Pitfalls of Predictive Modeling

By Ira Robbin

Abstract:

This paper provides an accessible account of potential pitfalls in the use of predictive models in property and casualty insurance. With a series of entertaining vignettes, it illustrates what can go wrong. The paper should leave the reader with a better appreciation of when predictive modeling is the tool of choice and when it needs to be used with caution

Keywords: Predictive modeling, GLM

1. INTRODUCTION

There are many success stories featuring use of Predictive Models in Property and Casualty Insurance applications, but what does not get so widely reported are the failures: mistakes that range from subtle misinterpretations and minor miscues to unvarnished disasters. What is a Predictive Model? Some use the term in a generic way to refer to any model used to make predictions. However, others use the term in a more restrictive sense to refer only to Generalized Linear Models (GLM) and related methodologies. That is the approach taken in this paper. This article will focus on the use of the GLM family 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.

1.1 If You Build It, They Will Come

Predictive Modeling has experienced an incredible surge in popularity over the last decade. This is due not just to the marketing appeal of the "Predictive Modeling" label, but more fundamentally 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. There has been a rush to jump on the bandwagon and start building models. It is in this context that models and their results have sometimes been accepted uncritically and recommendations supposedly dictated by a model have been treated as if issued by a Delphic oracle, not subject to question. The view of the author is that the models are quite useful and often are the tool of choice, but the actuary needs to be aware of the pitfalls in their construction and use.

1.2 Existing Literature

There already are papers about pitfalls in the use of Predictive Models. In their paper, Werner and Guven ([9]) warn against the "failure to get full buy-in from key stakeholders" and admonish analysts for not doing appropriate up-front communications. They discuss how to explain results to non-technical audiences. Kucera ([3]) gave a presentation on pitfalls that listed the challenge of getting senior management buy-in and the danger of "treating predictive modeling as a black box". He also highlighted the problems that exist getting reliable data and the need for the IT resources to be available to implement models. The author agrees with most if not all that is said by Werner and Guven as well as by Kucera. However, their main focus is about pitfalls that could prevent a presumably sound model from gaining acceptance or that could forestall sensible model-based recommendations from being implemented. In contrast, the focus in this paper is to make modelers more clearly aware of some of the real pitfalls in the construction and use of models. These are pitfalls that could lead the unwary analyst to make a foolish or useless recommendation or lead a gullible company to implement an unprofitable strategy.

1.3 Organization of the Paper

The discussion will begin in Chapter 2 with a definition of Predictive Modeling, contrasting it with other modeling approaches used to make predictions. The basic framework for Predictive Modeling will be explained and this will lead to a summarization of the factors that determine when it will be effective.

Chapter 3 will turn to insurance applications. It will survey a range of proposed and actual uses and examine successes and challenges.

Chapter 4 consists of a series of vignettes illustrating what can go wrong.

2. PREDICTIVE MODELING

2.1 What is a Predictive Model?

The term, "Predictive Model", is itself subject to some debate. Some use it in a generic sense to refer to any model used to make predictions. However, the usage in this paper will be that the term, "Predictive Model", refers only to a Generalized Linear Model (GLM) or other related model. This is intended to exclude catastrophe (CAT) simulation models and Econometric time-series models. These other models use different approaches to solve different types of problems. The author believes use of one term for all tends to muddle critical distinctions. In particular, GLM applications seldom explicitly consider time as a separate factor, whereas Econometric time-series models are fundamentally about the evolution of variables over time. As another point of contrast, consider GLMs are used to estimate the relativities of expected loss between different groups of customers and seldom consider the aggregate loss, but CAT models are designed to estimate portfolio loss distributions from large events given the exposure concentration of the portfolio.

Beyond the conceptual differences, in practice it was the author's experience at large Commercial Lines insurers and reinsurers that CAT modelers, Econometric analysts, and Predictive Models worked in different departments using different software[1]. While this could change in the future, what the author has seen is that Predictive Modelers generally construct and run GLMs, but not CAT or Econometric analyses.

GLMs do not "predict" the future as much as they describe relations between different fields of data that existed at the time the data were gathered. GLM predictions of the future are thus forecasts predicated on the implicit assumptions that those relations will continue into the future. Only a more explicit treatment of what lies ahead can produce an explicit opinion about whether such assumptions are reasonable in any specific case.

The GLM terminology has been around since the 1970's when Nelder and Wedderburn ([6]) unified several existing linear modeling techniques in a "generalized" construct.

2.2 GLM Predictions

Construction of a GLM entails using data on attributes of individuals[2] to estimate the value of the outcome variable for each of those individuals. The process of modeling involves selecting structural relations and fitting parameters to give the best parsimonious fit of the predictions to the actual outcomes. Once the GLM has been constructed and the parameters determined, one can then employ it to use information on some of the attributes of another individual not in the original data set and "predict" the outcome for that individual.

This notion of prediction has nothing necessarily to do with peering into the future, but it

[1] It is the author's experience that employers and recruiters adhere to these distinctions in terminology. In particular there are separate ads for CAT modeling and Predictive modeling positions.

[2] An individual member of the population could well be an insurance policy or an insurance claim. It could also refer to a corporation, a country, a sports team or other collective entity and the population sample is a subset of all such entities.

is useful nonetheless. For example, suppose I already have a GLM that predicts the amount a consumer will spend on the purchase of low-fat yogurt. If I learn you have 3 or more pair of red shoes, downloaded 5-8 songs from iTunes last month, have a credit score between 700-740, bought a low emissions car last year, and weigh between 100 and 190 pounds, then with that pre-existing 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 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, and
- Subgroup membership of individuals

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

2.3 Modeling Process

2.3.1 Explanatory Variables

GLM predictions of individual outcomes are based on the values of various input variables, often called explanatory variables. The input variables could be continuous or categorical. A continuous variable can be recast as a categorical one by using a set of ranges. The ranges need not be of uniform size or have the same number of members. For example starting with the weight of an individual consumer as a continuous variable a set of five weight ranges could be defined as shown in Table 1.

While the frequencies do not have to be equal, it is desirable to avoid ranges with percentages so small that they have few representatives in the population. It will be difficult to pin down and prove statistical significance for the coefficients associated with such sparsely populated ranges.

Table 1

Weight Range	Frequency
Less than 100 lbs.	25%
100-150 lbs.	30%
150-190 lbs.	20%
190-220 lbs.	15%
Over 220 lbs.	10%
Total	100%

Beyond ranges, one has the license to transform continuous inputs or outputs in a variety of ways that may dramatically improve the fit. Variables can be squared, raised to higher powers, and polynomial functions can be used. Log transforms are also commonly employed on continuous data such as insurance claim severity. 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.

2.3.2 Goodness-of-fit Statistics

The modeler will examine goodness-of-fit statistics to determine how good a fit has been achieved. One such statistic is root mean square error (RMSE) where mean square error is the average square difference between the model result and the actual outcome. This can be improved by adjusting for differences in the expected variance: an error of a given magnitude that is large relative to the expected variance should penalize the model more than the same magnitude error when a much larger variance in to be anticipated. A generalization of the RMSE called the Deviance captures these effects.

When the outcome data are assumed to have a conditional parametric form, the best fit parameters for a given parametric structure can be found by Maximum Likelihood (MLE) techniques. For example, claim counts in each cell might be assumed to be conditionally Poisson and the structure of the model would express the Poisson parameter for a cell as a function of the explanatory variables. An error bar around the MLE parameters can be found using more advanced formulas such as the Rao-Cramer bound.

Another statistic called the lift measures how much better the model does at prediction than going with random chance. For example, using random chance to assign survey participants to quintiles of low-fat yogurt consumption would result in only 20% being correctly assigned on average. Suppose using a pre-existing model, one was able to boost the percentage of correct assignments to 50%. Then the lift is 2.5 (= .50/.20).

Another important measure of model fit is the R^2 statistic. This measures the relative proportion of the total variance of the outcome that is explained by the model. An R^2 of unity implies all variation has been captured by the model, but that may mean the model is tracking the noise in the data and confusing it with systematic effects.

More generally Goodness-of-fit for purposes of predictive modeling involves more than reproducing the given outcomes. It entails fitting that will lead to a model that is useful in predicting outcomes for different sets of data. So while the most basic fitting statistics measure only how closely a model fits the data, more advanced measures try to minimize the noise by implementing an Occam's razor philosophy of modeling. This is done by using a measure of fit that penalizes use of additional variables. One such measure is the Akaike Information Criterion (AIC).

This very brief introduction to fitting statistics is not definitive or complete. It is meant to present a few commonly used metrics and make the reader aware there is a natural tension between fitting too closely and making good predictions.

2.3.3 Variable Selection

A key challenge is to select a good set of explanatory variables to use in the model. As noted by Sanche and Lonergan ([7]), "When a modeling project involves numerous variables, the actuary is confronted with the need to reduce the number of variables in order to create the model". The tendency to load up a database with variables that are highly correlated with one another should be avoided. Later in the process the near-duplicates will need to be thrown out. For example, the age of a driver is likely to be highly correlated with the number of years a driver has been licensed. Having both of those variables in a model adds duplicate information and leads to coefficients that are unstable.[3] Even if the explanatory variables are not highly correlated, the model with more variables is not necessarily the better model: it may actually have less predictive power due to overfitting.

2.3.4 Overfitting

Overfitting means that a model has too many explanatory variables and the model is too complex when it does not need to be so. Some of the variables are extraneous in explaining the results because they are simply tracking the random fluctuations of the fitted data and they could be actively misleading when used to make predictions on new data. One clear sign of overfitting is the presence of one or more coefficients that are not statistically significant.

Overfitting can also be present even if all coefficients are statistically significant on the given data set. In this case, the extra variables are modeling eccentricities in the particular data set at the cost of reducing the predictive power of the model on other data sets.

2.3.5 Training Sets versus Testing Set

A frequently used "best-practice" is to train the model on a subset of the data, the training set. Then its accuracy is tested by looking at the predictions it makes on the hold-out data, the data not used in the fitting. The hold-out data is also called the testing set.

Analysis of the accuracy of fits on the testing set is important. It can quickly reveal overfitting. Testing accuracy on hold-out data is a procedural protection against the tendency to add variables that effectively model the noise. More advanced procedures such as cross-validation[4] also address this problem of estimating how well the model ought to work when applied to, but not refit to, other data.

2.4 Significance and Sample Size

2.4.1 Not Enough

It is critical that there be enough points in the sample to pin down the coefficients with a sufficient degree of statistical precision. A larger sample size might be required all else being

[3] As stated by Sanche and Lonergan [7], "The parameter estimates of the model are destabilized when variables are highly correlated between each other."

[4] See Hastie, Tibshirani, and Friedman [1].

equal if the outcome variable is relatively noisy. An outcome variable with many zeros and few large values is an example of a relatively noisy outcome. Such outcomes are common in insurance applications. Another factor that could increase the sample size is if the coefficient for a variable is of small magnitude. In that case, more trials are needed to achieve a level of confidence to reject the null hypothesis that the coefficient is zero.

2.4.2 Too Much

On the other extreme, with large sample sizes almost all differences are statistically significant. However, the differences are often not terribly relevant. The reason is that in nature null hypotheses are rarely exactly true, but are more often approximately true. The modeler should be wary of letting in variables that even though statistically significant lead to differences that are insignificant in practical terms. For example, a 0.1% difference in yogurt consumption between those having street addresses with even numbers versus those having street addresses with odd numbers may be statistically valid at the 95% level of confidence when the data base is very large, but it is of no practical consequence. Such differences also have a way of disappearing when the model is fit to new data and their presence may be regarded as a possible sign of overfitting.

Further, hypothesis tests in large samples should be conducted at substantially smaller significance levels so as to retain good power. Using the same significance levels that one would use in small samples fails to balance the costs of the two error types.

2.5 Bad Data

2.5.1 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, another 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. Removing the outliers is usually the preferred route as it costs little and often leads to only minor increases in the theoretical error. However, the modeler should investigate further if there are an unduly large number of outliers or if they cluster about any particular values.

2.5.2 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% of the sample population refused to supply information on their credit scores, we could randomly assign credit scores by sampling the remaining 80%. 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. In that case, it may be worthwhile to develop a model that has "missing" as its own category.

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 may 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% of its records show a credit score of 678.

2.5.3 Misunderstood Data

The modeler should do the necessary investigation to ensure each of the variables is consistently defined across the entire data base. Problems of definition can easily crop up within a large company operating in several locations or when there are several different data suppliers. With multinational companies, one often encounters currency data that was originally drawn from mixed currency datasets but which was later converted to a common currency. It is sometimes better to keep separate databases and develop separate models for each currency, but if the data is kept together the analyst should find out and verify just how the currencies were converted. A finding that consumers from country X spend twice as much on yogurt as consumers from other countries could be true or it could be an artifact of the way currencies were converted.

Categorical variables are prone to being defined differently by different data suppliers unless care is exercised to ensure the definition is implemented consistently. For example, for some of the data suppliers a "low emissions passenger vehicle" might have included diesel engine sedans since most have low emissions of carbon dioxide and carbon monoxide. Other data suppliers might have seen they have high levels of nitrous oxide emissions and put them in a different emissions class. In another example, when a company adds a new variable, it is often surprising how difficult it is for all company personnel and agents to administer it consistently. Who is a "good student" eligible for a good student discount? Does it include the first year second semester junior college student who got "A"s the first semester after getting "C"s in high school?

The modeler should try to find and fix all such inconsistencies in definition. Any remaining concerns about the data should be disclosed as they may impact how much reliance to place on the model in the real world.

2.5.3 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% level, then with one hundred variables that are all statistically significant on a particular set of data, we might expect one of them has achieved its significance through luck. What that means is that its apparent significance is a false positive. 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.

2.6 Biased Samples

The validity of extending predictions from a model based on one set of data 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 on-line 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.

Care also needs to be exercised in selecting the training set so that it is random and representative of the whole population of interest. If for example the training set is the first 1,000 surveys submitted to the company and 500 of these were collected by a survey firm that offered $5 off tickets to a NASCAR event for those completing the yogurt preference survey, our resulting model might fail on the testing set in which those who attended a NASCAR event in the last year make up 10% of the population.

In insurance applications some may treat one set of accident years (AYs) as a training set and another set of AYs as the testing set. This may provide some insight about the model, but it risks confounding time series effects with Predictive Model variable effects.

2.7 Predictions of the Future

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 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.

2.8 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.

2.9 Art versus Science

Given the same set of data, would seventy different but capable modelers, like the translators of the Septuagint, come up with nearly the same set of variables and transformations and structural equations and thus arrive at nearly the same predictions? Or would we get seventy different training set designs, seventy different models and seventy different predictions? This would be highly undesirable. Instead it would be hoped that any capable practitioner could select a reasonable set of variables and transformations, use an appropriately random training set, and arrive at a model that would be roughly similar to one produced by another competent modeler. However, at this point there is little in the way of proof. But given the common training foundations and the sharing of best practices, it is the author's opinion that different competent Predictive Modeling teams will arrive at roughly the same answer. Yet a more careful manager might want to give one set of data to two separate teams deliberately kept apart from each another. A large divergence in their answers would indicate the need for caution and further investigation.

.

3. PREDICTIVE MODELING IN PROPERTY CASUALTY INSURANCE

3.1 Personal Lines

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. Those lines are well-suited for Predictive Modeling. There are a large number of policyholders and extensive reliable information on their attributes.

There is also extensive data on the losses. The number and size of the claims for each policyholder are known over many policy periods. There are enough losses and the losses are usually small enough that any real effects come through and are not overwhelmed by noise.

Proper analysis of all this data promises a potentially large payoff: a company with a better model than its competitors might be able to find the most profitable niches, ones its competitors are unaware of. Just as valuable, it can better avoid unprofitable classes.

3.1.1 Credit Scoring and Telematics

Predictive models in Personal Auto have also been implemented using Credit Scoring and Telematics data. 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 data on how an insured vehicle is actually being driven.

US 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.[5] McCarty ([4]) argues 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, the author believes most of the public finds the connection too tenuous: if I take out a new credit card at a retail store and get 20% 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: 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. In the view of the author, a purely statistical connection, without a plausible causal explanation, seems much less compelling to the public.

Though the use of Telematics for Personal Auto rating is fairly new, it appears to the author to have achieved greater consumer acceptance than Credit Scoring. Several factors may explain this. First is the obvious point that acceptance of a telematics device is often coupled with a price discount. A second appealing feature is that it is the customer who makes the decision to accept a Telematics device. Finally, the author speculates that many consumers find it logical and fair that their rates should be adjusted based on data on how their vehicle is being driven.

Telematics may in fact 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 telematics device may be a biased sample of those who are extremely safe drivers to begin

[5] See McCarty [4].

with.

3.2 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 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 Casandra-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 that have large relative errors versus model predictions, or conversely, on data that is too regular and well behaved to be believable.

3.3 Commercial Lines Pricing

Attempts have been made to extend Predictive Modeling pricing applications to Commercial Lines and successes have been reported in modeling BOP and other small Commercial Package policies (CMP).[6] Other successes have been reported in pricing Workers Compensation, Medical Malpractice, and various E&O (Errors and Omissions) lines using Predictive Models. However, as noted in [10], “Compared to personal lines data, commercial lines data poses an even greater challenge during the development of pricing models.” Data is often not available in as much detail as it is for personal lines risks. Another problem is that different sublines and classes may use different exposure bases.[7] The problems only get more challenging with regard to the large risk and specialty markets. When there are a large number of plausibly relevant explanatory variables and a relatively small number of risks, the model will be prone to overfitting. The uniqueness of insureds in some specialty classes makes it hard to model in a consistent framework. Greater variability in claims severity also introduces

[6] See Walling [8].

[7] See Yun et al [10].

noise that obscures real effects and thus limits how well GLMs work in Commercial Lines.

However whenever a large enough body of homogeneous data can be gathered for a Commercial Lines business, Predictive Modeling should provide valuable insights. Usually this entails focusing on a population the small entities. These could be small firms or franchisees within a larger firm. The key is being able to collect data of sufficient granularity and consistency on attributes of a large number of entities and couple that with data on the losses they generate. For example Error and Omissions (E&O) liability for lawyers, actuaries, and real estate brokers to name a few segments could likely be modeled effectively. Program business produced and underwritten by a Managing General Agent (MGA) may also be amenable to Predictive Modeling. Examples could include programs for youth sports leagues, fishing tour operators, volunteer fire departments, senior center transport services, and so on. The MGA might have very detailed information on customers, data in enough detail to support Predictive Modeling.

Overall, there are possibilities in developing Predictive Modeling applications in Commercial Lines, but the possibilities are more limited than in Personal Lines. Users should beware of attempts to develop a model when the requisite data is not available.

3.4 What Would Success Look Like?

Many Predictive Models have been acclaimed as successes, but the assertions in some cases may be overblown. For pricing, claims analysis, or any other application where an existing procedure is used, the basis for comparison should not be whether the model performs better than a random guess but whether it outperforms the existing methods. A good R-squared, significance, and good lift versus random chance do not say the Predictive Model is better than a standard method. Computing lift versus the existing algorithm could show whether the predictive model is actually better at making predictions.

3.5 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%. However for larger and larger Commercial Casualty risks, credibility increases till it reaches 100%. 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 non-medical professional Errors and Omissions.

This underscores the challenge of extending Predictive Modeling pricing applications beyond Personal Lines and small Commercial Lines risks. If we are going to afford 100% credibility to the actual loss experience of a large risk, then what is the point of doing a detailed Predictive Model?

4. PROPERTY AND CASUALTY PITFALL EXAMPLES

It is time to see how these issues lead to mistakes in hypothetical scenarios. These were constructed to be exaggerated versions of what actually has happened or could happen in practice.

4.1 Pitfall Example: Thinking a Predictive Model Predicts the Future

Joe, the Chief Pricing Actuary of a medium size 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 rating formula based on the Predictive Model. 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.

4.2 Pitfall Example: 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 four times as many accidents as cars of any other color. On average red cars cost the company more than $800 a year 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. "Of course giving these customers free paint jobs would make them better drivers", he said.

4.3 Pitfall Example: Hidden Variable

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

4.4 Pitfall Example: Tell Me Something I Don't Know

Alyssa headed a team of statisticians at a consulting firm developing a Predictive Model for Hospitals 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 $ amount of insurance loss was correlated with the number of beds and the bed occupancy %.
- 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. The exposure base for Hospitals Medical Malpractice is Occupied Bed Equivalents (OBE). Adjustments are made to reflect the distribution of OBE by specialty. Increased limits factors are used to charge for higher limits of coverage. Perhaps if the results had been presented as an investigation, validation, or refinement of existing approaches, the audience might have been more accepting.

4.5 Pitfall Example: How Often is Too Often?

Edward hired a team to do predictive modeling for his company's Homeowners business. Each year he directed his team to develop a new and improved Predictive Model of Residential loss costs. Models were developed for each peril separately. Each year new models were developed by refining classifications or adding new variables not in the previous ones. Pricing tools were implemented based on the new models.

Edward confidently told his CFO to expect continuously rising profits, but that did not happen. Each new model produced increases for some policyholders and decreases for others. After a few years of introducing new models, the company had lost 50% of its original policyholders: roller-coaster rate changes had driven away many long-time 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 rate 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.

4.6 Pitfall Example: 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 social media and telecommunications meta-data of this sort. Her management did some surveys and came back unconvinced. Surveys showed that many of the customers who had accepted Telematics 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 sizeable 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, even though small, movements in premium. These movements were inexplicable to the customer and would have to remain so. The company could not attempt any detailed explanation without revealing the workings of its complicated proprietary statistical model. Not that the average consumer would understand or accept why their auto insurance premium went up because they had fewer "Likes" that month.

Thinking fast, Priscilla recast her model as a marketing tool. The model would be one input directing ads to online customers. Customers whose rates were higher than Predictive Model indications would be targeted with more ads. A year later sales were up and loss ratios were down.

5. CONCLUSION

Predictive Models have sailed in to the Property and 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. They also work well for small standard commercial risks. When the modeling process selects a set of transformed explanatory variables all having statistically significant weights and the analyst has avoided overfitting, 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 that might be good targets for fraud investigations. They can also inform underwriting and marketing activities far more accurately than traditional approaches.

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.

REFERENCES

[1] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning*, Springer, 2001.

[2] E. Frees, G. Meyers, and D. Cummings, “Predictive Modeling of Multi-Peril Homeowners Insurance”, Casualty Actuarial Society E-Forum, Winter 2011-Volume2, p. 1-34.

[3] J. Kucera, “Predictive Modeling: Pitfalls and Potentials”, CAS Annual Meeting presentation, 2005.

[4] K. McCarty, "Testimony of Kevin M. McCarty, Florida Insurance Commissioner, Florida Office of Insurance Regulation and Representing the National Association of Insurance Commissioners, Regarding: The Impact of Credit-Based Insurance Scoring on the Availability and Affordability of Insurance, May 21, 2008" Subcommittee on Oversight and Investigations of the House Committee on Financial Services, 2008.

[5] G. Meyers, “On Predictive Modeling for Claim Severity”, Casualty Actuarial Society Forum, Spring 2005, p. 215-253.

[6] J. A. Nelder and R.W.M. Wedderburn, “Generalized Linear Models”, Journal of the Royal Statistical Society, Vol 135, No. 3, 1972, p. 370-384.

[7] R. Sanche and K. Lonergan, “Variable Reduction for Predictive Modeling with Clustering”, Casualty Actuarial Society Forum, Winter 2006, p. 89-100.

[8] R. J. Walling III, “Commercial Applications of Predictive Analytics”, Presentation at Casualty Actuarial Society Ratemaking and Product Management Seminar, 2010.

[9] G. Werner and S. Guven, “GLM Basic Modeling: Avoiding Common Pitfalls”, Casualty Actuarial Society Forum, Winter 2007, p. 257-273.

[10] J. Yan, M. Masud, and C. Wu, “Staying Ahead of the Analytical Competitive Curve: Integrating the Broad Range Applications of Predictive Modeling in a Competitive Market Environment”, Casualty Actuarial Society E-Forum, Winter 2008, p. 1-15.

Abbreviations and notations

CAT, Catastrophe
GLM, Generalized Linear Model

Biography of the Author

TI**ra Robbin** currently holds a position in Economic Capital Modeling at TransRe. He has previously worked at AIG, Endurance, Partner Re, CIGNA PC, and INA in several corporate, pricing, and research roles. He has written papers and made presentations on a range of topics including risk load, capital requirements, ROE, credibility, reserve risk, and price monitoring. He has a PhD in Math from Rutgers University and a Bachelor degree in Math from Michigan State University.

Disclaimers

Opinions expressed in this paper are solely those of the author. They are not presented as the express or implied positions of the authors’ current or prior employers or clients. No warranty is given that any formula or assertion is accurate. No liability is assumed whatsoever for any losses, direct or indirect, that may result from use of the methods described in this paper or reliance on any of the views expressed therein.

Acknowledgments

The author gratefully acknowledges the contributions of David Clark and Susanne Sclafane. Their insightful commentary and constructive reviews were critical in helping the author draft and improve this paper. Christopher Steinbach also deserves honorable mention for inspiring some interesting examples.