2014 Workshop on Mathematical Problems in Industry
Probability of Default Model Performance Measurement

Executive summary of results from the Standard & Poor’s team at the 30th Annual Mathematical Problems in Industry Workshop in Newark, NJ, June 23-27, 2014. Group participants: JiaMing Chen, Sunil Dhar, Dean Duffy, Yiheng Liu, Richard O. Moore, Michael Pedneault, Andrew Pole, Yuzhou Qian, David Rumschitski, Ting Wang, and Maxim Zyskin. This executive summary complements a presentation of results given at the MPI Workshop on June 27, 2014, available online at http://web.njit.edu/~rmoore/MPI2014. A full report will be delivered electronically to William Morokoff and Liming Yang, Standard & Poor’s, in addition to being posted on the website above.

The MPI2014 Standard & Poor’s team recommends:

1. **Accounting for how the data set influences the ability of standard performance measures to accurately rate models of company default likelihood.** Statistical measures to compare historic data on company defaults with model predictions based on likelihood or goodness of fit, or graphical measures such as the area under curve (AUC) or accuracy ratio (AR), are sensitive to the distributions of defaulters and non-defaulters that are available in the data. In particular, if these distributions are well separated, all models look good and if they overlap significantly, all models look bad. The team proposes quantifying this inherent observability of the data using the Kullback-Leibler divergence (KL) to measure an effective distance between the sample density conditioned on default and the sample density conditioned on non-default. To quantify each data sample the team used both exact analytical expressions for its KL and a sample-based method that does not require computation of the underlying distributions. The team proposes using the KL distance in one of the following ways:

   - to provide a separate statistic expressing confidence level in the model performance measure used (e.g., AR);
   - to provide an adjusted model performance measure that appropriately discounts AR when KL is large, indicating very observable data, or augments AR when KL is small, indicating indistinguishable data; or
   - to use the KL of the data set to construct an interval, e.g., \([AR - \epsilon, AR]\) with \(\epsilon = \frac{1}{\kappa} \frac{KL-1}{KL+1} > 0\) (high observability) or \([AR, AR-\epsilon]\) with \(\epsilon < 0\) (low observability), that replaces a single model performance rating with a range of performance ratings in which the true model performance lies. Analysis of multiple data sets would produce many intervals, the intersection of which could be taken as the final interval expressing model quality. Requiring that such an intersection be nonempty is one possible criterion for determining the empirical tuning parameter \(\kappa\).
2. **Accounting for correlations in default probabilities.** Using regression to a logistic model with artificial default data, the team showed that inclusion of covariates significantly improved agreement between model output and the data.

3. **Improving goodness-of-fit measure for regression models when discrete and continuous covariates are present.** A standard statistical test used to accept or reject a proposed discrete predictor variables regression model as appropriate to a given data set is the scaled sum of squared differences between data and model output, known as the Pearson $\chi^2$ test. If the regression model contains only continuously varying explanatory variables, a related test known as the Hosmer-Lemeshow test is appropriate for model evaluation. Neither of these tests is appropriate to mixed discrete and continuous explanatory variables for regression evaluation, however, and the team consequently proposes use of a recently developed mixed-parameter test developed by Pulkstenis and Robinson.