Predictive Modeling Pitfalls: When to Be Cautious

By Ira Robbin

There are many success stories featuring predictive models, but what does not get as widely reported are the failures: mistakes that range from subtle misinterpretations and minor miscues to unvarnished disasters.

The way the term is used, a predictive model is not simply any model used to make a prediction, but instead refers to the class of methods that includes generalized linear regression (GLM) and data mining.

This article will focus on the use of predictive models in property and casualty insurance and will illustrate several pitfalls. Many of the pitfalls have nothing to do with technical aspects of model construction but rather with false assumptions about the data or misapplications of model results.

Predictive Modeling

What Is a Predictive Model?

The term "predictive model" refers to a generalized linear model (GLM) or other related model. It does not include models such as catastrophe simulation models and econometric time series models, which are also used to make predictions. The GLM terminology has been around since the 1970s, when John Nelder and Robert Wedderburn unified several existing linear regression techniques in a "generalized" construct.

If You Build It, They Will Come

Predictive modeling has experienced an incredible surge in popularity over the last decade, not just because of the marketing appeal of the "predictive modeling" label but largely due to the rise of big data.

The increased availability of large datasets, cheap data storage capacity and computers capable of quickly processing large amounts of data make it feasible to apply GLMs to gain new insights and potentially reap competitive advantages.

GLM Predictions

Given data on the attributes of individuals, a GLM can be constructed to use information on some of the attributes of an individual to predict one of the other attributes. This notion of prediction has nothing to do with peering into the future, but it is useful nonetheless.

For example, if I know you wear red shoes, downloaded five to eight songs from iTunes last month, have a credit score between 700-740 and bought a fuel-efficient hybrid car last year, then with a GLM I might be able to predict you are five times as likely to have purchased low-fat yogurt last week than another person chosen at random from the database. Or I could predict that a person with your attributes spent an average of $3.28 per week last year on low-fat yogurt. With another variation, I could also compute a score based on your attributes and on the basis of that score assign you to the second of five quintiles of low-fat yogurt consumption.

The predictions are not foolproof—I could be wrong. Despite what all of your other attributes might lead me to believe you may have an aversion to yogurt and would not be caught dead buying any.

As this example demonstrates, GLMs can be used to make predictions about:

- **Relativities between individuals** in the population with respect to some dependent outcome variable.
- **Expected average outcomes** for each individual.
- **Subgroup membership** of individuals.

Applications abound in marketing, advertising, politics and other areas.

Modeling Process

GLM predictions of individual outcomes are on the values of various input variables, often called explanatory variables. The input variables could be continuous or categorical.

Examples of continuous variables are items such as annual incomes, heights, weights, which can take on a wide ranges of values—152.8 pounds, or 175.25 pounds, or 212.7 pounds, for instance.

A category variable, in contrast, is one that takes on a fixed countable number of values.

Marital status, with values such as married, single, divorced, or widowed, for example.

A categorical variable, in contrast, is one that takes on a fixed, countable number of values. Marital status, for example, with values such as married, single, divorced or widowed.

A continuous variable can be recast as a categorical one by using a set of ranges. Consider heights of women, which might be reclassified as: category 1-petite, for women under 5 feet 2 inches tall; category 2-average, for those over 5 feet 2 inches tall but under 5 foot 10 inches; category 3-tall, for those over 5 foot 10.

Further, in a GLM, a modeler has the license to transform continuous inputs or outputs in a variety of ways that may dramatically improve the fit. The degree by which the model predictions beat random chance is quantified in a statistic called the ***lift*** of the model.

The modeler uses diagnostics to figure out, on a statistically sound basis, just how much weight to give each transformed input variable in making the prediction. Some "explanatory" variables may be ignored in the estimation. Intuitively these are either independent of the dependent variable being predicted and so have no explanatory power, or are so closely related to a mix of other input variables that they are extraneous. The remaining variables in the model should all have weights statistically different from zero.

One sign of overfitting is when some of the coefficients are not statistically significant. In an overfitted model, minor changes in the data could lead to dramatic changes in the model coefficients when refit to slightly modified data and overly large movement in the predicted outcomes for some individuals.

Significance and Sample Size

A key aspect of model construction is to select a good set of explanatory variables. The tendency to load up a database with variables that are likely to be highly correlated with one another should be avoided. Later in the process the near-duplicates will need to be thrown out.

It is also critical that there be enough points in the sample to pin down the coefficients with a sufficient degree of statistical precision. The more variables there are, the larger the sample needs to be to achieve this. In addition, the more volatile or noisy the outcome, the larger the sample needs to be to separate a signal from the noise.

Outliers

The modeler may throw out (or cap) some unusually large input values or outcomes as "outliers." For example, someone in the training set may have purchased 1,200 cups of non-fat yogurt wholesale last week for resale at their diner; another may have grown bored with the interview and typed in 999; a third may have misunderstood the question to be how many cups could they eat in a competitive eating contest and answered 150; and another may have bought 120 cups last week for their college sorority.

Some of the outliers are errors; some are bad data; some are correct but in a context different from the one assumed by the modeler; and still others are extreme yet legitimate values. It is hard to decide which without burrowing into the data, but it is usually prohibitively costly or otherwise impractical to do so.

Missing Data

Often there are individuals on whom the data is incomplete. We know some of the attributes for these individuals but not all. The question the modeler faces is whether to throw out all such records in the database or attempt to fill in missing fields with likely values.

For example, if 20 percent of the sample population refused to supply information on their credit scores, we could randomly assign credit scores by sampling the remaining 80 percent. We could go further and use correlations that exist in the group with more complete data to possibly do a better job of filling in the missing data.

However, in concept, these fill-in approaches only work if the data is missing at random. If the attribute of having a missing credit score is strongly correlated with low-fat yogurt preference, then filling in a value for the credit score eliminates a possible source of information.

Note this sense of filling in data is distinct from the auto-complete or auto-correct algorithms that start with text a user has entered and attempt to correct and complete the word. Such algorithms have large dictionaries of words (and spelling and typing mistakes often made by users) to compare against the text entered, and each additional symbol entered narrows down the search.

Related to the problem of missing data is the problem of data that is not missing but should be. Sometimes those collecting data will fill in missing fields with default values. This might happen, for instance, if the company collecting the data admonishes the data collectors to fill in all fields. Strange clusters can result. For example, we might find that all data from ACME Data Collection Services LLC is complete, but 20 percent of its records show a credit score of 678.

Feeling Lucky

With a large enough set of significant variables, we run into the increasing possibility that at least one variable doesn't truly belong and was let in only by the luck of the draw. Intuitively, if statistical significance has been defined at the 99 percent level, then with 100 variables that are all statistically significant on a particular set of data, we might expect one of them has achieved its significance through luck. Of course, we can bump up the level of significance, but each such increase requires a larger sample size to declare variables are significant at that level.

Biased Samples and Predictions of the Future

The validity of extending predictions from a model to a different set of data rests on the critical assumption that the sample used to train the model was an unbiased sample. In many cases, however, the sample is really a sample of convenience—data that a company has on its existing customer base, for example.

Self-selection also frequently introduces bias: Those who reply to a survey are often different from the general population. The relevance of bias may depend critically on the application. If we know our sample of yogurt purchase preferences was obtained from people who responded to an online survey that provided a coupon for $5 worth of free yogurt for finishing the survey, we might find the results biased and misleading if we use it to launch a marketing campaign to lure new customers who have never tried yogurt.

To use a predictive model to make predictions of the future, one is implicitly or explicitly assuming the future will be sufficiently like the past so that model predictions remain valid. This may not be such a bad assumption as far as predicted relativities are concerned. However, predictions of the future value of an absolute monetary amount should be viewed with caution.

We might grant that a prediction that a consumer in the second quintile of yogurt consumption this year is likely to be in the same quintile next year. However, additional econometric assumptions are needed to arrive at a prediction that the person will spend an average of $3.79 a week next year, up from the predicted $3.28 per week spend this year.

Correlation, Causality and Hidden Variables

Statistical analysis on its own can only show whether an input is correlated to the output variable. This does not imply a causal relation. No matter their degree of statistical significance, a deliberate change made in the inputs will not necessarily produce a change in the output. Owning a pair of red shoes may be a significant variable in predicting the purchase of low-fat yogurt, but giving red shoes to those that did not have them will not necessarily make them any more likely to make such a purchase.

Often there are hidden variables that directly impact both the input and the output. In such a situation, a change in the input is telling us something about a change in the hidden variable, which by extension is also causing a change in the output variable. If we consciously alter the explanatory input variable, we eliminate its relation to the underlying hidden variable and thereby eliminate its predictive power.

Predictive Modeling in Property/Casualty Insurance

Personal Lines Applications

Predictive modeling in property/casualty insurance has been most widely used in pricing, underwriting and marketing personal insurance products such as personal auto and residential. Applications in those lines are well suited for predictive modeling, since there are a large number of policyholders and extensive, reliable information on their attributes. The number and size of the claims for each policyholder are known over many policy periods. There are enough claims and the losses are usually small enough that any real effects come through and are not overwhelmed by noise.

There is also a clear reward and a potentially large payoff: A company with a better model than its competitors might be able to find the most profitable niches, ones of which its

competitors are not aware. It can also better avoid unprofitable classes.

Predictive modeling can be (and has been) used not only with new types of data but also with data already gathered under existing rating and underwriting systems. Over the years, actuaries had developed multidimensional cross-classification algorithms that minimized error and were unbiased. It is a question of some debate how much better GLMs are versus these traditional actuarial methods. What is clear though is that it is much easier to test new variables and explore transformed variable alternatives with a GLM.

Predictive models in personal auto have also been used to implement credit scoring and telematics. Credit scoring takes items on an individual's credit report and computes a score that is used in underwriting and pricing. Telematics uses a remote device to gather actual data on how an insured vehicle is being driven.

U.S. state regulators and the general public have adopted increasingly negative views on the use of credit scoring, and many states have laws restricting its use. The use of credit scoring appears to unfairly penalize the poor and has a disparate adverse impact on racial minorities, recently divorced people, new immigrants and those adhering to religions that discourage borrowing.

Beyond that, most of the public finds the connection too tenuous: "If I take out a new credit card at a retail store and get 20 percent off my purchases that day, why should I pay an extra $50 for car insurance two months later?"

In contrast, many accept the plausibility of territorial rating differentials, where one county has higher costs than another due to greater traffic density or more expensive medical care and repair costs. So when the statistics bear that out, there is an attitude of acceptance. A purely statistical connection, without a plausible causal explanation, seems much less compelling.

Telematics has achieved greater acceptance because it is voluntarily accepted by the consumer and because it is measuring how the vehicle is being driven. Consumer acceptance is also facilitated by the promise of a discount.

In addition, telematics may be reducing claim costs, as drivers operate their vehicles more safely knowing the computer is recording their every move.

On the other hand, those accepting a telematic device may be a biased sample of those who are extremely safe drivers to begin with.

Claim Predictions

Predictive models are also being used in claims applications, for example, using attributes of a claim to predict its likelihood of blowing up. Applications go beyond personal lines claims. Successes have also been reported developing predictive models for commercial general liability and workers compensation claims.

One particular use is to identify claims that will be given special handling if the model predicts they are potentially troublesome. Such claims might be transferred to more experienced adjusters and extra funds could be provided to pursue private investigations and discovery processes with more diligence.

Assuming the special handling is effective at tempering ultimate claim costs, the predictive model will have made predictions that are inaccurate. However, the Cassandra-like predictions are useful. When compared with the actual values, they may convincingly demonstrate the savings the company has achieved by its intervention.

Another application is to target fraud investigations on claims with values with large relative errors versus model predictions, or conversely on data that is too regular and well behaved to be believable.

A Bridge Too Far?

Attempts have been made to extend the predictive modeling pricing applications to commercial lines, and some believable successes have also been reported with small commercial general liability businesses and in commercial auto.

Successful applications have also been reported in pricing workers compensation, medical malpractice and various E&O (errors and omissions) lines. However, not enough has been publicly shared about these to be convincing. The uniqueness of risks in these lines, the large number of relevant attributes and the relatively small number of such risks all pose challenges in extending predictive model pricing applications into the large risk and specialty markets.

What Would Success Look Like?

Many predictive models have been acclaimed as successes, but the claims in some cases seem overblown. For pricing applications, the basis for comparison should not be whether the model performs better than a random guess but whether it outperforms existing methods. A good R-squared, significance, and good lift versus random chance does not say the predictive model is better than a standard method.

Is Experience the Best Teacher?

The standard actuarial algorithm uses an overall manual rate modified by several classification rating factors to arrive at an initial manual rate for a risk. This is then modified by an experience mod based on the actual historical experience of the risk. The credibility of the experience dictates how much we would rely on it, and actuaries have spent years refining different approaches to credibility.

Many predictive modeling pricing applications are effectively focused solely on producing more accurate and refined classification rating factors. In such models, prior loss experience is not considered an explanatory variable useful for predicting future loss costs.

It is true that for small personal lines risks, most actuaries have found actual risk experience has low credibility, usually less than 10 percent. However, for larger and larger commercial casualty risks, credibility increases until it reaches 100 percent. Loss rating at lower limits is often used to provide a base for estimated loss costs for a wide range of liability coverages, including medical malpractice and some nonmedical professional errors and omissions.

This highlights a difficulty in extending predictive modeling pricing applications beyond personal lines risks and smaller commercial lines risks. If we are going to afford 100 percent credibility to the actual loss experience of a large risk, then what is the point of doing a detailed predictive model?

Pitfall Examples

It is time to see how these issues lead to mistakes in hypothetical scenarios.

Pitfall 1: Thinking a Predictive Model Predicts the Future

Joe, the chief pricing actuary of a medium-sized personal lines writer, laid off a few of his conventional actuaries and hired some statistical modelers. They developed an excellent predictive model that was used to derive new relativities for private passenger auto customers.

The model achieved a finer level of segmentation than before. It highlighted profitable and unprofitable niches. Joe proposed a new underwriting strategy and rate formula based on the predictive model, which his company implemented.

Joe promised the CFO that profits would rise.

A year later, the CFO was quite disappointed when profits fell. While Joe and his team were focusing more and more on predictive models, the skeleton crew devoted to traditional actuarial matters was inadequate to cover trends, legal rulings and loss development. They had failed to spot exploding cost level trends and adverse judicial rulings in several key states.

Other companies had seen this and boosted prices, leaving Joe's company as one of the best bargains in those markets. Premium volume rose in the unprofitable states, and Joe's company was left with a large book of unprofitable business, albeit one with accurately calculated price relativities between risks.

Pitfall 2: A Car of a Different Color

Stuart developed a GLM for personal auto rating for his company. He added some additional variables found in the customer's insurance application but not used in the traditional rating formula. One new result he found was that red cars got into twice as many accidents as cars of any other color. On average, red cars cost the company $800 a year more than non-red cars.

Stuart came up with a brilliant scheme. The company would give a one-time $100 rebate to policyholders with red cars and would pay to have those cars painted a different color. Since the cost of the paint job was $400, the total cost to the company would be $500, but that

would be more than offset by the predicted saving of $800. So the company would be making over $300 in the first year per car.

When the company senior vice president first heard the idea, he couldn't stop laughing for half an hour.

Pitfall 3: Correlation Does Not Imply Causality: Hidden Variables

Jane used a GLM to model the legal expense on general liability claims. Her model showed that legal expense costs on claims handled by Claim Office A were 20 percent higher than those handled by Claim Office B.

Based on her advice, the company, to minimize costs, shifted many claims over to Claim Office B. Next year, her management was not amused when legal expense shot up at Claim Office B and declined at Claim Office A. Overworked adjusters overloaded with cases had hired more outside counsel and supervised that outside counsel less diligently.

The underlying cause of the difference had all along been driven by the relative case load.

Pitfall 4: Tell Me Something I Don't Know

Alyssa headed a team of statisticians at a consulting firm developing a predictive model for hospital medical malpractice losses. The team had no actuaries or underwriters. She and her team garnered statistics from numerous medical centers. They developed a predictive model and announced its completion with great fanfare.

Alyssa and her team presented the model to a meeting of insurance brokers, underwriters and actuaries. The model had a high R-squared and all remaining variables were significant. The new insights she announced were:

- The dollar amount of insurance loss was correlated with the number of beds and the bed occupancy percent.
- Loss varied by specialty: NICU had relatively high losses.
- Hospitals with higher limits had more severe losses.

The audience was not impressed. The big new model told them nothing they did not already know.

Pitfall 5: How Often Is Too Often?

Each year Edward developed a new predictive model of residential loss costs. Each model refined classifications or added new variables not in the previous one. Pricing tools were implemented based on the new models.

Edward confidently predicted continuously rising profits, but that did not happen. Each new model produced price increases for some policyholders and decreases for others. After a few years of introducing new models, the company had lost 50 percent of its original policyholders—roller-coaster rate changes had driven away many longtime customers. The company went after new risks to maintain volume, but they were not as profitable as predicted by the model.

The pitfall here is not that the model was wrong, but that the existence of modeling organizations can sometimes drive a need to develop new models each year. Company executives need to weigh the improvements in accuracy that a new model may bring against the possible loss of customers from overly frequent implementation of changes.

Caps on changes to individual policyholders may be a way to a more profitable strategy.

Further, there may be differences between new and renewal customers. A model trained on renewal customers may not be accurate for new ones due to sampling bias.

Pitfall 6: Big Data Variable Explosion: Are We Ready?

Priscilla convinced her management to subscribe to an expensive data service that provided quite detailed information on a large sample of potential customers. Priscilla's statistical team scoured hundreds of variables in search of models that could identify customers with low accident frequency.

After half a year of diligent effort, they had some solid models and interesting results. Some of the best performing models indicated loss costs were statistically well correlated with the number of hours spent on the Internet, the number of tweets in a month, the number of pizza delivery requests over the last year and the number of Facebook "Likes" last week.

Priscilla proposed a rating approach with month-to-month refinements based on voluntary monitoring of customer telecommunication data of this sort. Her management did some surveys and came back unconvinced. Surveys showed that many of the customers who had accepted telematic devices that monitored their driving would not accept the more intrusive monitoring needed to implement the models proposed by Priscilla's team. Only a small minority would agree to such extensive monitoring, and their expectation was that they would receive sizable rate decreases.

Equally disturbing, in looking over how prices moved based on a sample of historical data from select risks, the survey review team noticed a number of cases of seemingly bizarre though small movements in premium. These movements were inexplicable to the customer and would remain so. The company could not attempt any detailed explanation without revealing the workings of its complicated proprietary statistical model. Would the average consumer understand or accept that their auto insurance premium went up because they had fewer "Likes" that month?

Pitfall Lessons

Predictive models have ridden into the property/casualty insurance industry on the wave of big data and have rightly earned a place in the analyst's toolkit.

They work best in personal lines where there is a sufficient volume of reasonably reliable and complete data and also work well for small standard commercial risks when the data is available.

When the modeling process selects a set of transformed explanatory variables all having statistically significant weights, predictive models are unexcelled at producing accurate individual pricing relativities. They also have many useful applications in claims analysis, identifying factors that are correlated with high severity or spotlighting outliers.

But they are not causative models and, depending on the variables used, they may produce results of a purely statistical nature that are not easy to explain. They don't have built-in econometric or trend components, and they are not catastrophe simulation models.

It is questionable whether they can ever do a better job than experience rating for large risks unless they also incorporate actual loss experience and reflect trend and development. Even when they produce answers better than a standard method, how they are implemented can make all the difference. So when grandiose claims are made about a predictive model, it is wise to be cautious and look ahead to avoid potential pitfalls.

Disclaimer

The opinions expressed are solely those of the author and are not presented as a statement or description of the views and practices of any employer or client, past or present. The author assumes no liability whatsoever for any damages that may result, directly or indirectly, from use or reliance on this article or any part thereof.

ABOUT THE AUTHOR

Ira Robbin *has a Ph.D. in Math from Rutgers University and a bachelor's degree in Math from Michigan State University. He currently holds a position with AIG as a Property/Casualty reserving actuary and has previously held positions with P/C Actuarial Analysts, Endurance, Partner Re, CIGNA PC and INA working in several corporate, pricing and research roles. He has written papers and made presentations on risk load, capital requirements, ROE, price monitoring, excess of aggregate pricing, Solvency II, CAT pricing and other topics.*