ARIZONA
DEPARTMENT OF INSURANCE
& FINANCIAL INSTITUTIONS

Predictive Modeling Checklist

Provide the following information when **making a filing to introduce a new or modify an existing** predictive model, to include but not limited to: Generalized Linear Model (GLM), GBM, Decision Trees and Random Forest.

If filing a usage based insurance, telematics or similar type of behavior model, please refer to the Telematics Checklist found on SERFF

Please send any questions regarding this checklist to: propcas@difi.az.gov- include Predictive Modeling Checklist in the subject line.

NOTE: File the predictive model under the applicable type of insurance (TOI) and filing type Predictive Model. Any associated rates, rating factors or changes to base rates associated with the predictive model are to be filed separately as Rate, or Rate/Rule filing types.

In a descriptive written narrative, provide the following information. Provide requested charts and graphs and Excel workbooks where applicable.

1. MODEL DEVELOPMENT

a. What is the Predictive Model intended to model? (e.g. Frequency and Severity, Loss Ratios, Pure Premium)
b. If not purchased from an Rate Service Organization (RSO) or an Advisory Organization, provide the modeler's qualifications and credentials, include the entire modeling team as well as the members of company leadership involved. If purchased from an RSO/Advisory, provide a contact or the SERFF filing number that the model was filed under.
c. Is this model a re-fresh or a re-build of a previously filed model? If so, also provide the SERFF number/s for the prior version/s. In addition to addressing questions in sections 1-4 refer to additional questions in section 5) Revised Model.
d. If any external models were used in the building of this model, reference those models and the appropriate SERFF filing number.
e. Address if this predictive model is intended to interact with an existing model, or is a sub-model.
f. Reference the modeling software used and why it was selected.
g. Provide a description of the training, testing, and validation data used.
h. Describe actuarial standards employed, including any specific ASOPs.
i. Provide an academic reference for the model type used.
j. Provide a correlation matrix or heat map in Excel for the variables used in the model.
k. How was the model tested for stability? Was the risk of overfitting data considered and how was it mitigated?
l. Describe the methods used to validate the goodness of fit such as lift charts,

statistical tests, and holdout samples.

m. How were the outputs deciled, mapped and placed into different tiers/buckets? Provide the loss ratio history or experience for each “bucket”.
n. How often will the model be validated against new data?
o. Provide Variance Inflation Factors
p. Provide 100 records of sample Modeling Data in Excel.
q. Provide Residual Plots.
r. Provide Lorenz Curves.

2. DATA

a. Provide the scope of the data: years, number of records/claims, companies, geographies (zip code, census block), nationwide or Arizona specific data. If nationwide data was used, provide the percentage distribution of each state used in the dataset.
b. Describe the sources of the data used to build the model and why they are appropriate for the model.
c. How is Arizona data utilized and what credibility is given to it?
d. Provide policy counts and exposures by state for the data included and excluded in the model for the same time period.
e. Describe any adjustments or modifications made to the data (for example, trending, development, exclusion of catastrophe losses, capping).
f. Identify and explain how outliers were treated during the modeling process.
g. Was any non-insurance data used? If so, from what source or company was this data obtained and validated?
h. What percentage of data was missing and/or required data cleansing?
i. What adjustments were made for missing or duplicate data? Provide support to demonstrate that the adjustments used were appropriate for the model.
j. Describe the methods used to compile, filter, or merge data from different sources. What logical tests were performed to ensure the compilation did not result in lost data?
k. Provide a guarantee that the modeling data will be retained for at least seven (7) years.
l. Provide volume metrics by year for each data subset.
m. Provide univariate Actual vs. Predicted plots by variable.

3.VARIABLES

a. Provide a description of the variable selection process. Provide a Data Dictionary with all modeling variables and acronyms used listed and defined. Include names/types and definitions/uses of each predictor variable and all other variables. Provide the rating algorithm to display the order of the variables used.
b. Are the variables being utilized loss related, expense related, or related to risk in some other way? Explain.
c. Provide a rational explanation for each modeled variable that discusses why it would plausibly impact insurance risk.

d. With each variable listed that was used, provide the weight given to each when determining premium.
e. For variables that were considered and not used, what were the processes and measures used for removing and/or introducing variables in the model?
f. If an outside vendor was used to obtain information/data for rating and underwriting, provide a list of the variables each vendor provided data for.
g. Provide betas and p-values (Model Output) in an Excel file.
h. Provide Chi-Squared tests or F nested model tests by variable.
i. Provide iterative Akaike Information Criterion (AIC) metrics showing how AIC changes as variables are added to the model.

4. RATING IMPACT

a. What effects, if any, will this model have on tiering and underwriting?
b. How will this model be integrated into the rating program when rating new policyholders versus renewing policyholders.
c. How are customers able to identify and correct any errors in their data? Where does the company provide instructions to the customer regarding this process?
d. Provide a table in Excel that contains the following elements as applicable: modeled relativity, current factor, model indicated factor, confidence intervals, and selected factor.
e. Provide explanations for selected rating factors that deviate from the indicated relativities and provide supporting information for the decision.
f. Provide a histogram comparing the pricing impact of the current versus proposed rating program.
g. Provide a list of the states in which this *version* of the model has been filed? Provide a list of the states with the appropriate SERFF tracking numbers.
h. Provide five (5) different examples in Excel of how premium will be computed with the integration of this model into the rating program. Ensure the examples display a wide variance of policyholders.

5. REVISED MODELS

a. How does this model or revision improve upon the prior model?
b. Provide double lift charts analyzed by the company throughout the model revision process.
c. What adjustments or selections were made to the model to derive the final proposed model? What material changes were made? What variables are being removed or introduced?
d. Provide redline copy of tables displaying revised relativities and rating factors.
e. Provide histograms of rate changes impacting the current book of business.