

February 4, 2022

Kris DeFrain, FCAS, MAAA, CPCU

Director, Research and Actuarial Services

National Association of Insurance Commissioners Central Office

**Re: Best Practices for Regulatory Review of Random Forests**

Dear Ms. DeFrain,

Thank you for the opportunity to comment on changes to the white paper “Regulatory Review of Predictive Models” that will address Random Forest models.

Below are suggested revisions to the white paper. Revisions are shown in blue font.

**Section A.3.d**

|  |  |  |  |
| --- | --- | --- | --- |
| **Section** | **Information Element** | **Level of Importance** | **Comments** |
| A.3.d | Determine how missing data was handled. | 1 | This is most relevant for variables that have been “scrubbed” or adjusted. The regulator should be aware of assumptions the modeler made in handling missing, null, or “not available” values in the data.  For example, it would be helpful to the reviewer if the modeler were to provide a statement as to whether there is any systemic reason for missing data. If adjustments or recoding of values were made, they should be explained. It may also be useful to the regulator if the percentage of exposures and premium for missing information from the model data are provided. This data can be displayed in either graphical or tabular formats.  The modeler should describe the way the tree fitting process handled missing values. The modeler should specify if missing values are treated before running the tree model or if they are allowed to be handled by the tree model.  When creating predictions on new datasets (such as hold out datasets), tree-based models may have different approaches for handling missing data or categorical levels not encountered in the training data for a predictor variable. The modeler should specify the process utilized when this occurs. |

**Comments**: We suggest revising section A.3.d to expand the commentary on situations where the handling of missing value may be relevant.

**Section B.1.h**

|  |  |  |  |
| --- | --- | --- | --- |
| **Section** | **Information Element** | **Level of Importance** | **Comments** |
| B.1.h | In conjunction with variable selection, obtain a  narrative on how the company determined the  granularity of the rating variables during  model development. | 3 | The narrative should include discussion of how credibility was considered in the process of determining the level of granularity of the  variables selected.  ~~Minimum data volume constraints can be applied to a tree based model, such that the trees will not create a split that would result in terminal nodes with volume below a set amount. The modeler should comment on how the threshold was chosen.~~ |
| ~~B.1.j~~ | ~~If adjustments to the model were made based on~~  ~~credibility considerations, obtain an explanation of~~  ~~the credibility considerations and how the~~  ~~adjustments were applied.~~ | ~~2~~ | ~~If there was no minimum data volume threshold applied to the trees, or if the threshold was very small, obtain an explanation of any post modeling adjustments the modeler made to address the~~  ~~credibility considerations and how the adjustments were applied.~~ |
| New B.3.4 | Obtain parameters that determined the volume of data in each tree node and a narrative of how parameters were chosen. | 1 | Minimum data volume constraints can be applied to a tree-based model, such that the trees will not create a split that would result in terminal nodes with volume below a set amount. The modeler should comment on how the threshold was chosen.  If there was no minimum data volume threshold applied to the trees, or if the threshold was exceedingly small, obtain an explanation of any post modeling adjustments the modeler made to address the credibility considerations and how the adjustments were applied. |
| C.4.a | Determine what, if any, consideration was given to the credibility of the output data. | 2 | The regulator should determine at what level of granularity credibility is applied. If modeling was by coverage, by-form, or by-peril, the company should explain how these were handled when there was not enough credible data by coverage, form, or peril to model. ~~The company should comment on the minimum data volume requirement at each node before splitting.~~ |

**Comments**:

We recommend that the comment on “minimum data volume” be removed from **B.1.h** and create a new section **B.3.4** requesting that the minimum data volume in a leaf be provided as a level 1 request due to the basic nature of this information.

We also recommend removing section **B.1.j** and adding the commentary of **B.1.j** to the newly created **B.3.4** section on the minimum data volume hyperparameter.

**C.4.a** contains discussion of data volume and we believe commentary on data volume can be removed from **C.4.a** since it is discussed in other sections.

**Section B.3.a**

|  |  |  |  |
| --- | --- | --- | --- |
| **Section** | **Information Element** | **Level of Importance** | **Comments** |
| B.3.a | Obtain a complete data dictionary, including the  names, types, definitions, and rationales for each  variable. | 1 | Types of variables might be continuous, discrete, Boolean, etc.  ~~Tree based models do not have offset or control variables, as all variables are treated the same way in the trees.~~  Identify any variable used as an offset or control in the random forest model and the offset factor that was applied for each level of the offset variable.  For any variable(s) intended to function as a control or offset, obtain an explanation of its purpose and impact. Also, for any use of  interaction between variables, obtain an explanation of its rationale and impact. |

**Comments**:

Offsets and control variables can apply to tree-based models. Offsets may be applied to the starting prediction for a given record before a tree-based model is built. With proper treatment, control variables can also exist in tree-based models if there are variables used in the model building process that are not part of the final rating plan.

**Section B.3.d**

|  |  |  |  |
| --- | --- | --- | --- |
| **Section** | **Information Element** | **Level of Importance** | **Comments** |
| B.3.d | Obtain plots describing the relationship between each predictor variable and the target variable.  Obtain a rational explanation for why an increase in the observed relationship between each predictor  variable should increase or decrease and the target variable (frequency, severity, loss costs, expenses,  or any element or characteristic being predicted). | 1 | Partial dependence plots, accumulated local effects plots, or shapley plots will help improve model interpretability. The plots should be  accompanied by commentary on why the visualized relationship ~~is~~ may be reasonable for variables of concern. Considering possible causation may be relevant, but proving causation is neither practical nor expected. If no rational explanation can be provided, greater scrutiny may be appropriate.  For example, the regulator should look for unfamiliar predictor variables and, if found, the regulator should seek to understand the connection relationship that variable has to increasing or decreasing the target variable.  The regulator should also consider that interpretability plots for tree- based models need to be reviewed with other considerations in mind. For example, partial dependence calculations assume independence with other variables in the model. |

**Comments**:

We suggest removing the request that every plot be accompanied by commentary in the initial filing. Given that a loss model is built for each loss type and each model will contain tens of variables, this would require commentary on several hundred plots. We believe that asking for commentary on plots related to variables of concern would be more appropriate.

We also suggest adding some commentary to illustrate that each type of interpretability plot is imperfect and no plot should be completely relied upon.

**Section B.3.f**

|  |  |  |  |
| --- | --- | --- | --- |
| **Section** | **Information Element** | **Level of Importance** | **Comments** |
| B.3.f | Obtain variable importance plots. Obtain a  description of how variable importance was  calculated. | 1 | Variable Importance Plots for tree-based methods highlight  which variables contributed most to the model. There are  multiple ways to calculate variable importance, and variable importance can be used to create an intuitive understanding of model operation. ~~Variables with the lowest importance measures should be prioritized when reviewing predictor variables for significance.~~  Credibility can be addressed through proper hyperparameter selection. Variables with highest importance should be prioritized when reviewing the model for appropriateness. Variables with lower importance may be evaluated objectively through their correlation with other variables in the model, and subjectively through how they may be interacting with other variables in the model to identify a subset of risks. |

**Comments**:

We recommend changing the focus on low importance variables. In tree-based models, carefully selected hyperparameters should prevent the model from splitting erroneously on non-credible variables. For example, the minimum leaf count or minimum improvement threshold for a split should demonstrate sufficient credibility for the segment being identified. This commentary may erroneously guide a model reviewer to request the removal of a variable that demonstrates material signal through complex interactions with other variables.

**Section B.4.e**

|  |  |  |  |
| --- | --- | --- | --- |
| **Section** | **Information Element** | **Level of Importance** | **Comments** |
| B.4.e | Obtain evidence that the model fits the training data well and for the overall model. | 2 | The regulator should ask for the company to provide exhibits or plots that show the fitted average makes sense when compared to the observed average for variables of interest. Regulators would ideally review this comparison for every variable, but time constraints may limit the focus to just variables of interest. ~~Variables of interest should include those with a low importance measure according to~~  ~~diagnostic tests, variables without an intuitive relationship to loss, or variables that may be a proxy for a protected class attribute.~~ Variables of interest should include those with a high or medium importance measure according to diagnostic tests. Variables with low importance or without an intuitive relationship to loss may be evaluated objectively through their correlation with other variables in the model, and subjectively through how they may be interacting with other variables in the model to identify a subset of risks. |

**Comments**:

We recommend changing the focus from variables with low importance to variables of high and medium importance. Variables of low importance provide low predictive power to the model and are therefore only mildly affecting any segment. Variables with low importance may only be meaningful in the tails of their distribution, and goodness of fit for most of the variable’s range may be immaterial. Variables with lower importance may be better evaluated objectively through their correlation with other variables in the model, and subjectively through how they may be interacting with other variables in the model to identify a subset of risks.

**Section B.4.h**

|  |  |  |  |
| --- | --- | --- | --- |
| **Section** | **Information Element** | **Level of Importance** | **Comments** |
| B.4.h | Obtain a narrative on how potential concerns with  overfitting were addressed. | 2 | Tree based models such as Random Forest models are notorious for over-fitting. The company should provide a narrative on how overfitting was addressed. The company should provide lift charts on ~~training data and~~ testing data that is separate from the training data. |
| New B.3.5 | Obtain a narrative of the process to select all hyperparameters for the Random Forest. Detail how this process addressed potential overfitting in the model | 2 | Hyperparameter tuning can be done in a variety of ways. The rigor of the tuning process should reflect the risk of overfitting on the specific dataset. |

**Comments**:

Overfitting in tree-based models should be addressed through the hyperparameter selection process. We recommend that B.4.h request only a one-way lift chart against holdout data to demonstrate that the model is not overfit, and that an additional section B.3.5 be added to request a narrative of the hyperparameter tuning process and how this has addressed overfitting.

**Section B.4.j**

|  |  |  |  |
| --- | --- | --- | --- |
| **Section** | **Information Element** | **Level of Importance** | **Comments** |
| B.4.j | Obtain 5-10 sample records with corresponding  output from the model for those records. | ~~2~~ 4 | The company should provide 5-10 sample records with corresponding input variable values, the prediction from each  component tree in the model, and the final ensemble model  prediction. The company should describe how the final model  prediction aggregates the individual tree model predictions. |

**Comments**:

We suggest changing the level of importance of this item to level 4. A narrative describing how the predictions of each tree are combined for a final model prediction is essential information and is already requested in section **B.2.e**. However, we do not feel the request for sample records and sample calculations is part of a normal model review unless there are concerns about the model.

**Section C.10.d**

|  |  |  |  |
| --- | --- | --- | --- |
| **Section** | **Information Element** | **Level of Importance** | **Comments** |
| C.10.d | Obtain complete documentation of all component  trees and how the individual predictions are  aggregated together into a final prediction. | ~~1~~ 4 | The company should provide either tree diagrams for each component tree or comprehensive if-else statements that would replicate the logic of the trees. The company should state how the individual component tree predictions are combined into a final  prediction. |

**Comments**:

We suggest changing the level of importance of this item to level 4. We do not feel the request for each tree is part of a regular model review unless there are concerns about the model. Tree diagrams are also of limited usefulness by a reviewer since there are generally hundreds of trees in one model and it is unlikely that a reviewer would have the time to review each tree.

**General Comments**

We recommend additional commentary to clarify that this guidance is for Random Forests and may not be appropriate for all other types of tree-based models. For example, while examining individual trees for a Random Forest may be enlightening, examining individual trees for a Gradient Boosting Machine (GBM) is much less intuitive as each tree’s interpretation depends on the results of the possibly substantial number of trees before it. Other such differences exist and are out of scope for this commentary. We also recommend that the committee provide additional guidance specific to other types of tree-based models such as GBMs to avoid potential misapplication of this guidance on other tree-based models.

Once again, thank you for the opportunity to comment.

**Allstate Property & Casualty Actuarial Modeling Department**

For any questions, please contact:

Mike Woods, FCAS, CSPA, MAAA

Allstate Insurance Company

[mike.woods@allstate.com](mailto:mike.woods@allstate.com)