Predictive Analytics and Futurism

June 2017
Issue 15

Predictive Model Building 101

By Dorothy L. Andrews

Your boss has just given you a project to build a predictive model to identify highly profitable customers, and you have no idea where to start. The predictive modeling exercise begins with understanding the business problem and ends with validation of the model and dissemination of the results. However, there will be the ongoing task of monitoring the model for continued fit as new data emerges and the business changes. A lack of fit is a clear signal the model needs either a "refresh" or a "rebuild." You will need to overcome many obstacles in getting a model from the drawing board to company production systems. This article is intended as a guide to help you navigate through 10 modeling phases for building a predictive model and to provide you with some insights as to how to overcome obstacles you will likely encounter along the journey.

PHASE 1: DEFINE THE PROBLEM

The financial objectives of the organization should be a guiding light in defining problem statements your model will address. Management are more likely to allocate resources and sponsorship to your modeling project if the solution addresses "pains" that keep them up at night. If you cannot clearly articulate how your model is important to the continued health of the organization, it is unlikely management will leverage scarce resources to fund its execution. Make sure you define the problem in terms your stakeholders will understand. It is important that management who can eliminate obstacles that may hinder the successful implementation of your model project be included among your stakeholders.

It is important to demonstrate that the problem you wish to solve is observable, measurable and subject to classification on some metric. For example, observable characteristics of a highly profitable customer are the types and number of insurance products they own. However, merely owning a product is not sufficient. We need to measure characteristics such as policy retention, premium payment levels and cancellation/renewal behavior to refine profiles of highly profitable customers. Once profitability criteria are identified, then customers can be rank ordered on a scale from least profitable to most profitable. Management is then better positioned to remediate the least profitable and improve retention efforts to keep the most profitable and find more like them. It is important to keep the financial objectives of the company in mind as you develop your problem statement.

PHASE 2: DEMONSTRATE THE FINANCIAL IMPACT OF THE SOLUTION

Key stakeholders in your organization include members from the C-Suite and senior leaders in the actuarial, underwriting and information technology (IT) groups. Agents and brokers may also be stakeholders since they assist their clients with purchasing products using customer scores resulting from predictive models. For example, if agent portals are equipped to render a customer profitability score based on data entered by the agent, then agents may be motivated to produce the best score possible. Data controls will need to be in place to identify when possibly conflicting combinations of data may adversely impact a customer profitability score.

The proposed model should be of financial significance for each of your stakeholders. It is important to understand how the model will improve the financial position of the organization. The more significant the financial impact, the greater the likelihood your stakeholders will support the implementation of your modeling project.

PHASE 3: UNDERSTAND THE PRODUCTS

Model building begins with a solid understanding of the design and features of the products being modeled, how they are marketed and their distribution channel, and the accuracy of underlying administrative and other company data. Many company administrative systems lack adequate controls around the data entry of application and product attribute data. As a result, it becomes essential for the modeler to develop assumptions regarding missing and incorrectly specified data. This requires expert knowledge of the product's distribution, marketing, features and design. Such expert knowledge is also invaluable in understanding anomalous data elements. For example, if a particular product feature appears more frequently in your data set than it should, then it is important to investigate such an anomaly to determine its validity. The results of the

investigation can often become a teachable moment to improve the administration of application and product attribute data. The fewer assumptions needed to prepare data for modeling, the more reliable the results of the model to measure phenomena of interest to the company.

PHASE 4: IDENTIFY INTERNAL AND EXTERNAL DATA

A number of considerations are necessary when constructing modeling data sets from internal company data. Most financial data is transactional in nature, requiring extensive coding to summarize it, recognizing canceled and backdated transactions, in particular. Failure to recognize the cancellation of premium payments, for example, will lead to overestimating net premiums paid on a policy, impacting any derived metric based on premiums. Many companies still currently rely on legacy systems that require Job Code Language (JCL) and COmmon Business Oriented Language (COBOL) to extract data needed for modeling. Further, the number of programmers familiar with these languages is dwindling, putting a premium on sought-after resources for your project.

Changing IT platforms can be extremely expensive, but most companies recognize the need to make the transition to more relational architectures, and they are making the investment. These architectures need to be more flexible, however, to accommodate the codification of new data elements. For example, it is common that the only data element that captures height and weight is the adjuster note. The adjuster note is an example of an unstructured data element. It is free-flowing text entered at the discretion of the adjuster. These notes represent data-mining gold if you are studying the relationship of height and weight to the duration of workers compensation claims. Although text-mining tools are available to assist with the mining of adjuster notes, companies can gain greater leverage from their data by structuring the collection of data elements once it becomes clear they have predictive value.

Once internal data has been structured, appending external data can significantly increase the predictive power of predictive models. What external data should you include? Good question, because we have lots to choose from. Currently, models are including census data, geospatial variables, economic data and consumer attribute data marketed by companies like Acxiom to assist with customer segmentation. Companies recognize the need to market differently to Millennials than to Gen X'ers and Baby Boomers, and they are incorporating marketing data in their predictive models. Depending on the purpose of your model, it is very important to make sure model results based on internal and external data do not unfairly discriminate against policyholders. Regulators have, as one of their primary missions, to prevent unfair discrimination in the pricing and distribution of insurance products. They are becoming educated on advanced modeling techniques, and they especially scrutinize model variables for their unfair discriminatory power. Do yourself a favor and make sure your in-house counsel reviews your variables, especially if your models need to be disclosed in regulatory filings.

PHASE 5: ITERATIVE DATA SCRUBBING AND ANALYSIS

Modelers are fairly united in their view that most of the heavy lifting in building a predictive model involves scrubbing and analyzing the raw data and augmenting these data with relevant external data. Insurance company data, like that of others, is transactional by design. Every time a change is made to some aspect of a policy, a new data record is created in every company system where the change applies. The first step in constructing the modeling data set is the extracting of raw data from company systems and summarizing these data to an appropriate level and at an appropriate periodicity. For example, data may be summarized at a policy level for every quarter in the model study period. This means your data set contains a snapshot of the policy at every quarter end for the model study period. This is a programming task that is often achieved with the help of the IT department or, what is becoming more likely, by the modeling team to avoid delays often associated with IT project scheduling. When the modeling team takes on this task, it is paramount that control totals are identified to validate modeling data against to assess the accuracy of the programming results. External data is usually appended to the summarized data records.

Missing data and misspecified data are unavoidable in any data set, but if improperly resolved, the data set will likely bias your results in unwanted directions. Resolving missing and misspecified data requires a solid understanding of how the products being modeled are distributed, designed and marketed to develop assumptions and adjustments to transform "messy" data into usable data. Construct frequency distributions of the levels for each attribute variable and histograms for continuous variables as a first step. Discrete variables are often treated as attribute variables if there are a limited number of values in their range. External data can be missing and misspecified if out of date. For example, if policy zip code data is invalid, it may not be possible to append census data, such as average income or home values, two important attributes in life, health and P&C modeling.

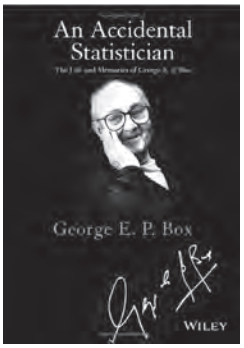
Misspecified data elements can be harder to detect. Examining frequency distributions can shed light on values that don't belong in a field. Conducting inspections on dependent fields is also another tool to identify misspecifed fields. Data dependency in this context means the values on one data elements limit the possible values on the dependent data element. The results of such inspections can be used to correct company processes responsible for misspecified data elements. Modelers should feel some responsibility to influence the correction of data anomalies companywide and not just for the modeling exercise. The modeler can use the results of analyzing missing and misspecified data to develop eligibility criteria for including records in modeling data sets. It is, additionally, important to quantify the financial impact of excluded records by some standard of materiality. Modelers may want to exercise more due diligence in correcting records determined to be financially material.

PHASE 6: MODEL VARIABLE DEVELOPMENT

The raw data records have now been cleaned, appended with external data and assessed for inclusion in the final modeling data set. However, the modeling data set may not be complete. Additional considerations include the development or grouping of levels on attribute variables, the derivation of new variables from the raw data, the treatment of variables either stochastically or deterministically and the identification and derivation of a target variable, if applicable. These are nontrivial considerations and a function of the purpose of the model.

It is important to understand the qualitative relationships among the variables in the data set to eliminate variable dependencies. Your predictor variables should be mathematically independent. Examine any correlations that may exist among your variables to avoid including variables that measure the same model effect. A principal components analysis is a useful technique for isolated uncorrelated variables. Examine the clusters of correlated variables to determine which **one** from each cluster to include in the final modeling data set. Correlated variables can lead to unstable parameter estimates and should be avoided in constructing the final modeling data set. Naturally, this extends to derived variables and the variables used to create them. A simple correlation matrix can assist with this identification. Univariate analyses are also useful to identify variables to include, but not the ultimate criteria by which to select model variables. Stepwise procedures additionally can help demonstrate the statistical importance of variables in the presence of other model variables and are yet another tool for finalizing the final set of modeling variables.

PHASE 7: MODEL CONSTRUCTION


An Accidental Statistician
The Life and Memories of George E. P. Box
George E. P. Box
WILEY

This is the phase of the project every modeler loves to reach. This phase involves selecting the "right" statistical model to fit the data. I want to caution modelers in thinking they have the "right" model when they are done with the exercise. In the words of Dr. George E. P. Box, "Essentially, all models are wrong, but some are very useful." Dr. Box founded the Statistics Department at the University of Wisconsin at Madison. He taught himself statistics while serving in the British Army. During that time he became very good friends with Dr. R. A. Fisher, considered to be the founder of modern-day statistics, and he went on to earn a Ph.D. in statistics from the University of London. He is considered to be "one of the greatest statistical minds of the 20th century."[1] Dr. Box co-invented the Box-Cox Power Transformation used in regression analysis with Dr. David Cox, noted for his contributions in the area of proportional hazards regression modeling. Please read Dr. Box's memoir, *An Accidental Statistician: The Life and Memories of George E. P. Box*. You will find it thoroughly captivating and inspirational.

The notion of a "useful" model should remind modelers that a more useful model may exist. Software packages are greatly simplifying the identification of "useful" models using just a few keystrokes. Once the modeling data set has been constructed, software packages are available that will run several kinds of statistical models against the data set and rank order the resulting models under a set of tests of statistical significance. These software packages require little to no program skills to run, but let's face it, running models falls in the 20 percent of the effort category of the "80–20 Rule" as applied to building a predictive model. The real modeling building takes place in transforming the data under a set of modeling assumptions and developing the criteria for selecting potential data variables, which is the 80 percent of the "80–20 Rule." The number of lines of programming code needed to program a generalized linear model (GLM), for example, is a mere fraction of the amount of code needed to build the modeling data set, unless your data is naturally perfect. Naturally perfect data is a modeler's dream, but seldom encountered.

A word of caution is in order in respect to some of these packages. While they may be child's play to use in terms of simplicity, interpreting model results should be left to a subject matter expert with a thorough understanding of statistics, the products being modeled and the business environment in which model results will be applied. Further, don't underestimate the need to clearly articulate model results to your stakeholders. It will be important to demonstrate how the model results solve the proposed problem in terms they understand so they may comment on the model. All your hard work will have been for nothing if you express you results in esoteric statistical jargon your business leaders can't understand, which may compromise the likelihood of its adoption by the company.

PHASE 8: MODEL VALIDATION AND TESTING

Most would agree that recognizing the "wrong" model is easier than qualifying the "right" model, if a "right" model is even possible to build. Model validation can help you assess whether your model is a reasonable representation of the phenomena under study. But remember, the model is only a representation of the "real thing" at a given point in time. It is **not** the "real thing." (Sounds like an ad for Coca-Cola, right?) The phenomena under study is constantly changing, while the models are always in catch-up mode in their predictive power. The greatest flaw of any model is the model risk they pose for organizations using them.

In a 1996 Goldman Sachs “Quantitative Strategies Research Note,” Goldman Sachs defined model risk as “the risk of loss by using a model to make financial decisions” and identified several forms of model risk. They identified the following types of model risk: 1) inapplicable model, 2) incorrect model, 3) correct model, incorrect solution, 4) correct model, inappropriate use, 5) badly approximated model, 6) software and hardware bugs and 7) unstable data. The reader is directed to this paper for the details of each type of model risk. However, the meaning of each type of risk should be fairly intuitive. The paper also goes into considerable detail enumerating the signs a model may be incorrect. For example, the modeler may not have considered important factors in the design of the model or the model may be correct only under ideal conditions, which rarely present themselves.

Insurance companies might borrow a page from banking to establish a formal model validation process for vetting company models. In banking, a model validation group is a group of interdisciplinary academics and banking professionals familiar with the company’s products and business functions that convenes to vet proposed models before they are presented to senior management. The model validation team, by design, is an interdisciplinary team of professionals who can assess the impact of the model on all aspects of a company’s operations, from its distribution channels to its marketing and underwriting departments and processes. The rigorous nature of the validation process is critical to mitigating model risk by identifying weaknesses in models and recommending remedies to increase the likelihood of their company adoption or recommending the nonadoption of models that could adversely harm the company financially. This can be an unpleasant experience for the modeler, but the continued health of the organization is the paramount concern to all involved in the model validation process.

PHASE 9: SYSTEM INTEGRATION

It probably does not come as a surprise that you will need to build a model to test the implementation of your predictive model by the company IT department. The testing of the implementation needs to include enough scenarios to ensure the model behaves as expected once in production. Otherwise, a very soundly constructed model could get a “bad rap” because IT implementation failed to properly operationalize it. In the testing of the IT implementation, don’t ignore even the smallest of discrepancies. A seemingly immaterial difference could yield unexpected results once a model goes into production and attempts are made to evaluate a combination of policy data not represented in one of your modeling test scenarios.

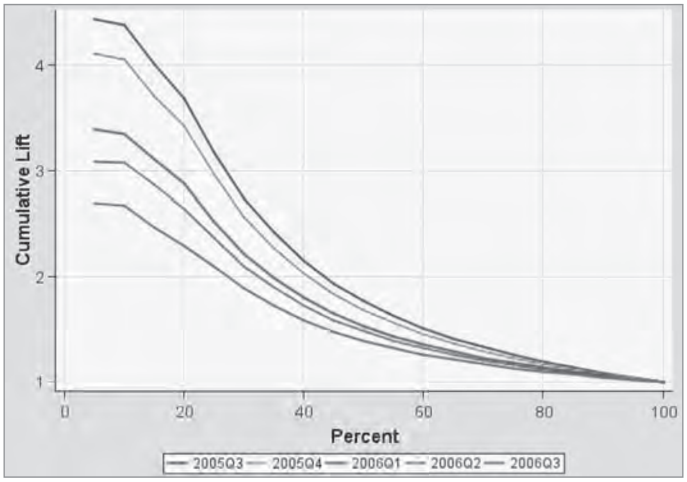
Production models should be tested for their ability to replicate the results of all test scenarios, which should include simple and complex test cases as well as boundary or extreme cases. It can’t be stressed enough that the importance of models accurately replicated the simple cases. Models quickly lose credibility with end users and senior management if they fail at replicating simple cases, casting doubt on results for more complex cases. End users then become engaged in scrutinizing model results rather than looking for emerging risks that may challenge the profitably of the organization. When underwriters, for example, spend an inordinate amount of time trying to disprove model results they don’t trust, they are engaging in the wrong kind of behavior for the organization. The simple truth is if they don’t trust the results, they are not going to use them to make underwriting decisions anyway. The time spent earning “scout badges” every time they disprove a result from the model could have been better spent on behalf of the organization looking for emerging risks. This is a prime reason the model validation exercise is so important. The better the interdisciplinary review of the model and the testing of its IT implementation, the higher the confidence level around the organization for the model and the greater its utilization in decision making.

PHASE 10: DEVELOP MONITORING METRICS

Monitoring metrics are used to assess the continued fit of the model as new data emerges and the business environment changes. If model results are not as expected and/or major distributional shifts from modeled data present in emerging data, then it is time to consider whether the model requires a “refresh” or a “rebuild.” Minor distortions may necessitate only a model refresh. A model refresh is performed by running the same model against an updated modeling data set to update model parameters. Major distortions necessitate a complete overhaul of the existing model, which includes developing an entirely new data set based on new model predictor variables. Some of the old variables may still apply, but the degradation of your model is a suggestion they are failing to capture new signal-affecting business metrics.

Chu et al. (2007) discuss many best practices for monitoring predictive models once they have been installed into production. They discuss developing performance thresholds and the automation of the periodic generation of performance metrics to identify when models are underperforming. A key performance degradation tool the authors discuss is the model degradation lift chart exhibited below. A lift chart measures how well predicted values line up with actual values. In this chart, the model is run quarterly to examine how the lift changes over time. One could run the analysis at a frequency greater than quarterly depending on the volume of new data likely to be available at that frequency. Gains charts and ROC curves are other types of visual aid that can be useful in identifying model degradation. Rerunning the model on new data at some desired frequency and measuring the changes in parameter estimates is also insightful in measuring the continued effectiveness of your model.

Model Degradation Lift Chart



Copyright © 2007. SAS Institute Inc. All rights reserved. Reproduced with permission of SAS Institute Inc., Cary, NC, USA

CONCLUDING REMARKS

In any organization, there are hunters (those who get the business), gatherers (those who prepare data related to the business) and scavengers (those who consume and analyze the data). Sound data is the foundation of a sound analysis. Senior management relies on analytics to make decisions that are in the best interest of the company. The processing of data for new and inforce business needs constant review and oversight from those who analyze company data. The data is most meaningful to those who consume it for analysis and decision making, and they are in the best position to inform the controls around its collection and accurate recording. Building a predictive model will waste the efforts of company talent and lead to faulty decision making if modeling data is flawed. Stay cognizant of the 80–20 rule: Modeling is 80 percent data construction and 20 percent statistical model construction. Short-changing the investment in data improvement will lead to suboptimal model building and decision making by senior management. ■

Dorothy L. Andrews, ASA, MAAA, is a consulting actuary with Merlinos & Associates, Inc. She can be contacted at *dandrews@merlinosinc.com*.

ENDNOTES

[1] Morris H. DeGroot, "A Conversation with George Box," Statistical Science, vol. 2, no. 3 (August 1987), 239–258.

REFERENCES

Box, George E. P. 2013. *An Accidental Statistician: The Life and Memories of George E. P. Box*. New York: John Wiley & Sons.

Chu, Robert, David Duling and Wayne Thompson. 2007. "Best practices for managing predictive models in a production environment," SAS Global Forum 2007, Paper 076-2007, 10 pp.

Derman, Emanuel. April 1996. "Model risk," Goldman Sachs Quantitative Strategies Research Notes, United Kingdom.

Larson, Anders. December 2016. "Creating a useful training data set for predictive modeling," SOA Predictive Analytics and Futurism Section Newsletter, Issue 14, pp. 32–34.

Rud, Olivia Parr. 2001. *Data Mining Cookbook: Modeling Data for Marketing, Risk, and Customer Relationship Management*. New York: John Wiley & Sons.