Cross-Tab Weighting for Retail and Small-Business Scorecards in Developing Markets

Dean Caire and Mark Schreiner (Microfinance Risk Management L.L.C.)

19 February 2013

Introduction

This paper presents a simple technique for building predictive models that we call "cross-tab weighting". The point weights in the scorecards come directly from cross tabulations of each individual risk factor with loan status ("good" or "bad"). Cross-tabulation is the only concept required to build and understand the models.

We use cross-tab weights to build application credit-scoring models for retail and small business segments. Their predictive power (measured by "AUC") compares well with more popular—and more complex—methods such as Logistic Regression. More important to developing markets, cross-tab weighting facilitates the transfer of skills for model development and maintenance to users with limited experience and knowledge of predictive modelling and statistics.

In this paper, we explain how to build and evaluate cross-tab weighted scorecards. We present examples for three portfolios from developing markets. We purposely use as little math as possible to facilitate communication of the key concepts to professionals with a wide range of specialties and educational backgrounds. Likewise, we use the term "cross-tab weights" because this directly and intuitively communicates to business users what the model is and how it works, unlike more traditional terms for this approach, namely "Naïve Bayes" (Antonakis and Sfakianakis, 2009) or the "Independence Model" (Aspey, Hinder, and Lucas, 2003).

This general approach reflects the belief that the main challenge in the implementation of predictive modeling in a business is not ensuring the predictive power of the scorecard but rather designing a process and a product that helps convince users to accept the organizational and operational changes involved (Edelman, 1992). Educating users and aligning their expectations with what scoring can do is central (MacNeill, 1999). The simpler the scorecard, and the more involved the users in its creation, the more likely it is that the scorecard will actually be used.

Cross-Tab Analysis

A cross tab (also known as a contingency-table analysis) is a two-dimensional table that shows the number of good and bad contracts for loans with different risk characteristics.

Please direct any questions to: dean_caire@dai.com

The first step in preparing cross tabs for credit-risk modeling is to define a "bad" loan and assign a good/bad status to each loan in a given set of portfolio data. The definition should describe a loan that the financial institution would prefer to avoid making in the future. For example, a "bad" loan for a bank is often one that reaches 90 days of arrears. In contrast, a microlender may define "bad" as reaching 30 days of arrears.

Table 1: Cross Tab of Gender of Borrower with Good/Bad Status

	Women	Men	Total
# Goods	52	40	92
# Bads	3	5	8
# Total	55	45	100
Bad rate (%)	5.5	11.1	8

Table 1 is a cross-tab for the risk factor "Gender of Borrower" with good/bad status. The top row labels the risk characteristics, called "bins". For gender, the two bins are "Women" and "Men". The "# Goods" is the number of good contracts in each bin, "# Bads" the number of bad contracts in each bin, and "# Total" the total number of contracts in a bin. The focus of our analysis is the row "Bad rate". For a given bin, the *bad rate* is defined as the number of bads (# Bads) divided by the number of total contracts (# Total). A higher bad rate indicates higher risk.

Cross-Tab Weighting

Cross-tab weighting simply means using the bad rate for a given bin as its points in a predictive model. For the example in Table 1, a male borrower would receive 11.1 points and a female borrower would receive 5.5 points, because in the model development data, 11.1 percent of contracts with men were bad and 5.5 percent of contracts with women were bad.

To keep things simple, we focus only on bad rates and the differences in bad rates across bins. We do not discuss concepts such as the odds ratio, although these could be calculated from the cross-tabs, as discussed below.

Building a Predictive Model with Cross-Tab Weighting

Single-factor models

In this paper, we will not discuss the very first and sometimes most challenging step in predictive modelling – preparing the data set – but assume that we have a clean data set from the database manager. In this case, the first step in building single-factor models is to create and analyze cross-tabs of good/bad status with each potential risk factor.

For categorical factors (for example, gender, industry, and legal form), this process has been illustrated in Table 1. For numeric factors (such as age or years in business) or financial factors (such as income or indebtedness), we usually start by sorting the values in the data into five bins so that 20% of loans are in each bin. We examine these "quintiles" for common-sense patterns of change in the bad rate, such as consistent increases or decreases. Then we use judgment—usually the combined knowledge and experience of a working group of users with deep domain knowledge—to adjust the cut-off points that define the bins so that the trends in bad rates make sense and fit the working group's expectations. In this way, the models are intuitively appealing to the end-user and avoid over-fitting the development data.

The differences in the bad rates and the distribution of contracts across bins in the cross tab can give a good idea of a factor's predictive power. All else constant, greater predictive power is suggested by greater differences in the bad rates and a more-even distribution of contracts across bins. If the bad rates are nearly identical for most bins, or if the vast majority of contracts are clumped in one or two bins, then the factor will probably be less powerful. We also evaluate the predictive power of each one-factor "cross-tab model" more formally with the Area Under the Curve (AUC) statistic. This statistic is the area under the Receiving Operating Characteristic (ROC) curve, and it is also a simple function of the Gini coefficient.

To summarize, single factor models should balance the goals of maximizing predictive power with ensuring that the relationships in the scorecard make sense to end-users

Multi-Factor Models

Once we have a set of one-factor cross-tab models, we can begin to build a multi-factor model. The two main goals in this stage are first to maximize predictive power and second to choose a set of factors that form a comprehensive risk profile and cover the same areas that lenders look at with traditional subjective/manual underwriting systems. These are the terms of the loan, borrower demographics, and the borrower's financial ability and willingness to repay the loan as evidenced by financial statements and credit history.

We start building the multi-factor model by choosing the single factor that looks best, based on the two goals above. With only this single factor, the model's AUC is the same as the first factor's AUC. Then we add a second factor, again based on the two goals above, and look at the AUC of the two-factor model. It should be higher, and it normally will be a good deal higher unless the two factors contain similar information. We continue choosing factors in this way to assemble a group of factors that create a comprehensive and powerful risk profile. Usually, AUC for the multi-factor model rises quickly as we add the first 5–10 variables and then begins to taper off as we add additional factors; this is the "flat maximum" phenomenon (Wainer, 1976; Dawes, 1979; Baesens et al., 2003). We do not follow any rules in terms of how many factors to include, only adherence to the two goals of maximizing predictive power with a set of factors that comprehensively assess risk as an end-user would in the absence of a predictive model.

Interpreting the Model

Using cross-tab weighting—that is, points set equal to the bad rates in each bin for each risk factor—the total points, or "score", for a borrower is the sum of the points received for each risk factor in the model. Borrowers are then ranked by total score, where lower scores indicate lower risk. To evaluate model power, we create a number of risk groups (usually seven or more) by initially dividing the scores into equal intervals. Given these risk groups, we evaluate model performance with a "Good/Bad Table". This is a cross tab between loan status (good/bad) and risk groups. Table 2 is the Good/Bad Table for a Tajikistan microlender's scorecard for loans to groups of borrowers (discussed more below).

Table 2: Cross Tab of Risk Group from Multi-Factor Model with Good/Bad Status

Risk Group	1	2	3	4	5	6	7	Total
Points in Range	9.2 to 11.7	11.8 to 14.2	14.3 to 16.8	16.9 to 19.3	19.4 to 21.9	22.0 to 24.4	24.5 to 27.0	
# Goods	1,694	11,850	4,937	1,906	1,216	161	10	21,774
# Bads	1	35	32	37	96	25	3	229
# Total	1,695	11,885	4,969	1,943	1,312	186	13	22,003
Bad Rate (%)	0.1	0.3	0.6	1.9	7.3	13.4	23.1	1.1

To sum up, cross-tab weighting is useful for teaching end-users how to build complete scorecards. Of course, the approach also follows other standard modelling practices, such as breaking the data into development and validation sets. The key point here, however, is that novice users from varied backgrounds can easily learn to create, interpret, and manipulate cross tabs to create models that are nearly as powerful as more complex approaches such as Logistic regression, the standard for most credit-risk scoring applications (Thomas, 2000). In the next section, we compare these two modelling techniques using three examples.

Cross-Tab Weighting Versus Logistic Regression

For a single categorical variable, cross-tab weighting and Logistic regression models will have exactly equal predictive power. In fact, Logit coefficients for a single variable model can be calculated directly from a cross-tabulation (Hosmer and Lemeshow, 2001). We present such an example for the cross tab Gender of Borrower in Table 1a below.

Table 1a: Cross Tab of Gender of Borrower with Good/Bad Status, Odds, Odds Ratio and Logit Coefficients

	Men	Women	TOTAL
# Goods	40	52	92
# Bads	5	3	8
# Total	45	55	100
Bad Rate (%)	11.1	5.5	8.0
Good Rate (%)	88.9	94.5	92.0
Odds	0.13	0.06	
Odds Ratio (Men as reference)	1	0.46	
Logit Coefficient	0	-0.7732	

Table 1a extends Table 1 to include four new rows. The “Good Rate” is the number of good contracts divided by the number of total contracts. The “Odds” are the “Bad Rate” divided by the “Good Rate” (or the number of bad contracts divided by the number of good contracts). The “Odds Ratio (Men as reference)” is the odds of a category divided by the odds of the reference category, where the category with the highest bad rate (in this case “Men”) serves as the reference category; in Table 1a, the odds ratio for Women is 0.06/0.13= 0.46. The “Logit Coefficient” is the natural logarithm of the odds ratio. For Women, this coefficient is ln(0.46) =-0.7732. For Men, it is ln(1/1) = 0.

For multi-factor models, it is not possible to use this “long-hand” approach to calculate Logit coefficients from cross-tabs; the better predictive power of a Logit model is due in part to how it accounts for correlations between the variables in the model. However, we have analyzed many data sets using both cross tabs and Logistic regression, and we consistently find—given a set of factors—that Logistic regression is only slightly more powerful. This means that cross-tab weighting may sometimes be a better choice because of its practical advantages and in particular, the model's simplicity and transparency that allows business end-users with limited modelling experience to fully understand how the model makes predictions.

Of course, cross-tab weighting is not the best approach in all cases. As shown below, Logit models are generally more powerful, and Logit modelling would not create any particular technical issues for an organization with an experienced modeller. The niche for cross-tab weights is in organizations such as smaller banks or microlenders where limited human resources place a premium on simple, powerful models that are easy to understand and maintain.

Empirical Examples

1. Loans to groups of microborrowers in Tajikistan

The first example uses data on 31,433 contracts that were issued and fully repaid (or defaulted) over a two-year period. The loans were issued to individuals in small groups who mutually provided each other with a joint-liability guarantee. About 1 percent of these contracts (327) were bad, defined as having reached 30 days in arrears. Model development used a random sample of 70% of goods and 70% of bads, with the remaining 30% of goods and bads used for out-of-sample validation.

A working group of the microlender's business, credit-risk, and database managers made cross-tabs for each potential risk factor. This group selected the scorecard's 14 factors by adding variables one at a time in order of their subjective preference (expertise) and considering each factor's additional contribution to AUC. Cross-tab weighting with a working group of users allows the scorecard to incorporate user expectations and constraints (Thomas, Banasik, and Crook, 2001), reflecting their preferences and experiences while minimizing the sharpness of the transition from manual underwriting to a hybrid system (manual plus scoring).

The 14 factors[1] give a comprehensive view of risk across the same key dimensions as the lender normally has used to make subjective/manual lending decisions:

- Loan terms (3 factors)
- Borrower demographics (6 factors)
- Borrower credit history (2 factors)
- Financial ability to repay (3 factors)

The Logit model we compare with the cross-tab model uses the same binned factors created for the cross-tab model. If we were to use more sophisticated techniques to handle non-categorical variables with a Logit model (such as modelling a continuous variable without

[1] We do not describe the individual model factors and weights in order to focus on our main point that the performance of simple cross-tab models is similar to that of Logit models.

bins or with splines), we would expect slightly better results. But our main point would remain the same: the much simpler cross-tab weighted models have similar predictive power.

Out-of-sample, the Logit's AUC of 0.85 is about 4 percent higher than the cross-tab weighted model's AUC of 0.82. In this case, performance favors the Logit model *if predictive power were all that mattered.* As we continue to emphasize, however, the simplicity of the cross-tab approach helps end-users understand and trust the model enough to actually use it and to monitor its performance over time.

A caveat on this Tajik example is that with only 1 percent of the sample being "bad", drawing a different 70% development sample can lead to large changes in the bad rates for some bins and factors. Nevertheless, the general trends hold. More important in practice is that the working group can observe and comprehend this by setting cross-tab weights for one 70% sample and comparing them to bad rates in a second 70% sample, even though they do not have any prior knowledge of the term or concept of "sampling variation".

2. Overdraft loans to small businesses in Bulgaria

The second example uses data from a small-business lender in Bulgaria. The data set had 1,434 contracts, 9.7 percent (140) of which were defined as "bad" because they reached 90 days of arrears.

The Bulgarian lender had already developed an expert scorecard for overdraft loans using only its experience and judgement to assign points to the selected risk factors. We used the performance data to set cross-tab weights for the same set of factors originally included in the expert scorecard. We adjusted the cut-off points of the bins, however, for quantitative factors based on patterns of bad rates in the data, even in cases where the patterns differed from those implied by the expert scorecard. At the same time, we did not adjust bin cut-offs if bad-rate patterns in the data could not be reasonably explained by the working group of business users. We then set points based either on the actual bad rates in the data (per the cross-tab method), or, where the data did not make sense to end-users, based on the risk relationships originally expected by expert judgment but with adjustments to the point scale to fit the 9.7-percent bad rate in the performance data.

This model's 17 factors form a comprehensive risk profile similar to the one the bank used in subjective/manual decision-making:

- Loan terms (2 factors)
- Non-financial factors (3 factors)
- Borrower credit history (2 factors)
- Financial ability to repay (5 factors)
- Risk mitigation (5 factors)

Applied to the entire performance data set, the lender's original expert scorecard had an out-of-sample AUC of 0.70. Using a 70-percent sample of good and bad contracts from the performance data to transform the expert scorecard into a cross-tab weighted scorecard, AUC for the remaining 30% of contracts was 0.82. A Logit regression of these same factors had an out-of-sample AUC of 0.85.

3. A Microlender in Latin America

In our Latin American example, the data set had 14,039 contracts, of which 2,187 (15.5 percent), were defined as bad because they reached 30 days of arrears. As usual, we divided the data 70/30 for development and validation.

The model profiles borrower risk with 37 factors including:

- Loan terms (6 factors)
- Borrower demographics (18 factors)
- Borrower credit history (8 factors)
- Financial ability to repay (5 factors)

In this case, the bins were selected using Logit to evaluate individual factors, and we later set cross-tab weights given these bins. Out-of-sample AUC for the Logit model is 0.71, versus 0.69 for cross-tab weights.

The cross-tab and Logit models in our three examples have similar decreases in AUC between development and validation samples, and the Logit models are consistently more powerful, with AUC between 1 and 5 percent higher (Table 3). Given that these three lenders use the models for decision support (as opposed to “auto-decisioning”) and portfolio-risk management, the slightly lower AUC of the cross-tab models is unlikely to have a material negative impact on the business.

Table 3: Summary of Empirical Examples

		AUC		
Sample	**Model**	**Tajikistan**	**Bulgaria**	**Latin America**
Development (70%)	Cross-tab	0.83	0.82	0.73
	Logit	0.87	0.83	0.75
	Difference (%)	5	1	3
Validation (30%)	Cross-tab	0.82	0.82	0.69
	Logit	0.85	0.85	0.71
	Difference (%)	4	4	3

Considering this similar performance, end-users may prefer the simplicity and transparency of cross-tab weighting. Seen from the project level, a simple, transparent model is more likely to be accepted and used thoughtfully in practice. Thus, the average value-added from a cross-tab project may be higher than for a Logit project, even though a Logit model, when used correctly, is more powerful.

Conclusion

Cross-tab weighted scorecards for retail and small business segments have predictive power that compares well with Logit. The simplicity and transparency of the cross-tab method facilitates fuller understanding of models for end-users and increase the likelihood that scoring projects will succeed and be managed prudently.

References

Antonakis A C and Sfakianakis M E (2009). Assessing Naïve Bayes as a Method for Screening Credit Applicants. *Journal of Applied Statistics* **36 No. 5:** 538–545.

Aspey J, Hinder J, and Alan L (2003). The New Basel Accord: Implications. *Rhino Risk* http://www.crc.man.ed.ac.uk/conference/archive/2003/presentations/lucas2.pdf, retrieved 31 March 2011.

Baesens B, Van Gestel T, Viaene S, Stepanova M, Suykens J A K, and Vanthienen J (2003). Benchmarking State-of-the-Art Classification Algorithms for Credit Scoring. *Journal of the Operational Research Society* **54:** 627–635.

Dawes R M (1979). The Robust Beauty of Improper Linear Models in Decision Making. *American Psychologist* **34, No. 7:** 571–582.

Edelman D B (1992). The Introduction of Credit Scoring into Branch Banking. In L.C. Thomas, J.N. Crook, and D.B. Edelman (eds) *Credit Scoring and Credit Control*. Oxford: Clarendon Press, pp 161–177.

Hosmer D W and Lemeshow S (2001). *Applied logistic regression*. John Wiley & Sons.

MacNeill A (1999). *Mortgage Scoring: Combining Judgemental and Empirical Decision Systems—The Rationale?* Credit Research Centre Working Paper No. 99/8, University of Edinburgh.

Thomas L C (2000). A survey of credit and behavioural scoring: forecasting the financial risk of lending to consumers. *International Journal of Forecasting* **16:** 149–172.

Thomas L C, Banasik J, and Crook J N (2001). Recalibrating scorecards. *Journal of the Operational Research Society* **52:** 981–988.

Wainer H (1976). Estimating Coefficients in Linear Models: It Don’t Make No Nevermind. *Psychological Bulletin* **83:** 213–217.