difference also shows up in the Top 4 measures seen in Figure ( 1 ); for one account only about 25% of cross model Shapley's lead to all 4 Top 4 AACs aligning while for the other over 50% do. This indicates that one of these observations is a kind of outlier in terms of the correlations between its attributes, with high variations across bootstrap models. However, for other observations, mean risk drivers stay relatively constant and the different bootstrap models tend to agree.

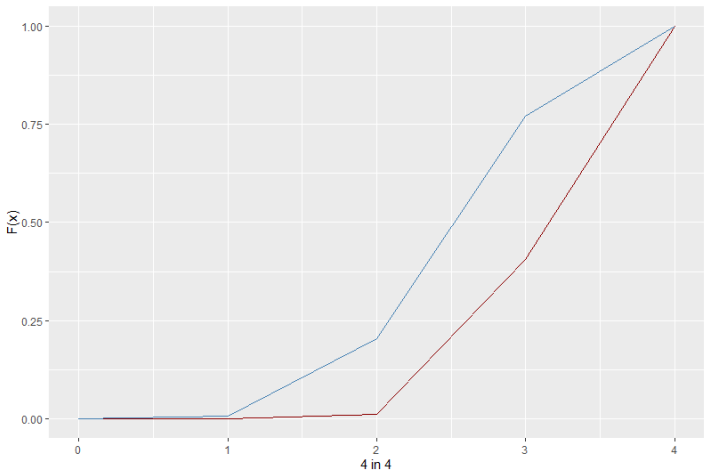
This highlights that there is substantial variation across models in derived AACs for the same AAC methodology as well as substantial differences in this variation across OOT observations. This latter point is why each observation's between-AAC method derived distance should be compared to its observation-specific ECDF as some observations have inherently greater variation in their predictions, risk drivers, and thus AACs due to how their attributes



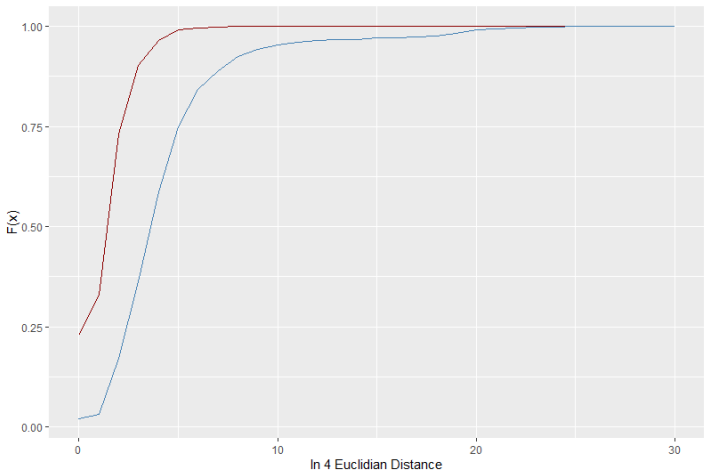

(a) Kendall Tau




(b) Euclidean Distance




(c) 4 in 4




(d) Top 4 Euclidean Distance

Figure 1: Between Model Variation in AACs

compare to the development sample.

### 4.2.2 Testing Results

Table (3) re-frames all of our distance measures for the credit card model and mortgage model based on our pseudo p-value approach. The table indicates the percentage of the OOT sample for whose between-method distance measures fall into the extreme tails of their between-model distance measures' distribution (we focus on the extreme 1/5/10% single sided tails). Here we see, starting first with the card model, that for all of our measures across the full set of AACs (Kendall Tau and Euclidian Distance) for all approaches the majority of observations' measures are 'significant' in in our use of the term; focusing on the 5% level the range of accounts indicating 'meaningful/significant' differences range from 83-100%. For our more restrictive measures focused on the top 4 AACs the results are more mixed. For Shap-MPL difference, *slightly more* than 25% of observations have top 4-in-4 measures in the extreme 5% tail and nearly 50% if we look at the Top 4 Euclidian distance; a very similar picture emerges from Shap-Mean measures as well. Thus for measures based on all AACs, the Mean and MPL approaches clearly lead to 'significantly' different results than Shapley, while if we restrict our focus to the top 4, there is enough variation even simply between models, that the results are less clear. For the Univariate approach however, this ambiguity is less pronounced as 66% have 4-in-4 measures in the extreme 5% tail of their between model distribution; similarly over 90% are in the 5% tail for the Top 4 Euclidean Distance measure.

Similarly for the mortgage model the Kendall Tau and Euclidean distances are clearly in the extreme tails of the ECDFs. And again, when focused on the more restrictive 4-in-4 we get somewhat more mixed results with roughly 50% or more of observations falling in the extreme *5%* tail for the Shap-MPL and Shap-Uni difference, and only around 25% for the Shap-Mean; however the Top-4 Euclidean distances are again more pronounced in this model than the cards model. Once again, this confirms that the Univariate methodology deviates the most from the other methodologies, while MPL and Mean tend to cluster together.

Table 3: Percentage of Accounts' Differences 'Significant' Between Shapley and Other AACs Methods

| | Credit Card Model | | | Mortgage Model | | |
|---|---|---|---|---|---|---|
| | 1% Sig | 5% Sig | 10% Sig | 1% Sig | 5% Sig | 10% Sig |
| Shap-MPL Tau | 0.81 | 0.90 | 0.94 | 0.99 | 1.00 | 1.00 |
| Shap-MPL Euc. Dist | 0.76 | 0.86 | 0.91 | 0.98 | 0.99 | 0.99 |
| Shap-MPL 4in4 | 0.19 | 0.29 | 0.37 | 0.42 | 0.46 | 0.49 |
| Shap-MPL Euc. in4 | 0.26 | 0.48 | 0.61 | 0.73 | 0.85 | 0.90 |
| Shap-Uni Tau | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |
| Shap-Uni Euc.Dist | 1.00 | 1.00 | 1.00 | 0.98 | 0.99 | 1.00 |
| Shap-Uni 4in4 | 0.53 | 0.66 | 0.74 | 0.61 | 0.65 | 0.68 |
| Shap-Uni Euc. in4 | 0.74 | 0.91 | 0.95 | 0.90 | 0.95 | 0.96 |
| Shap-Mean Tau | 0.78 | 0.87 | 0.91 | 0.96 | 0.99 | 0.99 |
| Shap-Mean Euc.Dist | 0.74 | 0.83 | 0.87 | 0.90 | 0.95 | 0.97 |
| Shap-Mean 4in4 | 0.21 | 0.33 | 0.41 | 0.22 | 0.26 | 0.29 |
| Shap-Mean Euc. in4 | 0.40 | 0.62 | 0.72 | 0.64 | 0.74 | 0.80 |

*Notes*: Numbers represent percent of raw differences that are in the X% tail of their between-model distribution of distances.

## 4.3 Differences vs. Risk

For our third analysis we compare our difference measures to the observations' predicted risk level, these are presented in Tables 5 and 6 for our credit card and mortgage model respectively. Here we see, again starting with the credit card model, a clear link between risk and our measures, with those less risky (closer to the accept region) having more dissimilar derived AACs (for Tau and 4-in-4 larger measures indicate similarity while for the Euclidean measures smaller ones do). Note here that the 4-in-4 regression results for the quantiles are somewhat misleading as there are only 5 possible results for the underlying measure (0-4) and so for this measure the only meaningful result is the mean regression. These results indicate that, beyond the previous results indicating these approaches lead to significantly different AACs, this effect is more substantial for those more likely to be on the border of being accepted and thus presumably those for whom meaningful and stable AACs are more important.

Results are similar for the mortgage model for the measures based on the full set of AACs (Kendall Tau and Euclidean Distance), though differ for those based on Top 4 for the Univariate and Mean approach. The 4-in-4 regression for Univariate and Means both have the opposite sign, and opposite signs also occur for the Top-4 Euclidean Distance at the lower quantiles. This reflects in part the restrictiveness of focusing on the Top 4. It also likely reflects aspects unique to the mortgage model, such as more factor variables, which reduces AAC variability for low risk customers.

Table 4: Regression Results: Distance Measures vs. Predicted Risk (Credit Card Model)

| | Mean | 10th Q | 25th Q | 50th Q | 75th Q | 90th Q |
|---|---|---|---|---|---|---|
| Shap-MPL Tau | 0.13 | 0.21 | 0.17 | 0.13 | 0.06 | 0.02 |
| | *(0.01)* | *(0.01)* | *(0.01)* | *(0.01)* | *(0.01)* | *(0.01)* |
| Shap-MPL Eucl. Dist. | -6.99 | -0.81 | -2.85 | -7.78 | -10.26 | -11.64 |
| | *(0.48)* | *(0.74)* | *(0.65)* | *(0.54)* | *(0.54)* | *(0.58)* |
| Shap-MPL 4in4 | 0.13 | | | | | |
| | (0.06) | | | | | |
| Shap-MPL Eucl. Top 4 | -3.89 | -0.00 | -1.11 | -3.20 | -5.63 | -8.77 |
| | *(0.33)* | *(0.22)* | *(0.25)* | *(0.32)* | *(0.48)* | *(0.66)* |
| Shap-Uni Tau | 0.14 | 0.15 | 0.15 | 0.14 | 0.14 | 0.14 |
| | *(0.01)* | *(0.01)* | *(0.01)* | *(0.01)* | *(0.01)* | *(0.01)* |
| Shap-Uni Eucl. Dist. | -7.03 | -7.94 | -7.03 | -7.29 | -7.13 | -6.82 |
| | *(0.35)* | *(0.58)* | *(0.50)* | *(0.46)* | *(0.51)* | *(0.63)* |
| Shap-Uni 4in4 | 0.91 | | | | | |
| | *(0.05)* | | | | | |
| Shap-Uni Eucl. Top 4 | -8.87 | -11.23 | -11.31 | -7.99 | -6.53 | -6.88 |
| | *(0.48)* | *(0.68)* | *(0.68)* | *(0.73)* | *(0.77)* | *(0.76)* |
| Shap-Mean Tau | 0.11 | 0.29 | 0.20 | 0.11 | -0.04 | -0.10 |
| | *(0.01)* | *(0.01)* | *(0.01)* | *(0.01)* | *(0.01)* | *(0.01)* |
| Shap-Mean Eucl. Dist. | -5.83 | 7.16 | 2.60 | -6.59 | -10.87 | -15.31 |
| | *(0.56)* | *(0.65)* | *(0.63)* | *(0.59)* | *(0.53)* | *(0.56)* |
| Shap-Mean 4in4 | 0.86 | | | | | |
| | *(0.06)* | | | | | |
| Shap-Mean Eucl. Top 4 | -6.03 | -3.27 | -3.74 | -5.79 | -7.70 | -9.10 |
| | *(0.42)* | *(0.36)* | *(0.40)* | *(0.52)* | *(0.69)* | *(0.68)* |

*Notes*: Numbers represent coefficients of regressing risk against distance measure, standard errors are in parenthesis.

Table 5: Regression Results: Distance Measures vs. Predicted Risk (Mortgage Model)

| | Mean | 10th Q | 25th Q | 50th Q | 75th Q | 90th Q |
|---|---|---|---|---|---|---|
| Shap-MPL Tau | 0.08 | 0.14 | 0.11 | 0.07 | 0.05 | 0.04 |
| | *(0.00)* | *(0.01)* | *(0.00)* | *(0.00)* | *(0.00)* | *(0.00)* |
| Shap-MPL Eucl. Dist. | -6.27 | -4.33 | -4.83 | -5.76 | -7.29 | -8.90 |
| | *(0.24)* | *(0.36)* | *(0.34)* | *(0.32)* | *(0.32)* | *(0.37)* |
| Shap-MPL 4in4 | 0.92 | | | | | |
| | *(0.03)* | | | | | |
| Shap-MPL Eucl. Top 4 | -4.10 | -0.67 | -1.69 | -3.76 | -6.81 | -10.15 |
| | *(0.15)* | *(0.07)* | *(0.08)* | *(0.11)* | *(0.16)* | *(0.34)* |
| Shap-Uni Tau | 0.06 | 0.07 | 0.07 | 0.08 | 0.05 | 0.03 |
| | *(0.00)* | *(0.01)* | *(0.01)* | *(0.01)* | *(0.00)* | *(0.00)* |
| Shap-Uni Eucl. Dist. | -5.46 | -4.38 | -6.02 | -7.16 | -5.44 | -4.89 |
| | *(0.31)* | *(0.33)* | *(0.34)* | *(0.41)* | *(0.60)* | *(0.79)* |
| Shap-Uni 4in4 | -1.26 | | | | | |
| | *(0.03)* | | | | | |
| Shap-Uni Eucl. Top 4 | -4.45 | 2.96 | 2.95 | -2.16 | -20.36 | -8.62 |
| | *(0.37)* | *(0.15)* | *(0.18)* | *(0.37)* | *(0.47)* | *(1.45)* |
| Shap-Mean Tau | 0.17 | 0.25 | 0.20 | 0.16 | 0.13 | 0.11 |
| | *(0.00)* | *(0.01)* | *(0.01)* | *(0.01)* | *(0.00)* | *(0.01)* |
| Shap-Mean Eucl. Dist. | -11.58 | -8.24 | -9.66 | -11.77 | -13.40 | -15.07 |
| | *(0.27)* | *(0.40)* | *(0.33)* | *(0.35)* | *(0.37)* | *(0.49)* |
| Shap-Mean 4in4 | -0.05 | | | | | |
| | *(0.02)* | | | | | |
| Shap-Mean Eucl. Top 4 | -8.15 | 1.19 | -0.56 | -1.72 | -15.91 | -21.71 |
| | *(0.28)* | *(0.09)* | *(0.09)* | *(0.14)* | *(0.94)* | *(0.32)* |

*Notes*: Numbers represent coefficients of regressing risk against distance measure, standard errors are in parenthesis.

## 4.4 Sensitivity

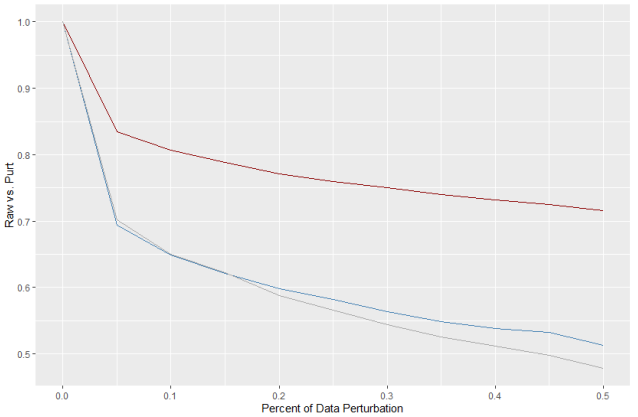
For our last analysis we investigate the sensitivity of the different AAC methodologies to small perturbations in the data (we note here again that the Univariate approach, by its design, is not affected by such perturbations). Figures 3 and 5 plot the Kendall Tau and Euclidean Distance and how it changes *within* observation as the data is perturbed for the Card and Mortgage Model.

Unlike in a linear model which would be projected to change in a relatively uniform way following a data perturbation, in an ML model, we might expect potentially sharp changes in predictions, risk drivers, and thus reason codes, due to nonlinearities and interactions. Hence, a low sensitivity to data perturbations is not necessarily a desirable feature in an AAC, but the "right" sensitivity is. The Shapley method is designed to take into account nonlinearities and interactions, and so we can expect the changes following the perturbation in Shapley-derived AACs to relatively accurately reflect the changes that arise due to the structure of the model. This is reflected in the red lines in figures (2) and (3) and can be thought of as a baseline for which to evaluate the other AACs.

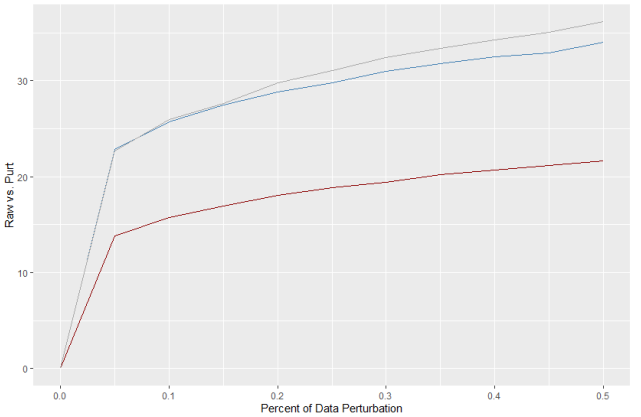
From the beginning, since the univariate method does not by definition move with any change in perturbation in the $(x_{-i})$ covariates, this can be thought of as an issue with this methodology. On the other hand, the picture with the MPL and and Mean approaches show the opposite issue - they are much more sensitive to perturbations in the data than Shapley across the board. We do notice some difference in the relation between the MPL and Mean breakdown between models with the MPL deviating further than the Mean approach for the mortgage model, this may reflect aspects of the mortgage model related to categorical variables.

These results are perhaps understandable as Shapley consists of an average of a variety of combinations of variables' perturbations, while MPL and the mean methods will only

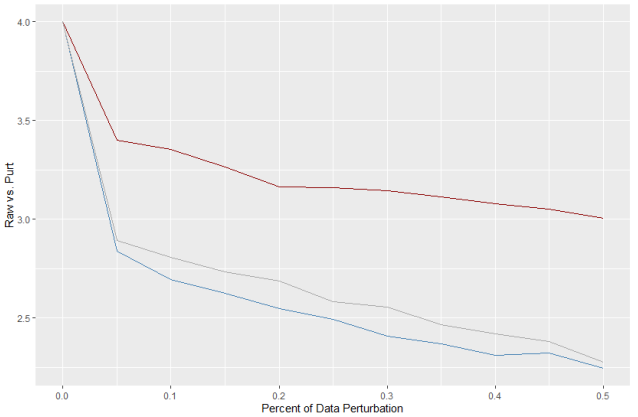
consider a single variable at a time, with a single variable's effect on output potentially changing drastically due to interactions. This lack of stability of these AACs when faced with these perturbations can be thought of as a failure of the univariate approach in understanding multi-dimensional relationships.



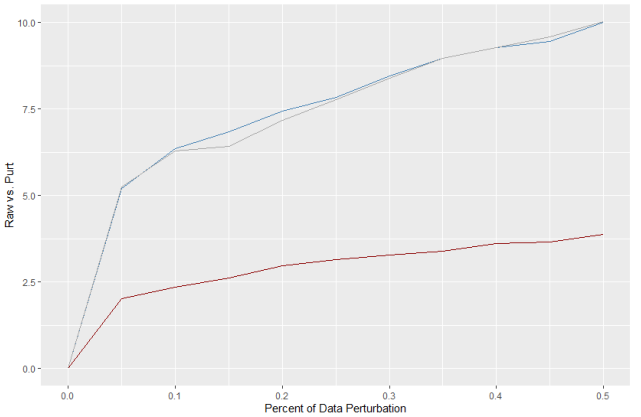

(a) Kendall Tau




(b) Euclidean Distance

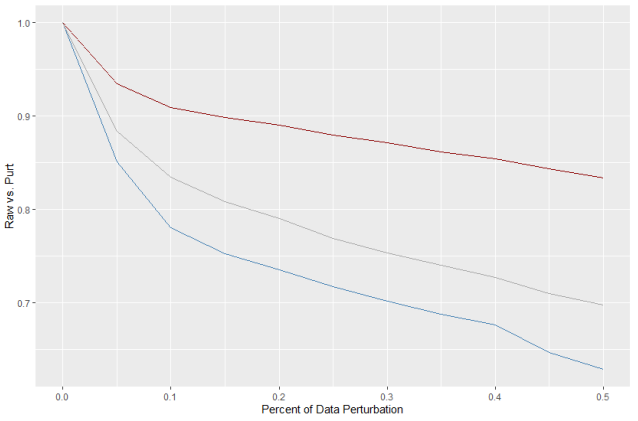



(c) Top 4 in 4

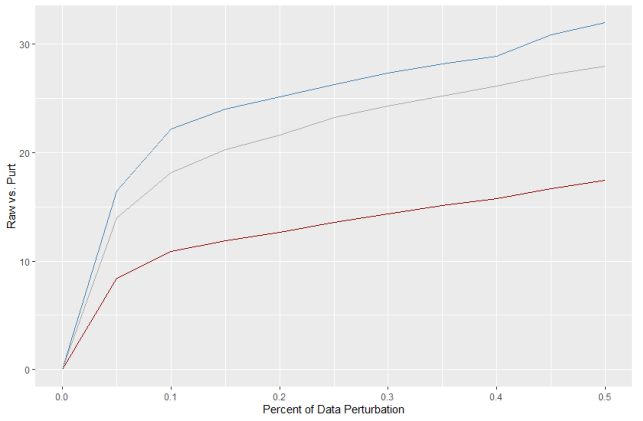



(d) Top 4 Euclidean Distance

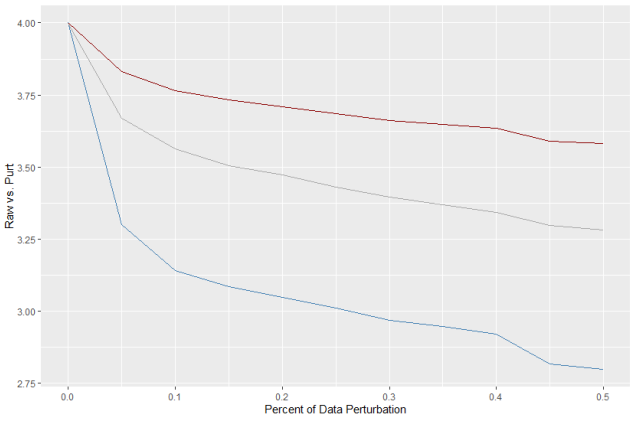
Figure 2: AAC Method Sensitivity (CC Model). Blue=MPL, Red=Shap, Grey=Mean



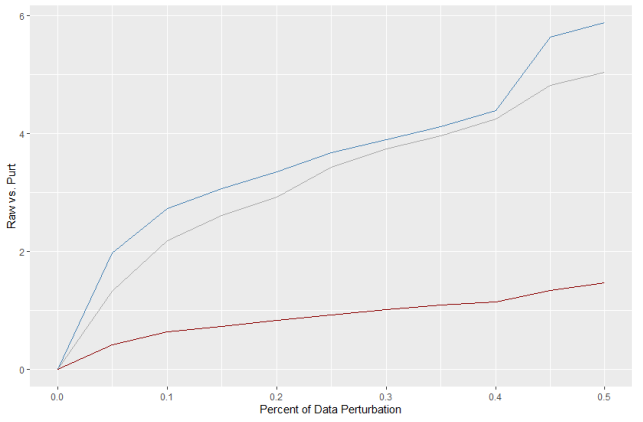

(a) Kendall Tau




(b) Euclidean Distance




(c) Top 4 in 4




(d) Top 4 Euclidean Distance

Figure 3: AAC Method Sensitivity (Mortgage Model). Blue=MPL, Red=Shap, Grey=Mean

# 5 Conclusion

We have investigated various AAC approaches and compared them and analyzed their results in several dimensions. We find, using Shapley as a ground truth due to its theoretical and axiomatic 'correctness that the alternative methodologies lead to clearly different AACs, and also that these differences are significant based on a synthetic placebo testing-derived pseudo p-value. Within this finding we find the Univariate approach to be the most dissimilar based on a variety of metrics, while the MPL and Mean methods are less dissimilar, and also comparable to each other due to similar methodologies between the two. We then find that for all of these methodologies, differences are most pronounced for the least risky customers on the border of the accept region, meaning the initial findings are most likely to affect those for whom AACs are most actionable. Lastly we find that Univariate method is invariate to small data perturbations that emphasize interaction effects, and so loses information relative to Shapley which is able to capture these ML-specific model features. On the other hand, MPL and Mean methodologies are much more unstable with respect to data perturbations than Shapley, indicating a potential lack of robustness of these methodologies due to their narrow univariate focus compared to Shapley.

# References


Abadie, Alberto, Alexis Diamond, and Jens Haimueller (2010), "Synthetic control methods for comparative case studies: Estimating the effect of california's tobacco control program." *Journal of the American Statistical Association*, 105, 493–505.

Abadie, Alberto, Alexis Diamond, and Jens Haimueller (2015), "Comparative politics and the synthetic control method." *American Journal of Political Science*, 59, 495–510.

FinRegLab (2022), "Machine learning explainability and fairness: Insights from consumer lending." *White Paper*.

Shapley, Lloyd S. (1953), "Stochastic games." *PNAS*, 39, 1095–1100.

Sirignano, Justin, Apaar Sadhwani, and Kay Giesecke (2018), "Deep learning for mortgage risk." *arXiv*.

Strumbelj and Kononenko (2014), "Explaining prediction models and individual predictions with feature contributions." *Knowledge and Information Systems*, 41, 647–665.

Sundararajan, Mukund, Ankur Taly, and Qiqi Yan (2017), "Axiomatic attribution for deep networks." *arXiv: Machine Learning*, arXiv:1703.01365.

# A Data and Models

In this paper, we analyze 2 XGBoost-based risk models, one of which being a retail credit card risk model, and the second being a mortgage risk model. Even before ML came into use, credit cards were in the vanguard in the replacement of manual underwriting with Automated Underwriting Systems (AUSs) due to the high volume of decisions necessary as well as the availability of large quantities of data. In the same vein, industry practitioners in credit card modeling expanded into machine learning relatively early on, due to its potential as one of the most fruitful areas of machine learning adoption. This is due to the availability of extremely large credit card datasets (both long and wide, with millions/billions of observations and thousands of variables) which could then be mined to extract the potentially complicated nonlinear and interacting elements of consumer card behavior. As ML adoption spreads throughout the industry, use cases for ML have also rapidly diversified, with ML models not only expanding into mortgage modeling, upon which we base our second model, but also in such diverse areas as fraud and marketing.[9]

Most credit risk models in the retail credit world are divided into underwriting/acquisitions-level scorecards, and account management models, with the primary differences being in the data and usage. Acquisitions-level models are primarily used for booking customers, and as they do not yet have internal account performance data, are usually restricted to using application-level predictors and credit bureau data. Account management/performance monitoring models on the other hand are estimated on customers that are already booked and also have seasoned account performance data, typically being used for things like line management, loan modification, and portfolio monitoring. This means that they can leverage observed customer behavior - indeed, the best guess of a customer's future behavior is in their past behavior. In fact, this behavior can exhibit nonlinear and highly interacting relationships and is potentially an ideal use case for machine learning. It is upon this consideration, as well as the richer data that can be used, that both of the models that we study are **account management/performance monitoring models.** It is important to note that the kind of data these models are estimated with is by necessity entirely consisting of booked loans since one cannot directly observe performance for loans that are rejected.[10] We now describe the details of model development and provide high-level information on the main risk drivers and shape of relationships in the two models.

For the credit card model, the data we use is a random sample of internal OCC supervisory data on credit card performance at major banks.[11] We choose our target variable to be the occurence of a 90 day delinquency of a customer within a year of the snapshot date, a typical performance window and target variable for a basic risk scorecard, and so our model is solving a binary classification problem. We select a snapshot of 150,000 general purpose cards with at least 12 months of performance from December 2013 and follow their performance for a year. As an out-of-time set upon which we estimate our reject reasons, we

[9]For an example of an application of Deep Learning to mortgage risk, see Sirignano et al. (2018).

[10]Several reject inference approaches can help correct this selection bias, but we generally do not have the data to conduct this. However, this issue is largely tangential to our analysis of AAC methods.

[11]This data is known as OCC Credit Card Metrics.

choose to take a snapshot of 150,000 general purpose cards, again with at least 12 months of performance. Note that there is some overlap in the accounts between the two sets, however, they are from non-overlapping times, so this should minimally affect the out-of-sample nature of these observations.

For the mortgage model, we use a similar approach, but leverage a GSE's publically available single family performance data for the analysis.[12] In this case, we choose our target variable to be 90 DPD within **two** years of the snapshot day, in deference to the typically slower dynamics of mortgage loans compared to credit cards. In addition, we also jointly estimate a prepay exit in a competing-risks style approach to take into account the possibility of the borrower refinancing or selling their home, so the model is a multiclass classification problem where we focus only on the PD side. Once again we take only loans with at least 12 months of performance and take a snapshot from January 2015, giving us 300,000 loans in the training sample. The OOT set for which reject reasons are calculated is from January 2017, also comprising a set of around 300,000 loans.

Both the datasets contain a typical rate of missing observations and outliers. We pre-process the data using missing value inference, assuming a missing-at-random (MAR) missing value structure, utilizing boosted trees to estimate missing variable $x_i$ given all other covariates $(x_n)_{n \neq i}$. Outliers are winsorized differentially on a variable-by-variable basis - more details are available upon request.

The functional form that we choose for both models is boosted trees, leveraging the open-source XGBoost algorithm. This is a widely adopted methodology in banking and other industries, and regularly is one of the top contenders in horse races between other methodologies, such as deep neural nets (deep learning), random forests, or SVNs. The disadvantage of this methodology is its non-smooth nature - indeed the relationship between covariates and predictions is by definition stepwise and commonly exhibits large discontinuous jumps from small movements in the covariate. These possibly unexpected jumps are a source of potential risk, and also make calculating some explainability statistics more difficult since there is no gradient for which to calulate them for. This is in contrast to, say, any kind of neural network approach which gives a smooth function as output. Due to the discontinuous nature of boosted trees, some practitioners will use them in an ensemble with neural networks. However, by and large banks are comfortable with using boosted trees in a standalone fashion. We follow this approach and use both standalone as well.

We estimate the credit card model with a joint hyperparameter and model selection routine, while in the mortgage model the already parsimonious number of features (34) only necessitated a hyperparameter search. In the credit card model, we start with a set of 472 covariates that mostly consist of trended customer internal performance statistics (such as balance payoff, utilization, past due amounts) and credit bureau attributes (such as external card utilization, age of oldest account, number of external tradeline delinquencies) along with some customer-reported characteristics such as borrower income. We utilize an early

[12] See `https://capitalmarkets.fanniemae.com/credit-risk-transfer/single-family-credit-risk-transfer/fannie-mae-single-family-loan-performance-data`

stopping function on 5-fold cross-validated AUC (area under the curve) to constrain the number of trees, and the other hyperparameters such as learning rate $\eta$, and depth of trees are selected using a simultaneous grid search that also selects for 5-fold cross-validated AUC on the training set. We set $\gamma$ to be .5 to penalize if a single variable is too influential to help mitigate overfitting. The approach for hyperparameter selection in the mortgage model is the same but we use an "mlogloss" target due to the multiclass classification problem that it is solving. For visualization of the searches, see figures (7) and (8).

For the remainder of our analysis, since we are analyzing reject reasons, we must construct a hypothetical rule for rejection. Setting any kind of accept/decline bound implies weighing the costs of accepting false positives (those we flag as risky potential defaulters who do not subsequently default) and rejecting false negatives (those we flag as low-risk customers that do subsequently default). Banks use multiple methods to set these bounds, and indeed oftentimes these bounds are not simple univariate rules but instead incorporate other risk scores or risk factors in a dual matrix framework. These bounds are set upon various profit and risk considerations and depend on the product strategy of the bank and their general risk appetite. For simplicity, we will keep to a simple univariate accept/decline bound on the model's predicted probability of default (PD). We choose our cut-off to be 0.1 PD, which approximately maximizes F-score for both models, although weights more towards prioritizing recall than precision - see figure (9) and (10) for visualization.

We can also give an approximation of the general relationships from features to predictors in what are known as Accumulated Local Effects, or ALE plots in figures (11) and (12). These attempt to understand the sensitivity of the prediction to a certain covariate $X_i$ in a realistic area of the feature space while at the same time holding the other $X_{-i}$ covariates constant, to avoid confounding the effects of correlated covariates changing with the actual effect of the covariate of interest to the prediction. This is accomplished by finding the effect of highly local perturbations of the $X_i$ covariate with the $X_{-i}$ covariates set at their mean conditional on $X_i$. We can also overlay these ALE plots with a sample of observations' Shapley contributions, which we will discuss in more depth in later sections. Conceptually, there is a parallel between ALE plots and Shapley values since they both calculate the effect of perturbing the variable of interest on the prediction. However, ALE plots explicitly assume the non-perturbed covariates are set at the conditional mean, while Shapley values $X_{-i}$ covariates are specific to the observation being examined, and are perturbed with respect to all possible combinations of the other covariates - see Section (2.1). In addition, the weighting between ALE and Shapley is slightly different as can be seen in the cases where data is relatively sparse in certain ranges of the data as in figure (12).

# B Figures and Tables


Figure 4: Credit Card Model: Joint variable selection and model estimation routine

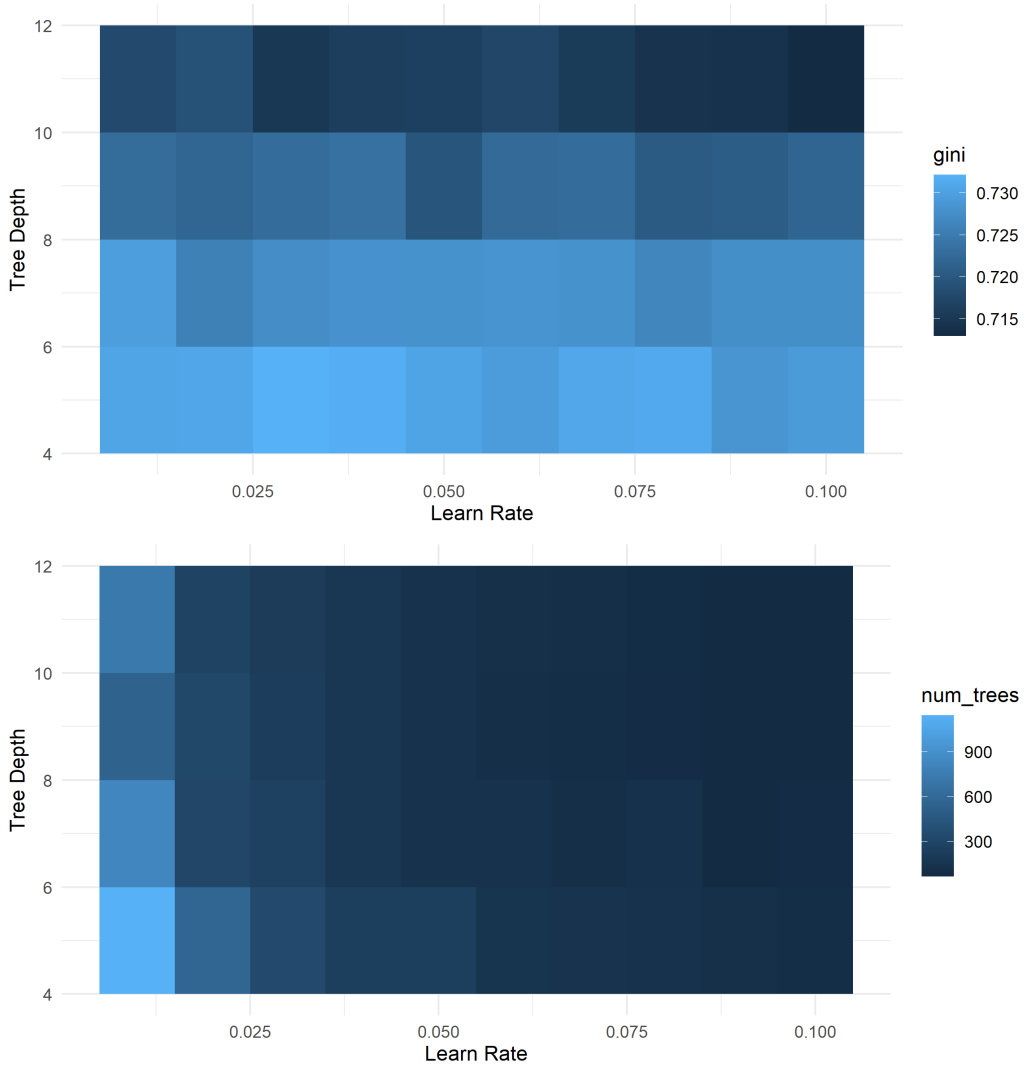



Figure 5: Credit Card Model: Hyperparameter search in final iteration

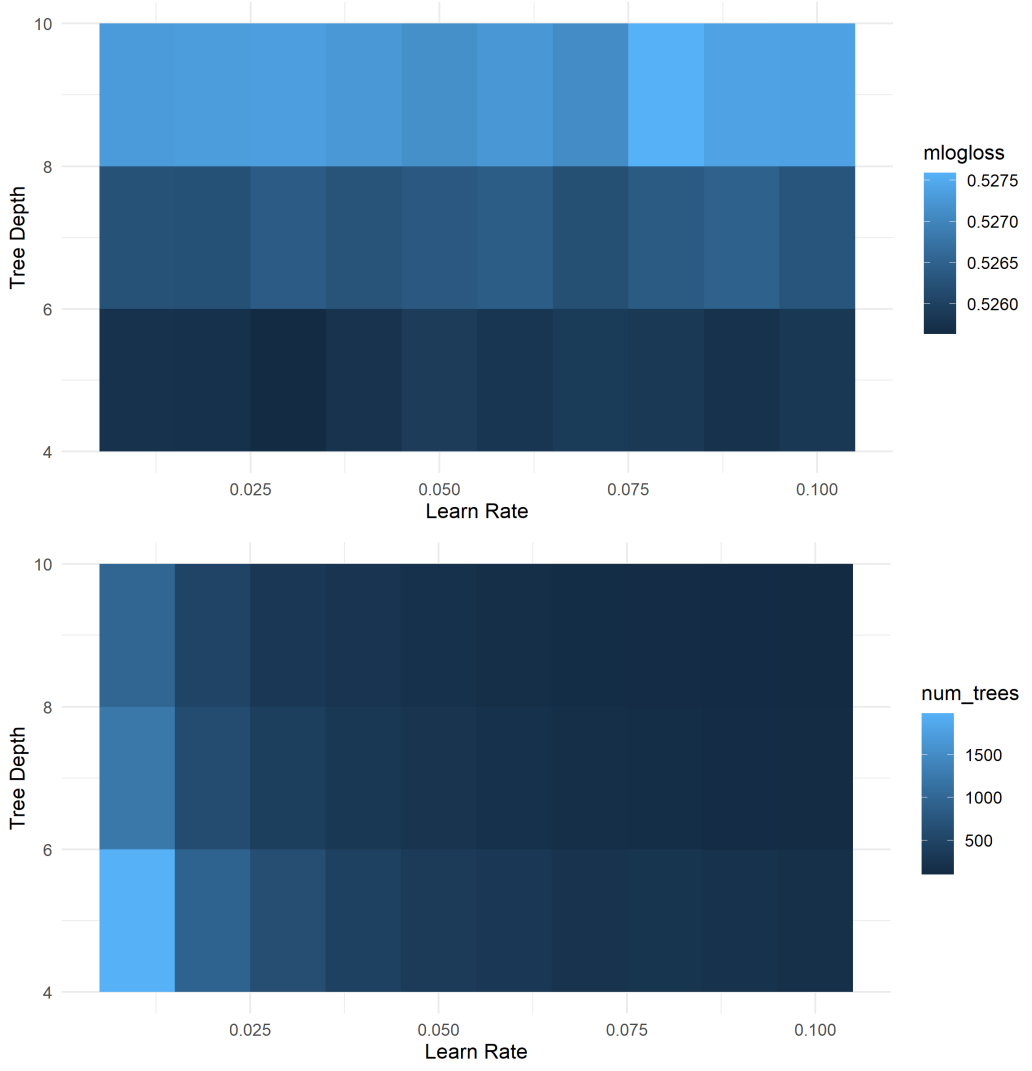



Figure 6: Mortgage Model: Hyperparameter search

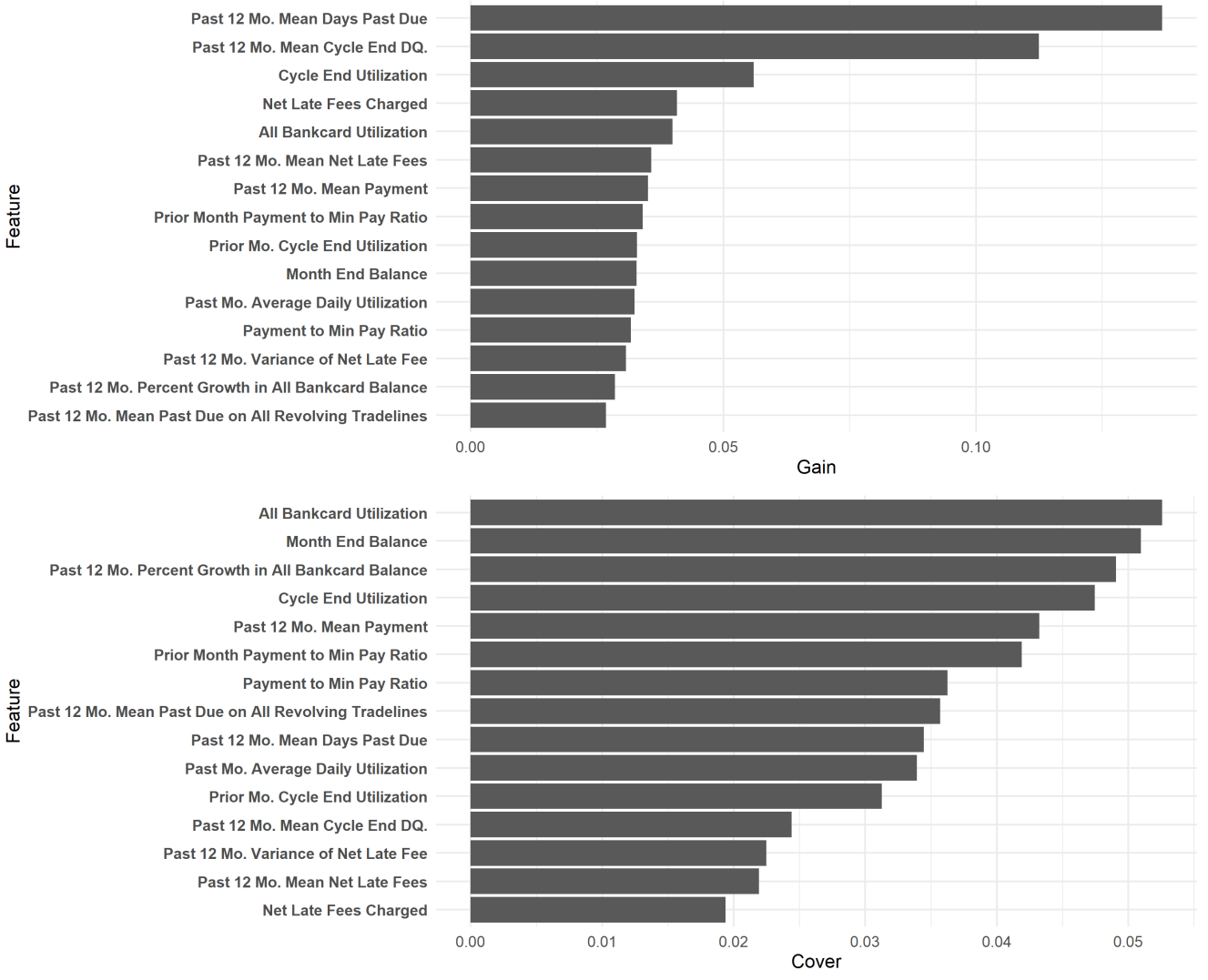



Figure 7: Credit Card Model: XGBoost Variable Importance Metrics

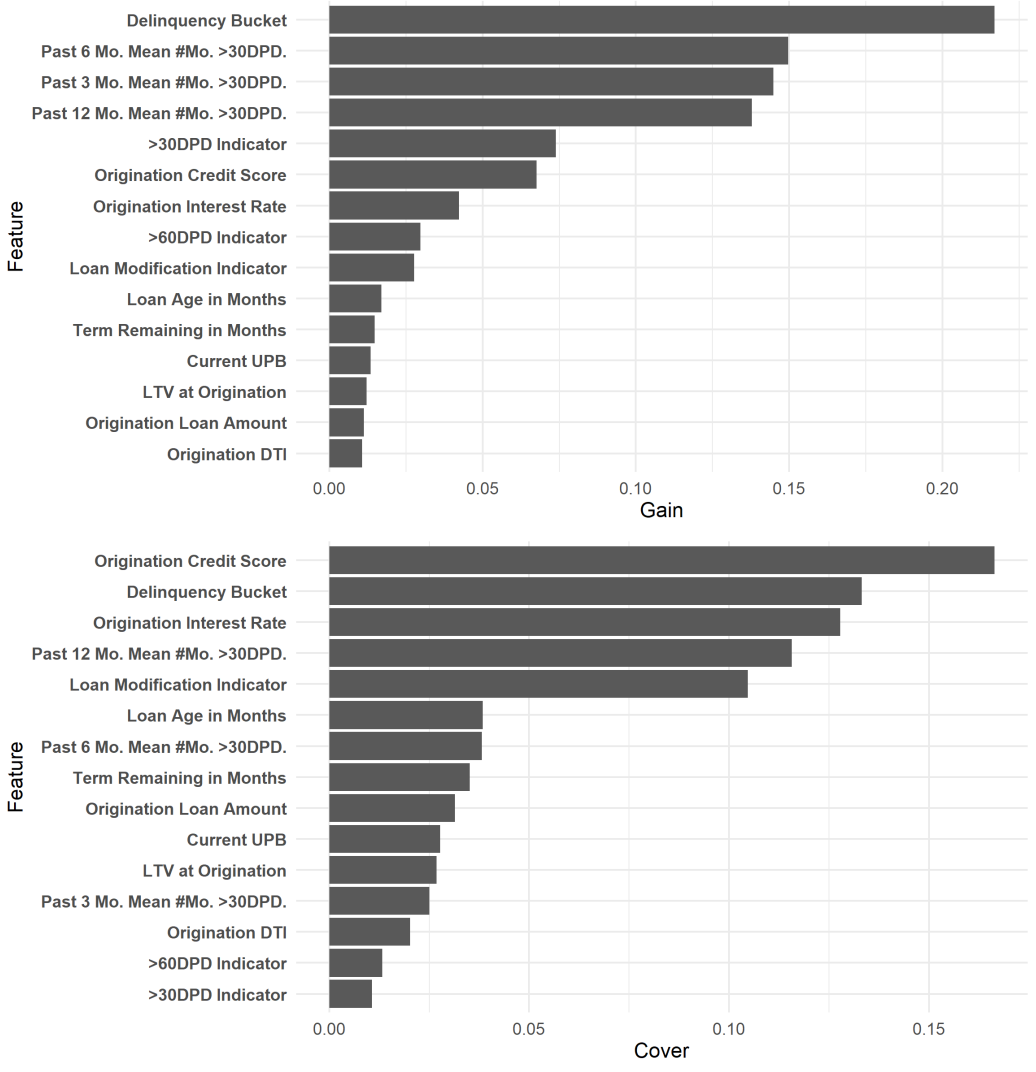



Figure 8: Mortgage Model: XGBoost Variable Importance Metrics

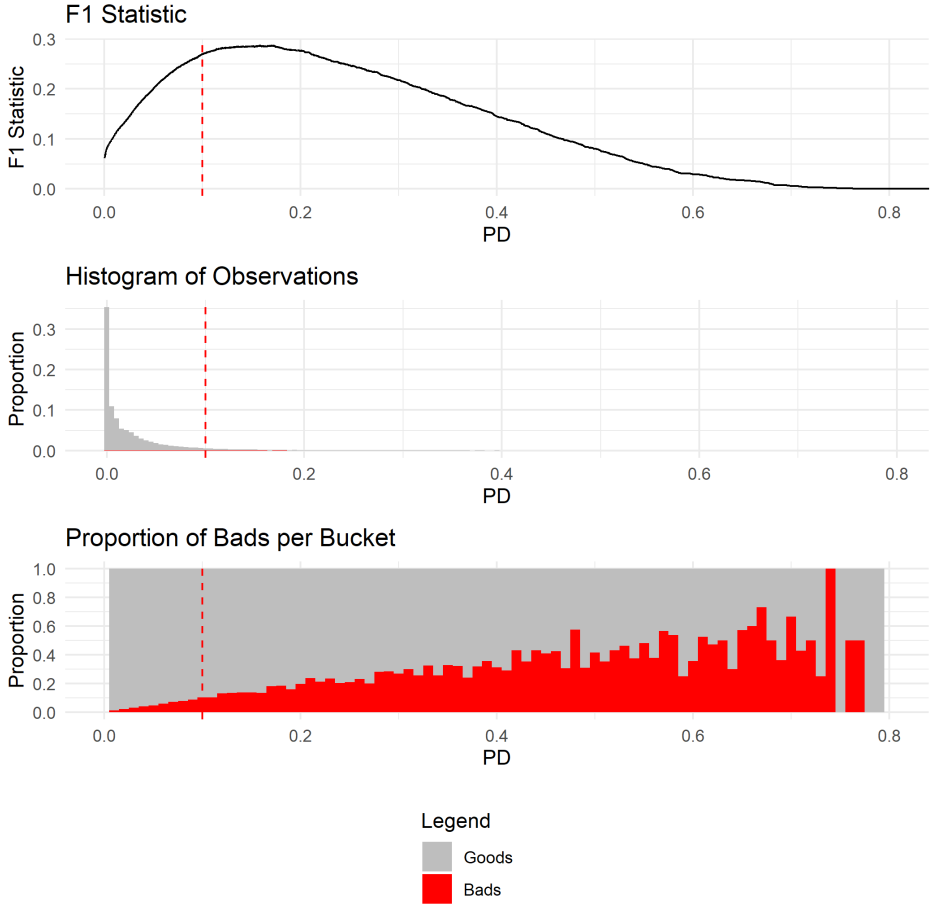



Figure 9: Credit Card Model: Performance statistics on OOT set. Cutoff set at .1

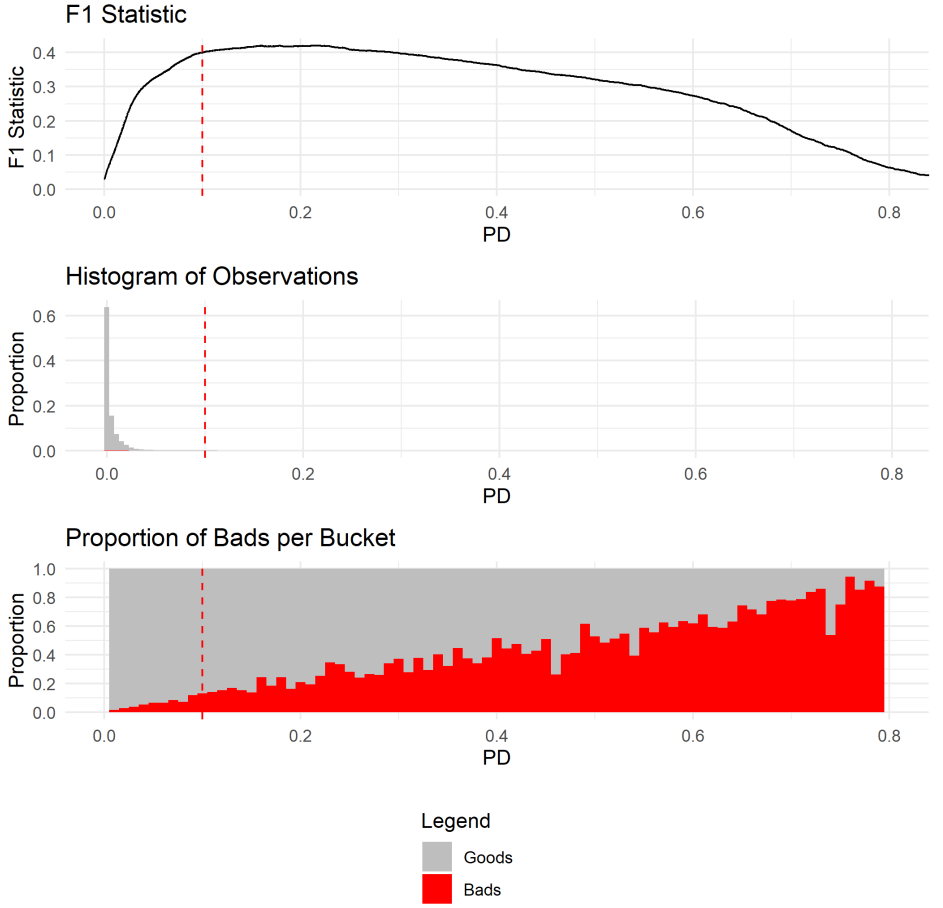



Figure 10: Mortgage Model: Performance statistics on OOT set. Cutoff set at .1

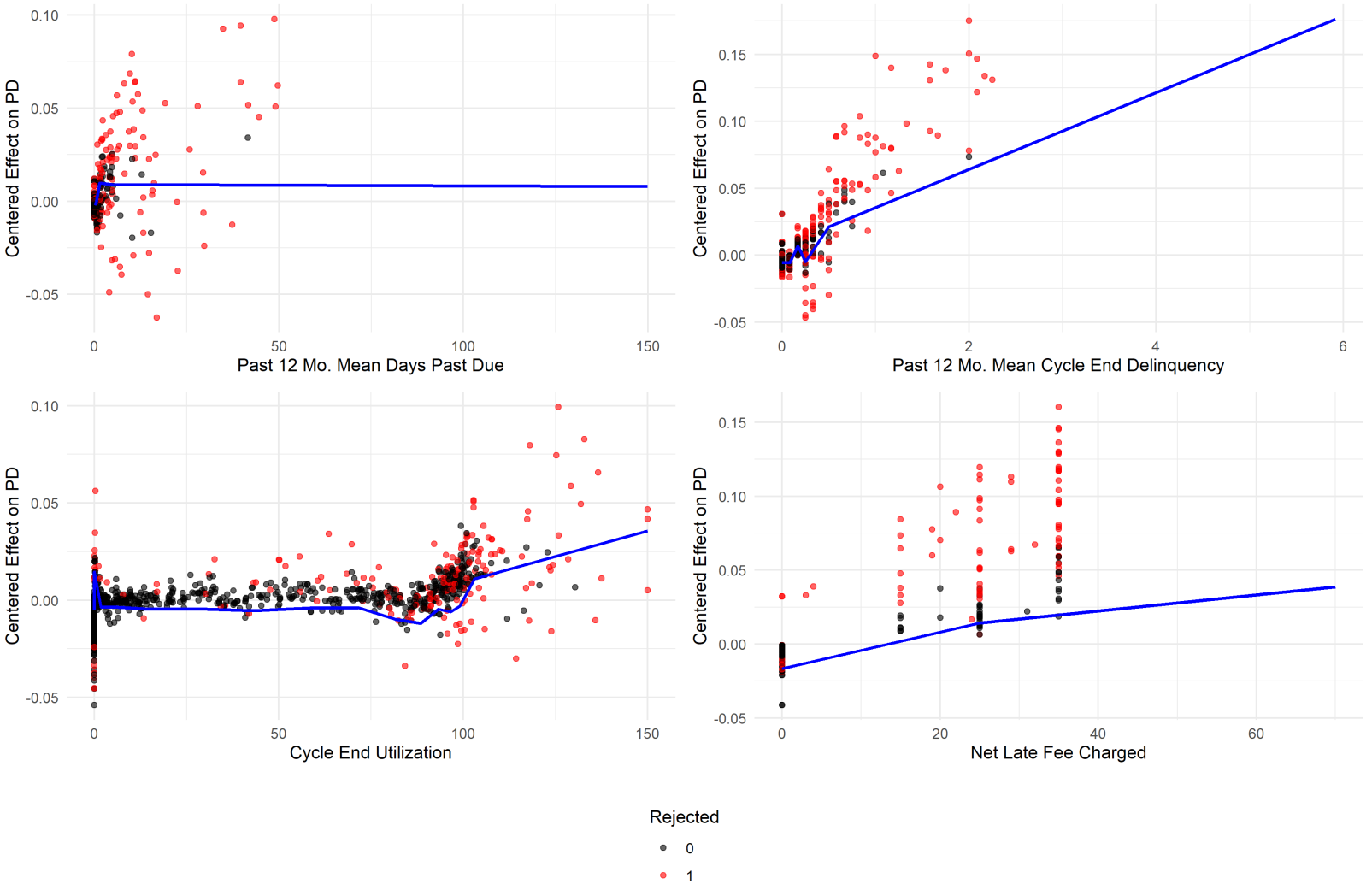



Figure 11: Credit Card Model ALE Plots for top 4 XGBoost "Gain" variables, overlaid with sample of Shapley values "trained" on training set. Rejected customers with PD> .1 highlighted.

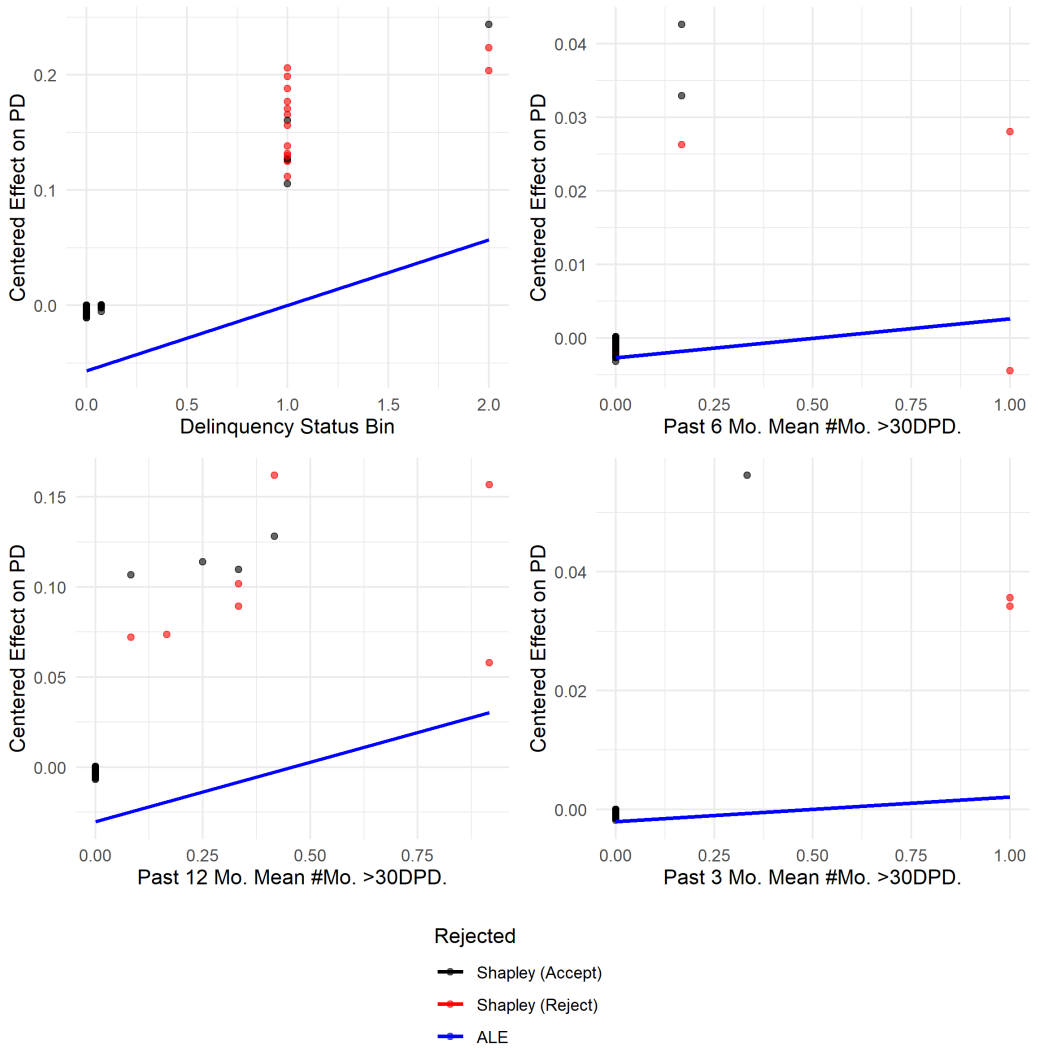



Figure 12: Mortgage Model ALE Plots for top 4 XGBoost "Gain" variables, overlaid with sample of Shapley values "trained" on training set. Rejected customers with PD> .1 highlighted.