MILLIMAN WHITE PAPER

Effective model validation using machine learning

Jonathan Glowacki, FSA, CERA, MAAA
Martin Reichhoff

Milliman

Model validation has been a key focus area for many financial institutions that rely on models for underwriting, pricing, reserving, and capital models. Traditionally, modelers used relatively transparent techniques such as regression, averages, and other statistical methods for forecasting.

However, recent developments in technology and open-source software have increased computational power, allowed for more effective data-processing algorithms, and created a surge in demand for more advanced predictive solutions. And the market has responded to this demand with machine-learning algorithms.

Once limited to the academic world, these predictive modeling techniques have the power to out-predict more traditional modeling methods. However, with this increased predictive power comes increased modeling risk. The processes that allow these new techniques to perform well also make them complex and opaque. With little transparency in the modeling process, it's easy for an untrained user to misuse machine-learning techniques, resulting in "black box" results with little understanding of the underlying relationships, making model validation all the more important.

A brief overview of machine-learning techniques

Machine-learning techniques are a family of algorithms that bridge the gap between applied statistics and computer science. Unlike traditional regression methods, machine-learning techniques are non-parametric, meaning they aren't bound by a functional form and there is no prior assumption on statistical distributions. In parametric techniques, like a linear regression, an increase in an independent variable (e.g., the variable that explains the outcome such as changes in home prices) must either lead exclusively to an increase or decrease in the dependent variable (e.g., the outcome you are estimating such as the number of borrowers that default). But in machine learning, the effect of an independent variable on the dependent variable can differ based on levels and interactions with other variables. The ability to capture these interactions across different variable values gives machine-learning techniques the ability to out-perform traditional regression techniques.

While there are many different machine-learning algorithms, they can be split into two different categories: supervised and unsupervised. These categories differ based on what criteria they use to classify data or predict outcomes.

In supervised learning, the machine-learning algorithm splits data based on a target variable. This variable can be either continuous or discrete, depending on the algorithm used. To split and classify data, the algorithm iteratively partitions data based on different variables to minimize a cost function. There are many different cost functions used, but in general, they measure the difference between predicted and actual values. Because of their potential to make accurate predictions, supervised learning techniques have numerous applications in risk management. Classification and regression trees, random forests, and support vector machines are all supervised learning techniques.

In unsupervised learning, there is no target variable that the data is split on. Instead, the algorithm works to group observations based on how similar they are to each other across given dimensions. Unsupervised techniques are often used in marketing campaigns to personalize ads for similar groups of consumers. Techniques like K-means clustering and Markov models are both examples of unsupervised learning techniques.

Potential pitfalls of machine learning

While machine-learning algorithms have the potential to improve business processes, predict future outcomes, and save money, they do come with their own risks. Below are three pitfalls commonly associated with machine-learning techniques.

OVERFITTING

Machine-learning techniques and classification algorithms are more susceptible to overfitting than traditional modeling methods. Overfitting occurs when a model bases predictions on spurious correlations within a data sample rather than on

genuine relationships in the population as a whole. While linear and generalized linear models are capable of overfitting data, machine-learning algorithms are more susceptible to this risk, largely because they aren't parametrically constrained. Without a functional form, these algorithms can use every relationship between variables in sample data to cluster or make predictions. And every sample has its own idiosyncrasies that aren't reflective of the true population. Classification and regression trees are especially susceptible to this risk as they can split data until perfect or near-perfect classification is achieved, resulting in many trivial partitions.

If a model that has been over fit is applied to new data from the same population, it has the potential to produce erroneous predictions. And with the weight many businesses give predictive analytics in decision-making processes, these erroneous predictions can have devastating effects.

REDUCED TRANSPARENCY

Machine learning also reduces the transparency of the modeling process. With more traditional regression techniques, it's easy to see how variables interact. To see the magnitude and direction of an effect, one has only to look at a coefficient. If the coefficient is positive, it implies a positive relationship between the independent variable of interest and the dependent variable, and vice versa. Most machine-learning techniques don't produce such readily interpreted results, though. A few, like simple classification and regression trees, output graphics that are fairly understandable. But others, like gradient boosting, random forests, and neural networks, function as black boxes of sorts. A user inputs data and model specifications, and the algorithm returns predictions. While it is possible to look through other outputs from the algorithm to understand how it arrived at its predictions, most users' understandings of these techniques aren't comprehensive enough to do this.

Without the ability to understand how a model arrives at outputs, users can miss warnings of flawed models. In traditional regression techniques, a glance at coefficient direction and magnitude, standard errors, and overall model fit allows users to get a rough idea of how well models are performing. If the majority of variable coefficients are insignificant, the model fit statistics are bad, or variables don't have the relationship expected, the user knows that something might be wrong with the model. The model could have been fitted on bad data, poorly specified, or used in an incorrect context. With machine learning, though, the lack of transparent output makes it more difficult to spot these types of flaws.

BASING RESULTS ON BAD SAMPLE DATA

A model will only perform as well as the data on which it is fitted allows. Machine-learning techniques are no exception to this rule. As mentioned above, the opacity of these algorithms makes it easy for users to overlook flaws in the modeling process, like basing predictions on bad data. And despite the abundance of available data brought on by technological advances, proper steps must be taken to ensure any samples used for modeling are representative of the population as a whole. Flawed sampling techniques can still produce data that don't represent the overall population. In addition to flawed sampling techniques, using old data can also present a risk in the modeling process. If a sample being used in the modeling process no longer represents the underlying process that produces it, the model will produce results that can't be safely used in the present day.

Model validation techniques

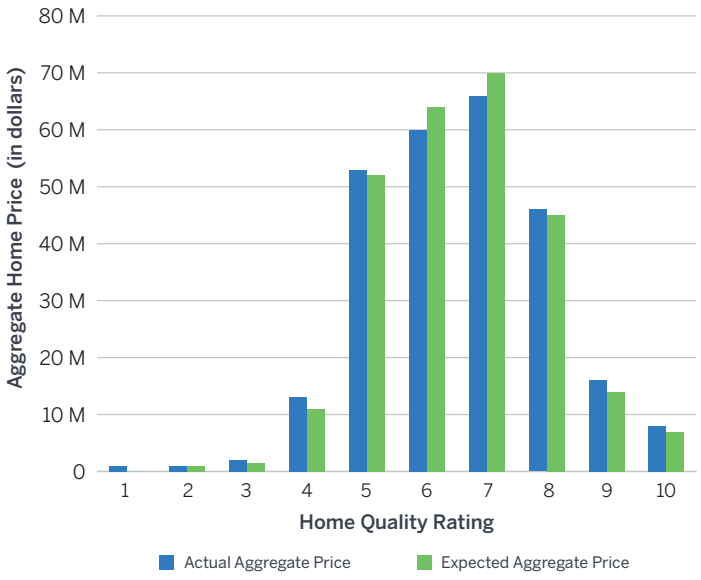
An independent model validation carried out by knowledgeable professionals can mitigate the risks associated with new modeling techniques. In spite of the novelty of machine-learning techniques, there are several methods to safeguard against overfitting and other modeling flaws. The most important requirement for model validation is for the team performing the model validation to understand the algorithm. If the validator does not understand the theory and assumptions behind the model, then they are likely to not perform an effective model validation on the process. After demonstrating an understanding on the model theory, the following procedures are helpful in performing the validation.

OUTCOMES ANALYSIS

Outcomes analysis refers to comparing modeled results to actual data. For advanced modeling techniques, outcomes analysis becomes a very simple yet useful approach to understanding model interactions and pitfalls. One way to understand model results is to simply plot the range of the independent variable against both the actual and predicted outcome along with the number of observations. This allows the user to visualize the univariate relationship within the model and understand if the model is overfitting to sparse data. To evaluate possible interactions, cross plots can also be created looking at results in two dimensions as opposed to a single dimension. Dimensionality beyond two dimensions becomes difficult to evaluate, but looking at simple interactions does provide an initial useful understanding of how the model behaves with independent variables.

Figure 1 is a chart depicting an outcome analysis comparing actual aggregate home price to expected aggregate home price across home quality.

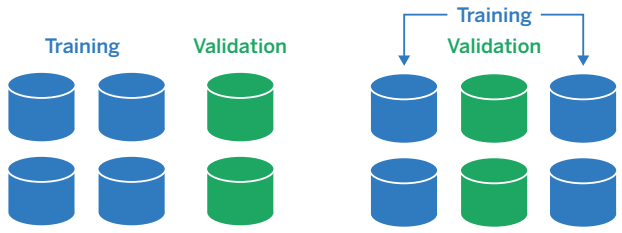
FIGURE 1: AGGREGATE ACTUAL TO EXPECTED HOME PRICE (BY HOME QUALITY)



CROSS-VALIDATION

Cross-validation is a common strategy to help ensure that a model isn't overfitting the sample data it's being developed with. Cross-validation has been used to help ensure the integrity of other statistical methods in the past, and with the rising popularity of machine-learning techniques, it has become even more important. In cross-validation, a model is fitted using only a portion of the sample data. The model is then applied to the other portion of the data to test performance. Ideally, a model will perform equally well on both portions of the data. If it doesn't, it's likely that the model has been over fit. Sample data is most commonly split at an 80-20 level, with 80% being used to fit or train the model and 20% being held out to test the model. More rigorous approaches to cross-validation also exist, including k-fold validation, in which the cross-validation process is repeated many times with different splits of the sample data.

FIGURE 2: K-FOLD VALIDATION



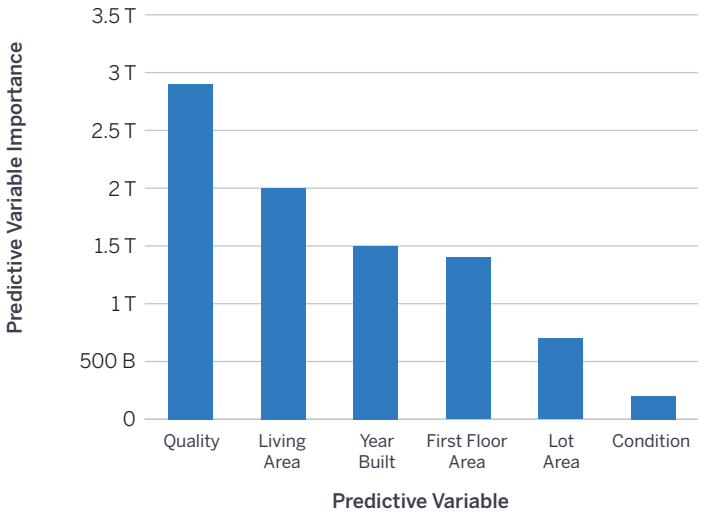
For model validation, cross-validation can be used to compare the stability of the final model and sensitivity test the predictions for the same exposures based on changing the training dataset. In practice, you can train the model to different sets of data, re-populate the outcomes analysis, and determine if the model is overly sensitive to the data used to train the model.

FEATURE IMPORTANCE CHARTS

While cross-validation can help prevent overfitting, it doesn't protect against faulty data. Ideally, if data is poorly sampled and not representative of the true population, its flaws will be detected and addressed during the data exploration process. But due to difficulties in defining and detecting bad data, it sometimes slips through this stage. When it does, there needs to be further safeguards in place to examine how a model is using the data to produce output. As mentioned before, the complexity of most machine-learning algorithms makes this task difficult. However, many implementations of machine-learning algorithms in open-source software can produce feature importance charts. These charts show which variables are the most important in producing output. An examination of one of these charts can tip off users when something is wrong with their model. If traditionally important variables are contributing little to the model, it could be a sign that the model is poorly specified or based on unreliable data.

Figure 3 is an example of a variable importance chart from a random-forest algorithm.

FIGURE 3: VARIABLE IMPORTANCE FROM RANDOM FOREST FIT



Variable importance tables are generally output from the model estimation process. For model validation, it is critical to evaluate the variable importance tables to evaluate if the most influential variables make intuitive sense and are consistent with results from other model approaches (e.g., regression).

Conclusion

As technology advances, machine-learning techniques will become more and more prevalent. These techniques have the potential to help improve business processes, better manage risk, and advance research. But these techniques are not without their own potential risks. The risks mentioned above underscore the importance of independent, external model validation. With proper model validation, users of machine-learning algorithms can be more confident in their results and less worried about potential risks.

Milliman

Milliman is among the world's largest providers of actuarial and related products and services. The firm has consulting practices in life insurance and financial services, property & casualty insurance, healthcare, and employee benefits. Founded in 1947, Milliman is an independent firm with offices in major cities around the globe.

milliman.com

CONTACT

Jonathan Glowacki
jonathan.glowacki@milliman.com

Martin Reichhoff
martin.reichhoff@milliman.com