LityxIQ

MODELING APPROACH

About Lityx Models

All models are produced using LityxIQ, a powerful, user friendly modeling platform developed by Paul Maiste Ph.D. Statistics NC State, President and founder of Lityx LLC. LityxIQ makes modeling easier and more accessible to a wider audience while also making the modeling process faster and the models more predictive. It utilizes best in class algorithms and approaches, and tightly integrates model building with model implementation for ease of deployment.

Lityx Models are built to be successful for business applications such as acquisition, retention, value, and risk modeling. The focus is on key business metrics, model fit and stability and not just the traditional statistical measures so common to other modeling tools and providers.

Some of the unique differences of a Lityx model built in PredictIQ include:

Model Selection Criterion - Analysis of candidate predictors and final model selection is based on The Akaike Information Criterion (AIC) where the “best” model is the one with the lowest AIC value. AIC is founded on information theory and picks the best model from a set of all possible models that provides the best fit to the data with the fewest number of predictors. This is important because traditional model selection rewards models with more predictors and leaves it up to the modeler to decide the right number at the risk of having too many or too few predictors and resulting in the models that do not perform well outside of the data they were built on.

Model Development Process - Model development involves two separate stages that in combination lead to accurate estimates of model performance while providing the best possible model back to the user. Stage 1 computes estimated final model performance metrics using the chosen validation technique such as holdout or cross-validation. Models are built during this stage using all model settings, but the resulting model(s) are used only to estimate in-market performance of the model. In Stage 2, the final model is trained using all the model settings and all available data. The resulting model is returned to the user in the form of model coefficients, trees, and other output relevant to the algorithm(s) selected. It is also the model that is used for subsequent scoring jobs. This approach makes optimal use of all available data and can lead to stronger models.

Variable Binning – This is the process of transforming a numeric predictor into a series of categorical ones as well as re-grouping and consolidating categorical predictors. There are many benefits to this.

- Increases model stability: some characteristic values rarely occur, and will lead to instability if not grouped together.
- Improves quality: grouping of similar attributes with similar predictive strengths will increase model accuracy.

- Non-linear relationships can now be modeled using a linear relationship which is important for logistic regression modeling
- Prevents model from including reversal patterns and extreme values.
- Prevents overfitting(too many variables) possible with numerical variables not binned
- A model with binned variables is easier to read and interpret

Insignificant Predictors – Most models that utilize a step-wise variable selection process make the decision to include or exclude a predictor based on the p-value of that predictor. A p-value tell us about the probability of a predictor being significant. However, the use of p-values can be misleading. Predictors that are collinear (have high correlation), will show up as non-significant and be left out of the model even though they may have important information to contribute. Because LityxIQ utilizes AIC criterion and binning you may see a single bin of a significant predictor show up as insignificant. A Lityx model is not seeking to make a purely "statistical" fit but rather one based on the overall fit and the size of the model. As a result some purely statistical tests and p-values that require purely statistical assumptions may show up as insignificant. These represent no risk to the performance of the model and in fact demonstrate the difference in the Lityx approach.

Concerns With Multicolinearity – You should not be concerned about this with a LityxIQ model for two primary reasons: (1) There exists an autocorrelation feature that drops all but one predictor among a group of highly correlated variables prior to model building. There are adjustments to the tolerance that can be made if you wish to be more conservative, but the default settings are .75 for numeric predictors and .99 for categorical predictors. (2) Multicolinearlity does not affect overall model performance only the coefficients of the variables involved. See Wikipedia comment on this.

As stated on Wikipedia:

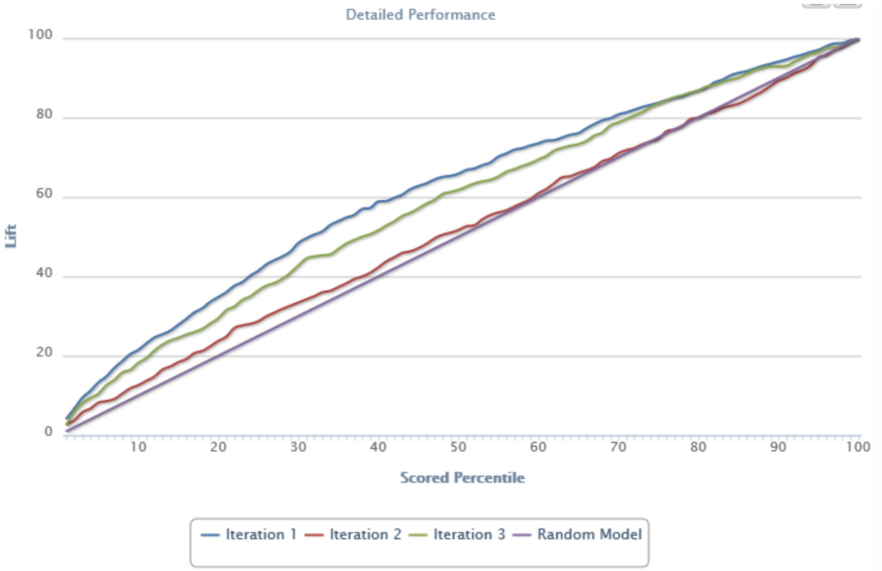
"Multicollinearity (also collinearity) is a statistical phenomenon in which two or more predictor variables in a multiple regression model are highly correlated, meaning that one can be linearly predicted from the others with a non-trivial degree of accuracy. In this situation the coefficient estimates of the multiple regression may change erratically in response to small changes in the model or the data. Multicollinearity does not reduce the predictive power or reliability of the model as a whole, at least within the sample data themselves; it only affects calculations regarding individual predictors. That is, a multiple regression model with correlated predictors can indicate how well the entire bundle of predictors predicts the outcome variable, but it may not give valid results about any individual predictor, or about which predictors are redundant with respect to others."

Variable Contribution – There are multiple ways to estimate the incremental contribution each variable in the model is making, but none are perfect. What we recommend is directionally accurate and representative.

RECOMMENDATION: For each variable in the model use the bin with the highest absolute value of the test statistic then determine the percent of total this represents after summing across all of the model variables. To clarify the model coefficients may include multiple entries for each variable due to binning so we are talking about only using the highest valued bin to represent that variables contribution.

One issue with using standardized coefficients for this type of thing is that they are always based on "incremental" predictive power… how much the variable improves prediction over and above everything else. That isn't the same as individual predictive power, more and more so if there is high colinearity.

Multiple Models "Bake-off" – LityxIQ is able to easily build multiple models simultaneously using algorithms such as logistic regression, CHAID, Cart, Neural Net, Probit, Random Forest etc. The benefit is no single algorithm works best for every business problem and dataset. By having access to multiple solutions you can compare the results side-by-side and go with the option that is providing the most value.


Detailed Performance
Lift
Scored Percentile
Iteration 1
Iteration 2
Iteration 3
Random Model