California Department of Insurance

Non-Catastrophe Model Form

For Predictive and Scoring Models

Rate Application – Model Disclosure Page #: _____ (e.g. 12.3)

The purpose of this form is to provide a complete set of basic information about all models employed in the development of the filed program. For the purposes of this form, a "model" should be understood in terms of its common meaning, and includes:

- Statistical modeling techniques, which relate input data to quantities of interest by making assumptions about the probability distribution of output data. GLMs are an example of these techniques.

- Machine learning techniques, which relate input data to quantities or labels of interest, typically by optimizing for some loss function or energy function of the input data. Examples include k-means clustering, neural networks, or gradient boosting.

The questions below apply to all predictive and scoring (or, "non-catastrophe")[1] models, including, but not limited to, rating models, and territorial models, . To the extent that a question is worded in a way that does not "fit" one of the models used in the filing, please nevertheless use appropriate judgment to provide the relevant technical details regarding that model.

Submit one model form for each non-catastrophe model listed on Page 12 of the Rate Application - Model Disclosure, regardless of model submission status (New, Revised, Refreshed or No Change). If no change is being proposed to the model, the model form associated with that model from a previous approved filing may be attached. Please identify the CDI file number of that previous approved filing.

[1] For purposes of this form, a scoring model for a catastrophe peril is considered a "non-catastrophe" model.

If multiple models are combined into one form, please explain the rationale. If sub-models are included in the same form as the main model, provide a description of how that sub-model is incorporated into the main model.

In the event that sub-models or a group of very similar models are included in one form, please provide responses in a logical ordering. For instance, if a frequency and a severity model are used, and one answer addresses the frequency model first followed by the severity model, please ensure that all other answers address the models in the same order. Alternatively, please group together all responses for the frequency model separately from all responses for the severity model.

1. General

1. Please provide a description of the purpose of the model. Identify how this model relates to other models in the filing.

2. Please provide a general description of the model's output, how the output will be used, and an example of the output.

3. Identify all methods of regularization employed in the model (e.g., prior distributions / Bayesian treatment, ridge regression, lasso regression, elastic net, boosting, etc.).

4. Identify major model assumptions (e.g., link functions or error distributions) and major model parameters or hyper-parameters (e.g., learning rates, regression penalties, number of trees, max tree-depth, number of clusters, Tweedie power parameters, etc.). For each major model parameter, explain how the parameter was determined.

5. Identify all software used for model development.

6. Identify any major sensitivities and dependencies within the model.

7. Is this model replacing a different model that was previously used for a similar purpose? For example, one vendor's (scoring, etc.) model is replacing a previously used model from another vendor. If so, please explain why this change is being made.

8. If this model revision reflects a change to a model submitted previously in a filing approved by the California Department of Insurance, provide an overview of what has changed in the model since the prior filing, including any changes to data, methodology, modeling assumptions, and output.

For a third-party vendor model, please answer the following questions:

9. Provide the names and version number as well as the vendor that produced model.

10. Is there a more recent version of this model available from the vendor? If so, why is the more recent version not being used?

2. Data

This form distinguishes between “input data” and “training data”. “Input data” is meant to refer to data entered into the fully trained model to produce outputs such as final rates, AALs, or scores. “Training data” is meant to refer to data used to initially develop and fit the model.

For training data, please answer the question(s) below:

11. Please identify the periods of time, geographic areas, lines of business, program segments and companies included in the data underlying the model. If the latest policy period is more than two years old, explain why more recent data was not used.

12. Provide a summary of data underlying the model by policy/accident year, policy form/program segment, and peril/coverage, and include the following (if the model is based on countrywide data, provide the information separately for California and countrywide excluding California):

    a. Exposure

    b. Premium

    c. Claim count

    d. Losses

    e. DCCE

    Provide a description for each of the above items (e.g., definition of exposure, closed vs. all claims, incurred vs. paid losses, manual vs. modified premium), specify any adjustments made to the raw data, and include a report reconciling raw data to the modeling data.

13. Provide a description/definition of catastrophe claims, explain whether catastrophe claims/losses have been removed from the data underlying the model(s), and include a discussion of the rationale for the treatment of catastrophe claims/losses in the data underlying the model(s).

14. How were data split for model development purposes? (e.g., data were split into a “training” set comprising 70% of available data and a randomly selected “test” dataset comprising the remaining 30% of available data.)

    a. How was a test set selected? (e.g., *at random* or *out-of-time sample*)

    b. How many exposures and how many claims were in each data set?

c. If the dataset used was countrywide, please also identify how many exposures and how many claims were from exposures in California.

15. Were any notional datasets or simulated datasets used in the model building process? If so, explain how these datasets were built and how they compare to actual policyholder data.

16. Once the final model form was determined, was the model re-fitted using 100% of available data to determine the final model weights/coefficients? If not, explain why not.

For both training and input data, please answer the question(s) below:

17. Explain the level of granularity in the data. For example, for geographic data, please identify whether the data is captured / aggregated at the exact address, ZIP code level, county, etc.; if telematics data[2] is used, please identify the time interval at which data is collected.

18. Identify any major data pre-processing steps (e.g., calculated variables, capping, development, trending, on-leveling, or filtering, removing extreme values, centering, normalizing, Principle Components Analysis ("PCA"), etc.)

19. Identify any data elements obtained from external data sources and identify the source of each such element.

For input data, please answer the question(s) below:

20. Identify all data elements in the input data.

21. Explain and justify any differences in data sources between training data and input data, e.g., for a specific variable, industry data was used to train the model but internal data was used to rate policies.

22. Identify the percentage of the records with missing values for each component of the input data and explain how missing values in the input data are handled.

3. Variables

23. Please identify and define the target variable / output of the model. If applicable, provide the formulae for computing the target variable of each model (e.g., if Workers Compensation Loss Ratio is the target variable, is loss developed to ultimate? trended? Is

[2] Data is required to be compliant with California Code of Regulation 2632.5(c)(2)(D)2.

premium Trended? On-leveled? Manual? Including schedule rating?) along with a description of the data used in the computation.

24. For each other[3] variable / input in each model:

    a. Provide the name of the variable

    b. Provide a description of the variable

    c. Identify any transformations applied to the variable

    d. Identify whether the variable is continuous, discrete, binary, binned (including the bins), or something else.

        i. If the variable is discrete, binary, or binned, provide the number and percentage of exposure at each level, as well as discussion of how the credibility of the data at levels created for the variable has been considered, when building the structure of the model.

    e. Identify the percentage of the records with missing values for each variable, and how the missing values have been handled.

    f. Identify whether the variable is modeled as an independent variable, a control variable, as an offset, or something else.

    g. Identify the source of the data for the variable if external.

    h. If any of the variables are components from a PCA analysis (or similar), also explain all constituent variables that are part of the component(s).

    i. If the model is a GLM, also provide the following items:

        i. The coefficient of the variable in the fitted model

        ii. The standard error

        iii. The amount by which excluding each independent variable from the full model changes a measure of goodness of fit as compared to the full model. Examples of measures of goodness of fit include AIC, DIC, Deviance as measured on held-out data, RMSE as measured on held-out data, AICc, WAIC, PSIS-LOO, cross-entropy, or some cross-validation measure.

[3] Other variables refer to variables other than the dependent / target variable. They would include independent variables, control variables, offset variables, etc.

iv. For any variable where excluding the variable improves the goodness of fit, an explanation why that variable is included in the model.

j. If the model is a machine learning model (e.g., a neural network, regression tree, generalized additive model (GAM), etc.), please also provide the following items:

i. Partial dependency plots showing the dependent variable as a function of each independent variable.

ii. A permutation variable importance plot showing the variables that have the greatest impact on the prediction output.

[Note: A table showing an example response for question 25 is provided in Appendix A. Responses need not follow the example exactly; it is only for illustrative purposes.]

25. How was it determined which variables should be included or excluded from each model as part of the model development process?

26. If any rating factors were derived from a separate analysis (i.e., outside of the model):

a. Provide support for the selected factors.

b. Were the dependent variables in the models offset for the externally-derived rating factors? If not, explain how it was ensured that exposure correlation with the non-modeled factors did not bias model results.

27. If applicable, discuss how the variables in the model do or do not interact with experience rating/schedule rating or other adjustments not included in the model(s).

28. Describe any analysis done on variable interactions.

4. Territory Models

29. Explain how the territory model relates to the other model(s) in this filing; i.e., are territory models developed based on data residual of non-territorial models? Vice versa? If neither, please explain.

30. Provide the standard deviation of pure premium for all policyholders, the within-territory standard deviation, and the between-territory standard deviation.

31. Describe any smoothing applied as part of the territorial modeling process, including technical descriptions of any algorithms employed, and a numerical example if applicable.

32. Provide choropleth maps showing territorial relativities in California, as well as zoomed-in maps of the San Francisco and Los Angeles areas. If smoothing was employed as part

of the territorial ratemaking process, please provide choropleths of unsmoothed relativities and smoothed relativities separately.

33. Do the years / regions / data pre-processing steps underlying the training data in the non-territorial model differ from those in the territorial model? E.g., the non-territorial data uses 3 recent years and the territorial model uses 5 older years.

34. If countrywide data was used, explain how differences in causes of loss are considered in the territorial analysis. For example, are the water losses in Wisconsin similar to the water losses in California?

35. Explain how the relativities for territories with no or a small number of exposures have been determined.

5. Model Output and Selection

36. Has the output of the model been adjusted, smoothed, or transformed in any way? If so, provide a detailed description of these adjustments, including a numerical example of each adjustment.

37. To the extent that the final selected outputs differ from the model indicated outputs (e.g., if selected rating relativities differ from indicated), please provide an explanation for the selection methodology and any judgment employed in the selection process. If smoothing or weighting was used, provide the explicit formulas employed. (Provide an Excel file if applicable.)

6. Model Fit

38. For each model, please provide a quantile-quantile plot on the test dataset showing:

    a. The model indicated value by decile

    b. The actual value by decile

    c. If different from (a), the selected value by decile

39. For each model, provide a quantile-quantile plot on the test dataset (as in question 38), r*estricted to California-only data*, showing:

    a. The model indicated value by decile

    b. The actual value by decile

    c. If different from (a), the selected value by decile

40. Explain any major discrepancies between the values in the plots requested above (e.g., large reversals, large discrepancies caused by selections, etc.) If necessary, provide additional plots as part of the explanation.

7. Impact on Policyholders from the Model

41. Provide a risk profile (rating attributes) of the policyholders receiving the greatest percentage rate increases and the greatest premium amount increase as a result of the application of the model.

42. Provide a profile of the policyholders receiving the greatest percentage rate decreases and the greatest premium amount decrease as a result of the application of the model.

43. What aspects of the model(s) or changes from prior models have had the greatest impact on policyholders in this filing?

44. Which variables have the greatest impact on premiums? Order the variables in decreasing importance.

Appendix A

Example Response to Question 3.25:

Name	Description	Transform	Type	Treatment	Source	Coefficient	SD	Change in AIC
Age	Age of Policyholder	Log	Independent	Continuous	Internal	1.16	0.45	+105
Roof_A	Age of Roof	Poly – 1st Degree Term	Independent	Continuous	Internal	1.21	0.33	+253
Roof_A	Age of Roof	Poly – 2nd Degree Term	Independent	Continuous	Internal	0.25	0.10	+82
PY_19	Policy Year 2019	None	Control	Binary	Internal	0.72	0.52	n/a
PY_20	Policy Year 2020	None	Control	Binary	Internal	-0.43	0.28	n/a
PY_21	Policy Year 2021	None	Control	Binary	Internal	-0.38	0.32	n/a
Sprinklers	Has sprinklers	None	Independent	Binary	Internal	-0.68	0.19	+10