Forecasting & Futurism

December 2013 – Issue 8

PREDICTIVE MODELING IN INSURANCE

Modeling Process

By Richard Xu

In the July 2013 issue of the Forecasting & Futurism Newsletter, we introduced basic techniques of statistical modeling that usually have applications in insurance. As there are increasing discussions about predictive modeling within the actuarial community, the Forecasting and Futurism Section Council members believe that we will only witness increasing applications of statistical modeling in the coming decade, and it will be beneficial to our readers to better understand predictive modeling and be well prepared for the modeling approach in every aspect of actuarial work. Starting from this volume, we decided to have a dedicated column on predictive modeling and discussion about various topics on how predictive modeling can be utilized to help actuaries to improve the effectiveness and efficiency of their work.

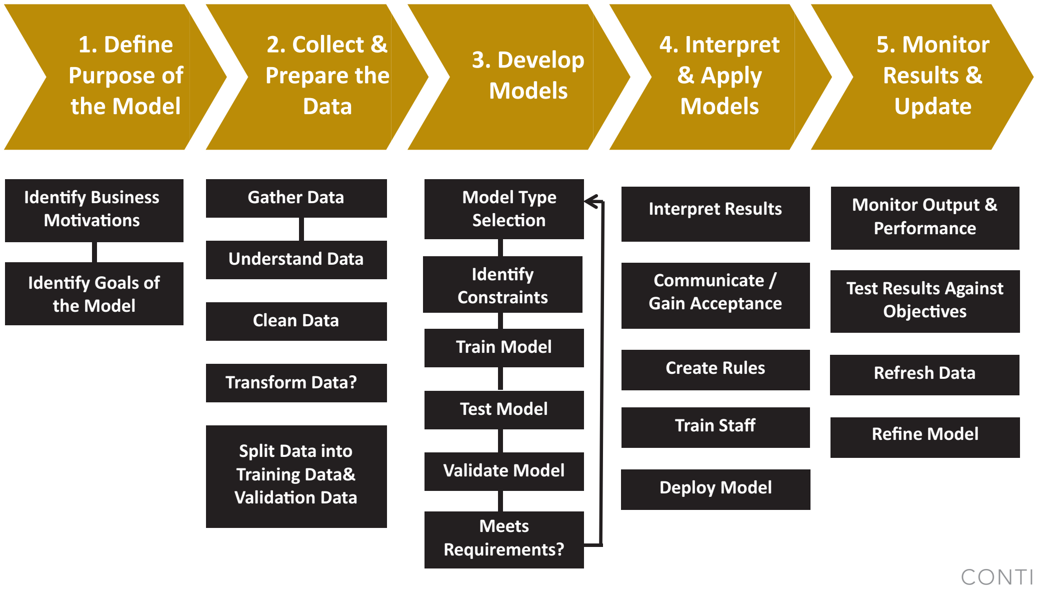
In this article, I will describe a general modeling procedure with an example. In the following newsletter, I plan to cover other topics, such as generalized linear models (GLMs) with their applications in insurance, classification and regression tree (CART) model with examples, applications of data clustering, other advanced models, data considerations, model considerations, etc. If there is any specific topic that our readers find of interest, we are happy to discuss it in this column.

---

To build a statistical model for in business is a very complicated process. It is nothing like an exercise in a classroom setting. There, a problem is usually well-defined, data are clean, and the focus is to learn modeling skills without consideration of practical implementation. In a real business world, actuaries may have to face many issues that have never been the topic in school. To achieve success in a modeling project, actuaries have to deal with these problems.

Although each modeling project is unique, there are certain features that are common to all projects. If a modeler can follow procedures that have been proven to be effective, some mistakes can be avoided and the project has more chance to succeed. Generally speaking, the following flowchart shows a best practice.

Modeling Procedure


1. Define Purpose of the Model
2. Collect & Prepare the Data
3. Develop Models
4. Interpret & Apply Models
5. Monitor Results & Update
Identify Business Motivations
Identify Goals of the Model
Gather Data
Understand Data
Clean Data
Transform Data?
Split Data into Training Data& Validation Data
Model Type Selection
Identify Constraints
Train Model
Test Model
Validate Model
Meets Requirements?
Interpret Results
Communicate / Gain Acceptance
Create Rules
Train Staff
Deploy Model
Monitor Output & Performance
Test Results Against Objectives
Refresh Data
Refine Model

CONTINUED ON **PAGE 28**

MODELING PROCESS … | FROM PAGE 27

Step 1: Define the Objectives

This may sound too naïve, but in reality there are many occasions in which a project starts without a clearly defined objective. As the project moves forward, the goal keeps changing and the team may lose focus and try to accomplish too much at one time. At the end, the project may fail to meet any business need.

Usually a project starts with business needs, where the objective may be vaguely known. Another possible scenario is the need to leverage the data that a company has already accumulated over time to gain a competitive advantage in the marketplace. In both cases, expertise is needed to define what could possibly be accomplished, from of the available modeling techniques and data sources. Without the needed experience, either the objective is not attainable by the current statistical model, the required data do not exist, or the model and data do not match.

Another important goal of this stage is to get the cooperation of the high-level management team. A successful predictive modeling project relies on the collaboration of many departments within an organization, and a change of company culture to a more data-driven approach is necessary. Without consensus from the management team, it is nearly impossible to have different divisions working together.

Step 2: Understanding Data and Business

Data is a critical component for an effective model. All relations and patterns that we can utilize for business have to come from data that a model is built on. It will never be overstated how important data is to statistical modeling. More often than not, a project could fail due to the lack of available data.

Generally speaking, an insurance company has accumulated a large quantity of data over the past decades. The data are usually good for the purpose of an experience study. However, current data usually lack the depth of information. Besides age, gender and location, insurers usually do not have much more information about their policyholders. This makes the modeling project nearly impossible for applications such as underwriting, fraud detection or a retention program. The solution is to find data from a third party, which may include credit score, financial information, motor vehicle records, etc. However, external data may bring up the privacy issue that many insurance companies try to avoid for risk of their reputation.

Equally important is the data quality. This could be another obstacle to a successful model. Actuaries are experts on data, and they will never be short of examples when there are many errors in data. In addition, data may come from different sources, and mismatched and missing values are always an issue with data.

It is always dangerous to build a model without full understanding of the data and the intended usage. That is often the difference between an experienced data scientist and brand new statistics graduate. It is important for the modeling team to work closely with the business and market experts to select a suitable model and variables, create derivative variables, or group data. All of these benefit from business understanding.

Step 3: Develop a Predictive Model

Rigorous mathematical training and decent business knowledge are prerequisites to being a good data scientist. On one hand, predictive modeling is built on statistical processes to find relationships in data, and solid understanding in modeling technique is a necessary component. On the other hand, the training from textbooks is far from enough to handle real projects. Thorough understanding of the insurance business and underlying business forces is equally important to building an effective model.

There are certain degrees of freedom to choosing types of models, but in reality the options are usually limited. Although there are many mathematic models available for potential in insurance (please see “Predictive Modeling” in last volume of Forecasting and Futurism Newsletter), you may find GLM discussed most frequently. One reason is its transparency and simplicity as it is a natural extension

of ordinary least squares (OLS) that all actuaries have had certain exposure to in their education. Other possible models include data clustering, CART model, etc. All of these have the one common feature, i.e., you can open the model and find business insights from the model. Another group of models—including neural network, random forest and support vector machine (SVM)—are black-box models. They are more powerful and effective in most cases, but lack of understanding leads to low acceptance in the . As actuaries are becoming more confident with modeling techniques, we may see more applications of these complicated models.

Once a model type is selected, to develop the model is more or less a process of selecting the right groups of variables, including their interactions, so that the model can best explain the observed data. To identify the most predictive variables and their combination is not a trivial task. It involves both statistical criteria and business sense. Certain statistical methods are available to select variables in a systematic and sequential manner, such as a stepwise procedure.

After a preliminary model is built, we would like to see its performance. There are statistical criteria for the purpose, such as deviance or "information criterion" metric, e.g., Akaike information criterion (AIC) or Bayesian information criterion (BIC). One large concern in modeling is so-called "fitting," where a model is so complicated relative to data that the model actually takes up noise in the data set rather than the actual underlying relationships. Validation can address both of these two issues. In the validation process, the developed model is applied to data that have not been used for model fitting. If the results deteriorate much from modeling results, fitting is most likely the reason. Also, validation results are close to the performance when the model is applied in real business.

These procedures are iterative in nature. The modeling results and validation results have always been compared to the desired performance. If results cannot meet requirements, the modeling team may have to go back to either improve the model or choose another type of model. This process may have to repeat several times until the model is effective enough to meet requirements.

Step 4: Interpret and Implement Model

After a predictive model is created, the understanding of the model is an important step. This includes interpretation, business insights and communication. It is dangerous to apply a model without fully understanding it. We can open up the model and interpret the relationship between variables, and check it against business intuition. This could be viewed as another layer of model validation. In addition, we may discover new relationships from the model that are subtle or hard to grasp by human intuition but could be found by statistical algorithm.

Communication is an intrinsic part of modeling. Without a proper understanding of the model, no one will feel comfortable to use the model even if it is from an expert, which makes implementation impossible. Detailed documentation without technical jargon is needed for all stakeholders. Presentations are very helpful. All these will help to ease the difficulties in applying the model in business procedures.

Implementation may involve many departments in a company, depending on the type of applications. For a large predictive modeling initiative, it may change the business routine of many divisions such as product development/pricing, underwriting, administration, IT, etc. It is collaborative among all parties, and the modeling team needs to work closely with experts from different fields and find an optimal way to deploy the model.

There might be other obstacles in implementation—for example, to understand and adjust for any regulatory or company constraints related to the model or the variables used

CONTINUED ON **PAGE 30**

MODELING PROCESS … | FROM PAGE 29

in the model during the model implementation as well as development phase. The considerations may include data limitations, company cultural barriers, privacy concerns, IT constraints, issues of indirect discrimination, and the company's reputational risk. Just because it is legal to use certain variables in the model does not necessarily mean it is ethical.

Step 5: Monitor Results and Update

As part of risk management, after a model is implemented, monitoring the performance is not just necessary, but required. The evaluation system may include early detection of possible error, unintended consequences of the model, anti-selection, or an impact on other product lines. A good quality assurance and feedback loop is also necessary to make sure the model is accomplishing the original objectives.

If new experience data are available, models must also be periodically updated or refined in order to stay current. The frequency of model updates will depend on the type of model and data being used.

These steps are general procedures that a good predictive modeling team needs to follow. For a specific project, certain steps may be more important than others. For example, for an experience study, development of a model may take more time than all other steps, as objectives and data are clearly defined—especially if a traditional experience study has already been in place. If it is an underwriting model, each of above steps is as important as others. Ignoring any step may lead to a failed project.

AN EXAMPLE

The applications of predictive modeling in insurance could cover nearly every aspect of the business, from product development/pricing, experience study, underwriting, to sales and marketing, administration, claim management. It will be helpful to have an example for readers to understand.

Here is the application of predictive modeling to an experience study on lapse rates in the level period of a life term product. As explained above, the goal is obvious, i.e., to use a statistical method to study lapse rates. Since it is a multivariate approach, bias from univariate in a traditional study could be avoided. In addition, some interaction terms could be included to address correlations between variables. The usage of data is much more efficient, and issues like low credible data can be handled

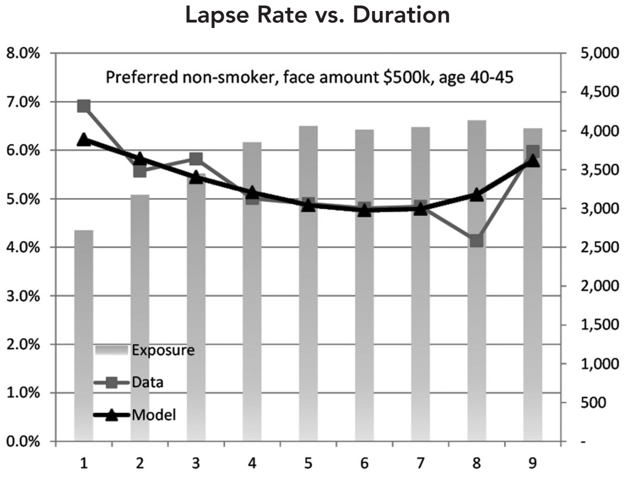
If a traditional experience study is already in place, data are readily available as the same data could be used for the predictive modeling. If not, the data understanding and cleaning are the same as a traditional study, and actuaries are experts in dealing with these issues, which are quite universal for actuarial study, not specific to statistical modeling.

A major task in the experience study is data modeling where statistical skills are heavily emphasized. From choice of model to variable selection, statistical training as well as experience is crucial. For this lapse rate study, a transparent model such as GLM is desirable as the model results can be compared to conventional wisdom. Naturally, a Poisson distribution with logarithm as the link function is used to model the lapse data. The variables available for modeling are limited to about a dozen variables. Statistical tools can be utilized to select which one would be in the model to explain lapse rate variance. However, in the process, business knowledge is equally important. This is where art and science come to play together. Statistical techniques alone may lead to a perfect model, but may have no value for business or have no connection to reality. The modeling process is iterative, and there might be back and forth discussion between the modeling team and other stakeholders, such as pricing, product development and valuation.

At the end, the lapse rate model will be in the following format:

$$R_{lapse} = R_{base} \times \prod_{i} F_i^S \times \prod_{i,j} F_{i,j}^C$$

where R_{lapse} is the formula-based lapse rate predicted by model, R_{base} is the base lapse rate, F_i^S is factor for variable *i*, and $F_{i,j}^C$ is the interaction term for variable *i* and *j* if it is needed. The major variables that are included in the model include duration, underwriting class, issue age and face amount. And a few interaction terms between these variables are also incorporated in the model. The final format of the model is consistent with current actuarial practice with multiplicative formulas.



After the model is finished, it can be evaluated for its performance in addition to statistical assessment. One useful comparison is the lapse rates predicted by model vs. observed values. For example, we can test each cell to see how good the fitness is by plotting the predicted vs. observed values. An example is shown in the above plot. In addition, we may gain business insights from the model. The match of the model to business experience could also boost confidence of stakeholders in applying the model.

When the model is finally utilized in business, it is not the end of the project. Instead, monitoring the performance of the model in real life is an ongoing effort to ensure the model is as effective as the model in development. When new data is available or new experience is accessible, update of the model is required, which could lead to a new round of modeling efforts.

CONCLUSION

Statistical modeling is potentially a double-edged sword. If applied correctly, it is a very powerful and effective tool to discover knowledge in data, but in the wrong hands it can also be misused and generate absurd results. Here is a hypothetical example. One day, you are surprised to hear on the news that "visiting a gas station more than twice a week leads to 18 times higher mortality," which is based on statistical modeling on 47,000 data points. It is understandable that mortality will be higher if you drive more miles, but the mortality is out of proportion to the possible mileage. It may be hard to argue with this "modeler" as he has a strong statistical model and sizable data to support his results. However, there might have been a fundamental flaw in the process of modeling and model interpretation. Instead of the number of trips to a gas station, other factors may be much more important that lead to elevated mortality. For example, these who frequently visit gas stations may be buying cigarettes instead of filling gasoline, or they may live in low-income neighborhoods, relying on the gas station for daily needs and have limited access to the health system. Correlation does not mean causation, and statistical models can only be powerful when they are properly constructed and interpreted.

Recently, there was quite a lot of media exposure about causation of crime in large cities by lead, such as in Forbes and *The Washington Post*. The theory states that popular use of lead in gasoline and paint in the 1950s and 1960s caused the high crime rates in large cities in the 1970s and 1980s. Although it has been scientifically proven that lead has a severe impact on childhood development by both experience and historical data, there is no study to show the impact on crime. The correlation does exist in historical data, but the high crime rates in the 1970s and 1980s may also be ex-

CONTINUED ON **PAGE 32**

MODELING PROCESS … | FROM PAGE 31

plained by other factors, such as baby boomers in their 20s and 30s during that period, or the after-effects of the Vietnam War. There are thousands of social and economic metrics, and thousands of correlations with crime. You have a very good chance to find a very high correlation with crime rate that is purely accidental. I am not here to say the observation is wrong (which actually could lead to a scientific study of the causation between lead and crime), but to point out a possible misuse of modeling that relies on an accidental correlation.

Richard Xu

Statistical modeling has broad applications in actuarial science as well as overall insurance business. To build an effective model, solid understanding of statistical knowledge is essential, but a good sense of business and enough data with high quality are also critical. Predictive modeling is about statistics, but more than that, it is about data and business. ▼

Richard Xu, FSA, Ph.D., is senior data scientist and actuary at RGA Reinsurance Company in Chesterfield, Mo. He can be reached at *rxu@rgare.com.*