January 13, 2013

StatSoft White Paper

The Predictive Data Mining Revolution in Scorecards:
Accurate Risk Scoring via Ensemble Models

Summary

Predictive modeling methods, based on machine learning algorithms that optimize model accuracy, have revolutionized numerous industries, from manufacturing through marketing to sales. However, credit risk scores in the financial industries are still built using traditional statistical models to derive “score-cards” where additive factors such as specific age, income, etc., make up the final credit score, and determine the final decision of credit worthiness. This paper describes successful approaches of how modern predictive modeling algorithms can be used to build better credit risk models and scorecards without sacrificing some of the critical benefits from applying traditional scorecards such as reason scores (for denial of credit).

How Traditional Scorecards Are Built

The process of building a credit scorecard usually follows three steps:

- Data preparation and predictor coding
- Model building using logistic regression, and model validation
- Final deployment

These steps are briefly described later. They are also described in some detail at the StatSoft Electronic Statistics Textbook, at http://www.statsoft.com/textbook/credit-scoring/.

A Common Scorecard

A final scorecard may have the following structure and look like this:
The most important columns in this spreadsheet are the first, second, and last ones. The first column identifies each predictor variable (e.g., Value of Savings).

The second column shows the specific interval boundaries or (combined) categories for categorical predictors. For example, Value of Savings < 140, Value of Savings between 140 and 700, and so on.

The last column shows the credit score “factor” to be used for applicants that fall into the respective interval. For example, an applicant with a Value of Savings between 140 and 700, a score value of 95 is added to the total credit score. If the Amount of credit is between 6,600 and 10,000, a score of 86 is added, and so on.

For each credit applicant, a total sum credit score can be computed based on the respective categories that person falls into; critical cutoff values are then used to determine the final credit decision.

The analytic workflow to get to the final scorecard follows a logical procedure to generate a table such as that shown above.

**Data Preparation, Predictor Coding**

After available data have been retrieved and “cleaned” (e.g., bad codes have been removed, missing data have been replaced, duplicates were deleted, etc.), the first step is to recode the value ranges for the continuous predictor variables such as age, average account balance, and so on into value ranges.
TheSTATISTICA™solution platform provides different methods to accomplish this coding, ranging from manual/visual methods (graphical methods) to fully automated optimal binning to maximize the information of the resulting binned variable for modeling credit default probability.

Model Building, Logistic Regression

The next step is to build a linear model that relates the coded categories to credit default risk. The statistical method that is used for this purpose is Logistic Linear Regression. The details of this method are described in the StatSoft Electronic Statistics Textbook (www.statsoft.com/textbook/).
In short, a simple best linear equation is estimated and scaled so that, for example, the odds of credit default double for every 10 points' decrease in the resulting predicted score.

Final Deployment

The final deployment of a scorecard can be accomplished “manually” by adding the scores for the different categories for each variable describing a credit applicant. Typically, in larger banks the scoring is, of course, done automatically through systems such as STATISTICA Enterprise Server™ or STATISTICA Live Score® for real-time scoring. Also, along with the final credit score, typically a final credit decision is returned by comparing the credit score to appropriately chosen cutoffs. Further, for rejected credit applications, the reasons for the rejection (e.g., those predictors whose respective scores were below the norm or average) can be provided to the applicant.
Advantages of Common Scorecards

The types of credit scorecards briefly described here have several convenient properties that explain their popularity.

Final credit scores are the simple sum of individual factors determined by classifying applicants into specific bins. It is easy to see what specific factors were most beneficial for a favorable credit decision or detrimental and leading to a rejection of a credit application.

In addition, because the basic statistical model is linear, it is easy to understand the “inner workings” or “behavior” of the credit risk model, in the sense that in most cases simple summaries can be derived such as “the higher the education of an applicant, the lower the credit default risk” and so on.

Predictive Modeling Approaches

So why are some progressive financial institutions and banks using other, perhaps more complex, approaches to credit scoring? The answer is simple: predictive modeling methods often result in more accurate predictions of risk. With more accurate predictions of risk, more credit can be extended to more applicants, while reducing or not increasing the overall default rate. Thus, the accuracy of the model to predict credit default has a direct impact on profits.

Accuracy of Scorecards, and Profitability

Shown below is a summary evaluation of different predictive modeling scorecards compared to the existing traditional scorecard for a large StatSoft customer.
This graph (based on actual data) shows the enormous impact that a high-quality, accurate model of credit risk can have on the bottom line:

Given the *Baseline* model, approving approximately 55% of applications for credit resulted in a little above 10% of eventual credit defaults. In this graph, the profitability of that *status quo* model and solution is pegged at 100% profit.

Moving to the right, different cut-off values for credit approval are shown for a powerful ensemble predictive model built via *STATISTICA*. Depending on where the credit dis/approval threshold is set, given a predicted probability of default, more credit can be approved (e.g., for almost 85% of all applicants in the *High Approval Rate* model), while either decreasing or maintaining a constant default rate.

The effect on the profitability is dramatic: over 80% greater profit resulted for all cutoff values considered. In short, and as expected, the more accurately a creditor can predict the risk of credit default, the more selective that creditor can be to extend credit only to those who will pay back the money with all interest as agreed.

**How Should Accurate Risk Models Be Built?**

In short, the answer is: by using algorithms that can find and approximate any systematic relationships between predictor variables and credit default risk, regardless if those relations are nonlinear, highly interactive (e.g., different models should be applied to different age groups), etc. Such algorithms are called “general approximators” because they can approximate any relationship—regardless of how complex it might be—and leverage those relationships to make more accurate predictions.

**Ensemble Models, Voting Models**

The algorithms that have proven to be very successful and that are refined in the *STATISTICA* platform are those applying voting (averaging) of predictions from different “tree-models,” or by **boosting** those models. **Boosting** is the process of applying a learning algorithm repeatedly to the poorly predicted observations in a training sample to represent correctly all homogeneous subgroups, groups of “different” and difficult-to-predict applicants, outliers, etc.

For example, in a recent international competition based on a credit risk scoring problem (the international 2010 PAKDD Data Mining Competition), the winning (most accurate) solution was computed using the *STATISTICA Boosting Trees* (stochastic gradient boosting) algorithm. The paper describing the winning solution is available from StatSoft.

The details of effective predictive modeling algorithms are often complex. However, efficient implementations of these algorithms on capable hardware, such as the implementations through the *STATISTICA Credit Risk* solutions, will solve even difficult problems against large databases in very little time.
The Dreaded “Black-Box” and How to Open It

There is little controversy around the notion that predictive modeling techniques as implemented in the STATISTICA solutions will outperform the accuracy of statistical modeling approaches in most cases. However, a common criticism that is raised in this context is about the concern that these models are “black-boxes,” making it nearly impossible to provide insights into the reasons for specific predictions, or the general mechanisms and relationships that drive the key outcomes such as credit default.

On the surface, these concerns are valid and perhaps sufficiently important to continue with the status quo and common scorecard modeling approaches. However, there are good solutions to address this issue of lack-of-interpretability and reason scores, and all of them are easily implemented through the STATISTICA solutions.

What-if (scenario) analysis. In most general terms, for every prediction that is made through a black-box predictive model, the computer can run comprehensive what-if analyses to evaluate what the credit default probability prediction would have been, if for each predictor different values had been observed. So for example, if an applicant for credit is 18 years old and credit is denied, the computer can quickly evaluate what the credit decision would have been if the applicant had been 20 years old or 30 years old.

Applying this basic approach systematically to every prediction or group of predictions will provide detailed insights into the inner workings of even the most complex models, and into the specific predictors and their interactions that are responsible for a specific credit decision.

Two-stage modeling. There are other and perhaps simpler approaches that can also be used to arrive at interpretable models and predictions. One can use the predictions of the most accurate black-box model as a starting point, and drill down into the model to identify the specific interactions and predictor bins that were important and drove the risk prediction (for example, using what-if scenarios or the various standard results around predictor importance available through STATISTICA). With that knowledge, better and more comprehensive common scorecards can then be built that include the relevant predictor interactions, pre-scoring segmentation of homogeneous groups of applicants, specific binning solutions, etc., identified through the first predictive modeling step.

Deploying Predictive Models

Risk models derived via predictive modeling tools are best deployed in an automated computer scoring solution such as STATISTICA Enterprise Server™ or STATISTICA Live Score®. In fact, the STATISTICA Enterprise Decisioning Platform® will manage the entire modeling lifecycle – from model development to single-click deployment for real-time or batch scoring with version control and audit logs for regulatory compliance.
Conclusion

Risk modeling is an area where the application of advanced predictive modeling tools implemented in the STATISTICA solution platforms can produce significant and rapid return on investment. The STATISTICA platform includes algorithms such as boosted trees, tree-nets (or random forests), and automatic neural network ensembles, to mention only a few of the very powerful and proven-to-be-effective implementations available.

When combined with the ease and comprehensiveness of features for model lifecycle management and integration with business rules and logic available in the STATISTICA Decisioning Platform, there is little reason not to adopt these methods to realize the revenue associated with more accurate risk prediction and to become more competitive in the complex financial services domain.

Further Reading:


For More Information:

StatSoft, Inc.
2300 East 14th Street
Tulsa, OK 74104
USA

www.statsoft.com

StatSoft Copyright 2013