Using Shapley Values to Enable Machine Learning-driven Adverse Action

February 2021

Joseph Hammond, Aaron McGuire, Xuetong Li, Vachagan Darbinyan

2nd Order solutions

Executive Summary

Currently, model explainability has been oft-cited as a barrier in more widespread use of Machine Learning (ML) in lending. There is a tension in using more sophisticated credit modeling techniques between risk splitting power and the resulting complexity. A majority of the research is being done to better understand the variable relationships within the ML models to try to mitigate the complexity and make them more transparent. In the past several years, Shapley values have risen in prominence in the ML community because they quantify the impact of independent variables on the model predictions for individual observations.

Model transparency is critical for a bank's customers and applicants, model governance teams, and its Compliance staff. Lenders are required to send out Adverse Action notices when declining loans or line increases, explaining to the customer the reasons for the decline. When those declines are driven by ML models, Shapley values can be a tool to increase explainability and reduce complexity.

In this paper, we lay out best practices we have developed to better leverage Shapley values in enabling ML model deployment in credit issuance. Shapley values can be a boon for lenders by better enabling the legally required Adverse Action notices. Through a variety of analyses and comparisons to existing Adverse Action methodology, we demonstrate how Shapley values can be used to satisfy the Adverse Action requirements and break down a common barrier that lenders face in adopting more sophisticated ML models.

Contents

Background

Adverse Actions Notifications

The Fair Credit Reporting Act (FCRA) requires communication to customers when a negative action is taken in regards to an application for credit. These actions include declining of a loan or credit card action, denials for credit line increase, raising the APR on a credit card, etc. These notifications serve two primary purposes. First, the customer must understand why the negative action was taken, and second to ensure that the decisions are not discriminating based on protected classes. These reasons need to be specific to the actual applicant. They cannot be broad generalizations, but must cite the actual variables which were pivotal in taking the negative action. Examples of properly written adverse action notifications include statements such as “balance on credit cards is too high” or “too many loans in delinquency”; they must be easy to understand and accurately represent the negative attribution the bank’s model has identified about the customer. Since most of the underwriting of loans has moved to statistical models to decide what actions to take, these models have to provide the necessary Adverse Actions. That requires a deep understanding of the model, both in terms of what variables are in the model and the underlining relationship those variables have with risk. There are four specific aspects the model needs to satisfy:

- The model must provide the reasons Adverse Action occurred. These are the variables in the model that determined the person was too risky; this means one must be able to isolate which variables have had the largest impact on the customer’s poor model score.
- The model must provide the source of the reasons by isolating the factors scored. The reasons need to be only those considered in the model; one cannot include a source not used in the model (e.g. if FICO is not used for decisioning, one cannot reference FICO for Adverse Action).
- The model must provide a specific description of the factor that caused the Adverse Action. This is the verbiage associated with the model’s findings; it’s the human-readable version of a model summary.
- The model must report the true top variables for the individual. Different individuals will have different reason since the underlining data varies from person to person.

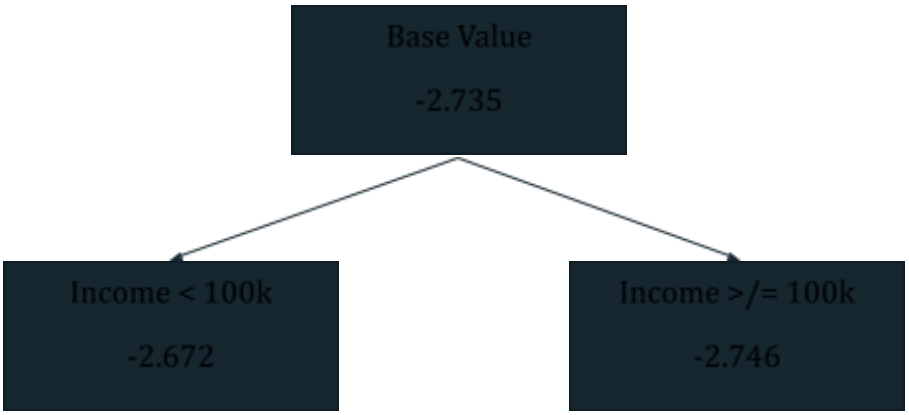
Some examples of Adverse Actions reasons might be “Too many recent inquiries” or “Utilization on credit cards is too high”. Lenders need to be able to parse out these factors on their model. In traditional logistic regression models where all the variables had a simple coefficient, getting at these was incredibly simple. It was the coefficient times the variable value. More complex tree-based models have improved risk prediction but have lend to more complex models. Obtaining precise model explanations is key to fulfilling the regulatory requirements. There is demand for more complex systems to explain what occurs within the models.

Shapley Values

The technical definition of Shapley values is the average marginal contribution of a feature value across all possible coalitions. There is a fair amount to unpack in that statement. It is important to note that the Shapley values are computed on the individual level per variable in the model. Let's start with what "all possible coalitions" means first. We can use a simple model as an example that has three variables –income > 100k, if the applicant is delinquent on any loan, and if there was an inquiry for credit in the past 3 months. All of these are binary indicators. To get the Shapley value for income, we need to get all other possible combinations of variables. In the case, it is the following

- No variables
- Account delinquent
- Inquiry in the last month
- Account delinquent and inquiry in the last month

The second part is the marginal contribution. For income >100k, we simply run the model on all coalitions with and without the income variable and subject the results. Here is even simpler example on a one variable model. The numbers are the logit predictions out of a binary classification problem.



Again, in this model there is a binary indicator for income >/=100k. To calculate the Shapley value, simply get all the coalitions and subtract the results. There is only one coalition here: no other features. The model output with income >/=100k is -2.746 and the model output without it is the base value of -2.735. Taking the difference of the two gives us to a -0.11 Shapley value. The negative means someone with an income >/= 100k is a lower risk than someone without. The calculations get more complex as one adds in more complex models, but the essence of the calculation is the same. Using tree-based models allows for exact Shapley calculations since there are finite number of outputs. Getting exact values is necessary for Adverse Action reasons.

There are three conditions that Shapley values satisfy. These are important to understand.

- **Local accuracy** - States the sum of the feature attributions is equal to the output of the function we are seeking to explain. Add up the individual Shapley values to get the prediction for an observation.
- **Missingness** - States features that are already missing are attributed no importance. There is no value from data that does not exist.
- **Consistency** - States that changing a model so a feature has a larger impact on the model will never decrease the attribution assigned to that feature. This is something the creator of Shapley value proved and is not the case for other explanatory tools like gain value.[1]

Analysis Results & Implications

Model Build Overview

To do the research if Shapley values can be used for Adverse Action, we built a model that mimics an acquisitions model for credit cards. Specifically, a monotonically constrained gradient boosted machine (GBM) was built. A GBM was used since the exact Shapley values can be computed on a tree-based model. The variables were monotonically constrained, which creates consistent relationships for Adverse Action reasons. For example, one would want the impact of number of inquiries to increase the likelihood of charge-off as that variable increases. This ties back to the verbiage on the Adverse Action letters. With a monotonic variable, language like "too many inquiries" can be used. Otherwise, there are instances in the model where "too few inquiries" could be the reason for decline. This explanation is completely counter-intuitive and additionally would drive customers to get more credit inquiries for approval. Adverse Action notices are typically thought of as what the customer needs to improve on their credit file.

Figure 1: Model parameters and performance

Parameters	
Learning rate	0.05
Number of Trees	300
Max Depth	7
Min Leaf Size	100
Model AUC	**0.738**

We used 350,000 observations for building and 150,000 observations for validation that had both bureau data and a performance target appended. The target variable was a binary 1/0 on whether or not the account charged-off. It was pared down to 25 final variables with hyperparameter tuning and arrived at the following model. As seen in Figure 1, the AUC performs well with the selected hyper-parameters.

One of the primary ways to understand the dynamics of tree-based model is through feature importance. There are several ways to calculate feature importance. Traditionally, methods like gain value, which is

[1] See the original paper on shapley values https://arxiv.org/pdf/1802.03888.pdf

the reduction in loss at the tree, splits for all variables. Shapley provides another method, which is simply the sum of absolute Shapley values.

***Figure 2**: Variable importance by GBM Gain value*

Variable	Contribution
Number trades never 30 or more days past due	15.6%
Number public record and trade line derogatories	14.2%
Months since oldest trade opened	11.3%
Percent of trades never delinquent	9.4%
Ratio balance to high credit	7.5%
Number inquiries	5.4%
Months since oldest bank revolving trade opened	4.9%
Number satisfactory trades 18 months or older	4.6%
Number trades 30 or more days past due in 6 months	4.2%
Number other finance trades	2.9%

***Figure 3**: Variable importance by Mean Absolute Shapley value*

Variable	Contribution
Number public record and trade line derogatories	9.3%
Months since oldest trade opened	8.0%
Ratio balance to high credit	6.9%
Number inquiries	6.6%
Number trades never 30 or more days past due	6.4%
Aggregate balance - all, excluding mortgage	5.8%
Percent of trades never delinquent	5.6%
Average balance of all trades	5.5%
Number other finance trades	5.2%
Months since oldest bank revolving trade opened	4.2%

Per the charts above, the top variables are different by the two methods with the gain value skewing to higher value on the most important features. Even though the model is exactly the same, the underlining explanation varies. The previously discussed consistency within the Shapley value is driving the difference. Gain values only look at individual components with the trees while Shapley values take into consideration the entire tree structure. This leads to consistent and more reliable feature contribution.

Comparison to Existing Methods

Another method typically used for model explainability is partial dependence plots (PDPs). PDPs are created by taking the data used to build the model and for an individual variable replacing that variable value in the entire data and scoring the model. The average of the scores is the impact. This is done through the range of values for the variable. These allow seeing the complete relationship of the variable in the model. Below are the PDPs for the top two variables in the model.

***Figure 4**: Partial Dependence plot for "Number of public record and trade line derogatories"*

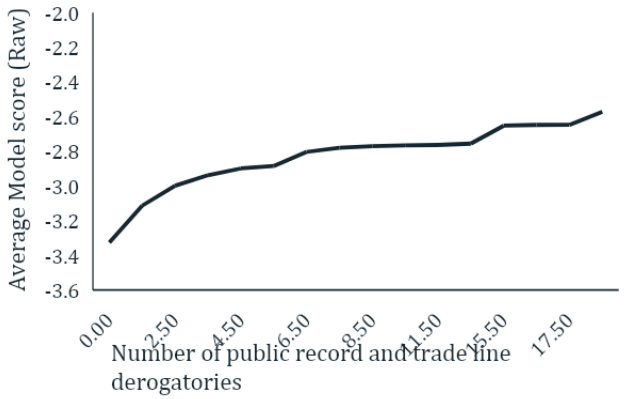
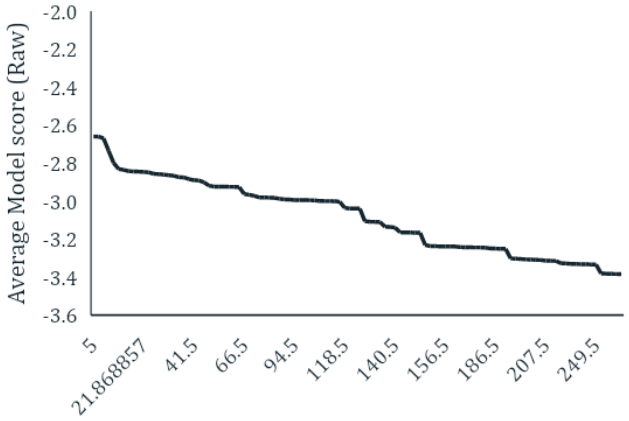


***Figure 5**: Partial Dependence plot for "Number of months since oldest trade opened"*



As demonstrated above, the relationship is monotonic and in the expected direction. To reiterate, monotonicity was enforced in the model. In fact, PDPs can be used to create Adverse Action reasons. This monotonic behavior provides a basis for the appropriate language while giving individual predictions.

Shapley dependence plots can also be created. These are slightly different than the PDPs with only the actual Shapley values being populated and averaged at the different points. There is no synthetic data created for these.

***Figure 6**: Shapley Dependence plot for "Number of public record and trade line derogatories"*

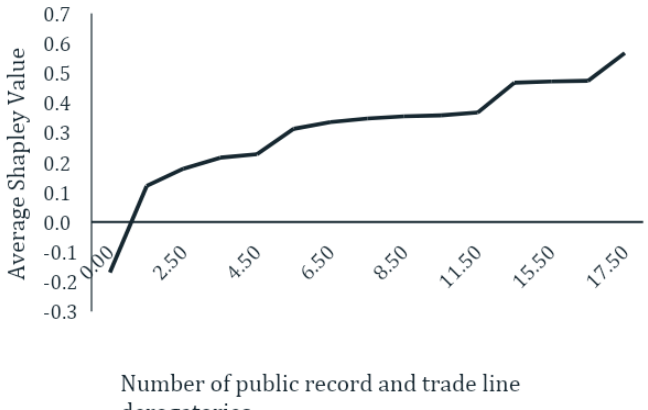
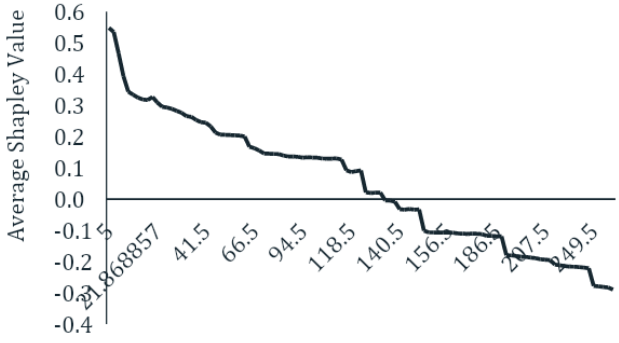


***Figure 7**: Shapley Dependence plot for "Number of months since oldest trade opened"*



There are two primary callouts:

- The two methods find similar relationships within the model. They are not exactly the same, but roughly the same shape.
- The monotonic relationship does not hold when looking at the Shapley dependence. This is driven by how the Shapley dependence is created. We are not looking at the effect on the model holding all other variables consistent and changing one value to understand the relationship.

Monotonic Relationships

We created several graphs that we are calling Shapley Individual Conditional Expectation (ICE) plots to find out if the Shapley values were monotonic when looking at the data in the same manner as PDPs. As we saw in the Shapley dependence plots, those relationships are not monotonic even though monotonicity was enforced in the model. This is due to the fact the Shapley dependence plot are not the correct way to analyze monotonicity. To correct this, we need to replicate the PDP methodology for Shapley values. Essentially moving one variable at a time and measuring the impact. These are the Shapley ICE plots.

ICE plots are PDPs with only one observation in the dataset. In fact, PDPs are the average of all ICE plots in a dataset. ICE plots are important with respect the Adverse Action because we need to show that all observations in the build sample are monotonic with respect to the individual variables. Below are sample of ICE plots and Shapley ICE plots for 5 observations on the top two variables.

***Figure 8**: ICE plot and Shapley ICE plot for "Number of public record and trade line derogatories"*

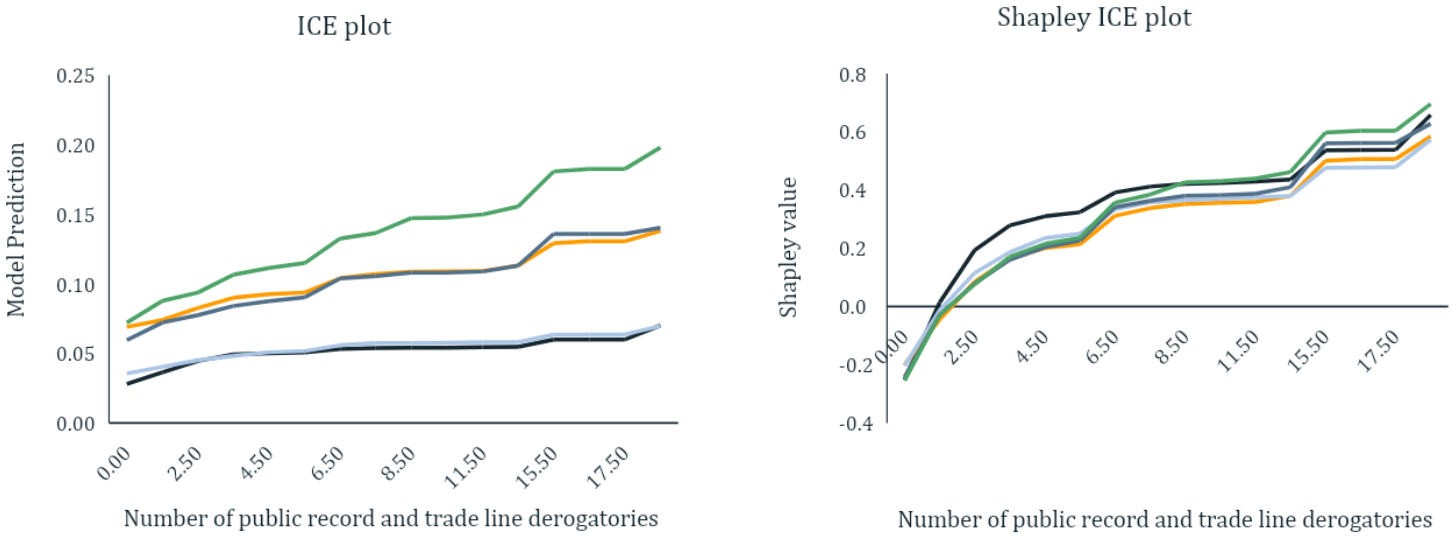


***Figure 9**: ICE plot and Shapley ICE plot for Number of months since oldest trade opened*

We checked all observations in the dataset and the Shapley values are monotonic for all observations. We want to call out these graphs are only useful to showing the monotonic behavior. The main criticism of PDPs and ICE plots is that it does not handle correlation among variables and can lead to unreasonable combinations of variables. Shapley values handle that correlation. We build these Shapley ICE plots to specifically ensure monotonicity for Adverse Action reasons.

Adverse Action Reasons

We performed an analysis using the existing PDP method for Adverse Action against using Shapley values. When giving Adverse Action notices, an issuer will often provide four reasons to the customer. These are the top four variables in the model that increased the prediction on the likelihood of charge-off. Figure 10 shows the number of times the Shapley value and PDP methodology agreed with the top reasons. 97.1% of the time three or four of the top four variables match across methodologies and zero times with no match.

***Figure 10**: Distribution of number of times the Shapley value and PDP methodology agreed with the top reasons.*

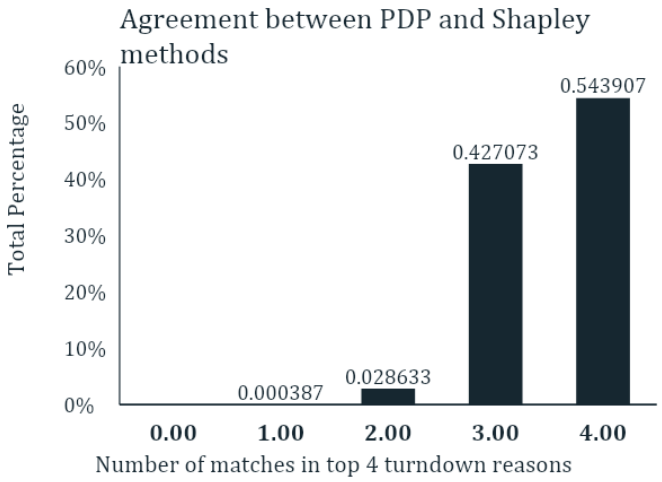
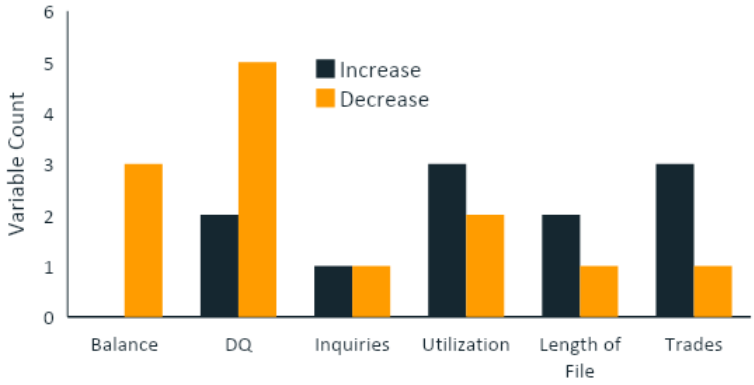


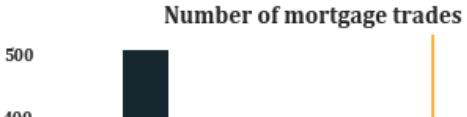
***Figure 11**: Comparison of shapley and PDP methodology. What type of variables increased or decreased relative to PDP*



Additionally, we can look at the type of variables that Shapley values prefer versus the PDP method. In Figure 11, we categorized the input variables and whether the variable increased or decreased in the top four reasons.

As shown, the Shapley values prefer utilization, length of file, and trade variables while balance and delinquency variables are preferred by PDP methods. The underlining reason

***Figure 12**: Single PDP value (orange) and Shapley value distribution of a subset of data.*



why this occurs has to do with the details around what data is used to make the calculations. Figure 12 shows the different data points in the model on a subset of data. The line shows all these observations will have same value for the PDP Adverse Action method while the Shapley value method has a distribution. Since all the variables have similar views this causes changes in the rank ordering of variable importance on the individual observations.

Conclusions

Model explainability is a large issue facing banks as they look to further harness the power of Machine Learning models. It's critical to explain decline reasons well to declined applicants, and well as unpack the key model variables for internal compliance and model governance teams. Model explainability is especially important for developing legally required Adverse Action notices. Shapley values, when combined with monotonic GBMs, can satisfy the requirements for the Adverse Action process. This is due to three reasons.

- The Shapley values are created on the individual basis showing the model impact on that specific customer.
- Impact on variable-by-variable basis is computed which allows finding the most impactful variables.
- Constraining a GBM with monotonicity also constrains Shapley values to monotonic behavior. This permits intuitive and useful customer communication on why the Adverse Action took place.

Both the PDP and Shapley value methodologies satisfy the Adverse Action requirement. Both can be used for model compliance. Shapley values have the benefit of a bit more granularity on the model impact, allowing lenders to comfortably employ more complex models and drive to better lending outcomes.