Paper SAS4290-2020

Model Governance for AI: Rethinking Our Approach to Validating and Using Black Box Models

Naeem Siddiqi, SAS Institute Inc

ABSTRACT

Model governance, including model validation, is a well-practiced discipline. The traditional approaches have served the industry well over the years. These typically involve activities such as backtesting, performing methodology and documentation reviews, benchmarking models, and performing qualitative assessments. A central part of this governance process is in *model explainability*, where knowing the exact variables and their weights is considered key. In the emerging world of artificial intelligence (AI) and machine learning models, it is often very difficult to understand model components, and traditional approaches might not be ideal. Such new models, known as *black box models*, require a re-think on how we approach model governance. This paper proposes a new framework for model governance that does not depend on knowing specific details about what is inside the model. Such an approach can enable institutions, particularly regulated financial ones, to be able to use black box models with confidence.

INTRODUCTION

The deployment of AI, particularly Machine Learning (ML), has been a hot topic in almost all business sectors in recent years. In the banking industry, one of the most common use cases cited for these are in the areas of credit risk and credit scoring. ML models are being used for issues such as segmentation analysis, feature engineering, benchmarking, income estimation, asset prediction, AML, KYC, and fraud. However, the actual operational usage of ML and AI in credit scoring for lending at present is not prevalent, at least not among banks that are regulated.

Some of the main challenges cited for using ML and AI models include:

- an inability to satisfactorily explain complex models, and how their outputs are generated (model explainability)
- governance, fairness, and transparency issues that cause perceived resistance from both internal and external parties around the usage of such models
- increasing customer concerns on data privacy, and being able to understand and challenge decisions made by banks

In order to deal with these issues, many organizations are undertaking research work to explain or interpret the more complex modeling algorithms. Complete model explainability has been a key part of model governance, and an important one, historically. However, with the increasing usage of more complex models in the risk domain, this policy, as well as other rules around governance of models, may need to be revisited. Recognizing that the modeling itself is only part of a much larger process, a more holistic approach for model explainability must be considered, one that takes into account all the steps in the model development cycle.

In fact, what is needed is a satisfactory approach to model governance that assumes that the model itself is not explainable − and then creates a process to get a high level of

confidence in such models. Where the traditional model governance approach requires complex models to fit into a process designed for simpler models, instead a new process is created that is designed to deal with complex models. This updated approach serves the general purpose of model governance validation, which is to attain a high confidence in the models.

This paper presents a high-level framework that can be used to allay some of the main concerns around model interpretability and governance.

MODEL GOVERNANCE CONSIDERATIONS

Let us begin by defining the process of model development across 3 broad areas (understanding that model development is a complex process that has many more tasks):

- data management: The process of acquiring data, performing exploratory work on it, data cleansing, variable transformation, and so on, that takes place before any model fitting is done.

- model fitting: The process of taking the above data and applying predictive modeling algorithms on it. This is the area where the usage of black box models and their interpretability causes the most issues.

- model output testing: The process of analyzing the predictions and other output of the model to make sure it is robust.

A comprehensive approach to model governance would cover all 3 areas to provide end users with increased confidence around their use of black box models. Let's consider each of these model development areas in sequence.

DATA MANAGEMENT

In cases where the model itself is very complex and cannot be explained, you can increase confidence in the model by understanding and validating, its input variables.

Data management validation includes data quality validation and causality validation.

Data Quality Validation

Data quality validation is a process wherein analysts ensure that the data:

- is clean
- is free from significant biases
- does not contain any variables that directly or indirectly can cause legal or ethical issues
- does not have dubious or counter intuitive causality

For example, development data sets sometimes contain variables such as "likes" or "follows" of various media outlets. These are obtained from social media accounts, and the use of these variables can cause problems. For example, in multicultural countries such as the UK, Canada, US, and Australia, immigrants may choose to like and follow ethnic media, or religious ones. If these variables enter a model, you may be making decisions based on race, religion, ethnic origin and language, which would contravene laws in many countries.

In addition, online data can also be biased as very often advertisers drive traffic using targeted segments that are based on gender, location, previous history, and other demographic factors. This can introduce bias into the samples being used in a given model. A simple way to test this can be to check for correlations with gender, race, ethnicity, and other segments of concern. If all the input variables are deemed clean and free from the issues cited above, a modeler can accept that all of them can be viable candidates for the

model. This should allay some concerns even if the content of the model is unknown.

It should be noted that removing some of these obvious problematic variables does not eliminate the issue altogether as there may still be other variables that can indirectly cause legal or ethical problems. (For example, the use of merchant codes for credit card transactions.)

Causality Validation

Causality validation involves performing various tests on the data to ensure that input variables have explainable relationships to output (target) variables. This can be done via variable ranking algorithms as well as Weight of Evidence based groupings. The latter is useful in that it shows the nature of the relationships, as well as the statistical correlations, which can lead to buy-in from business users.

This exercise is best done alongside business end users such as risk managers and model validation staff. The point of causality validation is to ensure that variables used can be explained in understandable business terms.

It is important to remember that algorithms in ML often employ complex interactions and relationships that cannot be captured by the simple bivariate analysis. However, the causality validation process can at least demonstrate that the input data set being used does not contain any data that has counter-intuitive or unexplainable relationships to the target, or is displaying significant sample bias. For example, known high risk segments that should demonstrate negative performance, but the data shows the opposite. If all the data used in the model is of reasonable strength and can be explained, this would again allay concerns around the usage of black box models.

Reviewing data quality and understanding some causal relationships should lead to an input data set that has been accepted by all stakeholders, including model development, model validation and business users. An analyst might not know which variables will end up in the final model, however, there would at least be some confidence that whatever has gone in is of good quality, is unbiased, is free of legal or ethical concerns, and has strong explainable relationships to the target being modeled.

MODEL FITTING

Model fitting is a step in which properly validated data is used to build models. Some modeling techniques (such as regression) are more explainable than others (such as neural networks). In cases where direct explanation of the model is not possible, there are various techniques available that provide proxies, including:

- Surrogate models: More transparent models (for example, points-based scorecards) are built to predict the *outcome* of black box ones. The variables in these scorecards are taken as proxies for what may be in the complex one. Note that the point of this exercise is not to build a model to predict the target variable but to mimic the black box model. The creation of these models can be done at a global or local level. This has an advantage of being easy to understand and explain. Another such technique is Local Interpretable Model Agnostic Explanations (LIME).
- Sensitivity analyses: Inputs into black box models are perturbed or changed and the impact is then measured in various ways, including changes to the prediction and error values. For example, the loan to value can be increased by 10% and the change in expected probability of default can be measured. This can provide not just an indication of whether the variable is in the model or not, but also its approximate weight. Common methods include Feature Importance based on variable perturbation, Partial Dependence plots, Individual Conditional Expectation plots, and

the use of Shapley Values.

All the techniques described above rely on approximations and have their own advantages and disadvantages. In particular, the use of simple models to explain a much more complex one can be problematic. However, in the absence of 100% certainty, analysts can use a combination of several methods to obtain some sense of the variables that may be in the model, and their relative contributions/weights. For example, many banks prefer using partial dependence plots as they are simple to interpret and explain to business users compared to the more complex methods.

In addition to looking at the variables, analysts are of course, expected to follow all the usual model governance around out-of-sample validations as well as reviewing model fit statistics.

MODEL OUTPUT TESTING

After checking the data and model, the third part of the process is to look at the output and perform some simple tests on it. The results of these tests are meant to give a sense of comfort around the robustness of the black box model developed. Some of these tests are standard practice for most models developed in banks. Output testing should include the following processes.

Benchmarking: The process compares the results from the black box models with those from simpler, more explainable ones. These can include:

- comparisons of model fit statistics such as Area Under the ROC Curve, KS, Gini, AR, and so on
- distribution of cases or populations by predicted probabilities or score bands
- measures of false and true positives/negatives depending on the underlying business case for the model

Many banks also produce comparisons of expected and actual results by time. This type of benchmarking can help gauge the stability of the models. For example, if the ML model over- or under-predicts the actual by a larger margin during cycles compared to traditional models, this may be an issue that requires further examination.

Stability: This process involves checking the overall population stability of both types of models across time (the Population Stability Index (PSI) is a common measure for this). ML models should not display significantly greater instability compared to simpler ones. This issue is particularly important in the case of credit models that are expected to be in use for at least 12 months, and often for 2 to 3 years in the case of business models. For example, if the PSI shows greater swings than linear models for the same population changes, the ML models may be unstable. You can also perform stability checks for model inputs or variables. In this case, where exact model variables are unknown, the variables identified by methods such as sensitivity analysis or partial dependence can be used as proxies.

Backtesting: This process involves checking the performance of the black box and benchmark models on historical data, as well as the stability of the predicted values against their actual for the same set of historical data. This backtesting process can be extended to include impacts on capital/approval rates or other relevant measures over time. For example, in credit risk it is important to ensure that the model does not result in large swings in such measures. If the results show that ML models have a greater variance from actuals over time, the forecasts may be unstable and erratic. This is in addition to the 'out of sample' validation done as part of the modeling exercise.

Compliance: This process involves reviewing the regulatory compliance of black box models. In some countries, such as the US, there are regulations prohibiting discrimination

in lending on the basis of factors such as gender, ethnicity, religion, and others. The final model can be tested for such compliance using existing processes to make sure that there is no negative impact.

The above list of model output testing activities is not meant to be exhaustive. Institutions should include other analytics that would serve to give them confidence in the performance of the models.

CONCLUSION

The traditional approaches to model governance have served us well over the past decades. However, applying the same exact rules to AI and ML models may not be appropriate, as the nature of these models are quite different. This requires a change in perspective. Instead of asking the ML models to fit into existing requirements, those involved in the governance and review of models must create a new paradigm and new processes that enable their organizations to benefit from the power of these algorithms.

The approach outlined here can help to create confidence in cases where the model algorithm is a black box and cannot be explained. It bypasses the question of "what is in the model" by adding additional validation and by carefully examining all the processes surrounding the model, both before and after the actual model fitting. In using this approach, it should be possible to therefore use such models with greater confidence.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Naeem Siddiqi
SAS Institute
Naeem.siddiqi@sas.com