Gaining trust in predictive models—an actuarial perspective
Validation of predictive models—an actuarial perspective

Contents

Page no.

Introduction 2

Model selection methodology 3

Data 11

Model interpretation and model tracking 13

Implementation 14

Conclusion 16
Introduction

With the promise of machine learning and AI also comes challenges to successful auditing and evaluation. One challenge in particular is the ability to validate these new models—the more complex the modeling becomes, the more difficult they are for reviewing, communicating, and controlling model risk. Many companies model risk management teams struggle to keep up with the pace of change and evolving complexity of models; we are seeing real-world examples of where predictive analytics have gone wrong and heightened regulatory expectations are beginning to intersect with the rise of these models.

The explosive growth of computing power over the past several decades has opened an entire new world of possibilities for actuarial predictive models. Predictive analytics are not new concepts to the insurance industry; the shift in the market has not been the techniques, but rather the technology that supports it. Innovation in predictive modeling and machine learning has yielded a rich and diverse landscape of modeling tools which actuaries can use to produce more data-driven predictions.

Actuarial model validation must keep pace with the additional complexities from the variety of models which now fall under its scope. While traditional methods like unit testing and static validation continue to be useful, additional methodology such as using metadata to track data flows over time or variable significance tests evaluating model structure will become essential for model validation as well.

Within the article below, we have attempted to provide an overview of all activities which should be considered by an actuary when validating a predictive model—a comprehensive review of the subject could fill many textbooks. This paper will describe several key components of model selection which actuaries should consider when validating a predictive model and will also provide an examination of various methodologies that can be implemented in the validation process. We hope that this article will spark your interest in model validation and introduce new perspectives for approaching not only validation of predictive models but also validation of traditional actuarial models.

Model selection methodology

For the purposes of this article, we define model selection as the selection of the actual modeling technique, components, and functional form to be used in the model.

For traditional actuarial models, the focus of the model selection stage of the validation process would focus on the choice and reasonableness of assumptions along with how the functionality of the model reflects the problem it is aiming to solve and any limitations in data. In these cases, the actual modeling technique would likely be less open to scrutiny and there would be limited value in designing evaluations for the purpose of choosing the best model.

When validating a predictive model, however, the choice of model is often the primary focus. It is essential for the validator to be comfortable with the choice of model and for the models to be able to justify their selection. Given the wide variety of possible options when choosing a predictive model or machine learning algorithm, the validator must understand the reason for the selection of a particular model and be able to make an assessment as to whether the model selected was appropriate for its purpose.

When choosing the model, the actuary would be expected to compare alternative solutions and clearly articulate why the given model was selected. In order to understand this model selection process, it’s essential for the validator to understand how various models were ranked, and to make an assessment as to whether the ranking methodologies were correctly applied. For instance, the modeler may be willing to forego accuracy in one area of the model in order to more precisely predict outcomes in another. The validator would need to assess whether the tradeoffs are acceptable and understand how the ranking and evaluation methodologies could lend themselves to understanding the desired outcome.

The core components of model selection are

In the next section, we will discuss these concepts in more detail, which actuaries should consider and understand when evaluating model selection.
Model selection methodology

Hold-out analysis

Hold-out analysis is used to evaluate how the model performs on unseen data i.e. data that has not been used to calibrate the model parameters. Conceptually, the approach is to remove (or “hold-out”) part of the training/calibration data and compare it to predictions from the model trained/calibrated on remaining data. To assess the accuracy of the prediction, the validator can use performance measures like mean squared error for regression or confusion matrix for classification.

Hold-out analysis requires modelers to split data into training and validation sets, which is commonly referred to as the train/validation split. The training data is used to fit the model and the validation data is used to evaluate the model performance. The test data should be used to provide an unbiased evaluation of the final model fit.

A common pitfall to avoid is “peeking”, also known as data leakage. Peeking occurs when the model is trained and tested on the same data; as the model has already “peeked” at the data, the results may unfortunately be too good to be true.

Two challenges when the modeler splits the data are deciding the best proportion of train/test/validation split and sampling bias.

A 60/20/20 split is a common division however the modeler should also review a series of different splits and observe the relative performances to understand the sensitivity of this decision.

Validators should review the data gathering process and the sampling methodologies to ensure there is no sample bias. Any steps in the data gathering process that could produce a test population which differs from the general population should be understood and considered. Top-sampling—where the top X% of the sample dataset is selected for training—should be avoided. If the data had been sorted in a particular way, the training data could leave out large groups with significant characteristics. Instead, random sampling is preferred, where each datapoint has an equal chance of inclusion in the training set.

Other techniques to avoid sampling bias include K-Fold Cross Validation and bootstrapping. In order to perform a K-Fold Cross Validation, the modeler will perform the following steps:

1. Split the data into training/validation and test data.
2. Divide the training/validation data into k subsets.
3. In each validation run, one of the k subsets is used as the validation set to estimate error and the other k-1 subsets are the training set.
4. The process will repeat k-1 times until each subset gets to be the validation set exactly once.
5. The error estimation is averaged over all k trials to get total effectiveness of our model.

The technique above significantly reduces sampling bias as all the data is being used for fitting and in the validation set.
Variable significance tests assess the relationship between the predictor variables and the dependent variable. It can be used to reduce the complexity of the model, since insignificant variables may impact the predictive capability in the future as mentioned in the section above, and it can also help validator interpret the model by identifying the more significant factors. Common techniques include: statistical tests, decision trees, and maximal information coefficients.

The p-value is often seen in regression analysis. The p-value for a coefficient tests the null hypothesis that the coefficient is equal to zero (no effect). A low p-value (< 0.05) indicates that you can reject the null hypothesis. In other words, a predictor that has a low p-value is likely to be a significant addition to the model because changes in the predictor value lead to changes in the dependent variable. Conversely, a larger p-value suggests that changes in the predictor are not associated with changes in the response.

In a decision tree model, the variable that the model decides to split on near the root node is significant. In the example below, family history and body mass index (BMI) are significant in predicting Life Insurance Risk Classes.

Multicollinearity could be a problem when the validator assesses the relationship between the predictors and response. “Multicollinearity,” or “collinearity,” is jargon for “highly correlated predictor variables.” For example, within a mortality study, issue age and attained age would be highly correlated, and only one of the variables should be in the final model. If both variables are in the model, the p-values for either predictor can be misleading. The addition of the predictor “issue age” would not be significant because the attained age is highly correlated. Therefore, the validator needs to consider assessing multicollinearity. One common method to detect multicollinearity is to create a correlation matrix with all the variables included in the model. If two variables have a correlation close to ±1, then often they will be highly correlated. Multicollinearity can also be observed by plotting the variables against each other using a scatter plot— oftentimes relationships can be inferred from patterns in the data.

MIC is a measure of the strength of the linear or non-linear association between a predictor variable and response variable. It takes values between 0 and 1, where 0 means statistical independence and 1 means a completely noiseless relationship. The validator can rank the variables by their MICs to identify the significant variables. For example, duration is more significant than moneyness in the MIC table below.

<table>
<thead>
<tr>
<th>Variable</th>
<th>MIC</th>
<th>MIC rank</th>
</tr>
</thead>
<tbody>
<tr>
<td>Duration</td>
<td>0.36</td>
<td>1</td>
</tr>
<tr>
<td>Moneyness</td>
<td>0.29</td>
<td>2</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

In this illustration, age is more significant than weight.
The goal of a successful machine learning model is to use training data in order to make accurate predictions on data the model has never seen. Both the modeler and the validator must be comfortable that the prediction on training data is accurate whilst also avoiding overfitting.

Overfitting occurs when a model incorporates too much noise or random fluctuations in the training data to the extent that it negatively impacts the performance of the model on new data. This typically happens when the model is too complex due to excessive use of variable transformations and higher power terms. The modeler needs to apply techniques to limit and constrain how much detail the model learns.

One potential technique to assess overfitting is the Akaike information criterion (AIC). AIC measures the fit of the model whilst simultaneously penalizing the complexity of the model. AIC rewards how well the model fits the training data, whilst simultaneously apply a penalty when a large number of parameters are required. This penalty therefore discourages complex models with a high number of parameters that add limited value. The formula for AIC is shown to the right. Complex models with increasing parameters will often improve the fit to training data but this does not make the model optimal.

AIC = 2k – 2ln (L)

L = L (φ) = maximum value of the likelihood function of the model}

Besides AIC, the validator can assess models using a range of potential techniques.

- The AUC (Area Under Curve)
- BIC (Bayesian Information Criterion)
- Brier score
- Confusion matrix metrics

Validators are encouraged to use multiple methods, especially when the scores among models are close.
 Validators are encouraged to use multiple methods, especially when the scores among models are close. (Continued)

### Performance Metrics

<table>
<thead>
<tr>
<th>Metric</th>
<th>Calculation</th>
<th>Selection</th>
<th>Graph Illustration</th>
</tr>
</thead>
<tbody>
<tr>
<td>Brier score</td>
<td>Mean squared difference between the predicted probability and the actual outcome. The forecast for rain was 90%, and it did rain. The Brier score is (0.9−1)^2 = 0.01</td>
<td>Smaller scores indicate better forecasts</td>
<td><img src="image" alt="Brier Score Graph" /></td>
</tr>
</tbody>
</table>

### Confusion Matrix

Confusion matrix metrics are often used to rank supervised learning models for categorical predictions. Each prediction can be translated into either a “positive” or “negative” result. The confusion matrix compares the predictions for positive or negative results against the true values which were known from the testing dataset.

Model with low false positive and false negative is better

- **Actual value**: A, B
- **Predictive value**: A, B
- **True Positive (TP)**, **False Positive (FP)**, **False Negative (FN)**, **True Negative (TN)**

We discuss these issues in the following section.

---

One of the primary factors in model selection will be the data.

The available computational capacity of cloud computing enables predictive models and machine learning algorithms to consider many more data sources than actuaries have typically worked with in the past. Validators will need to understand relationships between the datasets and assess the relevance of peripheral datasets to the model’s prediction target.

The validator must consider the appropriateness of the model for the given data. For instance, if the available data lends itself to human analysis for producing a set of “correct answers,” it is often a good idea to select a supervised learning model such as linear regression. On the other hand, if little is known about the data, it is often better to select a model to “explore” the data, such as a neural network in order to identify trends and relationships in the data that may not have been evident to humans.

Many tools of traditional model validation continue to be quite useful, for instance static validation for ensuring correct inputs are used. However, the responsiveness of predictive models to the shape of the data necessitates several additional tools, such as a stability index or K-S statistic which can be used to evaluate changes in the distributions of inputs to help determine if the model needs to be recalibrated.

Modelers will often tie various data sources into a Base Table, which can be used as a comprehensive dataset for testing and validating the models. The validator would need to assess the completeness of the data, review applicability of various data points, and make assessments as to whether decisions made by modelers regarding imputation of missing datapoints, dropping or clamping outliers, and mapping other datasets for simplification purposes will distort outcomes or will provide reasonable results.

The key components for considering data while validating a predictive model include assessing:

- **Data Lineage**
- **Data Stability**
- **Definitions in the Model**
- **Technical Requirements of the Model**

We discuss these issues in the following section.
Data validation methodology

Definitions in the model

Regarding data systems, the actuary will need to review definitions and ensure that they remain consistent throughout the model environment. Changes between model iterations and data pulls also need to be assessed to ensure that data definitions have not been changed, and that any significant changes in field meanings and mappings are thoroughly documented and understood.

Technical requirements of the model

In assessing the inputs to the model, the actuary will also need to assess whether the variables follow the technical requirements of the model. The actuary will need to examine the available data and related model outputs by applying expert judgment. One method that the actuary can use to assess whether the inputs relate as expected to the outputs is calculating correlations between the various input variables and the model outputs. If the actuary observes differing correlations between the various input variables and their respective model outputs from various model runs, it is possible that the model is using data inputs inappropriately.

Data lineage

Actuaries need to ensure that data origin (also known as "data lineage") is well documented. Some methods to accomplish this task include:

- Making use of metadata, which can be gathered at various points in the modelling process
- Ensuring that controls are set up, executed, reviewed, and recorded
- Relationships between various data sources are well documented and understood.

Data stability

Various tests can also be conducted to understand how data inputs change over time. This is important because a model may need to be recalibrated or discarded if the model inputs differ too significantly from the original sample which was used to initially calibrate the model.

One popular tool to assess changes in data inputs is the Kolmogorov-Smirnov test. To conduct this test, the cumulative distribution functions of the inputs to be compared from different samples are plotted on the same graph and the maximum difference is noted. A large maximum difference would indicate that the populations are relatively similar.

In order to assess selection bias, the actuary should evaluate whether the data used for training the model is representative of the larger population as a whole, or of the desired population. Predictive models will make inferences based on the given data—if the data used to train the model has characteristics which differ fundamentally from the data which will be later processed, the model may no longer be valid and results may be uninterpretable.

A traditional method for tracking model effectiveness is backtesting, where the trained model is run using actual input data and the predicted results are compared against actual results. This continues to be an effective tool for validating predictive models. However, due to the importance of the various types of data to the predictive models, it becomes important to monitor for concept drift, which is where patterns in prior datasets no longer hold true.

The core components of model interpretation and model tracking are:

- **Interpretability of the model**
  - **Bias**
  - **Model tracking over time**

Once the validator is comfortable with the model selection and data inputs of the predictive model or machine learning algorithm, the validator would need to assess model interpretation and model tracking. In some cases, there are limits to the validator’s abilities to interpret the results beyond taking the results as a given. For instance, in a neural network, the many levels of processing can obscure the validator’s expectations of results given the observed inputs. As the results may not be intuitive, the validator would need to leverage tools for assessing error and accuracy in order to make comparisons between the selected model results and results from other models tested.

The validator needs to consider model biases, the most common of which is selection bias. In order to assess selection bias, the actuary needs to ensure that the data used for training the model is representative of the larger population as a whole, or of the desired population. Predictive models will make inferences based on the given data— if the data used to train the model has characteristics which differ fundamentally from the data which will be later processed, the model may no longer be valid and results may be uninterpretable.

A traditional method for tracking model effectiveness is backtesting, where the trained model is run using actual input data and the predicted results are compared against actual results. This continues to be an effective tool for validating predictive models. However, due to the importance of the various types of data to the predictive models, it becomes important to monitor for concept drift, which is where patterns in prior datasets no longer hold true.

In the process of considering model tracking, the validator needs to assess and understand how model results change as experience unfolds. The validator would likely need to assess changes in data to determine whether datasets have changed so much that the model is no longer applicable. In the Data section, we introduced the K-S (Kolmogorov-Smirnov Test) as a tool to track changes in distributions over time. This tool can again be used to effectively detect concept drift after a model has been deployed. As a result, the validator may need to apply concept drift measurement methodologies above and beyond those used by the validators to understand the relationship between input changes and output changes as experience evolves.
Next we discuss several common practices which actuaries should consider and understand when evaluating Model Interpretation and Tracking while validating a predictive model.

**Model interpretation**

There are several key points which the actuary can consider when validating the predictive model's interpretability. First, the actuary should be able to understand the outputs, given the inputs. Depending on the model, the parameters chosen for the final model should also give an indication of the model results. For instance, the coefficients of a linear regression indicate the direction and magnitude of the predictors' impact on the prediction.

Second, the actuary should be able to understand how the model's predictions would change given changes in the underlying data. For instance, consider a clustering model trained on a population of 90% preferred plus and 10% standard to predict mortality. The actuary should be able to understand how the model's predictions might differ if the underlying population were to shift to 10% preferred plus and 90% standard.

Third, and most obvious, the model should give predictions which make sense. A predictive model which predicts negative mortality clearly is not conceptually sound.

Finally, the actuary should consider evaluating the similarity between the deployed data and training data by looking at summary statistics, making scatter plots to compare variables, and using distribution comparison tools like K-S tests and stability indexes (see below).

**Model tracking**

Methods for tracking model performance after deployment include rerunning performance tests such as confusion based metrics to see if performance suffers over time, assessing output distribution changes over time using a stability index, or assessing input distribution changes over time also using a stability index.

The purpose of the stability index is to detect concept drift through changes in the underlying distribution. The stability index measures changes in the model's predictions by comparing the relative proportions of predictions that fall into particular prediction buckets. A drastic shift in proportions could indicate that the underlying distributions have changed.

Even after the actuary becomes comfortable with model selection, data, and model interpretation and tracking, the actuary still needs to address possibilities of operational risk. This is especially challenging for predictive models due to the heavy use of various programming languages like Python and R, computational requirements, the wide variety of possible models and model evaluation methods, and the possibility that models will be developed by 3rd parties.

It is a validator’s role to assess the skillset of the individuals responsible for development of predictive models, and so validators should have predictive modelling skillsets themselves. Actuaries also need to be familiar with the various algorithms typically used in predictive models and need to understand circumstances which would warrant selecting one model over another.

Finally, the systemic risk around 3rd party predictive modeling development is another aspect which actuaries need to consider in their evaluation of predictive models.
Conclusion

Model validation can still be viewed as an emerging field in actuarial science. Model validation of a predictive model adds levels of complexity from the disparate data sources to the need for sophisticated algorithms, to the point that actuarial model validation will need to mature further to account for the additional fluidity and complexity that comes from predictive models.

Model validation for predictive models will require actuaries to develop new skillsets and learn new concepts. Actuaries need to have an understanding of the reason for selecting one model over another, as well as an understanding of how to evaluate the relative performance of the models. Another new challenge is understanding the appropriateness of each model for available data, as well as understanding how various data sources link to the predictive model. Interpreting these models and understanding when they need to be recalibrated or canned altogether is another need. Actuaries will need a comprehensive understanding of the end to end process of model development, calibration, and assessment in the new age of predictive modeling and machine learning.
Contact us

Laura Gray  
Principal  
Advisory  
T: 678-458-4305  
E: lgray@kpmg.com

David Alison  
Director  
Advisory  
T: 773-449-8350  
E: davidalison@kpmg.com

Wayne Tao  
Senior Associate  
Advisory  
T: 626-991-9296  
E: guanzhongtao@kpmg.com

Some or all of the services described herein may not be permissible for KPMG audit clients and their affiliates or related entities.

kpmg.com/socialmedia

The information contained herein is of a general nature and is not intended to address the circumstances of any particular individual or entity. Although we endeavor to provide accurate and timely information, there can be no guarantee that such information is accurate as of the date it is received or that it will continue to be accurate in the future. No one should act upon such information without appropriate professional advice after a thorough examination of the particular situation.

© 2021 KPMG LLP, a Delaware limited liability partnership and a member firm of the KPMG global organization of independent member firms affiliated with KPMG International Limited, a private English company limited by guarantee. All rights reserved. The KPMG name and logo are trademarks used under license by the independent member firms of the KPMG global organization. NDP211393-1A