Technical Note – EFL Modeling Methodology

July 2012

The purpose of this note is to detail the statistical modeling techniques used by EFL. We first outline the shortcomings of the traditional modeling approaches for credit scoring in the context of psychometric-based scorecards, namely small samples, large numbers of explanatory variables, and multi-country data. We then present a modified approach: EFL's Bayesian Hierarchal Model, and explain both its conceptual advantages and its technical specification. Finally, we present some results illustrating the superiority of this approach.

WHY A UNIQUE APPROACH IS NEEDED

Traditionally, credit scorecards are built using accumulated historical loan applications and pre-existing secondary data. There have been thousands upon thousands of people who have completed fields on loan application forms such as entering their age, gender, education, industry, and a host of other traditional questions. Those individuals have also provided identity numbers linked to databases that in some cases contain records of past loan performance, credit bureau information, court judgments, and so on. These large datasets of many observations, combined with relatively few explanatory variables, give model builders significant statistical degrees of freedom to build scorecards. The most common technique used on this data are logit regressions using traditional maximum likelihood estimation (see Credit Scoring Toolkit by Anderson 2007).

Unlike traditional credit scoring based on archival data from years and years of past applications, selection based on psychometric tools requires gathering additional new information. This is because unlike building a model based on typical socio-demographic characteristics, psychometric questions have not been asked on past applications nor are clients answers present in large bureaus, and therefore psychometric information represents new data that must be collected.

There are two approaches that could be taken to collect this information: administering the new questions to samples of existing clients and comparing it to their repayment history, or administering the new questions to new applicants and comparing it to their subsequent repayment performance over time. 'Bads' (defaulting clients) typically make up a small percentage the overall client pool. So when collecting data on new applicants, a very large number must be tested before you are able to collect a sufficient number of bads for statistical significance. For example, if a minimum of 200 bads are needed and the typcial default rate for new clients is 4%, this requires testing 5000 new applicants. Depending on the flow of new

applications this could take months or years, added to which is the time needed to wait for these 5000 loans to mature.

The approach of testing existing clients has some advantages. First, unlike the case of testing new applicants, for existing clients the bads can be identified and over-sampled. Instead of testing 5000 new applicants to reach 200 bads, one could stratify and sample 200 bads and 200 goods. Moreover, using past repayment performance means the data can be analyzed as soon as it is collected, rather than waiting until the loans mature. But, the downside to this approach is that defaulting clients have a fractured relationship with the financial institution, may be subject to collections and legal action by the bank, and are therefore less likely to agree to participate by answering the new questions.

Therefore, no matter the approach to collecting this new data, when attempting to create scorecards that leverage psychometric data (or any other type of data which must be newly collected from applicants rather than pulled from old application forms or secondary sources) there will be downward pressure on sample size.

DEALING WITH SMALL SAMPLES: INDICES AND POOLING

Since each individual bank is limited in its ability to collect large datasets to build scorecards based on psychometric questions and other new questions, pooling together data across multiple countries and market segments is one way to overcome this challenge. Though it may be difficult to accumulate a sufficient number of 'bads' in one particular market or segment, collecting this information across multiple countries and segments can give a larger sample size.

However, this pooling of data across countries and markets comes at a cost. The statistical relationships between questions and the likelihood of default could differ from one country to another. This is particularly the case for psychometric questions, as meaning can be affected by language and culture. Though EFL's past work has shown a remarkable cross-country stability of the predictive power of these questions against default, it is also true that predictive power is shown to typically double or more when models are customized to country and market-specific data.

There is a tradeoff therefore between the benefits of a larger sample size and the benefits of a more customized model. Customization gives better predictive power but requires smaller samples which can harm predictive power, as by definition the greater the customization, the smaller the group being modeled. Under the traditional modeling approach, the options for combining local and global data are to either completely pool cross-country data, or else make a model exclusively with country-specific data. It does not allow for a smooth transition between the two in order to remain on the efficient frontier of the tradeoff between more data and greater customization. Instead it largely ignores country-specific information until a critical mass is collected, and thereafter it does not benefit from global data.

Another way to deal with the lack of degrees of freedom in building psychometric scorecards is to reduce the number of explanatory variables by combining items (individual questions) into psychometric indices using pre-existing formulas. When used for psychological or traditional assessment purposes, psychometric tests use a large number of questions, or items, which are combined into a particular index score. For example, 10 questions may be asked, and then those 10 answers turned into a general index measuring a psychometric dimension like 'ambition'. These indices may be interesting and relevant from a clinical or research perspective, but are not ideal for credit scoring. Using these indices presupposes combinations of items into an index that is neither context-specific nor based on actual data on entrepreneurial default risk. For this reason when measuring psychometric indices for research, selection, or clinical purposes in psychology, extensive adaptation and norming of the indices is necessary. Creating predictive models at the item rather than index level would be much more appropriate for credit scoring, but is not possible under traditional logit regression unless extremely large samples are available- there are just too many items.

EFL Bayesian Hierarchical Logit

A Bayesian Hierarchical Logit model is uniquely well-suited to solve the small sample size problem, taking advantage of item-level rather than index-level data and pooling customized country-specific data with global data. At the lowest level, the model is similar to a classical logit, but with more flexibility than is traditionally allowed. Intuitively, the model assumes that the outcome varies by country and by market segment within each country, and that the relationship between covariates and the outcome vary by country. Rather than using pre-existing aggregations of items or only a subset of them, a Bayesian Hierarchical Model's resistance to overfitting allows each item in the questionnaire to have a unique effect.

To allow this level of flexibility, the hierarchical model imposes a second-level prior on the unknown parameters. This second-level prior shares information across parameters, depending on the amount of information available from different sources. For instance, in the final equation the estimated item-level effect is a weighted combination of the average global effect and the effect estimated from the data available at a particular country and segment. As more data arrives, more weight is placed on the latter. Likewise, the prior on coefficients shrinks the estimate effects towards a common effect estimated for all coefficients. The hierarchical model partially pools information to avoid overfitting.

Moreover, to balance between a country-specific and global model, the second-level prior assumes each coefficient comes from a common distribution across countries. This allows the model to smoothly transition from a global model to a country and market-specific model as data arrives. The rate of this transition depends on the estimated similarity in the model across countries and the available data. For items that behave similarly across countries, the global model, which pools all information and is therefore more precisely estimated, dominates. But for covariates with significant heterogeneity across countries, the country-specific coefficient begins to dominate quickly as data arrives. This partial pooling of country-specific data with a global model of default risk is a central benefit of Bayesian models and significantly improves hold-out sample predictive performance, particularly when limited arrears data is available for new implementing countries.

TECHNICAL SPECIFICATION

Let y_i be a binary variable indicating whether loan i is in arrears for specified period, for instance more than 90 days. As in a logit model, the probability of default p_i and observed default y_i are modeled as random process following

$$p_i = \text{logit}^{-1}(\alpha_{\text{partner[i]}} + \gamma_{\text{branch[i]}} + \sum_{j=1}^{J} \mathbf{x}_{\text{ij}}^{'} \boldsymbol{\beta}_{\text{j,partner[i]}})$$

$$y_i \sim \text{Bernoulli}(p_i)$$

This model is a mixed or vary-intercept and varying-slope logit. The probability of default depends on a partner and branch effect, α_c and γ_b, and J sets of controls, indexed by j, whose relationship with default risk, β_{jc}, varies by partner c.

Because both the slopes and intercepts of the model vary across partners, and because of the large number of covariates, the number of unknown parameters is large relative to the available data. Estimating this model using classical methods, such a marginal maximum likelihood, would therefore significantly over-fit the data. To address this, the EFL model uses a hierarchical prior and Bayesian estimation to impose a structure on the parameters that borrows information across partners and covariates.

The EFL Hierarchical Logit includes a level-2 model:

$$\alpha_c \sim N(\mu_\alpha, \tau_\alpha)$$

$$\gamma_b \sim N(0, \tau_\gamma)$$

$$\beta_{jkc} \sim N(\mu_{\beta jk}, \tau_{\beta j})$$

$$\mu_{\beta jk} \sim N(\lambda_j, \eta_j)$$

The model is completed with independent, weakly-informative, Normal and inverse-Gamma priors on the remaining parameters, μ_α , τ_α , τ_γ , $\tau_{\beta j}$, λ_j and η_j.

This hierarchical structure has several key features. The τ_α and τ_γ precision parameters capture how similar partners and branches are in terms of average default risk. They govern how quickly high default rates for a particular partner or branch will outweigh the global estimates of default risk. Next, $\mu_{\beta jk}$ captures the global effect of variable k in group j on default risk. While the actual effect, β_{jkc}, varies by partner c, the global effect does not. The global effect dominates when limited information is available at the partner level. As data arrives, the model specializes to better fit the partner but balances this specialization with global information. The rate of transition is governed by $\tau_{\beta j}$, which captures how much variation across countries is typical for covariates in group j. Finally, λ_j and $\eta_j$govern how much variation there is across coefficients of particular type j. If the global precision η_j is large, then item-level responses are strongly shrunk towards the common effect λ_j. This allows the use of item-level data rather than arbitrary aggregates, while guarding against overfitting. All these parameters, except the top most prior parameters, can be estimated from the data using Bayesian methods.

The primary goal is to estimate the default risk p_i. This is captured by posterior distribution of default risk, $p(p_i|\,\boldsymbol{y},\mathbf{X})$, which integrates over all unknown parameters and conditions on the observed data. Given the model above, the posterior distribution is

$$p(p_i|\,\mathbf{y},\mathbf{X}) \propto \int \text{logit}^{-1}(\alpha_{\text{partner[i]}} + \gamma_{\text{branch[i]}} + \sum_{\text{j=1}}^{J} \mathbf{x}_{\text{ij}}^{'}\,\boldsymbol{\beta}_{\text{j,partner[i]}})$$

$$\cdot \prod_{i=1}^{N} p\left(y_i \middle|\, \mathbf{X}_{\text{i}}, \alpha_{\text{partner[i]}}, \gamma_{\text{branch[i]}}, \boldsymbol{\beta}_{\text{partner[i]}}\right) \cdot \prod_{c=1}^{C} p(\alpha_c|\mu_\alpha,\tau_\alpha) \cdot \prod_{b=1}^{B} p(\gamma_b|\,\tau_\gamma)$$

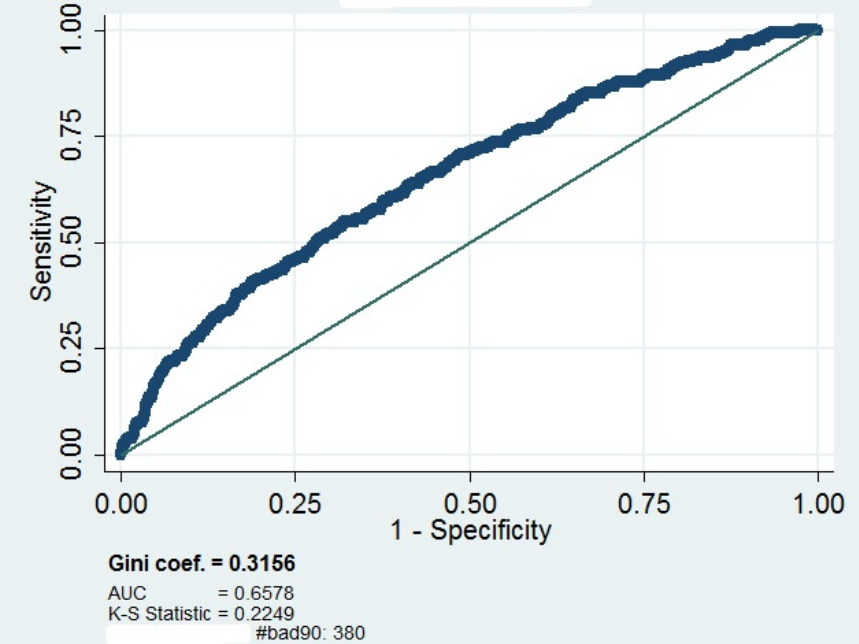
$$\cdot \prod_{j=1}^{J}\prod_{k=1}^{K_j}\prod_{c=1}^{C} p(\beta_{jkc}\,|\,\mu_{\beta jk},\tau_{\beta j}) \cdot \prod_{j=1}^{J}\prod_{k=1}^{K_j} p(\mu_{\beta jk}\,|\,\lambda_j,\eta_j)p(\mu_\alpha,\tau_\alpha,\tau_\gamma,\tau_{\beta j},\lambda_j,\eta_j)d\boldsymbol{\theta},$$

which follows from Bayes rule and basic rules of probability. This model can be estimated using Markov Chain Monte Carlo methods (see, for instance, Bayesian Data Analysis by Gelman et al. 2003).

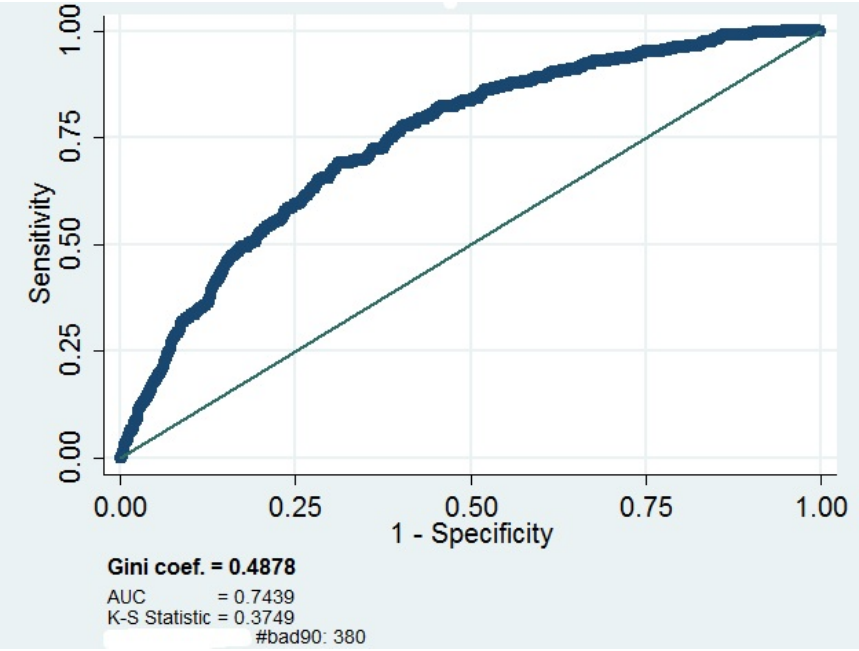
RESULTS

Here we illustrate the improved performance from this modeling approach, using data from one active EFL country.

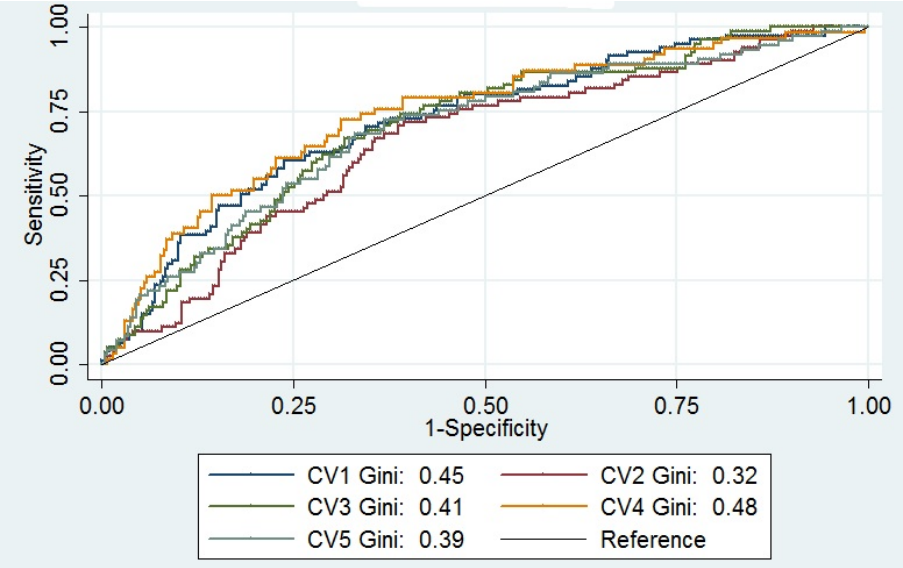
First, we show the AUC curve on a model built using traditional logit regression, with items reduced to indices so that the model can converge to a solution, and based only on local country data. Below is the ROC curve, with area under the curve (AUC) and Gini displayed. This sample includes 380 bads and 2160 goods. The in-sample Gini coefficient is 0.32 (AUC of 0.66).



Taking the same data, but building the model using the EFL Bayesian Hierarchical Logit allows us to both move to the item level rather than use indices, as well as to partially pool these 2540 local observations with another approximately 9000 observations from other countries. The resulting ROC curve is shown below.



The EFL Bayesian Hierarchical Logit achieves a Gini of .49 in-sample, an increase of over 50% from the traditional Logit. Given that this model is run on only 2540 observations (380 bads) and features over 200 explanatory variables, there would typically be a large overfitting effect from traditional estimation. However, below we show that even when the model is built on a randomly selected 80% and tested on the 20% hold-out, the results remain superior to the traditional method's in-sample result. The traditional method would also degrade out of sample increasing this difference further. Performing this build-test experiment five times yields an average Gini coefficient of 0.41.



Results from other countries show the same result: the EFL Bayesian Hierarchical Model produces more accurate predictions and more stability out-of-sample. It maximizes the amount of information being used by pooling cross-country data and smoothly customizing model coefficients to each country as country-specific results emerge. Moreover, it uses the individual application question responses rather than combining them into psychometric indices using generic formulas. In addition to offering superior predictions, this provides greater flexibility in terms of scorecard optimization, because individual questions can be included or dropped based on their statistical power, rather than having to include or drop the entire block of questions used for a particular index. Finally, not using indices bypasses the necessity for cultural adjustment of scale norms (and the regulations governing those adjustments).

For more information, please contact info@efinlab.com.