ABSTRACT

More and more banks are firming up on their capabilities to quantify the risks they face in order to implement strategies to minimize the effect of those risks. In particular, credit risk modeling is being done to help assess the credit quality of borrowers. Constructing a quantitative credit risk model entails a step-by-step, often iterative, process from data preparation to model implementation.

This paper examines some of the statistical aspects of the model-building process and proposes new optimality criteria and procedures for credit risk modeling. For instance, in scorecard development, a method is proposed to bin a continuous predictor such that the bin-transformed variable maximizes the separation between credit defaulters and nondefaulters, while maintaining a monotone relationship with the target variable (the default indicator). Also, modified validation measures are proposed to optimize model validation and selection while reflecting realistic score cutoffs. Other proposed statistical optimalities in different phases of model building are outlined.
Part 1: Introduction

The Basel Accords recommend that regulators require banks to quantify and mitigate the risks they face. More specifically, the internal ratings-based (IRB) approach to Basel II allows banks to develop their own internal credit risk rating systems. As more and more banks are adopting this approach, credit risk modeling and validation techniques are being studied and utilized to assess the credit quality of financial borrowers.

This paper proposes a number of statistical optimalities that may be used in credit risk modeling. Model building entails a series of steps from data preparation to model scoring using a variety of statistical machine learning techniques and validation measures. This paper takes a look at some aspects of the model-building process and suggests new and modified strategies to help optimize credit risk models.

To put the proposed strategies into perspective, we outline the following typical general steps in credit risk modeling:

1. Prepare, rectify, and transform the modeling data
2. Perform exploratory data analysis (EDA)
3. If a scorecard is to be developed, recode continuous input variables into intervals (binning process)
4. Construct several candidate models using one or more modeling techniques and/or variable selection strategies
5. Select final model using validation measures
6. Perform “reject inference” if possible
7. Validate final model
8. Perform model scoring, monitor, refine as needed

The topics on statistical optimalities in this lecture are organized into parts. Part 2 presents an optimal binning algorithm for use in credit scorecard development. Part 3 discusses the popular validation measures in model selection and validation and proposes modifications to these measures to reflect realistic banking practices. In Part 4, a modeling exploration and benchmarking approach is suggested to help optimize credit risk models.
Part 2: Optimal Binning Algorithm

A scorecard is a convenient way to assign a quantitative assessment measure to a borrower. It also allows the scorecard user to gain insight regarding risk predictors in the scorecard. Because of the widespread use of scorecards, the process of binning continuous variables has been an important step in credit risk modeling.

There are various methodologies for binning continuous variables for scorecard development. The usual preferred strategies involve using the target variable (the default indicator) to supervise the binning algorithm. An example of such supervised binning uses the so-called weight of evidence (WOE) and the associated information value (IV). This method, like many others, is often coupled with an interactive approach to ensure a monotone relationship between the bin-transformed variable and the target variable.

This paper proposes a supervised learning algorithm to optimally bin a continuous predictor while automatically maintaining a monotone directional relationship between the bin-transformed predictor and the target default indicator. The idea is that the transformation:

- preserves the directional relationship between input and target variables
- achieves a monotone relationship between the transformed input and the default indicator (or more formally, the probability of default)
- attains the maximum separation between defaulters and nondefaulters among all bin transformations satisfying the (i) and (ii)

Let $X$ denote the predictor variable we want to bin, and let $D$ be the default indicator. It is assumed that $X$ has wide range of values, so that it makes sense to discretize it. Also note that the event of interest is default ($D = 1$), so that the probability predictions are for $P(D = 1)$.

**Directional Coefficient**

The first stage in the binning process is to determine the general directional relationship between $X$ and $D$. For this, we may use any appropriate correlation measure. Or we may use the following simple algorithm.

1. Sort according to $X$ and divide its range using a number of cutoff values (say percentiles).
2. For each cutoff in the range of $X$, compute and the proportion $Prop_A$ of defaults above the cutoff, and the proportion $Prop_B$ of defaults below the cutoff.

3. Compute the \textit{directional coefficient} as:

\[
DC = \frac{\#(Prop_A > Prop_B) - \#(Prop_A < Prop_B)}{\#(\text{Cutoffs})}
\]

The general directional relationship between $X$ and $D$ is given by the sign of $DC$: If $DC > 0$, then the directional relationship between $X$ and $D$ is positive. If it is zero, then there is no monotone relationship between $X$ and $D$.

This step can serve as a screening stage to determine which predictors have a potential directional relationship with the target variable. Recall that the scorecard is a device of choice in credit risk assessment, and so even if a variable has a strong but, say, U-shaped (or inverted U) relationship with the probability of default, it cannot be used for the scorecard. The algorithm above can be used to filter the variables to exclude those that don’t have a monotone relationship with the default indicator.

\textit{Splitting into Two Bins}

The proposed algorithm is able to subdivide a continuous variable into a specified number of bins. The underlying approach, however, hinges on optimally dividing into two groups at a time. Hence, we present a here an optimality criterion to choose the best splitting value that divides a set of values into two bins.

Let

\[
\lambda = \text{sgn}(DC)(Prop_A - Prop_B)
\]

where $\text{sgn}(DC)$ is the sign of the directional coefficient, i.e.:

\[
\text{sgn}(DC) = \begin{cases} -1 & \text{if } DC < 0 \\ 0 & \text{if } DC = 0 \\ 1 & \text{if } DC > 0 \end{cases}
\]

and, as defined above, $Prop_A$ is the proportion of defaults above the cutoff and $Prop_B$ is the proportion of defaults below the cutoff.

Note that $\lambda$ is positive when the factors are both positive or both negative, that is, when the terms are in agreement. Hence, a splitting value is able to capture the directional relationship (whether positive or negative) if $\lambda$ is positive.
Note also that the range of $\lambda$ is \(-1\) to \(1\). It takes the maximum value \(1\) if either

$$\text{Prop}_A = 1, \text{Prop}_B = 0, \text{and } sgn(DC) = 1$$

or

$$\text{Prop}_A = 0, \text{Prop}_B = 1, \text{and } sgn(DC) = -1.$$  

This is the ideal value of $\lambda$, and so the closer $\lambda$ is to \(1\), the better it can discriminate between defaulters and nondefaulters while preserving the directional relationship between $X$ and $D$. A zero value means either that $DC = 0$ ($X$ is not monotone with $D$) or that $\text{Prop}_A = \text{Prop}_B$ (the cutoff does not discriminate between defaulters and nondefaulters). A negative value implies that the proportions $\text{Prop}_A$ and $\text{Prop}_B$ are in conflict with the direction implied by $DC$, and so the cutoff cannot preserve the directional relationship. The goal is then to find that cutoff that maximizes $\lambda$.

**Optimal Binning**

The criteria above is then used to answer the original goal of optimally splitting the $X$ range into a specified number of intervals, while maintaining the directional relationship between $X$ and $D$. The following binning algorithm is proposed.

1. Compute the directional coefficient $DC$.
2. Specify number of bins
3. Provide initial set of cutoff values (e.g. equal spacing)
4. Combine the lowest two groups and find the splitting value that maximizes $\lambda$.
5. Using the new cutoff, combine the second-lowest and third-lowest groups and find the splitting value that maximizes $\lambda$.
6. Continue up to the highest two groups.
7. Repeat steps 4 to 6 until convergence.

For each pair of adjacent bins, the optimal cutoff yields the greatest separation between defaulters and nondefaulters according to the $\lambda$ criterion while preserving the general directional relationship between $X$ and $D$. The algorithm is also robust to the choice of initial cutoffs.
Part 3: Modified Validation Measures

An important part of model building and monitoring is model validation. In credit risk modeling, the goal of validation is to assess a model’s ability to discriminate between defaulters and nondefaulters. In the model development process, one use of validation is in model selection, wherein candidate models are evaluated using a validation measure, and the model with the optimal value (highest or lowest depending on the measure) is chosen as the scoring model. Popular validation measures in credit risk assessment include:

- the cumulative accuracy profile (CAP) and the associated accuracy ratio (AR) statistic
- the receiver operating characteristic (ROC) and the ROC index (or the AUC – area under the curve)
- the Kolmogorov-Smirnov (KS) statistic
- misclassification rate

Note that the CAP/AR and the ROC/AUC are equivalent in the sense that the chosen model will be the same whichever measure is used.

The misclassification rate is the proportion of misclassified cases. Though this measure is easy to understand, the problem is that it is based on only one cutoff probability. That is, after the default probabilities have been computed, they are compared to a cutoff, and then the cases are classified. Hence, it does not give a broad measure of the accuracy of the model across the range of possible cutoffs. Also, the cutoff probability may be changed at a later time, perhaps at the implementation stage.

On the other hand, the CAP/AR (and the ROC) is based on a broad set of cutoff probabilities and so it provides a more comprehensive measure of the discriminatory power of the prediction model. However, in actual scoring implementation, only one cutoff probability is used, and so the best model based on CAP may not be the best one based on this cutoff probability.

**Trimmed CAP/AR.** Hence, a criterion is proposed to reflect the realistic cutoff probabilities in using the CAP and the AR (or equivalently, the ROC and the AUR). In this proposed “trimmed CAP/AR” approach, only a limited set/range of cutoffs is used. For instance, depending on the level of aggressiveness of a bank, they can use cutoff ranges that reflect their risk-aversion stance. This proposed approach can be viewed as middle ground between the misclassification rate and the CAP.
To illustrate the idea surrounding the trimmed CAP/AR, consider first the following CAP curves of two competing scoring models A and B.

The two CAP curves criss-cross when superimposed, but Model A has a bigger area compared to Model B. This gives a higher AR value for Model A, which would then be the statistical choice between the two models.

However, suppose that the bank, either traditionally or in accordance with their defensive risk tolerance, normally chooses, or intends to choose, only the top 20% to 30% in terms of the credit quality score (based either on PD or scorecard). In this case, they would choose Model B. To see why this is so, consider the same CAP curves but focus only on the 0.2 to 0.3 fraction range:

The trimmed CAP/AR is larger for Model B, and so this would be the model of choice, taking into consideration the usual range of cutoff values of the bank. The discriminatory power of the model is assessed, not using a single cutoff, but a range of cutoffs that the bank usually uses, or anticipates to use. A similar approach for the ROC can be made.
The KS statistic is the maximum vertical difference between the cumulative distribution of the defaulters and that of the nondefaulters. It is also the maximum vertical distance between the ROC of a candidate model and that of the random model. Consider the KS statistics from the same 2 models in the above example.

From the above comparison, Model A has the higher KS statistic and would be the preferred model.

However, suppose again the same argument that the bank usually has a reasonable range of cutoff scores, say, implying the top 20-30% in terms of credit score. In this case we have the following “restricted” KS statistic (the maximum vertical distance in the restricted range).

The restricted KS statistic is higher for Model B than for Model A, indicating that Model B is better at discriminating between defaulters and nondefaulters, considering only the range of cutoff values that the bank usually uses or anticipates to use.
There are many modeling techniques that can be used to predict a target variable of interest. In credit risk assessment, especially in predicting the probability of default, the modeling choice is logistic regression. The main reason for this is its interpretability, aside from the fact that the method is well-established and has good predictive power when specified correctly. The transparency of the model allows for evaluation by government regulators such as the central bank.

However, possible nonlinearities and interactions among predictors may cast doubt about the optimality of the logistic regression model in predicting default.

A proposed approach is to first use a random forest to have initial insights as to which candidate predictors are able to discriminate defaulters from nondefaulters. Since a random forest is a combination of many decision trees (the predictions, that is), it inherits the advantages of decision trees in that it can capture nonlinearities and predictor interactions. Moreover, variable selection is built-in, and there is no need worry too much about non-numeric recoding or the usual data issues such as missing values and outliers. The drawback of a random forest is that it is not interpretable, although this aspect is not needed in this proposed approach.

The top predictors chosen by the random forest algorithm can then be the starting point for variable selection in logistic regression analysis. Furthermore, because of its predictive ability, the random forest model can serve as a benchmark to which we can compare the logistic regression model. That is, if the performance of regression (as assessed by the validation measures) is still far from that of the random forest, then it is an indication that the regression model can still be improved by perhaps transforming some variables, introducing polynomial terms, and/or trying out interaction terms. If the validation measures indicate that the logistic regression model is (now) at par with the random forest, then this would add confidence to the optimality of the regression model.