Development and Validation of Credit-Scoring Models[1]

Dennis Glennon[2] Nicholas M. Kiefer[3] C. Erik Larson[4]
Hwan-sik Choi[5]

September 12, 2007

[1] *Disclaimer: The statements made and views expressed herein are solely those of the authors and do not necessarily represent official policies, statements, or views of the Office of the Comptroller of the Currency or its staff or of Fannie Mae or its staff.* *Acknowledgement: We are grateful to our colleagues for many helpful comments and discussions, and especially to Regina Villasmil, curator of the OCC/RAD consumer credit database.*

[2] U.S. Department of the Treasury, Office of the Comptroller of the Currency, Risk Analysis Division.

[3] Cornell University, Departments of Economics and Statistical Sciences, 490 Uris Hall, Ithaca, NY 14853-7601, US. email:nmk1@cornell.edu; US Department of the Treasury, Office of the Comptroller of the Currency, Risk Analysis Division, 250 E. Street, SW, DC 20219, and CREATES, University of Aarhus, Building 1322, DK-8000 Aarhus C, Denmark.

[4] *ceriklarson.com*

[5] Cornell University, Department of Economics, and Fannie Mae, Business Analytics and Decisions

Abstract

Accurate credit-granting decisions are crucial to the efficiency of the decentralized capital allocation mechanisms in modern market economies. Credit bureaus and many financial institutions have developed and used credit-scoring models to standardize and automate, to the extent possible, credit decisions. We build credit scoring models for bankcard markets using the Office of the Comptroller of the Currency, Risk Analysis Division (OCC/RAD) consumer credit database (CCDB). This unusually rich data set allows us to evaluate a number of methods in common practice. We introduce, estimate, and validate our models, using both out-of-sample contemporaneous and future validation data sets. Model performance is compared using both separation and accuracy measures. A vendor-developed generic bureau-based score is also included in the model performance comparisons. Our results indicate that current industry practices, when carefully applied, can produce models that robustly rank-order potential borrowers both at the time of development and through the near future. However, these same methodologies are likely to fail when the the objective is to accurately estimate future rates of delinquency or probabilities of default for individual or groups of borrowers.

JEL Classification: C13, C14, C52, G11, G32

Keywords: Logistic regression, CHAID, specification testing, risk management, nonparametrics, validation.

1 Introduction

The consumer credit market in the United States has grown rapidly over the last two decades. According to the Federal Reserve Board's *Statistical Release on Consumer Credit* (FRB (2006)), the total outstanding revolving consumer credit in the United States was $860.5 billion and increasing at an annual rate of 4.9 percent as of September 2006. Of course, the lion's share of this total represents debt in the form of credit card balances carried by consumers. More than 1 billion credit cards are in circulation in the United States; fully 74.9 percent of all families have credit cards, and 58 percent of them carry a balance. The Federal Reserve's triennial *Survey of Consumer Finances* in 2004 showed the average and median credit card balance of those carrying a balance was $5,100 and $2,200 respectively (see Bucks, Kennickell, and Moore (2006).) Given the continuing growth of the consumer credit market, efficient decision making is more important than ever both socially (for efficiency) and privately (for profitability).

Facing this growth, financial institutions have been pressed to develop tools and models to help standardize and automate credit decisions. From an economic point of view, increasing the efficiency of credit allocation has the effect of directing resources toward their most productive applications, increasing productivity, output, growth and fairness. From the financial institution's point of view, a small improvement in credit decisions can provide a competitive edge in a fiercely contested market, and lead to increased profits and increased probability of survival. Further, retail credit decisions are numerous and individually small, and it is costly to devote the time of loan officers to each application.

A simple economic model serves to introduce the conceptual framework. Suppose the revenue from serving a non-defaulting individual account over a fixed period is θ, the probability of an individual defaulting is π, and the loss given default is λ (defined here as a positive). Then the expected profit from this account over the period is $(1-\pi)\theta - \pi\lambda$. In this case, loans are profitable only if $\pi < \theta/(\theta+\lambda)$. As a practical matter, banks would apply this decision rule by ranking applicants according to their estimated value of π and extend loans to those applicants with the smallest default probabilities (as funds are available) up to the critical value $\pi^* = \theta/(\theta + \lambda)$. Of course, there is a lot missing in this formulation of the decision rule, including the existence of error in the estimation of π and how that might vary across applicants.

In a typical application, credit performance measures and borrower characteristics are calculated as functions of the sample data. These measures are then used to develop statistical credit-scoring models, or *scorecards*, the output of which are forecasts of credit performance for borrowers with similar characteristics. For example, a model might generate a predicted performance measure as a function of the applicant's use, in percent, of existing credit lines (often referred to as a *utilization rate*). A lender will typically use this performance predictor as an input into the underwriting decision process. A simple decision rule would be to approve an application only if the estimated performance measure exceeds a critical value. A more sophisticated application might use the performance measure to establish the terms of any credit offered.

Kiefer and Larson (2006) provide an overview of conceptual and statistical issues that arise during the process of developing credit-scoring models. Bierman and Hausman (1970); Dirickx and Wakeman (1976); Srinivasan and Kim (1987); Thomas, Crook, and Edelman (1992); Thomas, Edelman, and Crook (2002); Hand (1997); and others, outline the development of scorecards using a range of different mathematical and statistical techniques. A recent research conference with industrial, academic and supervisory participants sponsored by the Office of the Comptroller of the Currency (OCC), the primary supervisor of nationally chartered banks in the United States, had a full program of papers on specification and evaluation of credit-scoring models. This literature reflects substantial advances but not consensus on best practices in credit scoring.

In this paper, we demonstrate a range of techniques commonly employed by practitioners to build and validate credit scoring models using the OCC Risk Analysis Division (OCC/RAD) consumer credit database (CCDB). We compare the models with each other and with a commercially developed generic bureau-based credit score. The CCDB is unique in many ways. It contains both *tradeline* (account) and summary information for individuals obtained from a recognized national credit bureau, and it is sufficiently large to allow us to construct both a holdout sample drawn from the population at the time of development and several out-of-sample and out-of-time validation samples. The database also allows for one to observe the longitudinal performance of individual borrowers and individual accounts; however, models exploiting this type of dynamic structure generally have not been developed or used by lenders and other practitioners. Such dynamic models are consequently not within the scope of this paper.

Our model development process illustrates several aspects of common industry practices. We provide a framework in which to compare and contrast alternative modeling approaches, and we demonstrate the strengths and weaknesses of alternative modeling techniques commonly used to develop a scoring model. We focus on a limited number of sample and modeling issues that typically arise during the model-development process and that are likely to have significant impacts on the accuracy and reliability of a model.[1] It is not our purpose to identify an exhaustive set of modeling approaches, illustrate what we have observed in place at any single institution, or build models that compete with those currently available in the market.

One significant objective of our work is to illustrate aspects of *model validation* that can − and we believe should − be employed at the time of model development. Model validation is a process that is comprised of three general types of activities: (1) the collection of evidence in support of the model's design, estimation, and evaluation at the time of development; (2) the establishment of on-going monitoring and benchmarking methods by which to evaluate model performance during implementation and use, and (3) the evaluation of a model's performance utilizing outcomes-based measures and the establishment of feedback processes which ensure that unexpected performance is acted upon. The focus of this paper is on the first

[1] There are other legitimate ways of addressing issues of sample design, model selection, and validation beyond those outlined below. Moreover, we believe newer and better techniques continue to be developed in the statistical and econometric literature. For those reasons, we emphasize that there are alternatives to the processes outlined below that can and, under certain circumstances, should be used as part of a well-developed and comprehensive model development process.

of these activities: the compilation of developmental evidence in support of a model. However, as a natural part of the model development process, which involves benchmarking alternative models and identifying of appropriate outcomes-based measures of performance, we do touch upon some of the post-development validation activities noted in (2) and (3). Finally, we show that there are limitations to the application of a model developed using a static sample design as a risk measurement tool. A model that performs well at ranking the population by expected performance may still perform poorly at generating valid default probabilities required for pricing and profitability analysis.

In Section 2 we describe the data development process employed to create the OCC/RAD consumer credit database. Section 3 then outlines the methods used to specify and estimate our suite of models and the calibration process used to construct our scores. Section 4 describes methods that we employ to benchmark and compare the performance of the scores within the development sample and in various validation samples from periods subsequent to that of the development sample. Section 5 summarizes our findings.

2 Sample Design of The OCC/RAD Consumer Credit Database

Each of the three major U.S. credit bureaus – Equifax, Experian, and TransUnion – maintain credit files for about 200 million individuals. Approximately four and a half billion pieces of data are reported to the bureaus each month by grantors of consumer credit and collectors of public records. The bureaus are faced with the daunting task of collecting this information on an ongoing basis and using it to update the consumer credit histories in their repositories.

As the primary supervisor of nationally chartered banks, the OCC has a broad set of interests and issues that it would like to analyze using data on consumer credit performance. These include evaluating various credit scoring methods in use by banks, developing new methods, and identifying and documenting national or regional trends in credit utilization by product type. To this end, the OCC purchased a large multi-year extract of individual and tradeline data from one of the three national credit bureaus and used it to contruct the CCDB.

2.1 Bureau-Based versus Institution-Specific Models

Practitioners and researchers alike typically base their analysis and modeling on samples of data drawn from one or more of the credit bureaus, historical data drawn from their own portfolio, or a combination of both. Sample designs will vary with the intended use of the data; however, the primary consideration in the specification of any sample design will be the population to which the results of the modeling of analysis are to be applied.

Most financial institutions that purchase research samples of credit bureau data do so in order to analyze and build models that describe the credit behavior of their current or likely future customers. In these cases, the sample design might be limited to selecting a sample of the bank's current or prior customers, or alternately to selecting a sample of individuals with a generic credit score greater than some pre-specified value (under the assumption that future customers will look like those from the past.) In contrast, large-scale developers of generic bureau-based credit scores are interested in having these scoring tools robustly predict performance for a broad spectrum of the consumer credit-using population and consequently will want a broader, more nationally representative sample on which to base their work. In many ways, the design of the CCDB and the development of the models in this paper more closely parallel that of the later group.

2.2 Unit of Analysis

For our models, the unit of analysis is the behavior of an individual rather than that of any one tradeline. This reflects a common industry practice of using bureau data to construct credit scores for individuals rather than to develop tradeline-specific scores for each of an individual's accounts (a more common application of custom scorecards). In credit-scoring model building, it is also commonplace to develop summary measures of an individual's credit profile across tradelines – for example, the construction of a variable measuring aggregate bankcard balance or the computation of a generic credit score – and to use this *attribute data* in custom scorecard construction.

As a result, the existing CCDB consists of unnecessary tradeline level data and attribute data for sampled individuals. While some common attribute data were obtained directly from the bureau at the time of sampling, we have the ability to use the tradeline data to construct additional attributes as necessary. It is useful to think of the CCDB as consisting of two component databases: an individual-level database with attribute information, and a matching tradeline-level database with detailed account information for every account of each sampled individual.

2.3 Temporal Coverage

Sample designs differ in their breadth and unit of analysis and in terms of their temporal coverage. Common modeling practice in the development of credit scoring tools has historically utilized *cross-sectional* sampling designs, when a selection of consumer credit histories is observed at time t, and payment behavior is tracked over k future time periods (k is often typically defined as 24 months). Scoring models are developed to predict performance over the interval $[t, t + k]$ as a function of characteristics observed at time t.

In contrast, the study of the dynamic behavior of credit quality requires observations over multiple periods of time for a fixed set of analysis units that have been sampled in a base year (i.e., a longitudinal or panel data design). In both instances, data has to be extracted with sufficient detail to allow the tracking of performance, balances, line increases, etc., by tradelines (i.e., by lender) for each unit over time.

Under a longitudinal sample design, annual extracts represent updated (or refreshed) observations for each of the observations in the sample. To facilitate the objectives of illustrating existing cross sectional methods and allowing for experimentation with longitudinal-based analysis, the CCDB has a unique structure. The database has been constructed so as to incorporate a "rolling" set of panels, as well as an annual sequence of random cross sectional samples. Rather than simply identifying a base period sample and then tracking the same individuals though time, as might be the case in a classic panel, the CCDB seeks to maintain the representative nature of the longitudinal data by introducing supplemental parallel structure individuals at various points in time, and by developing weights relating the panel to the population at any point in time. Further details are presented in the following sections.

2.3.1 Cross-Sectional Sampling

The initial sample consists of $1,000,000$ randomly selected individual credit reports as of June 30, 1999. Nine hundred fifty thousand of these individuals were randomly sampled from the sub-population of individuals for whom the value of a generic, bureau-based score (GBS) could be computed (the scoreable population), while $50,000$ individuals were sampled from the unscoreable population. The allocation of the sample between scoreable and unscoreable populations was chosen in order to track some initially unscoreable observations longitudinally though subsequent time periods. Because the unscoreable segment represents roughly 25 percent of the credit bureau population, a purely random sampling from the main credit bureau database would have yielded too many unscoreable individuals.[2]

2.3.2 Longitudinal Sampling

Given the required cross-sectional size and the need to observe future performance when developing a model, it was also determined that the sample should include performance information through June 30, 2004 − the terminal date of our data set. The 1,000,000 observations from the June 30, 1999 sample make up the initial "*core*"set of observations under our panel data design. The panel is constructed by updating the credit profile of each observation in the core on June 30th of each subsequent year. In Figure 1 we illustrate the general sampling and matching strat-

[2] Unscoreable individuals include those who are deceased or who have only public records or very thin credit tradline experience.

egy using the 1999 and 2000 data; counts of sampled and matched individuals are presented in Tables 1 and 2.

In general, the match rate from one year's sample to the following year's bureau master file is high. Some of the scoreable individuals sampled in 1999 became unscoreable in 2000, again due to death or inactivity, and some of the previously unscoreable became scoreable in 2000 (for instance, if they had acquired enough credit history). Of the $1,000,000$ individuals sampled in 1999, $949,790$ individuals were found to be scoreable as of June 30, 2000. As indicated in Table 2, this change resulted from 17,339 individuals moving from scoreable to unscoreable or missing, while $17,129$ individuals moved from unscoreable to scoreable.

Over time, the credit quality of a fixed sample of observations (i.e., the core) is likely to diverge from that of a growing population. For that reason, we update the core each year by sampling additional individuals from the general population and then developing "rebalanced" sampling weights which allow for comparison between the updated core and the current population. For example, we update the core in 2000 by comparing the GBS distribution of the $949,790$ individuals from the 1999-2000 matched sample (tabulated using 10-point score buckets from 300 to 900, the range of the GBS) to a similarly constructed GBS distribution for an additional $950,000$ individuals randomly sampled from the credit bureau's master file as of June 30, 2000. The relative difference in frequency by bucket between the two distributions was then used to identify the size of an "update sample" of individuals to add to the 1999-2000 matched sample. The minimum of these bucket-level frequency changes (i.e., the maximum decrease rate in relative frequency) was then used as a sampling proportion to determine the number of additional individuals that would be randomly sampled from the June 30, 2000, scoreable population and added to the core data set (i.e., the 1999-2000 matched file). For 2000, the "updating proportion" was determined to be 7 percent, resulting in the addition of 66,500 individuals from the 2000 scoreable population to the 1999-2000 matched scoreable sample on the CCDB. Use of this updating strategy ensures that the precision with which one might estimate characteristics at the GBS bucket level in a given year does not diminish due to drift in the credit quality of those individuals sampled in earlier years.

Sampling for years 2001-2004 proceeded along similar lines, with the results reported again in Tables 1 and 2. The individuals who were members of the CCDB panel in a previous year (i.e., the core) were matched to a current year's master file. Individuals who were unmatched or remained or became unscoreable in the current year were dropped from the CCDB panel and then replaced with another draw of 50,000 unscoreable individuals from the current year's master file. The GBS distribution from the panel was compared with that for a random cross section of individuals drawn from the current master file and a "updating proportion" was determined and applied to define an additional fraction of the random cross section to add to and complete the current-year CCDB panel.

3 Scorecard Development

3.1 Defining Performance and Identifying Risk Drivers

We follow industry-accepted practices to generate a comprehensive risk profile for each individual. We use as a starting point the five broadly defined categories outlined in Fair-Isaac (2006). We summarized our own examples of possible credit bureau variables that fall within each category and which are obtainable from our data set; these are presented in Table 3.

Scorecard development attempts to build a segmentation or index that can be used to classify agents into two or more distinct groups. Econometric methods for the modeling of limited dependent variables and statistical classification methods are therefore commonly applied. In order to implement these types of models using the type of credit information available from bureaus, it is necessary to define a performance outcome; this is usually, but not necessarily, dichotomous, with classes generally distinguishing between "good" and "bad" credit histories based upon some measure of performance.

In this paper, we choose to classify and develop a predictive model for performance of good and bad credits based upon their "default" experience. Bad outcomes correspond to individuals who experience a "default" and "good" outcomes to individuals who do not. It is our convention to assign a default if an individual becomes 90 days past due (DPD), or worse, on at least one bankcard over a 24-month performance period (for example July 1999 through June 2001). Although regulatory rules require banks to charge-off credit card loans at 180 DPD, it is not uncommon among practitioners to use our more conservative definition of default (90+ DPD). We experimented with a definition of default based on both a 12- and 18-month performance period. The results of our analysis are fundamentally the same under the alternative definitions of default.

3.2 Construction of the Development and Hold-Out (In-Time Validation) Samples

We develop our model using a conventional scorecard sample design. The refinement process that was applied to the CCDB and that resulted in the development samples is presented in Figure 2. A randomly selected, cross-section sample of $995,251$ individual credit files with valid tradeline data $-$ representing over 14.5 million tradelines $-$ is drawn from the CCDB database as of June 30, 1999. The sample includes $733,820$ individuals with at least one open bankcard line of credit that had been updated during the January through June 1999 time period.[3] We drop $19,122$ files with a bankcard currently 90+ DPD, choosing to model the performance of accounts that are no worse than 60 DPD at time of model development. A separate model for accounts that are currently seriously delinquent (i.e., greater than 60 DPD) could be developed (although we do not attempt to develop such a model in this paper.) An additional $37,436$ accounts are deleted because their future performance could not be reliably observed in our panel, leaving us with a sample of bankcard credit performance on $677,262$ individual credit records. We split this group randomly into two samples of approximately equal size and then develop our suite of models using a sample of $338,578$ individual credit histories. The remaining $338,684$ individuals

[3] A bankcard tradeline is defined as a credit card, or other revolving credit account with variable terms issued by a commercial bank, industrial bank, co-op bank, credit union, savings and loan company, or finance company.

are used as a holdout sample for (within-period) validation purposes.

To allow for the more parsimonious modeling of different risk factors (i.e., characteristics), and possibly different effects of common risk drivers, it is standard practice in the industry to segment (or split) the sample prior to model development. We have implemented a common segmentation by introducing splits based upon the amount of credit experience and the amount, if any, of prior delinquency. Credit files that contain no history of delinquencies are defined as *clean*, and those with a history of one or more delinquencies are defined as *dirty*.[4] Because individuals with little or no credit experience are expected to perform differently from those with more experience and thicker files, we create additional segments within the clean group made up of individuals with *thin* credit files (fewer than 3 tradelines) or credit files (more than 2 tradelines). On the other hand, we created two segments within the dirty group consiting of individuals with no current delinquency and with mild delinquency (60- DPD). Consequently, we identify four mutually exclusive segments: *clean/thick, clean/thin, dirty/current,* and *dirty/delinquent.*

In Figure 2 we report the number of individuals and the average default rate in each of the segments. The development sample has an average default rate of 7.19 percent. The clean and dirty segments have a default rate of 3.1 percent and 20.3 percent respectively. Our objective is to model the likelihood of default (i.e., 90+ DPD) for each segment using credit bureau information only.

3.3 Model Forms

There are several analytical modeling techniques that are discussed in the scoring literature and used in the industry to construct a scoring model. These include regression-based models (i.e., ordinary least squares, logit procedures), discriminant analysis, decision trees, neural networks, linear programming methods and other semiparametric and nonparametric techniques.[5] In practice, most scorecards are

[4] We define an observation as dirty if the individual has a history of delinquencies greater than 30 DPD ever, a public record, or collections proceedings against him or her.

[5] By design, discriminant analysis, linear programming, and tree methods use a maximum divergence (between good and bad performance) criterion for selecting the best combination of factors

developed using a regression-based model.

We consider and illustrate the differences between the three most commonly employed model forms. First, we consider a logistic regression. Logistic regressions are a form of generalized linear model characterized by a linear index and a logistic "link" function. Next, we develop a form of semiparametric model in which we retain the linear index from the parametric model specification but estimate the link nonparametrically. Although we generalize the link function from logistic to nonparametric, we retain the assumption that the link function is the same across segments. That is, we retain the assumption that there is a common relationship between the value of the index and the default probability, though we no longer require the logistic functional form. We experiment with further generalizations to different link functions across segment;, however, these generalizations are not especially productive, especially for the segments with smaller sample sizes. Finally, we compare these two regression forms with a fully nonparametric model developed using a decision-tree approach. This can be thought of as a further generalization in which both the index and the link are estimated nonparametrically.

3.3.1 Parametric models

The parametric specification is the logistic regression

$$p_i = E(y_i|x_i) = 1/(1 + exp(-\beta' x_i)) \text{ for each individual } i, \tag{1}$$

where $y_i \in \{0, 1\}$ is an indicator variable for non-default/default, x_i is a vector of covariates, and β is the vector of associated coefficients. The estimates $\hat{p}_i$ of the probability of default are derived from the estimated model

$$\hat{p}_i = 1/(1 + exp(-b' x_i)), \tag{2}$$

where b is the maximum likelihood estimator of β.

---

and factor weights for developing classification models. Regression and neural network methods use an error minimization criterion, which is well suited for constructing prediction models. However, regression models often perform well over multiple objectives.

If we define the index $Z = b'x$, then Z represents the estimated log-odds

$$Z = \ln(\hat{p}/(1 - \hat{p})). \tag{3}$$

3.3.2 Semiparametric Models

The semiparametric models use the estimated (parametric) index function to partition the sample into relative risk segments. We rank the sample by the estimated index from the logistic regression and then estimate the link function nonparametrically. Specifically, for this model the estimates of the default rate are equal to the empirically observed default rate within each segment.

We follow current industry practice and partition the sample into discrete segments, chosen so that each band contains the same number of observations, m. Given the sample size, we create 30 distinct segments. For each segment, the predicted probability of default is given by

$$\hat{p}_i = \overline{y}_{J_i}, \tag{4}$$

where

$$\overline{y}_{J_i} = \frac{\sum_{k=1}^{n} y_k 1\{J_k = J_i\}}{\sum_{k=1}^{n} 1\{J_k = J_i\}}. \tag{5}$$

and $J_i \in \{1, ..., 30\}$ denotes the segment J to which individual i belongs.

3.3.3 Variable Selection Methods

Variable selection for the parametric and semiparametric forms is accomplished though application of each of three alternative variable selection methods; we refer to the variable selection methods as Stepwise, Resampling, and Intersection.

Our Stepwise method starts with an intercept-only regression model and then searches for the set of covariates to find the one with the strongest statistical relationship with performance (forward selection). It repeats this process, searching within a multivariate framework for additional covariates that are predictive of the performance variable. As each new covariate is added, the algorithm tries to eliminate the least significant variables (backward selection). The forward selection stops when the remaining covariates fail to reach a level of statistical significance at the 5 percent level.

The Stepwise method, however, may result in over identifying, or overfitting, the model especially in large samples (Glennon (1998)). To reduce this tendency to overfit the regression model, our Resampling method is characterized by the repeated application of a stepwise selection procedure over sub-samples of the data. Covariates that most frequently enter the model over multiple replications are then combined into a single model estimated over the full development sample. Specifically, we first randomly select ω-percent of the data and then run a stepwise regression. Then, we repeat the resampling and stepwise regressor selection k times and choose the variables that appear most often in the k replications (variables that occurred in 10 or more of the replications). We use $k = 20$ and experiment with values for $\omega = \{20 \text{ percent}, 50 \text{ percent}, 100 \text{ percent}\}$. After some experimentation, we use the results from the 50 percent trial. We applied the stepwise and Resampling methods separately to each segment.

Finally, we define the Intersection method as the variable selection resulting from construction of the common set of covariates that appear in the Stepwise and Resampling methods. The Stepwise selection approach generates the largest, and the Intersection approach the smallest, set of covariates.

3.3.4 A Nonparametric Model

The fully nonparametric model form does not assume a functional form for the covariates. To implement our nonparametric specification, we use a tree method called CHAID (Chi-squared Automatic Interaction Detector) to cluster the data into multiple "nodes" by individual characteristics (attributes). The variable selection process searches by sequential subdivision for a grouping of the data giving maximal discrimination subject to limitations on the sizes of the groups (avoiding the best fit solution of one group per data point). The approach is due to Kass (1980).[6] The CHAID approach splits the data sequentially by performing consecutive Chi-square

[6] Various refinements have been made to Kass's original specificatgion; we implement CHAID using the SAS macro `%TREEDISC (SAS (1995))`.

tests on all possible splits. It accepts the best split. If all possible splits are rejected, or if a minimum group size limit is reached, it stops. Each of the final nodes is assigned with predictions that are equal to the empirical default probability, $\hat{p}_n$ for node n. By design of the algorithm, individuals within a node are chosen to be as homogeneous as possible, while individuals in different nodes are as heterogeneous as possible (in terms of $\hat{p}_n$), resulting in maximum discrimination. Note that the splitting of the development sample data into four segments which preceded the construction of parametric and semiparametric models was not undertaken prior to implementing the CHAID algorithm.

For the CHAID method we have to specify (1) the candidate variable list, (2) the transformation of continuous variables into discrete variables, and (3) the minimum size of the final nodes. We considered two different sets of candidate variables. Initially, we considered all available attributes and kept only those that generated at least one split. As an alternative, we used only those attributes that were identified using the Intersection method for variable selection outlined above. In the latter case, for each model segment (i.e., clean/thick, clean/thin, dirty/current, and dirty/delinquent), we take the intersection of the variables from the stepwise selection process with the variables appearing 10+ times in the 20 percent, 50 percent, and 100 percent Resampling methods, then combine the selected variables across the model segments by taking the union of those sets of variables.

As the CHAID approach considers all possible splits, it requires the splitting of continuous variables into discrete ranges. We chose the common and practical approach of constructing dummy variables to represent each quartile of each continuous variable. As a validity check on this procedure we also split the continuous variables into 200 bins. (Note that this process includes all intermediate splits from 4–199 as special cases).

To prevent nodes from having too few observations or having only one kind of account (good or bad), we set the minimum of observations in a node to be $1,000$. The CHAID rejects a split if it produces a node smaller than $1,000$. Therefore the size of the final nodes works as a stopping rule for the CHAID. Since this specification is rather arbitrary, we experiment with different node sizes ranging from 100 to $8,000$

observations.

3.4 Explanatory variables

In Tables 4 through 7, we report the variables selected using the Stepwise, Resampling (50 percent) and Intersection methods for each of the segments. Each table includes the set of variables selected using the Stepwise method, sorted by variable type (see Table 3), for that segment of the population. In the third and fourth columns of each table, we list the subset of variables identified using the Resampling and Intersection methods, respectively. The *worst status for open bankcards within the last six months*, the *total number of tradelines with 30+ DPD,* and the *total number of tradelines with good standing* are "individual credit history" variables that consistently show up as important explanatory variables. *Utilization rates for bankcards* and *for revolving accounts* are the more important "amount-owed" variables. The *age of the oldest bankcard tradeline* enters as a relevant measure of the "length of credit history," and "new credit activity" is measured using the *total number of inquiries within the last 12 months* and the *total number of bankcard accounts opened within the last two years.* Finally, *the total number of revolving tradelines active* was an important explanatory variable capturing the impact of the "type of credit used." It is clear from our results that a fairly small set of variables suffices to capture almost all of the possible explanatory power. In Table 8, we report the set of "splitting" variables identified under the CHAID selection method, again sorted by variable type.

3.5 Score Creation though Model Calibration

We transform the estimated $\hat{p}$ into credit scores, namely *Risk Analysis Division Scores* or *(RAD).* Credit scores are a mapping from the estimates $\hat{p}$ to integers. Scores contain the same information as $\hat{p}$ estimates but are convenient to use and easy to interpret. We follow industry convention and calibrate the RAD Scores (S) to a normalized odds scale using the following rules:

1. $S = 700$ corresponds to an odds ratio (good:bad) of $20:1$. Equivalently

$$\frac{(1-\hat{p}_{700})}{\hat{p}_{700}} = 20, \tag{6}$$

where $\hat{p}_{700}$ is the $\hat{p}$ value at $S = 700$, and

2. Every 20-unit increase in S doubles the odds ratio. The score values, S, are calibrated using the affine transformation:

$$S = 28.8539(Z + 21.2644), \tag{7}$$

where Z is as given in equation (3). We calculate eight different RAD scores, one from each of the three parametric (Stepwise, Resampling, and Intersection), three semiparametric (Stepwise, Resampling, and Intersection), and two nonparametric (all variables, and Intersection) models.

We also recalibrate the GBS so as to allow for comparison with the RAD scores. Since we cannot observe the predicted $\hat{p}$ associated with the GBS, we estimate it though a linear regression of the empirical log-odds in our sample against the score values. Data for the regression consists of empirical log-odds estimated for 20 different buckets of individuals sorted by the GBS and the associated bucket mean bureau values.

4 Evaluation of Scoring Model Performance

4.1 Performance Measures

We evaluate our models based upon two primary metrics of interest: discriminatory power and predictive accuracy. We consider two types of measures by which to assess scorecard performance: separation measures and accuracy measures. These are widely used in practice. Separation measures give the degree of separation between good and bad performance, and the accuracy measures gives the degree of difference between the predicted and realized default rates.

A popular separation measure is the Kolmogorov-Smirnov statistic (K-S value), defined by the maximum difference between two cumulative distribution functions (CDFs) of good and bad performers.

For the accuracy measure, we consider the Hosmer and Lemeshow Goodness-of-Fit Test (H-L). It is based on the difference between the realized default (or bad) rates $\bar{p}_j$ and the *average* of the predicted default rates, $\check{p}_j$, for individuals grouped into deciles ($j = 1, ..., 10$) (the deciles are constructed by sorting the sample by individual predicted default rate $\hat{p}_i$). The H-L statistic is defined as

$$HL = \sum_{j=1}^{10} \frac{(\bar{p}_j - \check{p}_j)^2}{\check{p}_j(1 - \check{p}_j)/n_j}, \tag{8}$$

where n_j is the number of observation in each of the j deciles. The H-L statistic is distributed as Chi-square with 8 degrees of freedom under the null that $\bar{p}_j = \check{p}_j$ for all j. Just to be clear, a good model should have a high value of the separation measure, K-S, but a low value of the accuracy measure H-L. It would be perhaps better to label the H-L as an an "inaccuracy" measure, as it is a Chi-squared measure of fit, but the contrary convention is long established.

4.2 In-Time Validation at Development

We first compare the performance of models differentiated by variable selection method (Stepwise, Resampling, Intersection), given model form (parametric, semi-parametric, nonparametric (CHAID)). Then we compare the performances of the different model forms. The performance of scorecard models was measured in the development samples and is presented in Tables 9 and 10. Table 9 shows the median RAD Scores by segment and by validation samples, while Table 10 shows K-S and H-L measures constructed from pooled-across-segment model predictions and outcomes. The models developed using the Stepwise variable selection method perform best at differentiating between good and bad accounts, although the difference between the parametric and semi parametric approach is very small. Overall, the nonparametric CHAID approach performed worse on the pooled data. Although all the models perform well at differentiating between good and bad accounts, none of them is particularly accurate as reflected in the low p-values for the H-L test. All but the semiparametric model generate predicted values that are statistically different from the actual default performance. Because the actual (development sample) performance is used, by design, to predict performance under the semi parametric approach, the H-L test is not applicable for the development sample and not very informative for the in-sample, hold-out validation data.

We also evaluate the accuracy and reliability of each model (i.e., by segment and model form) as stand alone models. Table 11 shows K-S and H-L measures from the parametric and semiparametric models for each segment.[7] Individually, the segment-specific models perform well at differentiating between good and bad accounts. As is commonly observed in practice, credit bureau-based models perform better on the clean-history segments of the population as reflected in the nearly 20 point difference in the K-S values between the clean-history and dirty-history segments across model form and variable selection procedures. It is interesting to note, however, that the parametric models are relatively accurate on the development and in-sample, hold-out data except for the clean-history/thick-file segment. That latter result is likely driving the accuracy results in Table 10, given the relative size of the clean-history/thick-file segment.[8] These results clearly show that a model can perform well at discriminating between good and bad accounts (i.e., high K-S value), yet perform poorly at generating accurate estimates of the default probabilities – a result that illustrates the importance of considering model purpose (i.e., discrimination or prediction) in the development and selection of a credit scoring model.

The K-S test evaluates separation at a specific point over the full distribution of

[7]By design, the actual performance within each decile (i.e., score band) from the development sample is used to generate the predicted values under the semiparametric method. For that reason (as noted above), the H-L test is not well designed for evaluating the accuracy of the semiparametric models. Therefore, we use the actual performance (i.e., default rate) derived from the pooled-segment analysis summarized in Table 10 as the predicted values in the calculation of the H-L values for each of the semiparametric models in Table 11.

[8]The more accurate model results for the semi parametric model on the development and in-sample, hold-out data are likely to be due to the construct of the tests, and therefore, must be interpreted carefully. Clearly an out-of-sample test will better reflect the true accuracy of the models constructed using this approach.

outcomes. In Figures 3 and 4, we plot the Gains charts for each of models. The Gains charts describe the separation ability graphically by showing the CDF for observations with "bad" outcomes plotted against the CDF for all sample observations (the 45-degree line serves as a benchmark representing no separation power). The parametric and semiparametric models and the GBS produce very similar graphs, while the CHAID models showed much weaker discriminatory power.

In Figures 5 and 6, we plot the empirical log odds by RAD score for each model, for both the development and hold-out samples, respectively. We compare the empirical log odds for each model against calibrated target values. The calibration target line is given in eq. (7). The graphs show that the models preform relatively well for score values below 750. Although the semiparametric models and the CHAID do not generate estimates for scores below 600 due to the smoothing nature of the models, we point out that the parametric model continues to perform well on the score range below 600. For scores between 760 and 780, the parametric and semiparametric models slightly overestimate default risk. For the score range over 800, the RAD models underestimate the default rates. These results suggests that the lack of overall accuracy of the model is being driven primarily by the imprecision in the estimates at the higher end (i.e., greater than 750) of the distribution: that portion of the score distribution, based on the median scores reported in Table 9, heavily populated by observations from the clean-history (both thick and thin) segments.

It is worth noting that the Resampling and Intersection models generate very similar levels of separation and accuracy measures using fewer covariates. Those results hold across both the development and hold-out samples. The decision tree approach (i.e., CHAID method), however, clearly generates models with lower discriminate power. That result is reflected in the five-point difference in the K-S values between the Stepwise parametric model and Intersection CHAID model in Table 10. It is difficult, however, to interpret the meaning of that result. Instead, we look to the relationship between the Gains charts in Figure 3. The Gains chart for the Stepwise parametric model is above the Gains chart for the CHAID-Intersection model. As a result, at each point on the horizontal axis, the Stepwise parametric model identifies a greater percentage of the bad distribution. For example, over the bottom 10 percent of the score distribution, the Stepwise parametric model identifies roughly 60 percent of the bad accounts, while the CHAID-Intersection identifies only approximately 48 percent. At that point, the Stepwise model identifies nearly 25 percent (i.e., 12/48) more bad accounts − a substantial increase over the CHAID model.

We have estimated the models using the widely, but not universally, accepted 90+ DPD definition for the outcome variable. It is interesting to ask whether the model would also do well at discriminating between good and bad accounts if default is defined at 60+ DPD, or evaluated over a shorter performance horizon (e.g., 18, 12, and 6 months). In Table 12, we summarize the observed performance over these alternative definitions of performance. A reliable model should order individuals by credit quality over a variety of bad definitions. In Table 13 we compare the K-S measures from eight different RAD models and the GBS, using both the 90+ DPD and 60+ DPD bad definitions. We find that a model's ability to differentiate between good and bad accounts is virtually the same as reflected by the K-S values across the development and holdout samples for all methods. As expected, the models perform better under the 90+ DPD definition. Nevertheless, the models seem to order observations well by credit quality for the alternative definitions. This topic is revisited below.

4.3 Out-of-Time Validation (Subsequent to Development)

Given the longitudinal characteristics of the CCDB data set, we are able to track the out-of-sample performance of our models through 2002. Table 14 shows the sample sizes and bad rates for the development and out-of-time validation samples. On a pooled-segment basis, overall default rates increase from 1999 to 2001, and then decrease in 2002. However, there is a significant improvement in the clean/thin segment over those years. Table 15 shows the median scores by segment, model form, and variable selection method for the development and out-of-time validation samples. To better illustrate the shift in the distributions over time, we report the box and whisker plot for the RAD Scores, by model form in Figure 7, and by segment and model form in Figures 8 through 11. There is an obvious upward shift in the full score distribution over time for all model forms in Figure 7, which reflects the general trend in the median values reported in Table 15. Although the score distributions for the dirty/current and dirty/delinquent segments are shifting down over time (Figures 8 and 9), the overwhelming shift up in the distributions for clean/thick and clean/thin (Figures 10 and 11) dominate the overall shift in the distribution of scores.

We update the model separation and accuracy measures reported in Table 10 for the out-of-sample periods 2000-2002 in Table 16 on a pooled-across-segments basis. We observe that the H-L measures become very large (and the p-values very small) in the out-of-time validation samples, indicating general lack of statistical fit for predictive purposes. None of the scoring models developed using conventional industry practices generated accurate predictions over time even though all the models maintained their ability to differentiate between good and bad accounts. These conclusions are supported by the out-of-sample results in Table 17. For each segment, the K-S values remained relative constant, or improved, over time; however, in all cases, the H-L statistics increased significantly. The significant increase in the H-L values across all model segments in Table 17 suggests that our simple cross-section model is under-specified relative to the factors that reflect changes in the economic environment over time.

As an additional test of the non-parametric approach, we reran the CHAID model with continuous variables discretized to 200 values, and compared the performance to the CHAID model based on quartiles. The CHAID based on all variables did substantially worse in terms of model accuracy in the out-of-time validation samples. The CHAID based on the Intersection selection performed about the same with 200 values as with quartiles for the 2000 and 2001 samples but substantially worse in 2002 in terms of model accuracy. Thus, there seems to be no real benefit from adding splits beyond quartiles for our continuous variables.

Figure 12 compares the empirical log-odds by different RAD scores for the 2002 validation samples. The plot clearly shows a deterioration in the predicted default rate over the range 650-750. The actual performance is worse than the predicted, and the RAD scores underestimate the default rates. The results for other years were very similar to 2002, and are not shown here.

Overall, out-of-sample analyses show that the separation power of the models is relatively stable over time; however, model accuracy decreases substantially. This result, combined with the observed increase in the average default rate over the full sample period except for the clean/thin segment (Table 14), implies that the models estimated on a cross section of data from 1999 will underpredict defaults over future periods. Moreover, it suggests that when the defaults are disaggregated into buckets, the higher-default buckets will tend to be underpredicted more than the low default buckets − a result observed in Figure 12. These results imply that models aimed at accuracy should be frequently updated, or that dynamic models, with some dependence on macroeconomic conditions, should be considered.

Figure 13 compares the Gains chart for each of the RAD scoring models using the 2002 validation samples. Other years showed very similar results. As in the development samples, the parametric and semiparametric models, and the GBS performed very similarly, and the CHAID models were worse than the others. Although the Gains charts for all parametric and semiparametric models are nearly overlapping, the Stepwise selection method produces models that discriminate slightly better (for both parametric and semiparametric forms). The Resampling selection method is nearly as good, followed by the Intersection method.

We compare the Gains charts for the development samples and the validation samples for each of the "preferred" models (the Resampling-based parametric model, the Stepwise-based semiparametric model, the CHAID with all variables) and the calibrated GBS in Figures 14 though 17. For all models and the GBS, the Gains charts are again nearly overlapping and support the general results of the comparison of K-S values over time.

4.4 Robustness of separation

As noted above, useful credit scores should be informative about different credit related events. Although the RAD scores are developed for the event of 90+DPD within 24 months, it is expected that they will be relevant for reasonable changes in the outcome definition. As with the within-period validation above, we consider different horizons (6, 12, and 18 months) as well as a different delinquency definition (60+ DPD). If the RAD scores generate reasonable separation for these other events, we consider them to be robust in terms of separation. For some individuals, performance data was missing over sub-portions of the 24-month observation period. If performance information was missing as of the observation month (e.g., 6th, 12th, or 18th month), the observation was labeled missing in Table 12. As a result, we excluded individuals with missing observations in shorter horizons in calculating separation measures.

The results in Table 18 show that the K-S measures for different definitions of default are relatively consistent over time under the alternative event horizons. Although the models perform better under a 90+ DPD definition of default, they perform reasonably well under a 60+ DPD definition. If we compare across models, parametric and semiparametric models showed the best separation being slightly better than the calibrated GBS. The CHAID model consistently performs slightly worse at separating good from bad accounts. These results show that the RAD scores are very robust and informative in the separation metric for the delinquency events we considered.

5 Conclusion

We developed a credit-scoring model for bankcard performance using the OCC Risk Analysis Division consumer credit data base and methods that are often encountered in the industry. We validated and compared a parametric model, a semiparametric model, and a popular nonparametric approach (CHAID).

It is worth pointing out that data preparation is crucial. The sample design issues are important, as discussed, but simple matters such as variable definition and treatment of missing or ambiguous data become critical. This is especially true in cases where similar credit attributes could be calculated in slightly different ways. Evaluating these data issues was one of the most time-consuming components of the project.

With the data in hand, we find that careful statistical analysis will deliver a useful model, and that, while there are differences across methods, the differences are small. The parametric and semiparametric models appear to work slightly better than the CHAID. There is little difference between the parametric and semiparametric models. We find that within-period validation is useful, but out-of-time validation shows a substantial loss of accuracy. We attribute this to the changing macroeconomic conditions. These conditions led to a small change in the overall default rate. This change reflects much larger changes in the default rates of the high-default (low-score) components of the population. Thus accurate out-of-time prediction of within-score-group default rates should be based on models which are frequently updated, or which have variables reflecting aggregate credit conditions. On the positive side, the separation properties of the models seem quite robust in the out-of-time validation samples. This suggests that it is easier to rank individuals by creditworthiness than to predict actual default rates.

There are many additional models in each of the categories, parametric, semiparametric and nonparametric, which could be considered. We have taken a representative approach from each category. Our models are similar to those used in practice. Our results suggest that the performance of models developed using simple cross sectional techniques may be unreliable in terms of accuracy as macroeconomic conditions change. The results suggest that increased attention be placed on the use of longitudinal modeling methods as a means by which to estimate performance conditional on temporally varying economic factors.

References

Bierman, H., and W. H. Hausman (1970): "The Credit Granting Decision," *Management Sci.*, 16, 519–532.

Bucks, B. K., A. B. Kennickell, and K. B. Moore (2006): "Recent Changes in U.S. Family Finances: Evidence from the 2001 and 2004 Survey of Consumer Finances," *Federal Reserve Bulletin*, 92, A1–A38.

Dirickx, Y. M. I., and L. Wakeman (1976): "An Extension of the Bierman-Hausman Model for Credit Granting," *Management Sci*, 22, 1229–1237.

Fair-Isaac (2006): "Understanding Your Credit Score," http://www.myfico.com/CreditEducation/WhatsInYourScore.aspx.

FRB (2006): "Federal Reserve Statistical Release G.19, Consumer Credit, December," Discussion paper, Federal Reserve Board.

Glennon, D. C. (1998): "Issues in Model Design and Validation," in *Credit Risk Modeling Design and Application*, ed. by E. Mays, chap. 13, pp. 207–221. Glenlake Publishing.

Hand, D. J. (1997): *Construction and Assessment of Classification Rules*. John Wiley, Chichester, U.K.

Kass, G. (1980): "An exploratory technique for investigating large quantities of categorical data," *Applied statistics*, 29:2, 119–127.

Kiefer, N. M., and C. E. Larson (2006): "Specification and Informational Issues in Credit Scoring," *International Journal of Statistics and Management Systems*, 1, forthcoming.

SAS (1995): "%Treedisc Macro for CHAID (Chi-Squared Automatic Interaction Detection) Algorithm," Discussion paper, Cary, NC: SAS Institute Inc. August 2005 Revision.

Srinivasan, V., and Y. H. Kim (1987): "The Bierman-Hausman Credit Granting Model: A Note," *Management Sci*, 33, 1361–1362.

Thomas, L., J. Crook, and D. B. Edelman (1992): *Credit Scoring and Credit Control*. Oxford University Press, Oxford.

Thomas, L. C., D. B. Edelman, and J. Crook (2002): *Credit Scoring and Its Applications*. SIAM.

TABLES AND FIGURES

Table 1:
CCDB Sampling Design Counts

A	B	C	D	E	F	G	H	H
Year	Matched from Previous Year's Panel and Scoreable	Unmatched from Previous Year's Panel and Dropped	Random Scoreable Cross Section From Current Year Masterfile	Random Unscoreable Cross Section From Current Year Masterfile	Updating Proportion	Updating Random Sample	**CCDB Panel = B+E+G**	Total Current Year Masterfile Extracts = B+C+D+E
1999	***N.A.***	***N.A.***	***950,000***	***50,000***	100%	950000	**1,000,000**	1,000,000
2000	949,790	50,210	***950,000***	***50,000***	7%	66500	**1,066,290**	2,000,000
2001	1,015,469	50,821	***883,710***	***50,000***	7%	61860	**1,127,329**	2,000,000
2002	1,075,669	51,660	***822,671***	***50,000***	7%	57587	**1,183,256**	2,000,000
2003	1,130,033	53,223	***766,744***	***50,000***	5%	38337	**1,218,370**	2,000,000
2004	1,162,722	55,648	***731,630***	***50,000***	5%	36582	**1,249,304**	2,000,000

Table 2:
Transition to scoreable and unscoreable states

Transition Period	Base Year Panel Size	Individuals Transitioning to Unscoreable from Scoreable	Individuals Transitioning to Scoreable from Unscoreable	Net Transitions to Scoreable
1999~2000	1,000,000	17,129	17,339	210
2000~2001	1,066,290	14,150	14,971	821
2001~2002	1,127,329	11,988	13,648	1,660
2002~2003	1,183,256	9,846	13,069	3,223
2003~2004	1,218,370	8,837	14,485	5,648

Table 3:
Variables by Type

Payment History	**Name**
Worst status of open bankcards within 6 months	CURR
Total number of tradelines with 90 days past due or worse	BAD01
Dummy variable for the existence of tradelines with 90 days past due or worse	BAD11
Total number of tradelines with 60 days past due or worse	BAD21
Dummy variable for the existence of tradelines with 60 days past due or worse	BAD31
Total number of tradelines with 30 days past due or worse	BAD41
Dummy variable for the existence of tradelines with 30 days past due or worse	BAD51
Total number of bankcard tradelines with 90 days past due or worse within 12 months	BK03
Dummy variable for the existence of bankcard tradelines with 90 days past due or worse within 12 months	BK13
Dummy variable for the existence of installment tradelines with 90 days past due or worse within 12 months	IN13
Dummy variable for the existence of mortgage tradelines with 90 days past due or worse within 12 months	MG13
Dummy variable for the existence of retail tradelines with 90 days past due or worse within 12 months	RT13
Dummy variable for the existence of revolving retail tradelines with 90 days past due or worse within 12 months	RTR13
Dummy variable for the existence of auto lease tradelines with 90 days past due or worse within 12 months	AS13
Dummy variable for the existence of auto loan tradelines with 90 days past due or worse within 12 months	AL13
Dummy variable for the existence of revolving tradelines with 90 days past due or worse within 12 months	RV13
Months since the most recent 60 days past due or worse in bankcard tradelines of which the records were updated within 12 months	BK33
Worst status of bankcard tradelines with 60 days past due or worse and of which the records were updated within 12 months	BK43
Maximum of the balance amount, past due amount, and charged off amount of delinquent bankcard tradelines with 60 days past due or worse and of which the records were updated within 12 months	BK53
Total number of public records in the DB	PU01
Dummy variable for the existence of public records	PU11
Total number of tradelines with good standing, positive balance, and of which the records were updated within 12 months	GO01
Total number of closed tradelines within 12 months	NUM_Closed
Amounts Owed	
Aggregate credit amount of bankcard tradelines of which the records were updated within 12 months	BK27
Aggregate credit amount of installment tradelines of which the records were updated within 12 months	IN27
Aggregate credit amount of mortgage tradelines of which the records were updated within 12 months	MG27
Aggregate credit amount of auto loan tradelines of which the records were updated within 12 months	AL27
Aggregate credit amount of revolving tradelines of which the records were updated within 12 months	RV27
Dummy variable for the positive aggregate credit amount of bankcard tradelines	U11
Dummy variable for the positive aggregate credit amount of installment tradelines	U12
Dummy variable for the positive aggregate credit amount of mortgage tradelines	U13
Dummy variable for the positive aggregate credit amount of auto loan tradelines	U17
Dummy variable for the positive aggregate credit amount of revolving tradelines	U18
Aggregate balance amount of open bankcard tradelines of which the records were updated within 12 months	ABK16
Aggregate balance amount of installment tradelines of which the records were updated within 12 months	IN16
Aggregate balance amount of mortgage tradelines of which the records were updated within 12 months	MG16
Aggregate balance amount of auto loan tradelines of which the records were updated within 12 months	AL16
Aggregate balance amount of finance tradelines of which the records were updated within 12 months	ALN08
Aggregate balance amount of retail tradelines of which the records were updated within 12 months	ART08
Aggregate balance amount of revolving tradelines of which the records were updated within 12 months	RV16
Aggregate balance amount of open home equity tradelines of which the records were updated within 12 months	AEQ08
Bankcard utilization rate (Aggregate balance / Aggregate credit amount)	BK28
Dummy variable for zero bankcard utilization rate	BK28_0

Dummy variable for bankcard utilization rate=100%	BK28_100
Dummy variable for bankcard utilization rate>100%	BK28_101
Installment accounts utilization rate (Aggregate balance / Aggregate credit amount)	IN28
Mortgage accounts utilization rate (Aggregate balance / Aggregate credit amount)	MG28
Auto loan accounts utilization rate (Aggregate balance / Aggregate credit amount)	AL28
Open bankcard utilization rate (Aggregate balance / Aggregate credit amount)	ABK18
Revolving accounts utilization rate (Aggregate balance / Aggregate credit amount)	RV28
Average credit amount of bankcard tradelines with positive balance and of which the records were updated within 12 months	BK17
Average credit amount of installment tradelines with positive balance and of which the records were updated within 12 months	IN17
Average credit amount of mortgage tradelines with positive balance and of which the records were updated within 12 months	MG17
Average credit amount of retail tradelines with positive balance and of which the records were updated within 12 months	RT17
Average credit amount of auto loan tradelines with positive balance and of which the records were updated within 12 months	AL17
Average credit amount of revolving tradelines with positive balance and of which the records were updated within 12 months	RV17
Length of credit history	
Age of the oldest tradeline (Months)	AG04
Age of the oldest bankcard tradeline (Months)	BK04
Age of the oldest installment tradeline (Months)	IN04
Age of the oldest mortgage tradeline (Months)	MG04
New credit	
Total number of inquiries within 6 months	AIQ01
Total number of inquiries within 12 months	IQ12
Total number of bankcard accounts opened within 2 years	BK61
Total number of installment accounts opened within 2 years	IN61
Total number of mortgage accounts opened within 2 years	MG61
Dummy variable for the existence of new accounts within 2 years	NUM71
Types of credit in Use	
Dummy variable for the existence of installment tradelines within 12 months	D1
Dummy variable for the existence of mortgage tradelines within 12 months	D2
Dummy variable for the existence of retail tradelines within 12 months	D3
Dummy variable for the existence of revolving retail tradelines within 12 months	D4
Dummy variable for the existence of auto lease tradelines within 12 months	D5
Dummy variable for the existence of auto loan tradelines within 12 months	D6
Total number of credit tradelines (excluding inquiries/public records)	NUM01
Total number of bankcard tradelines	BK01
Total number of installment tradelines	IN01
Total number of mortgage tradelines	MG01
Total number of retail tradelines	RT01
Total number of revolving retail tradelines	RTR01
Total number of auto lease tradelines	AS01
Total number of auto loan tradelines	AL01
Total number of revolving tradelines	RV01
Total number of credit tradelines of which the records were updated within 12 months	NUM21
Total number of bankcard tradelines of which the records were updated within 12 months	BK21
Total number of installment tradelines of which the records were updated within 12 months	IN21
Total number of mortgage tradelines of which the records were updated within 12 months	MG21

Total number of retail tradelines of which the records were updated within 12 months	RT21
Total number of revolving retail tradelines of which the records were updated within 12 months	RTR21
Total number of auto lease tradelines of which the records were updated within 12 months	AS21
Total number of auto loan tradelines of which the records were updated within 12 months	AL21
Total number of revolving tradelines of which the records were updated within 12 months	RV21
Total number of bankcard tradelines with positive balance and of which the records were updated within 12 months	BK31
Total number of installment tradelines with positive balance and of which the records were updated within 12 months	IN31
Total number of mortgage tradelines with positive balance and of which the records were updated within 12 months	MG31
Total number of retail tradelines with positive balance and of which the records were updated within 12 months	RT31
Total number of auto loan tradelines with positive balance and of which the records were updated within 12 months	AL31
Total number of revolving tradelines with positive balance and of which the records were updated within 12 months	RV31

Table 4: Dirty/Delinquent Segment - Explanatory variables selected using Stepwise, Resample, and Intersection methods

Variables selected using the Stepwise method (sorted by variable type)	Variable Names	Significance Ranking in Selection Methods		
		Resample	Inter-section	Stepwise
I. Payment History				
Total number of tradelines with good standing, positive balance, and of which the records were updated within 12 months	GO01	1	1	1
Worst status of open bankcards within 6 months	CURR	3	3	3
Total number of closed tradelines within 12 months	NUM_Closed	10		6
Dummy variable for the existence of tradelines with 90 days past due or worse	BAD11	9		13
Worst status of bankcard tradelines with 60 days past due or worse and of which the records were updated within 12 months	BK43			15
Maximum of the balance amount, past due amount, and charged off amount of delinquent bankcard tradelines with 60 days past due or worse and of which the records were updated within 12 months	BK53			29
Dummy variable for the existence of installment tradelines with 90 days past due or worse within 12 months	IN13			31
II. Amount Owed				
Open bankcard utilization rate (Aggregate balance / Aggregate credit amount)	ABK18	4	4	4
Dummy variable for bankcard utilization rate>100%	BK28_101	7	8	7
Aggregate balance amount of retail tradelines of which the records were updated within 12 months	ART08			9
Aggregate balance amount of revolving tradelines of which the records were updated within 12 months	RV16			12
Average credit amount of bankcard tradelines with positive balance and of which the records were updated within 12 months	BK17	13		19
Auto loan accounts utilization rate (Aggregate balance / Aggregate credit amount)	AL28			21
Aggregate balance amount of finance tradelines of which the records were updated within 12 months	ALN08			23
Average credit amount of revolving tradelines with positive balance and of which the records were updated within 12 months	RV17			25
Dummy variable for the positive aggregate credit amount of auto loan tradelines	U17			26
Average credit amount of mortgage tradelines with positive balance and of which the records were updated within 12 months	MG17			27
Dummy variable for bankcard utilization rate=100%	BK28_100			30
III. Length of Credit History				
Age of the oldest bankcard tradeline (Months)	BK04	8	7	8
Age of the oldest installment tradeline (Months)	IN04			17
IV. New Credit				
Total number of bankcard accounts opened within 2 years	BK61	6	6	5

Total number of inquiries within 12 months	IQ12	11		10
Dummy variable for the existence of new accounts within 2 years	NUM71	14		18
Total number of installment accounts opened within 2 years	IN61	12		20
V. Type of Credit Used				
Total number of revolving tradelines with positive balance and of which the records were updated within 12 months	RV31	2	2	2
Total number of installment tradelines with positive balance and of which the records were updated within 12 months	IN31	5	5	11
Total number of credit tradelines of which the records were updated within 12 months	NUM21			14
Total number of mortgage tradelines with positive balance and of which the records were updated within 12 months	MG31			16
Total number of installment tradelines of which the records were updated within 12 months	IN21			22
Dummy variable for the existence of revolving retail tradelines within 12 months	D4			24
Dummy variable for the existence of retail tradelines within 12 months	D3			28

Table 5: Dirty/Current Segment - Explanatory variables selected using Stepwise, Resample, and Intersection methods

Variables selected using the Stepwise method (sorted by variable type)	Variable Names	Significance Ranking in Selection Methods		
		Resample	Inter-section	Stepwise
I. Payment History				
Total number of tradelines with good standing, positive balance, and of which the records were updated within 12 months	GO01	3	4	3
Total number of tradelines with 30 days past due or worse	BAD41	4	5	4
Total number of tradelines with 90 days past due or worse	BAD01	7	9	7
Total number of closed tradelines within 12 months	NUM_Closed	12	10	9
Total number of bankcard tradelines with 90 days past due or worse within 12 months	BK03	25		13
Dummy variable for the existence of mortgage tradelines with 90 days past due or worse within 12 months	MG13	18	21	16
Dummy variable for the existence of installment tradelines with 90 days past due or worse within 12 months	IN13	13	13	18
Dummy variable for the existence of bankcard tradelines with 90 days past due or worse within 12 months	BK13	24		20
Dummy variable for the existence of retail tradelines with 90 days past due or worse within 12 months	RT13			28
Dummy variable for the existence of revolving tradelines with 90 days past due or worse within 12 months	RV13	20	11	34
Dummy variable for the existence of revolving retail tradelines with 90 days past due or worse within 12 months	RTR13			36
Dummy variable for the existence of auto loan tradelines with 90 days past due or worse within 12 months	AL13			38
Dummy variable for the existence of tradelines with 90 days past due or worse	BAD11			43
Maximum of the balance amount, past due amount, and charged off amount of delinquent bankcard tradelines with 60 days past due or worse and of which the records were updated within 12 months	BK53			44
II. Amount Owed				
Open bankcard utilization rate (Aggregate balance / Aggregate credit amount)	ABK18	1	2	1
Dummy variable for bankcard utilization rate>100%	BK28_101	5	6	6
Average credit amount of bankcard tradelines with positive balance and of which the records were updated within 12 months	BK17	9	8	10
Average credit amount of mortgage tradelines with positive balance and of which the records were updated within 12 months	MG17	26	20	14
Aggregate credit amount of bankcard tradelines of which the records were updated within 12 months	BK27	14	15	17
Aggregate balance amount of open bankcard tradelines of which the records were updated within 12 months	ABK16	22	19	25
Average credit amount of revolving tradelines with positive balance and of which the records were updated within 12 months	RV17			31
Aggregate balance amount of revolving tradelines of which the records were updated within 12 months	RV16			37
Average credit amount of installment tradelines with positive balance and of which the records were updated within 12 months	IN17			40
Aggregate balance amount of open home equity tradelines of which the records were updated within 12 months	AEQ08			42
Dummy variable for the positive aggregate credit amount of revolving tradelines	U18			45

Aggregate balance amount of installment tradelines of which the records were updated within 12 months	IN16			46
Mortgage accounts utilization rate (Aggregate balance / Aggregate credit amount)	MG28			47
Dummy variable for the positive aggregate credit amount of installment tradelines	U12			48
III. Length of Credit History				
Age of the oldest bankcard tradeline (Months)	BK04	6		8
Age of the oldest mortgage tradeline (Months)	MG04	21		22
Age of the oldest installment tradeline (Months)	IN04	19	14	23
Age of the oldest tradeline (Months)	AG04			32
IV. New Credit				
Total number of bankcard accounts opened within 2 years	BK61	8	3	5
Total number of installment accounts opened within 2 years	IN61	16	16	11
Dummy variable for the existence of new accounts within 2 years	NUM71	23	18	21
Total number of inquiries within 6 months	AIQ01			39
Total number of inquiries within 12 months	IQ12	11		41
V. Type of Credit Used				
Total number of revolving tradelines with positive balance and of which the records were updated within 12 months	RV31	2	1	2
Dummy variable for the existence of revolving retail tradelines within 12 months	D4	17	17	12
Total number of revolving tradelines	RV01	10	7	15
Total number of installment tradelines of which the records were updated within 12 months	IN21	15	12	19
Total number of installment tradelines	IN01			24
Total number of credit tradelines (excluding inquiries/public records)	NUM01			26
Total number of bankcard tradelines with positive balance and of which the records were updated within 12 months	BK31	27		27
Total number of installment tradelines with positive balance and of which the records were updated within 12 months	IN31			29
Total number of mortgage tradelines of which the records were updated within 12 months	MG21			30
Total number of mortgage tradelines	MG01			33
Dummy variable for the existence of installment tradelines within 12 months	D1			35

Table 6: Clean/Thin Segment - Explanatory variables selected using Stepwise, Resample, and Intersection methods

Variables selected using the Stepwise method (sorted by variable type)	Variable Names	Significance Ranking in Selection Methods		
		Resample	Inter section	Stepwise
I. Payment History				
Worst status of open bankcards within 6 months	CURR	2	2	4
Total number of tradelines with 30 days past due or worse	BAD41			10
Total number of tradelines with good standing, positive balance, and of which the records were updated within 12 months	GO01			11
II. Amount Owed				
Revolving accounts utilization rate (Aggregate balance / Aggregate credit amount)	RV28	1	1	1
Dummy variable for bankcard utilization rate=100%	BK28_100	6	4	7
Dummy variable for bankcard utilization rate>100%	BK28_101	7		8
Open bankcard utilization rate (Aggregate balance / Aggregate credit amount)	ABK18			12
III. Length of Credit History				
Age of the oldest bankcard tradeline (Months)	BK04	5		9
IV. New Credit				
Total number of bankcard accounts opened within 2 years	BK61	4		2
Total number of inquiries within 12 months	IQ12	3	3	3
V. Type of Credit Used				
Total number of bankcard tradelines	BK01			5
Total number of bankcard tradelines with positive balance and of which the records were updated within 12 months	BK31			6

Table 7: Clean/Thick Segment - Explanatory variables selected using Stepwise, Resample, and Intersection methods

Variables selected using the Stepwise method (sorted by variable type)	Variable Names	Significance Ranking in Selection Methods		
		Resample	Inter-section	Stepwise
I. Payment History				
Total number of tradelines with 30 days past due or worse	BAD41			33
Dummy variable for the existence of tradelines with 30 days past due or worse	BAD51	3	5	5
Worst status of open bankcards within 6 months	CURR	2	3	3
Total number of tradelines with good standing, positive balance, and of which the records were updated within 12 months	GO01	6		4
Total number of closed tradelines within 12 months	NUM_Closed			19
II. Amount Owed				
Aggregate balance amount of open bankcard tradelines of which the records were updated within 12 months	ABK16	11		10
Open bankcard utilization rate (Aggregate balance / Aggregate credit amount)	ABK18	13	8	8
Auto loan accounts utilization rate (Aggregate balance / Aggregate credit amount)	AL28			36
Aggregate balance amount of finance tradelines of which the records were updated within 12 months	ALN08			24
Aggregate balance amount of retail tradelines of which the records were updated within 12 months	ART08			27
Average credit amount of bankcard tradelines with positive balance and of which the records were updated within 12 months	BK17	12		12
Bankcard utilization rate (Aggregate balance / Aggregate credit amount)	BK28			35
Dummy variable for bankcard utilization rate=100%	BK28_100	14		13
Dummy variable for bankcard utilization rate>100%	BK28_101	19		25
Aggregate balance amount of mortgage tradelines of which the records were updated within 12 months	MG16			34
Average credit amount of mortgage tradelines with positive balance and of which the records were updated within 12 months	MG17	15	6	23
Aggregate credit amount of mortgage tradelines of which the records were updated within 12 months	MG27			22
Aggregate balance amount of revolving tradelines of which the records were updated within 12 months	RV16	20		21
Average credit amount of revolving tradelines with positive balance and of which the records were updated within 12 months	RV17			29
Revolving accounts utilization rate (Aggregate balance / Aggregate credit amount)	RV28	1	2	1
III. Length of Credit History				
Age of the oldest tradeline (Months)	AG04	16	9	16
Age of the oldest bankcard tradeline (Months)	BK04	8	10	9
IV. New Credit				

Total number of inquiries within 6 months	AIQ01			31
Total number of bankcard accounts opened within 2 years	BK61	7	7	6
Total number of installment accounts opened within 2 years	IN61	10		11
Total number of inquiries within 12 months	IQ12	5	4	17
Total number of mortgage accounts opened within 2 years	MG61			32
Dummy variable for the existence of new accounts within 2 years	NUM71			37
V. Type of Credit Used				
Total number of auto loan tradelines	AL01			28
Total number of auto loan tradelines with positive balance and of which the records were updated within 12 months	AL31			20
Total number of bankcard tradelines with positive balance and of which the records were updated within 12 months	BK31	17		30
Total number of installment tradelines of which the records were updated within 12 months	IN21	18		14
Total number of installment tradelines with positive balance and of which the records were updated within 12 months	IN31	9		7
Total number of credit tradelines (excluding inquiries/public records)	NUM01			26
Total number of credit tradelines of which the records were updated within 12 months	NUM21			15
Total number of revolving tradelines of which the records were updated within 12 months	RV21			18
Total number of revolving tradelines with positive balance and of which the records were updated within 12 months	RV31	4	1	2

Table 8: Variables used at least once in CHAID splitting

Variable Names (sorted by variable type)	Variables	All Attributes	Inter-section
I. Payment History			
Total number of tradelines with 60 days past due or worse	BAD21	X	X
Dummy variable for the existence of tradelines with 60 days past due or worse	BAD31	X	
Total number of tradelines with 30 days past due or worse	BAD41	X	X
Dummy variable for the existence of tradelines with 30 days past due or worse	BAD51	X	X
Worst status of open bankcards within 6 months	CURR	X	X
Total number of tradelines with good standing, positive balance, and of which the records were updated within 12 months	GO01	X	X
II. Amount Owed			
Aggregate balance amount of open bankcard tradelines of which the records were updated within 12 months	ABK16	X	X
Open bankcard utilization rate (Aggregate balance / Aggregate credit amount)	ABK18	X	X
Average credit amount of bankcard tradelines with positive balance and of which the records were updated within 12 months	BK17	X	X
Aggregate credit amount of bankcard tradelines of which the records were updated within 12 months	BK27		X
Bankcard utilization rate (Aggregate balance / Aggregate credit amount)	BK28	X	
Dummy variable for bankcard utilization rate=100%	BK28_100		X
Installment accounts utilization rate (Aggregate balance / Aggregate credit amount)	IN28	X	
Average credit amount of mortgage tradelines with positive balance and of which the records were updated within 12 months	MG17	X	X
Average credit amount of retail tradelines with positive balance and of which the records were updated within 12 months	RT17	X	
Aggregate balance amount of revolving tradelines of which the records were updated within 12 months	RV16	X	
Average credit amount of revolving tradelines with positive balance and of which the records were updated within 12 months	RV17	X	
Aggregate credit amount of revolving tradelines of which the records were updated within 12 months	RV27	X	
Revolving accounts utilization rate (Aggregate balance / Aggregate credit amount)	RV28	X	X
Dummy variable for the positive aggregate credit amount of installment tradelines	U12	X	
Dummy variable for the positive aggregate credit amount of mortgage tradelines	U13	X	
III. Length of Credit History			
Age of the oldest tradeline (Months)	AG04	X	X
Age of the oldest bankcard tradeline (Months)	BK04	X	X
Age of the oldest installment tradeline (Months)	IN04	X	X
Age of the oldest mortgage tradeline (Months)	MG04	X	

IV. New Credit			
	IQ06	X	
Total number of bankcard accounts opened within 2 years	BK61	X	X
Total number of installment accounts opened within 2 years	IN61		X
Total number of inquiries within 12 months	IQ12	X	X
Dummy variable for the existence of new accounts within 2 years	NUM71		X
V. Type of Credit Used			
Total number of bankcard tradelines of which the records were updated within 12 months	BK21	X	
Total number of bankcard tradelines with positive balance and of which the records were updated within 12 months	BK31	X	
Dummy variable for the existence of installment tradelines within 12 months	D1	X	
Dummy variable for the existence of retail tradelines within 12 months	D3	X	
Dummy variable for the existence of revolving retail tradelines within 12 months	D4	X	X
Total number of installment tradelines of which the records were updated within 12 months	IN21		X
Total number of installment tradelines with positive balance and of which the records were updated within 12 months	IN31		X
Total number of mortgage tradelines	MG01	X	
Total number of credit tradelines (excluding inquiries/public records)	NUM01	X	X
Total number of retail tradelines with positive balance and of which the records were updated within 12 months	RT31	X	
Total number of revolving tradelines	RV01	X	X
Total number of revolving tradelines of which the records were updated within 12 months	RV21	X	
Total number of revolving tradelines with positive balance and of which the records were updated within 12 months	RV31	X	X

Table 9:
In-Time Validation: Median Scores for Various Models at Development

Segment	Scoring Model		Sample	
	Model Form	Variable Selection	1999 Dev	1999 Hold-Out
Dirty History and Presently Mildly Delinquent	Parametric	Stepwise	616	616
		Resampling	616	616
		Intersection	617	617
	Semi Parametric	Stepwise	598	598
		Resampling	599	599
		Intersection	600	600
	Calibrated Generic Bureau Score		636	637
Dirty History and Presently Current	Parametric	Stepwise	676	676
		Resampling	676	676
		Intersection	676	676
	Semi Parametric	Stepwise	673	673
		Resampling	672	672
		Intersection	672	672
	Calibrated Generic Bureau Score		675	675
Clean History and Thin File	Parametric	Stepwise	729	730
		Resampling	727	727
		Intersection	735	735
	Semi Parametric	Stepwise	733	733
		Resampling	727	727
		Intersection	735	735
	Calibrated Generic Bureau Score		722	723
Clean History and Thick File	Parametric	Stepwise	750	750
		Resampling	751	751
		Intersection	750	750
	Semi Parametric	Stepwise	759	759
		Resampling	758	758
		Intersection	760	760
	Calibrated Generic Bureau Score		747	747
All (Pooled)	Parametric	Stepwise	734	734
		Resampling	735	735
		Intersection	734	734
	Semi Parametric	Stepwise	738	738
		Resampling	737	737
		Intersection	735	735
	Nonparametric (CHAID)	All variables	725	725
		Intersection	724	724
	Calibrated Generic Bureau Score		734	734

Table 10:

In-Time Validation: Model Separation and Accuracy Measures at Development (Pooled Across Segments)

(Bad = 90+Days Past Due, or Worse, over the Following 24 Months)

Scoring Model		Statistic and Sample					
		Kolmogorov-Smirnov (K-S)		Hosmer-Lemeshow (H-L)			
Model Form	**Variable Selection**	**1999 Dev**	**1999 Hold-out**	**1999 Dev**		**1999 Hold-out**	
				(value)	**(p-value)**[2]	**(value)**	**(p-value)**
Parametric	Stepwise	64.0	64.0	74.2	(<.0001)	73.7	(<.0001)
	Resampling	63.7	63.8	70.0	(<.0001)	70.2	(<.0001)
	Intersection	62.7	62.9	69.3	(<.0001)	72.6	(<.0001)
Semi Parametric	Stepwise	64.0	63.9	NA[1]	NA	6.8	.5584
	Resampling	63.8	63.7	NA	NA	8.0	.4335
	Intersection	62.6	62.8	NA	NA	16.7	.0334
NonParametric (CHAID)	All Variables	58.3	57.9	2.1	.9778	360.4	(<.0001)
	Intersection	59.2	58.9	44.8	(<.0001)	145.9	(<.0001)
Calibrated Generic Bureau Score		62.4	62.6	194.1	(<.0001)	2506.3	(<.0001)

1. The H-L test does not apply. By design, the predicted outcomes under the semi parametric approach are equal to the actual outcomes.
2. The p-values are derived under null hypothesis H_0: $p_j=q_j$ for all j (see equation 8) under the assumption that the H-L $\sim \chi^2$ df=8.

Table 11:
In-Time Validation: Separation (K-S) and Accuracy (H-L) Measures for Parametric and Semi Parametric Models at Development, by Segment

Segment	Scoring Model		Statistic and Sample					
			Kolmogorov-Smirnov		Hosmer-Lemeshow			
	Model Form	Variable Selection	1999 Dev	1999 Hold-out	1999 Dev		1999 Hold-out	
					(value)	(p-value)	(value)	(p-value)
Dirty History and Presently Mildly Delinquent (n=13,302)	Parametric	Stepwise	40.3	38.9	16.5	.0358	22.1	.0047
		Resampling	39.3	37.2	12.5	.1303	23.9	.0024
		Intersection	37.5	37.0	6.1	.6360	10.8	.2133
	Semi Parametric[1]	Stepwise	40.3	38.5	489.8	(<.0001)	548.1	(<.0001)
		Resampling	39.1	37.3	526.8	(<.0001)	565.6	(<.0001)
		Intersection	39.1	37.3	592.7	(<.0001)	625.6	(<.0001)
Dirty History and Presently Current (n=67,814)	Parametric	Stepwise	42.9	43.2	30.8	.0002	26.9	.0007
		Resampling	42.4	43.0	26.0	.0011	27.3	.0006
		Intersection	41.4	42.0	31.1	.0001	32.8	.0001
	Semi Parametric	Stepwise	43.0	43.2	34.7	(<.0001)	25.8	.0011
		Resampling	42.7	43.2	46.1	(<.0001)	42.5	(<.0001)
		Intersection	42.2	42.6	52.7	(<.0001)	35.8	(<.0001)
Clean History and Thin File (n=15,132)	Parametric	Stepwise	58.2	57.1	8.3	.4047	27.3	.0006
		Resampling	57.4	56.1	11.5	.1750	16.7	.0334
		Intersection	54.3	54.4	48.7	(<.0001)	51.7	(<.0001)
	Semi Parametric	Stepwise	57.6	57.1	7.5	.4837	16.4	.0370
		Resampling	57.2	56.8	6.1	.6360	16.9	.0312
		Intersection	55.5	55.2	21.4	.0062	51.8	(<.0001))
Clean History and Thick File (n=242,330)	Parametric	Stepwise	60.2	60.1	84.9	(<.0001)	94.8	(<.0001)
		Resampling	60.0	59.9	74.6	(<.0001)	74.5	(<.0001)
		Intersection	58.7	58.9	67.6	(<.0001)	82.2	(<.0001)
	Semi Parametric	Stepwise	60.1	60.2	9.7	.2867	15.4	.0518
		Resampling	60.0	59.9	7.7	.4633	14.4	.0719
		Intersection	59.1	59.2	6.8	.5584	15.9	.0438

1. The predicted values were derived from the actual default rates in the decile range based on the pooled segment data in Table 10.

Table 12:

Empirical Bad Rates for Alternate Bad Definitions on the Development Samples

Bad Definition		N			Rate	
Event	Horizon (Months)	Bad	Good	Missing[1]	Bad	Good
90+ Days Past Due or Worse	24	24340	314238	0	7.19%	92.81%
	18	19665	318757	156	5.81%	94.19%
	12	14516	323679	383	4.29%	95.71%
	6	8019	329193	1366	2.38%	97.62%
60+ Days Past Due or Worse	24	30107	308471	0	8.89%	91.11%
	18	25204	313218	156	7.45%	92.55%
	12	19478	318717	383	5.76%	94.24%
	6	12192	325020	1366	3.62%	96.38%

1. Missing observations were generated if the lender failed to report performance as of the observation date 6, 12, or 18 months forward.

Table 13:
K-S separation measures for models built to alternate bad definitions on the development sample

Scoring Model			Bad Event Type and Sample			
			90+ Days Past Due or Worse		60+ Days Past Due or Worse	
Model Form	Variable Selection	Bad Event Horizon	Dev	Hold-Out	Dev	Hold-Out
Parametric	Stepwise	24	64.0	64.0	61.5	61.6
		18	65.8	65.8	63.1	63.4
		12	67.8	67.6	65.4	65.4
		6	71.8	72.2	68.6	68.7
	Resampling	24	63.7	63.8	61.4	61.6
		18	65.7	65.5	63.0	63.3
		12	67.7	67.5	65.3	65.3
		6	71.5	72.0	68.4	68.8
	Intersection	24	62.7	62.9	60.5	60.9
		18	64.6	64.8	62.2	62.6
		12	66.7	66.8	64.5	64.8
		6	71.0	71.7	67.8	68.2
Semi Parametric	Stepwise	24	64.0	63.9	61.6	61.7
		18	65.8	65.8	63.0	63.3
		12	67.8	67.6	65.3	65.3
		6	71.7	72.2	68.6	68.7
	Resampling	24	63.8	63.7	61.4	61.6
		18	65.7	65.5	63.0	63.2
		12	67.6	67.5	65.2	65.3
		6	71.6	72.1	68.4	68.8
	Intersection	24	62.6	62.8	60.5	60.9
		18	64.6	64.8	62.1	62.5
		12	66.6	66.7	64.4	64.8
		6	71.0	71.6	67.8	68.2
Non Parametric	All variables	24	58.3	57.9	56.6	56.4
		18	59.9	59.4	57.9	57.7
		12	61.6	61.0	59.6	59.4
		6	65.0	64.9	62.3	62.2
	Intersection	24	59.2	58.9	57.2	56.8
		18	60.9	60.4	58.4	58.1
		12	62.7	62.0	60.1	59.6
		6	65.8	65.7	62.8	62.0
Calibrated Generic Bureau Score		24	62.4	62.6	60.3	60.4
		18	64.6	64.4	61.9	61.9
		12	66.5	66.4	63.9	63.8
		6	70.2	70.7	67.2	66.7

Table 14:
Sample Sizes and Bad Rates For the Development and Out-of-Time Validation Samples
(Bad = 90 Days Past Due, or Worse, over the Following 24 Months)

Segment	Sample and Statistic							
	1999 Development		2000 Validation		2001 Validation		2002 Validation	
	Size	Bad Rate	Size	Bad Rate	Size	Bad Rate	Size	Bad Rate
Dirty History and Presently Mildly Delinquent	13,302	49.27%	29,252	56.25%	35,823	55.42%	39,523	53.79%
Dirty History and Presently Current	67,814	14.60%	133,399	18.02%	163,647	17.36%	174,290	15.79%
Clean History and Thin File	15,132	4.76%	39,032	4.38%	26,732	3.97%	30,984	2.73%
Clean History and Thick File	242,330	2.96%	549,635	3.61%	584,571	3.44%	611,924	3.27%
All	338,578	7.19%	751,318	8.26%	810,773	8.56%	856,721	8.13%

Table 15:
Out-of-Time Validation: Median Scores for Various Models Across Validation Samples

Segment	Scoring Model		Sample			
	Model Form	Variable Selection	1999 Dev	2000 Val	2001 Val	2002 Val
Dirty History and Presently Mildly Delinquent	Parametric	Stepwise	616	611	610	614
		Resampling	616	612	611	615
		Intersection	617	612	611	615
	Semi Parametric	Stepwise	598	598	598	598
		Resampling	599	599	599	599
		Intersection	600	600	600	600
	Calibrated Generic Bureau Score		636	631	631	631
Dirty History and Presently Current	Parametric	Stepwise	676	669	672	675
		Resampling	676	669	672	674
		Intersection	676	668	671	673
	Semi Parametric	Stepwise	673	666	673	673
		Resampling	672	666	672	672
		Intersection	672	666	672	672
	Calibrated Generic Bureau Score		675	670	673	673
Clean History and Thin File	Parametric	Stepwise	729	749	754	759
		Resampling	727	748	752	758
		Intersection	735	739	741	743
	Semi Parametric	Stepwise	733	759	765	770
		Resampling	727	748	764	769
		Intersection	735	741	741	751
	Calibrated Generic Bureau Score		722	737	740	747
Clean History and Thick File	Parametric	Stepwise	750	751	756	761
		Resampling	751	754	758	763
		Intersection	750	751	754	756
	Semi Parametric	Stepwise	759	759	770	772
		Resampling	758	764	769	771
		Intersection	760	760	766	766
	Calibrated Generic Bureau Score		747	751	754	756
All (Pooled)	Parametric	Stepwise	734	737	740	745
		Resampling	735	740	742	747
		Intersection	734	737	738	740
	Semi Parametric	Stepwise	738	738	740	746
		Resampling	737	746	746	748
		Intersection	735	735	735	741
	Nonparametric (CHAID)	All variables	725	730	729	733
		Intersection	724	729	726	730
	Calibrated Generic Bureau Score		734	739	740	743

Table 16:
Out-of-Time Validation: Model Separation and Accuracy Measures (Pooled Across Segments)

Scoring Model		Statistic and Sample											
		Kolmogorov-Smirnov				Hosmer-Lemeshow							
Model Form	Variable Selection	1999 Dev	2000 Val	2001 Val	2002 Val	1999 Dev		2000 Val		2001 Val		2002 Val	
						value	p-value	value	p-value	value	p-value	value	p-value
Parametric	Stepwise	64.0	65.4	66.1	65.8	74.2	(<.0001)	1294.8	(<.0001)	1773.3	(<.0001)	2144.7	(<.0001)
	Resampling	63.7	65.2	65.9	65.5	70.0	(<.0001)	1709.3	(<.0001)	2073.8	(<.0001)	2307.4	(<.0001)
	Intersection	62.7	65.0	65.7	65.2	69.3	(<.0001)	1185.3	(<.0001)	1408.5	(<.0001)	1050.4	(<.0001)
Semi Parametric	Stepwise	63.9	65.4	66.1	65.8	NA	NA	1233.3	(<.0001)	1679.5	(<.0001)	2051.5	(<.0001)
	Resampling	63.7	65.3	66.1	65.7	NA	NA	1402.1	(<.0001)	1713.2	(<.0001)	1915.8	(<.0001)
	Intersection	62.6	65.1	65.9	65.4	NA	NA	932.2	(<.0001)	1063.4	(<.0001)	649.6	(<.0001)
NonParametric (CHAID)	All Variables	58.3	59.3	60.0	59.9	2.1	.9778	3962.4	(<.0001)	3231.5	(<.0001)	2792.4	(<.0001)
	Intersection	59.2	60.6	60.9	60.9	44.8	(<.0001)	4163.6	(<.0001)	3758.9	(<.0001)	3053.0	(<.0001)
Calibrated Generic Bureau Score		62.4	64.9	65.6	65.1	194.1	(<.0001)	2506.3	(<.0001)	4267.6	(<.0001)	2970.0	(<.0001)

Table 17:
Out-of-Time Validation: Separation (K-S) and Accuracy (H-L) Measures for Parametric and Semi Parametric Models, by Segment

Segment	Scoring Model		Statistic and Sample												
			Kolmogorov-Smirnov				Hosmer-Lemeshow							
	Model Form	Variable Selection	1999 Dev	2000 Val	2001 Val	2002 Val	1999 Dev		2000 Val		2001 Val		2002 Val	
							value	p-value	value	p-value	value	p-value	value	p-value
Dirty History and Presently Mildly Delinquent	Parametric	Stepwise	40.3	40.2	40.7	40.7	16.5	.0358	349.7	(<.0001)	181.2	(<.0001)	500.9	(<.0001)
		Resampling	39.3	39.7	39.8	40.5	12.5	.1303	419.3	(<.0001)	227.6	(<.0001)	552.6	(<.0001)
		Intersection	37.5	38.5	38.9	39.5	6.1	.6360	313.7	(<.0001)	169.0	(<.0001)	386.9	(<.0001)
	Semi Parametric	Stepwise	40.3	39.5	39.6	40.8	489.8	(<.0001)	1991.3	(<.0001)	2079.4	(<.0001)	2348.7	(<.0001)
		Resampling	39.1	39.7	39.9	40.9	526.8	(<.0001)	2049.7	(<.0001)	2219.2	(<.0001)	2453.1	(<.0001)
		Intersection	39.1	39.7	39.7	40.7	592.7	(<.0001)	2277.3	(<.0001)	2499.5	(<.0001)	2464.0	(<.0001)
Dirty History and Presently Current	Parametric	Stepwise	42.9	43.2	42.6	42.0	30.8	.0002	113.9	(<.0001)	306.0	(<.0001)	310.2	(<.0001)
		Resampling	42.4	42.8	42.3	41.7	26.0	.0011	121.1	(<.0001)	264.1	(<.0001)	290.3	(<.0001)
		Intersection	41.4	41.8	41.3	41.0	31.1	.0001	127.9	(<.0001)	336.4	(<.0001)	275.4	(<.0001)
	Semi Parametric	Stepwise	43.0	43.0	42.5	42.0	34.7	(<.0001)	137.1	(<.0001)	348.3	(<.0001)	510.5	(<.0001)
		Resampling	42.7	42.9	42.5	41.9	46.1	(<.0001)	125.6	(<.0001)	276.3	(<.0001)	480.6	(<.0001)
		Intersection	42.2	42.5	41.9	41.3	52.7	(<.0001)	98.3	(<.0001)	230.3	(<.0001)	420.2	(<.0001)
Clean History and Thin File	Parametric	Stepwise	58.2	67.7	70.6	68.9	8.3	.4047	142.7	(<.0001)	111.3	(<.0001)	51.7	(<.0001)
		Resampling	57.4	67.2	70.0	68.9	11.5	.1749	157.9	(<.0001)	118.9	(<.0001)	73.2	(<.0001)
		Intersection	54.3	64.7	66.0	68.8	48.7	(<.0001)	188.2	(<.0001)	118.7	(<.0001)	130.6	(<.0001)
	Semi Parametric	Stepwise	57.6	67.5	70.6	69.6	7.5	.4838	132.0	(<.0001)	107.3	(<.0001)	36.9	(<.0001)
		Resampling	57.2	67.7	69.8	68.9	6.1	.6360	137.0	(<.0001)	107.5	(<.0001)	45.9	(<.0001)
		Intersection	55.5	65.1	67.5	68.9	21.4	.0062	168.2	(<.0001)	111.6	(<.0001)	104.2	(<.0001)
Clean History and Thick File	Parametric	Stepwise	60.2	61.2	63.2	63.2	84.9	(<.0001)	1342.1	(<.0001)	1850.5	(<.0001)	2635.1	(<.0001)
		Resampling	60.0	61.0	63.2	63.1	74.6	(<.0001)	1874.0	(<.0001)	2181.4	(<.0001)	2957.1	(<.0001)
		Intersection	58.7	60.6	63.1	62.8	67.6	(<.0001)	1199.3	(<.0001)	1283.7	(<.0001)	1338.6	(<.0001)
	Semi Parametric	Stepwise	60.1	61.2	63.4	63.3	9.7	.2867	1085.7	(<.0001)	1668.9	(<.0001)	2640.3	(<.0001)
		Resampling	60.0	61.1	63.3	63.3	7.7	.4633	1477.5	(<.0001)	1827.3	(<.0001)	2703.6	(<.0001)
		Intersection	59.1	60.9	63.3	63.0	6.8	.5584	830.0	(<.0001)	933.3	(<.0001)	965.4	(<.0001)

Table 18:
Out-of-Time Validation: Separation (K-S) Measures for Different Definitions of Default

Scoring Model			Bad Event Type and Sample									
			90+ Days Past Due or Worse					60+ Days Past Due or Worse				
Model Form	Variable Selection	Bad Event Horizon	Dev	Hold-Out	2000	2001	2002	Dev	Hold-Out	2000	2001	2002
Parametric	Stepwise	24	64.0	64.0	65.4	66.1	65.8	61.5	61.6	63.6	64.6	63.9
		18	65.8	65.8	66.9	67.6	67.6	63.1	63.4	64.8	65.9	65.6
		12	67.8	67.6	68.9	69.2	69.9	65.4	65.4	66.4	67.7	67.8
		6	71.8	72.2	73.0	73.3	73.8	68.6	68.7	70.5	71.0	71.1
	Resampling	24	63.7	63.8	65.2	65.9	65.5	61.4	61.6	63.5	64.4	63.7
		18	65.7	65.5	66.7	67.4	67.4	63.0	63.3	64.6	65.8	65.4
		12	67.7	67.5	68.8	69.1	69.6	65.3	65.3	66.3	67.5	67.6
		6	71.5	72.0	72.9	73.1	73.5	68.4	68.8	70.3	71.0	71.0
	Intersection	24	62.7	62.9	65.0	65.7	65.2	60.5	60.9	63.2	64.1	63.6
		18	64.6	64.8	66.6	67.1	67.1	62.2	62.6	64.5	65.4	65.2
		12	66.7	66.8	68.4	68.8	69.3	64.5	64.8	66.1	67.3	67.4
		6	71.0	71.7	72.5	72.7	73.1	67.8	68.2	70.0	70.6	70.9
Semi Parametric	Stepwise	24	64.0	63.9	65.4	66.1	65.7	61.6	61.7	63.6	64.6	63.9
		18	65.8	65.8	66.9	67.6	67.6	63.0	63.3	64.7	66.0	65.6
		12	67.8	67.6	68.9	69.3	69.8	65.3	65.3	66.5	67.7	67.8
		6	71.7	72.2	73.1	73.3	73.7	68.6	68.7	70.5	71.0	71.2
	Resampling	24	63.8	63.7	65.2	65.9	65.5	61.4	61.6	63.4	64.4	63.7
		18	65.7	65.5	66.7	67.4	67.4	63.0	63.2	64.6	65.8	65.4
		12	67.6	67.5	68.8	69.1	69.6	65.2	65.3	66.3	67.5	67.6
		6	71.6	72.1	72.9	73.0	73.5	68.4	68.8	70.4	70.9	71.0
	Intersection	24	62.6	62.8	64.9	65.7	65.3	60.5	60.9	63.2	64.1	63.6
		18	64.6	64.8	66.5	67.1	67.1	62.1	62.5	64.4	65.4	65.3
		12	66.6	66.7	68.4	68.8	69.3	64.4	64.8	66.0	67.3	67.4
		6	71.0	71.6	72.6	72.6	73.2	67.8	68.2	70.0	70.6	70.9
Non Parametric	All variables	24	58.3	57.9	59.3	60.0	59.9	56.6	56.4	58.2	59.0	58.9
		18	59.9	59.4	60.5	61.1	61.3	57.9	57.7	59.2	60.0	60.2
		12	61.6	61.0	62.2	62.2	63.0	59.6	59.4	60.5	61.3	61.7
		6	65.0	64.9	65.4	65.2	66.1	62.3	62.2	63.6	63.6	64.3
	Intersection	24	59.2	58.9	60.6	61.0	60.9	57.2	56.8	59.2	59.7	59.6
		18	60.9	60.4	61.8	62.3	62.5	58.4	58.1	60.1	60.7	60.9
		12	62.7	62.0	63.5	63.7	64.5	60.1	59.6	61.3	62.2	62.7
		6	65.8	65.7	66.3	66.6	67.1	62.8	62.0	64.3	64.4	64.9
Calibrated Generic Bureau Score		24	62.4	62.6	64.9	65.6	65.1	60.3	60.4	63.1	64.0	63.2
		18	64.6	64.4	66.4	67.1	66.9	61.9	61.9	64.2	65.2	64.8
		12	66.5	66.4	68.2	68.8	69.2	63.9	63.8	65.7	66.9	66.9
		6	70.2	70.7	71.8	72.2	72.5	67.2	66.7	69.1	69.9	69.8

Figure 1:
OCC/RAD CCDB Sample Design: 1999 & 2000

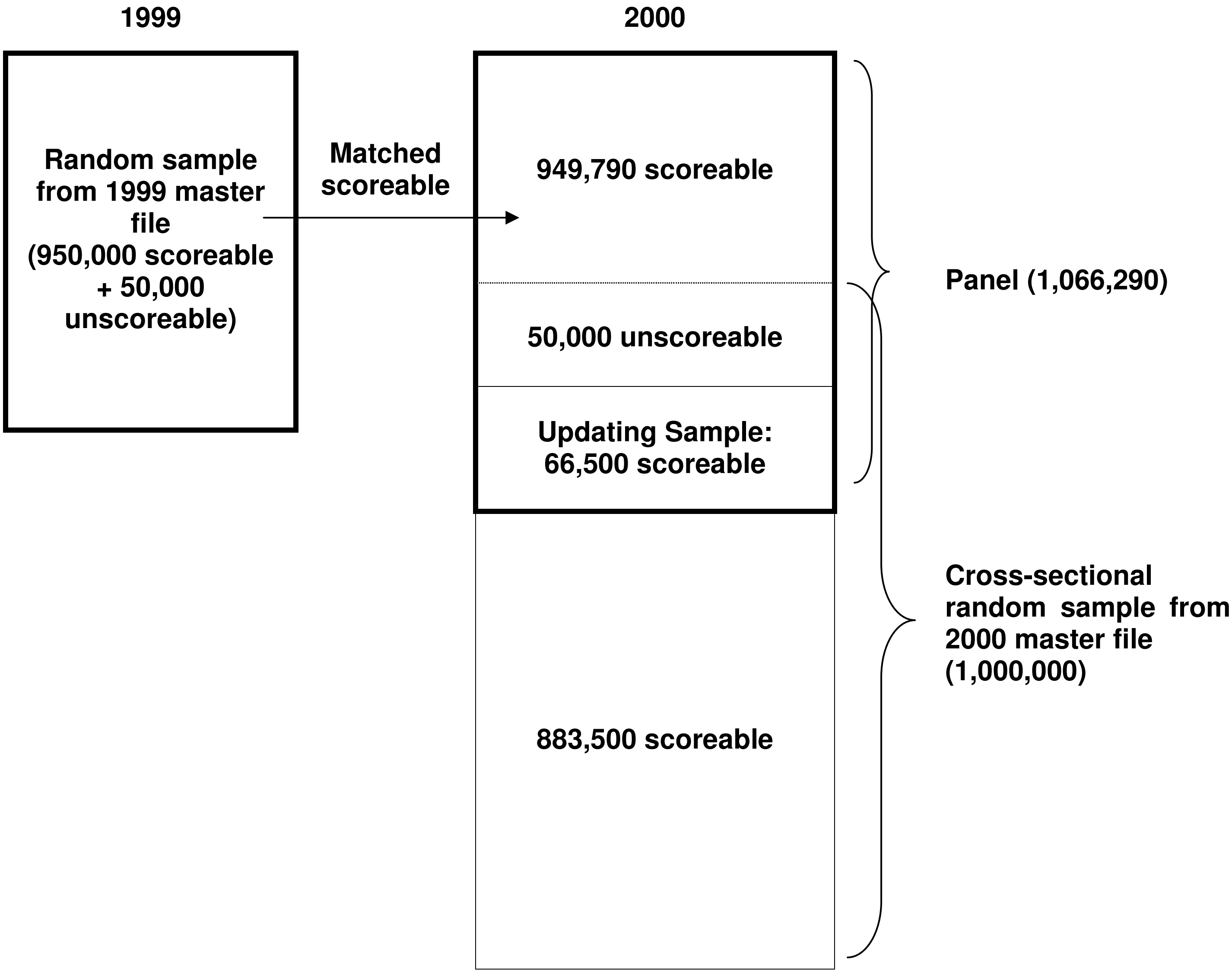


Figure 2:
1999 Development and 1999 Hold-Out Sample Construction, and Bad Rates
(Bad = 90+Days Past Due, or Worse, over the Following 24 Months)

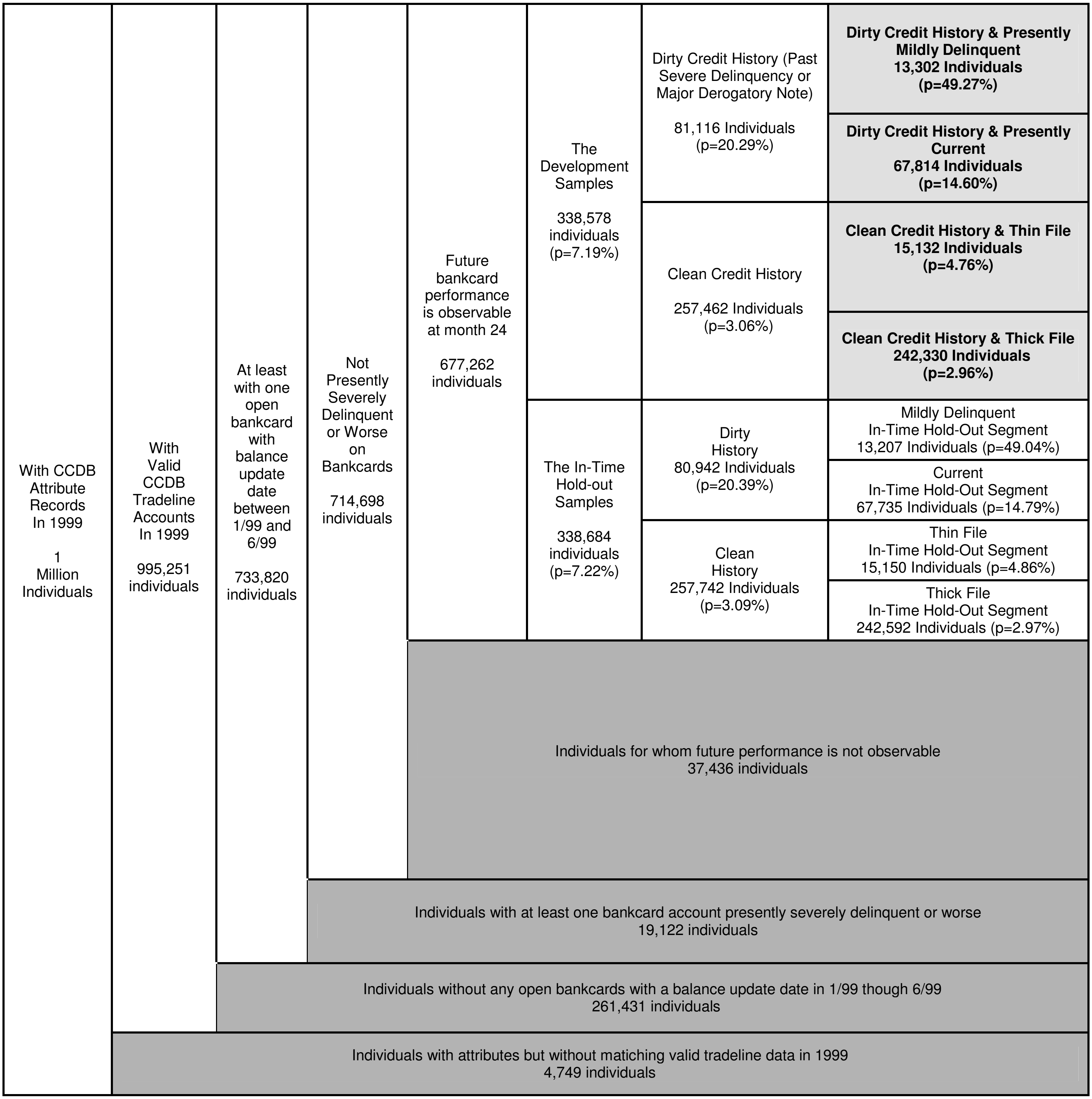


Figure 3:

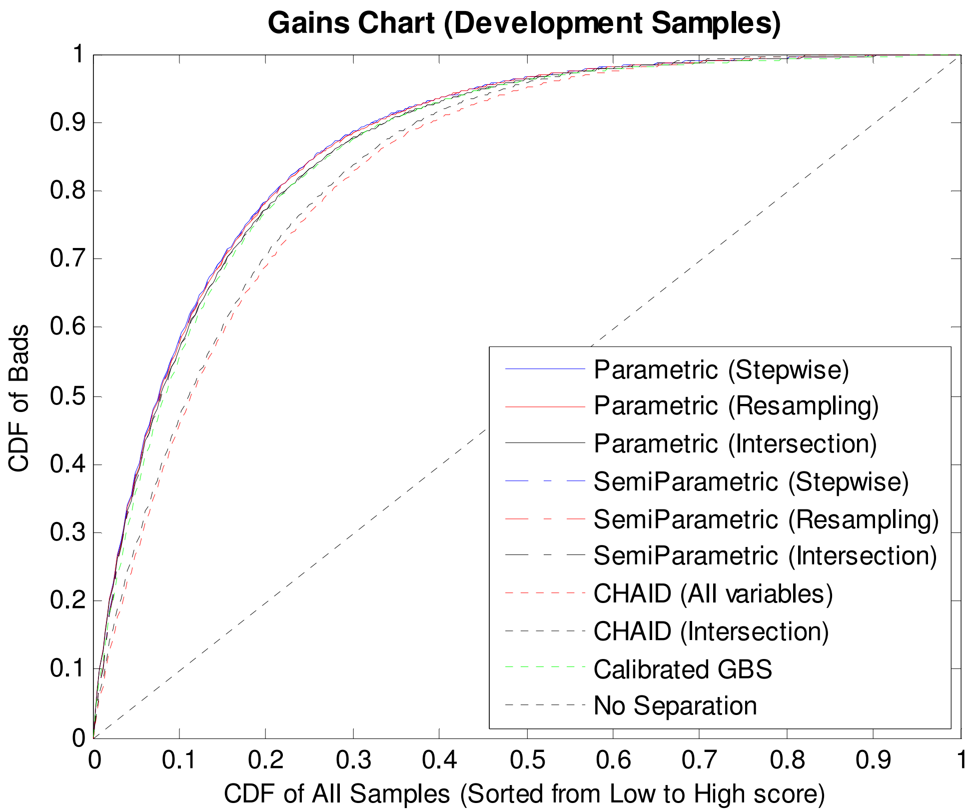

Gains Chart (Development Samples)
CDF of Bads
CDF of All Samples (Sorted from Low to High score)
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
Parametric (Stepwise)
Parametric (Resampling)
Parametric (Intersection)
SemiParametric (Stepwise)
SemiParametric (Resampling)
SemiParametric (Intersection)
CHAID (All variables)
CHAID (Intersection)
Calibrated GBS
No Separation

Figure 4:

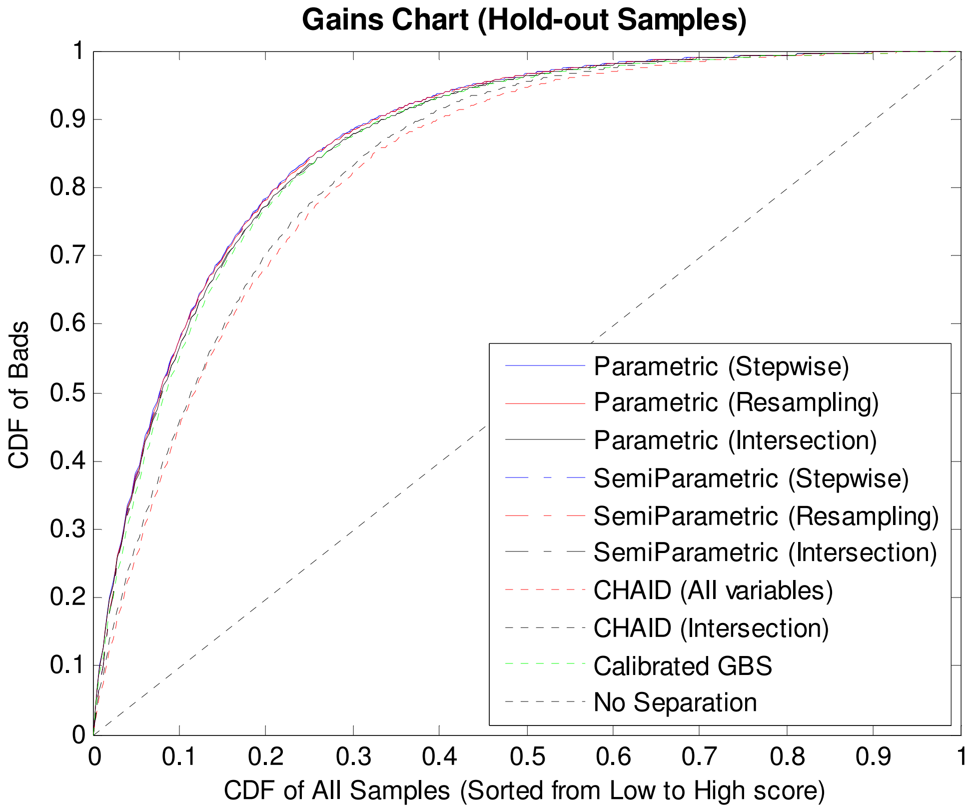

Gains Chart (Hold-out Samples)
CDF of Bads
CDF of All Samples (Sorted from Low to High score)
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
Parametric (Stepwise)
Parametric (Resampling)
Parametric (Intersection)
SemiParametric (Stepwise)
SemiParametric (Resampling)
SemiParametric (Intersection)
CHAID (All variables)
CHAID (Intersection)
Calibrated GBS
No Separation

Figure 5:

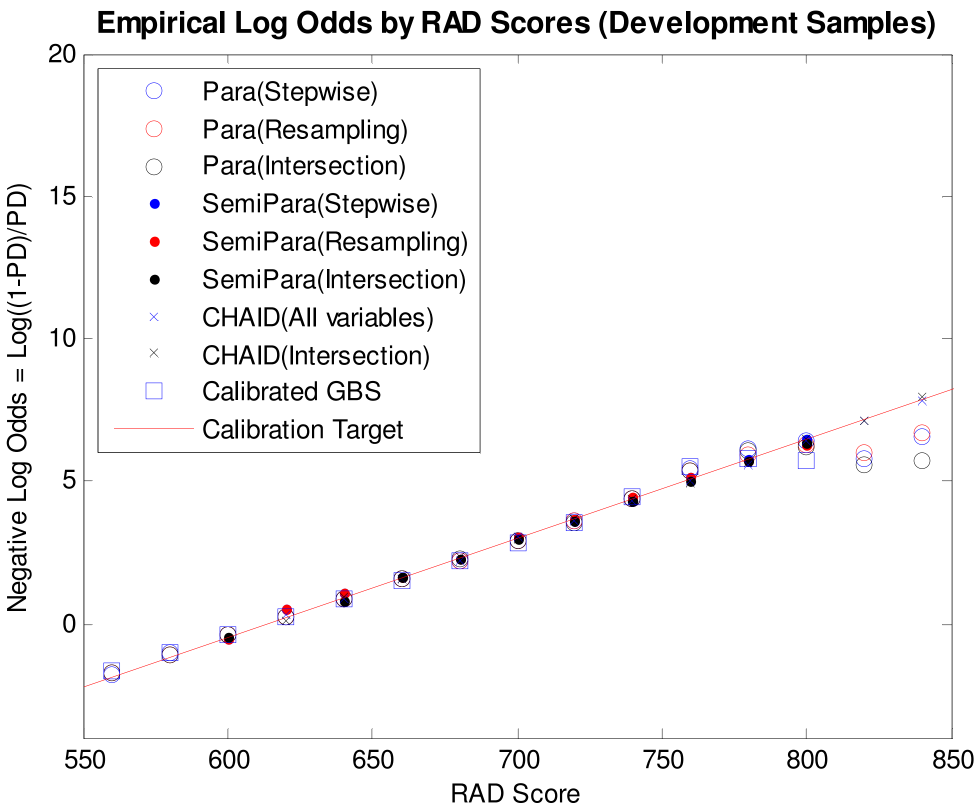

Empirical Log Odds by RAD Scores (Development Samples)
Para(Stepwise)
Para(Resampling)
Para(Intersection)
SemiPara(Stepwise)
SemiPara(Resampling)
SemiPara(Intersection)
CHAID(All variables)
CHAID(Intersection)
Calibrated GBS
Calibration Target
Negative Log Odds = Log((1-PD)/PD)
RAD Score
20
15
10
5
0
550
600
650
700
750
800
850

Figure 6:

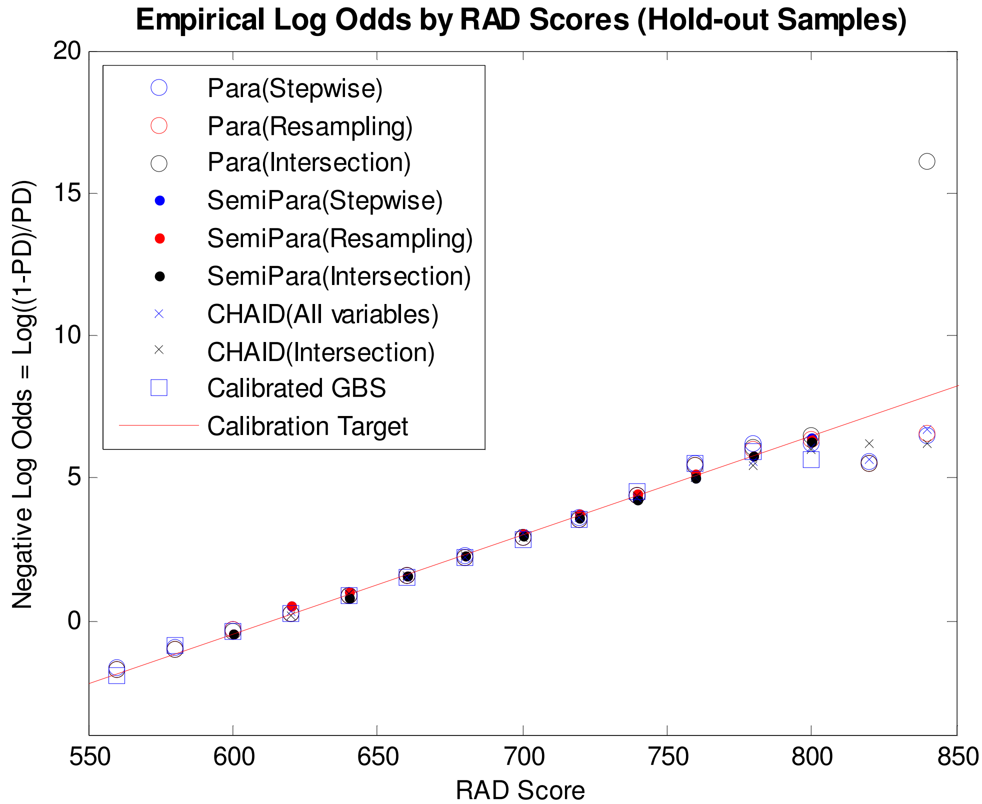

Empirical Log Odds by RAD Scores (Hold-out Samples)
Para(Stepwise)
Para(Resampling)
Para(Intersection)
SemiPara(Stepwise)
SemiPara(Resampling)
SemiPara(Intersection)
CHAID(All variables)
CHAID(Intersection)
Calibrated GBS
Calibration Target
Negative Log Odds = Log((1-PD)/PD)
RAD Score
20
15
10
5
0
550
600
650
700
750
800
850

Figure 7: RAD Scores, Full Sample, Development and Validation

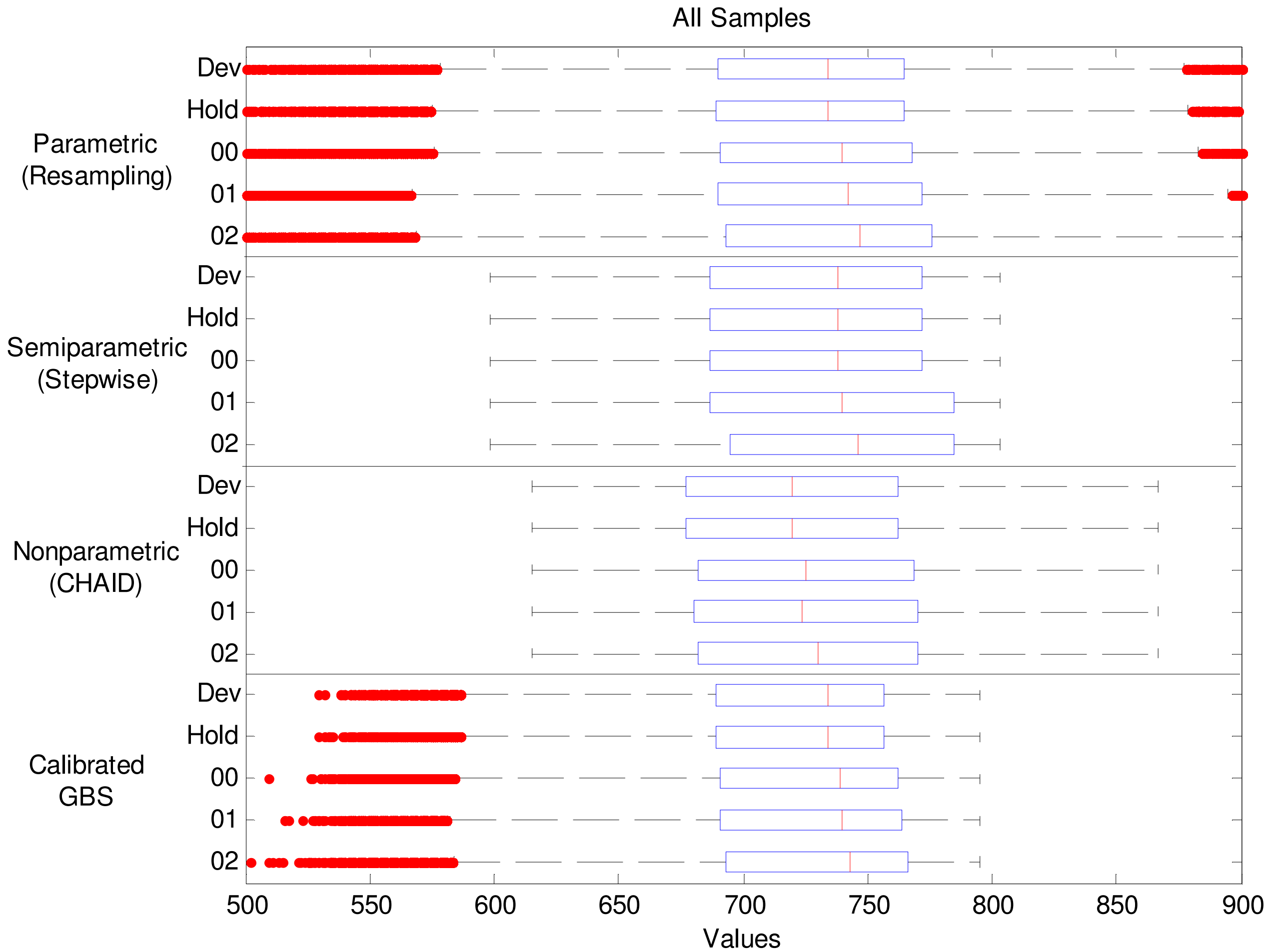


Figure 8: RAD Scores, Dirty/Delinquent Sample, Development and Validation

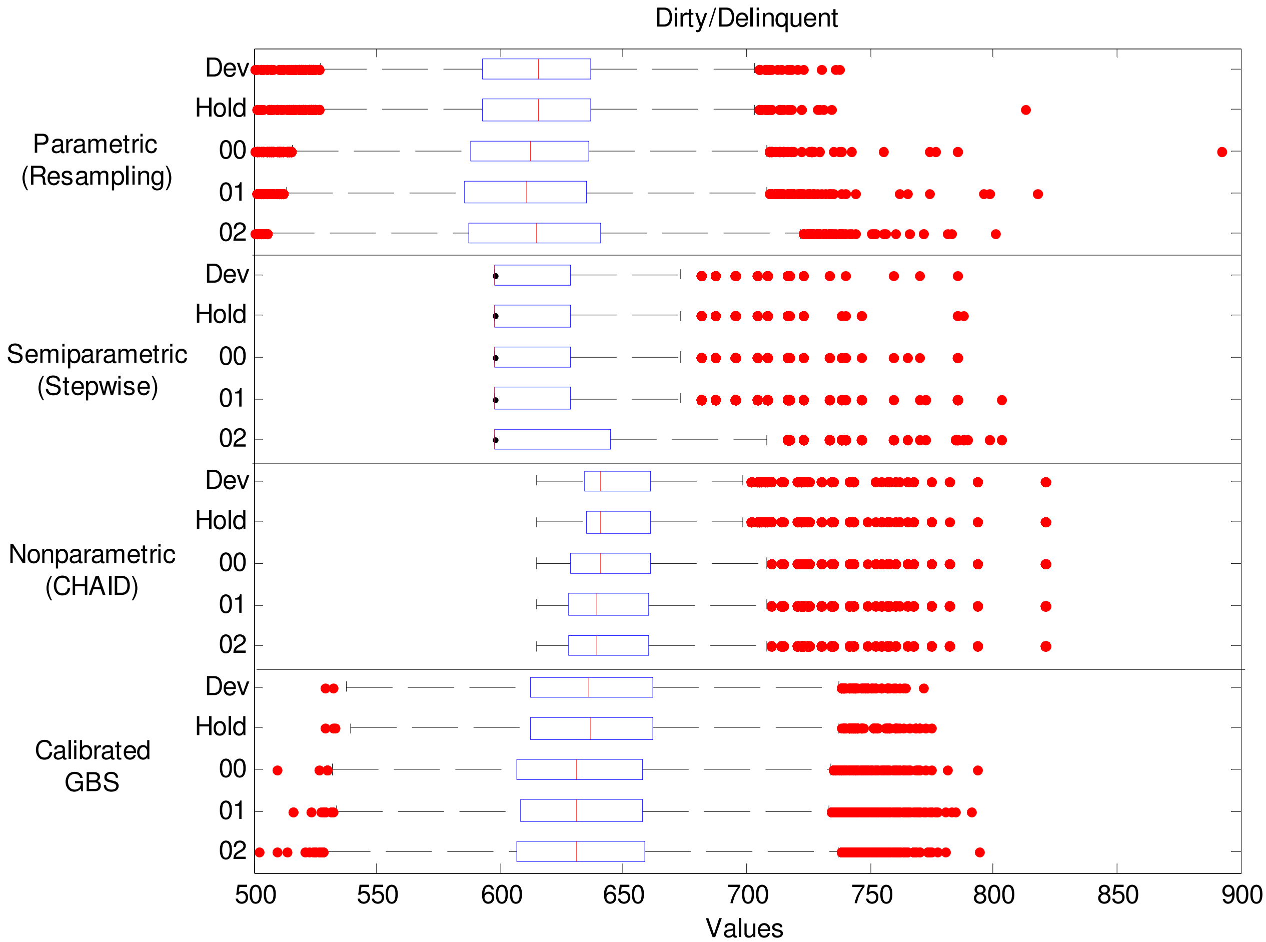


Figure 9: RAD Scores, Dirty/Current Sample, Development and Validation

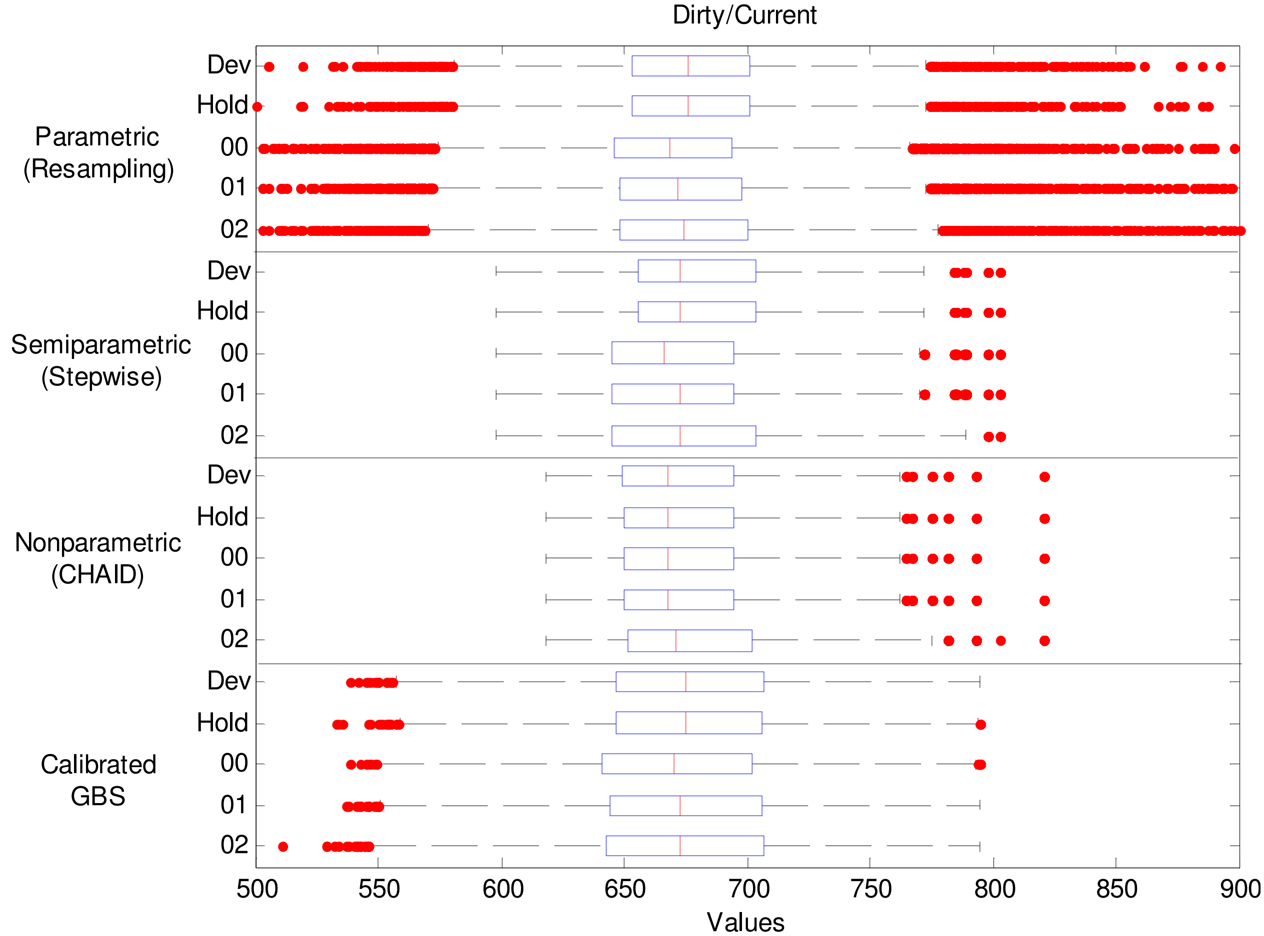


Figure 10: RAD Scores, Clean/Thin Sample, Development and Validation

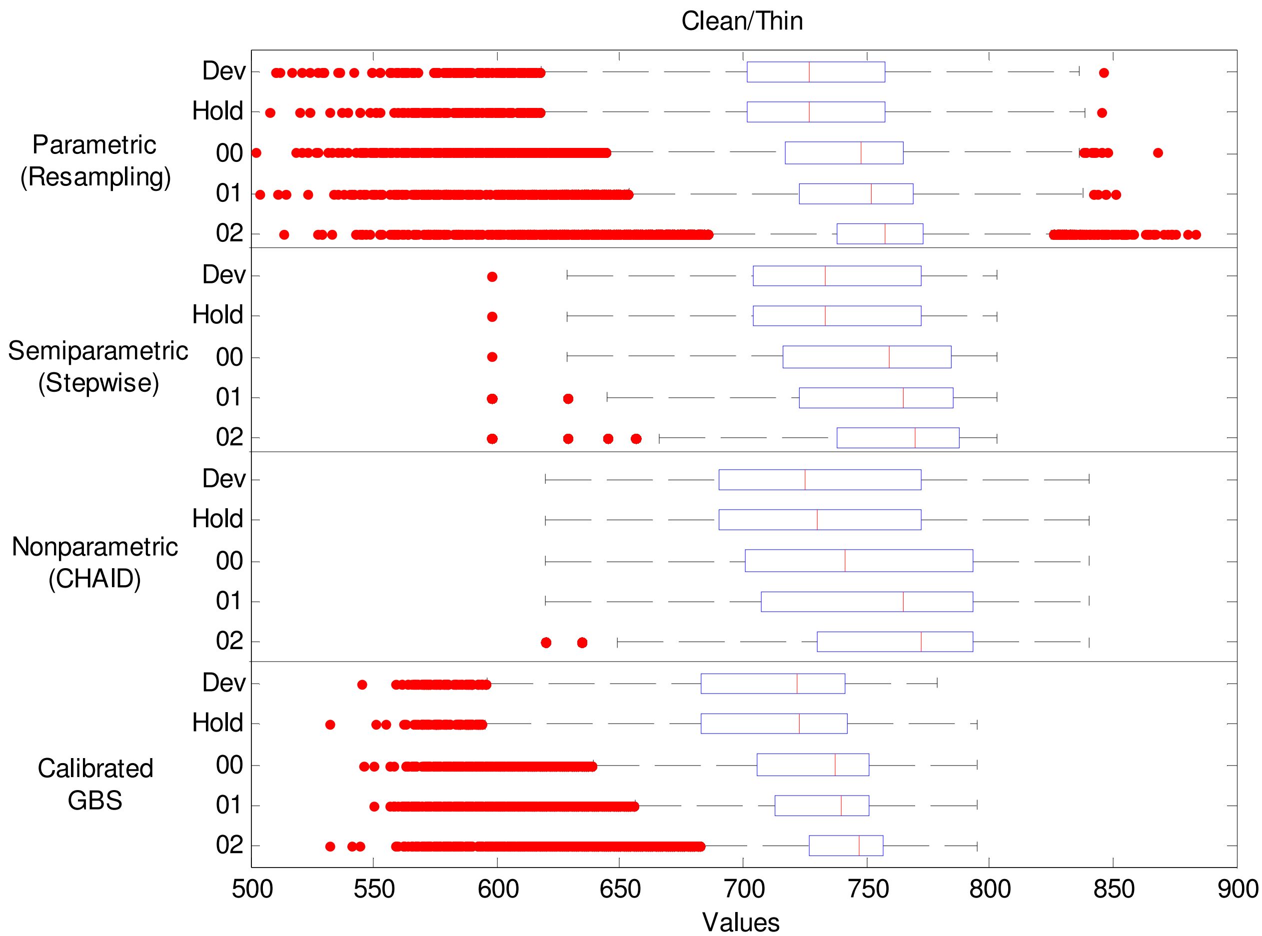


Figure 11: RAD Scores, Clean/Thick Sample, Development and Validation

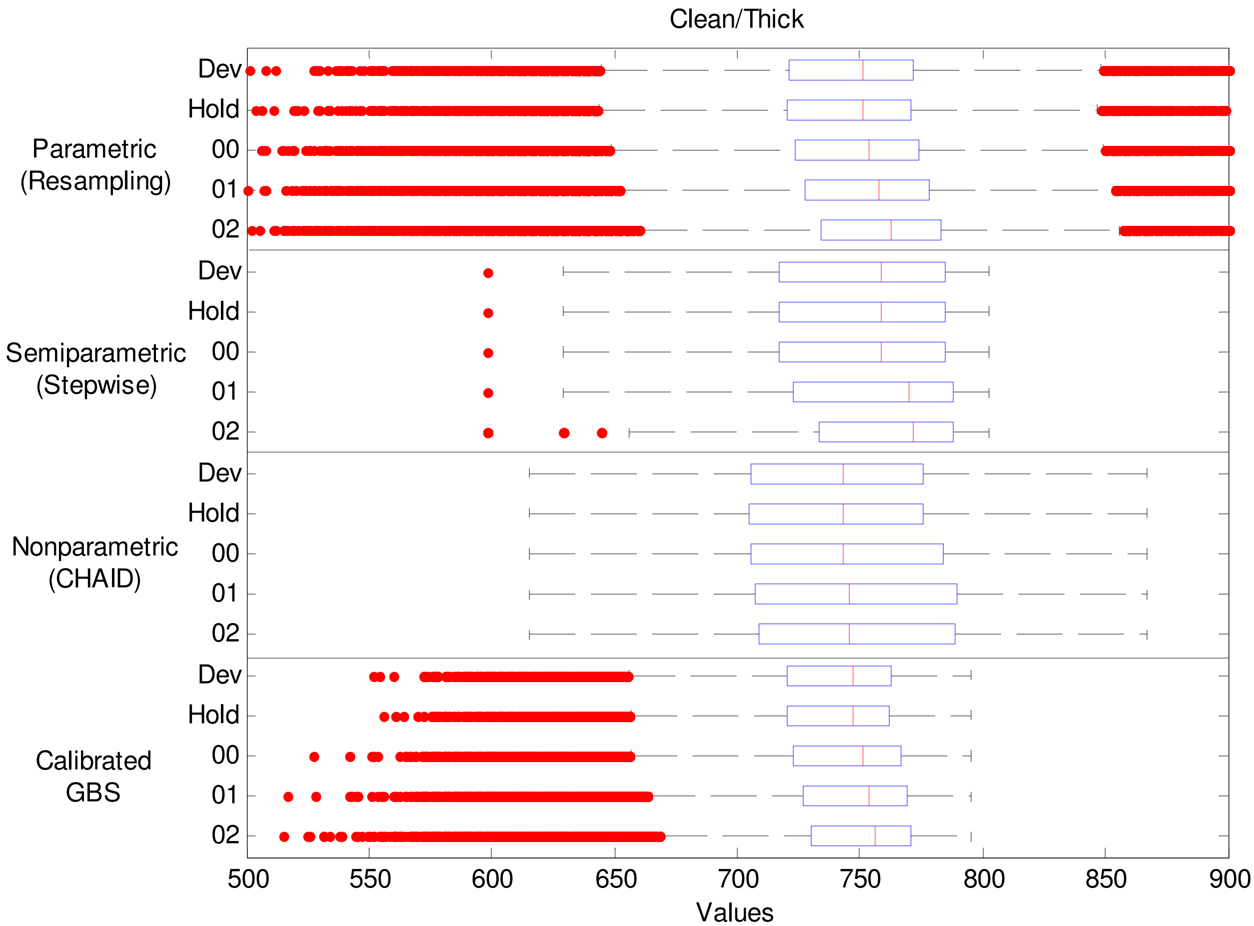
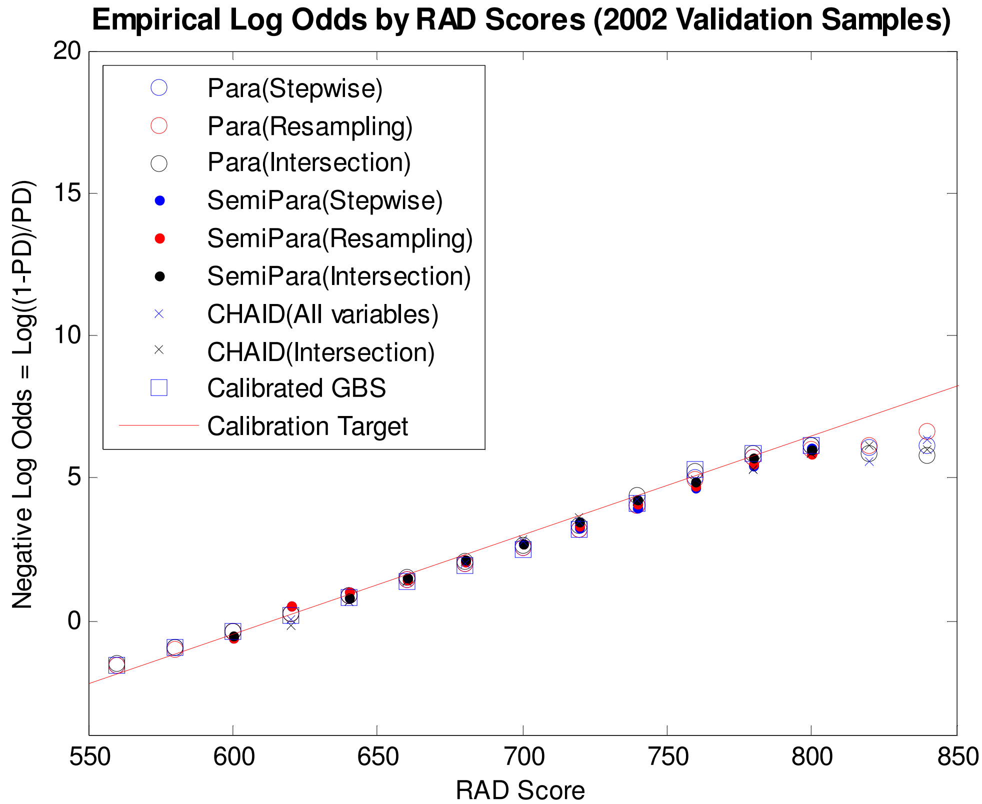


Figure 12:



Empirical Log Odds by RAD Scores (2002 Validation Samples)
Para(Stepwise)
Para(Resampling)
Para(Intersection)
SemiPara(Stepwise)
SemiPara(Resampling)
SemiPara(Intersection)
CHAID(All variables)
CHAID(Intersection)
Calibrated GBS
Calibration Target
Negative Log Odds = Log((1-PD)/PD)
RAD Score
20
15
10
5
0
550
600
650
700
750
800
850

Figure 13:

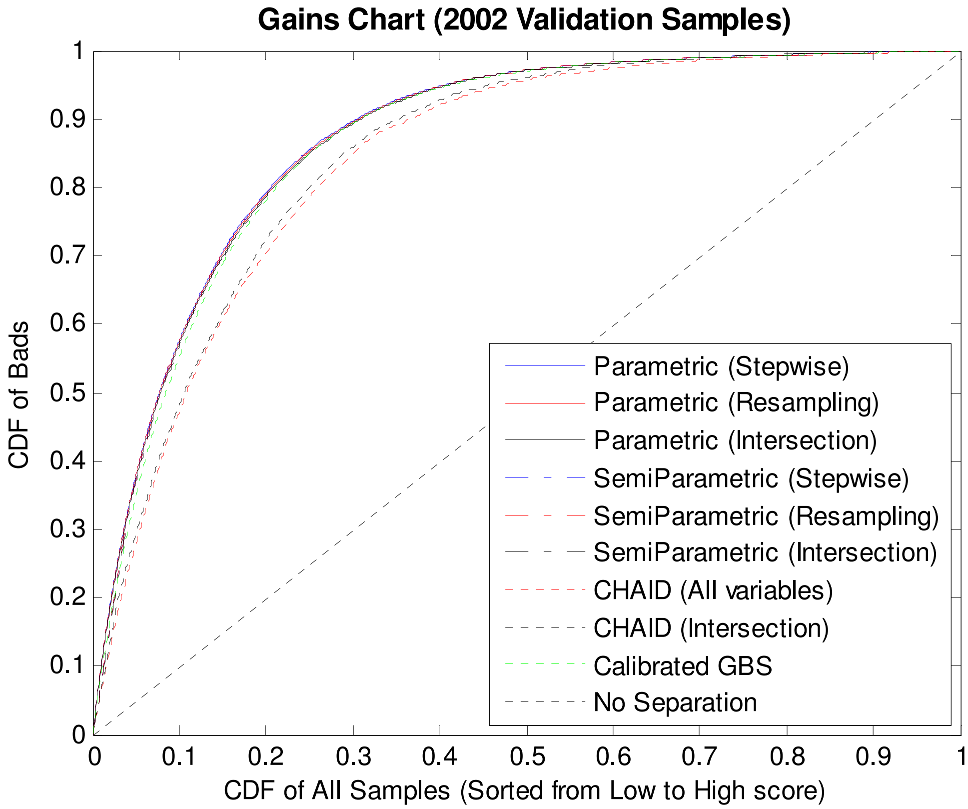

Gains Chart (2002 Validation Samples)
CDF of Bads
CDF of All Samples (Sorted from Low to High score)
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
Parametric (Stepwise)
Parametric (Resampling)
Parametric (Intersection)
SemiParametric (Stepwise)
SemiParametric (Resampling)
SemiParametric (Intersection)
CHAID (All variables)
CHAID (Intersection)
Calibrated GBS
No Separation

Figure 14:

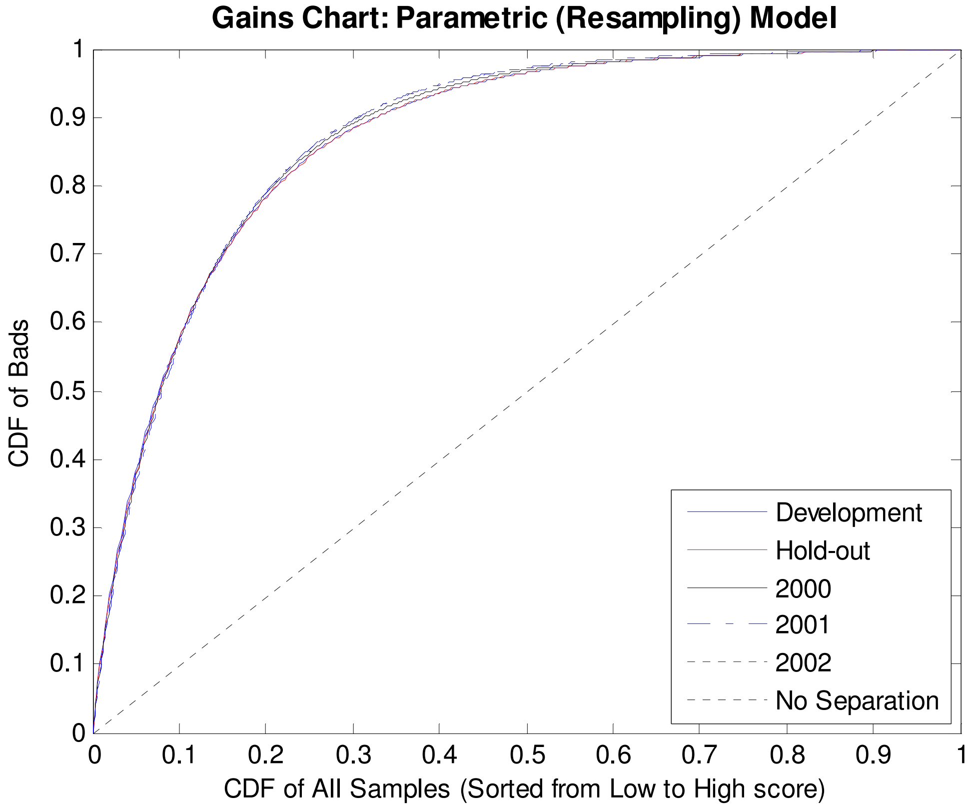
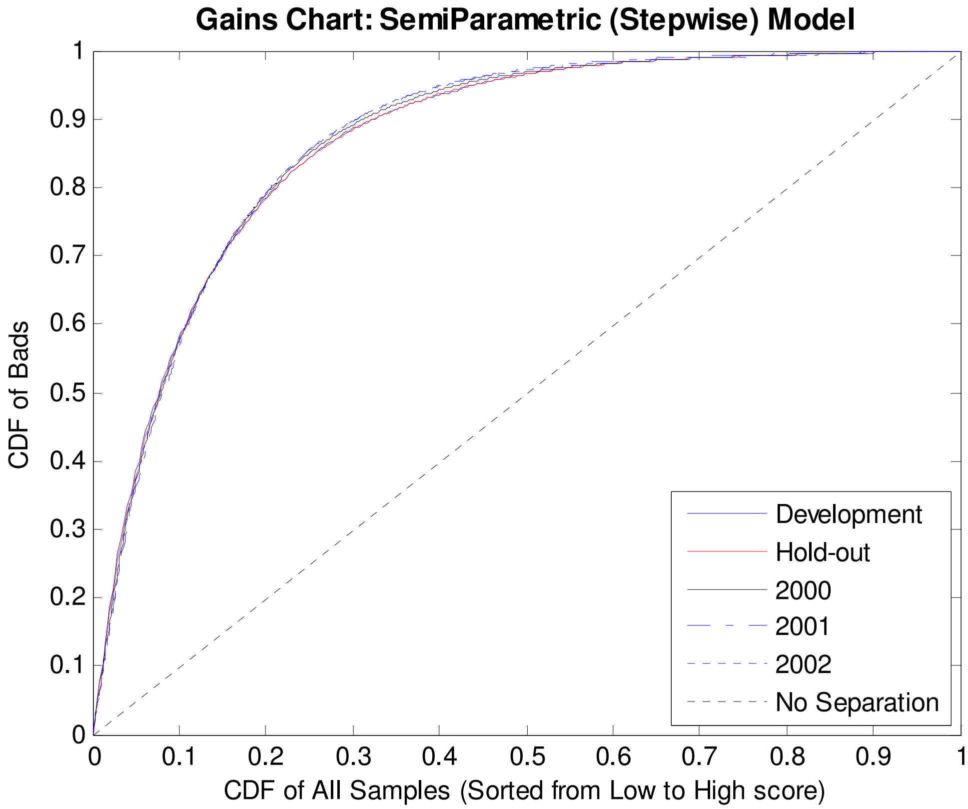

Gains Chart: Parametric (Resampling) Model
CDF of Bads
CDF of All Samples (Sorted from Low to High score)
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
Development
Hold-out
2000
2001
2002
No Separation

Figure 15:



Gains Chart: SemiParametric (Stepwise) Model
CDF of Bads
CDF of All Samples (Sorted from Low to High score)
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
Development
Hold-out
2000
2001
2002
No Separation

Figure 16:

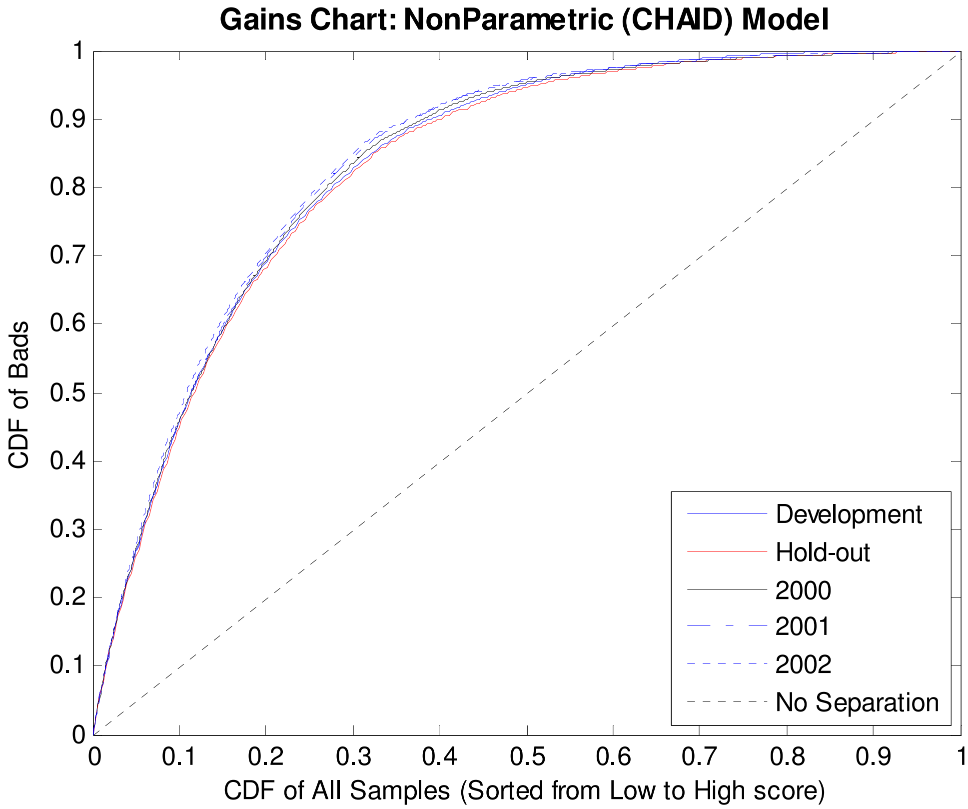
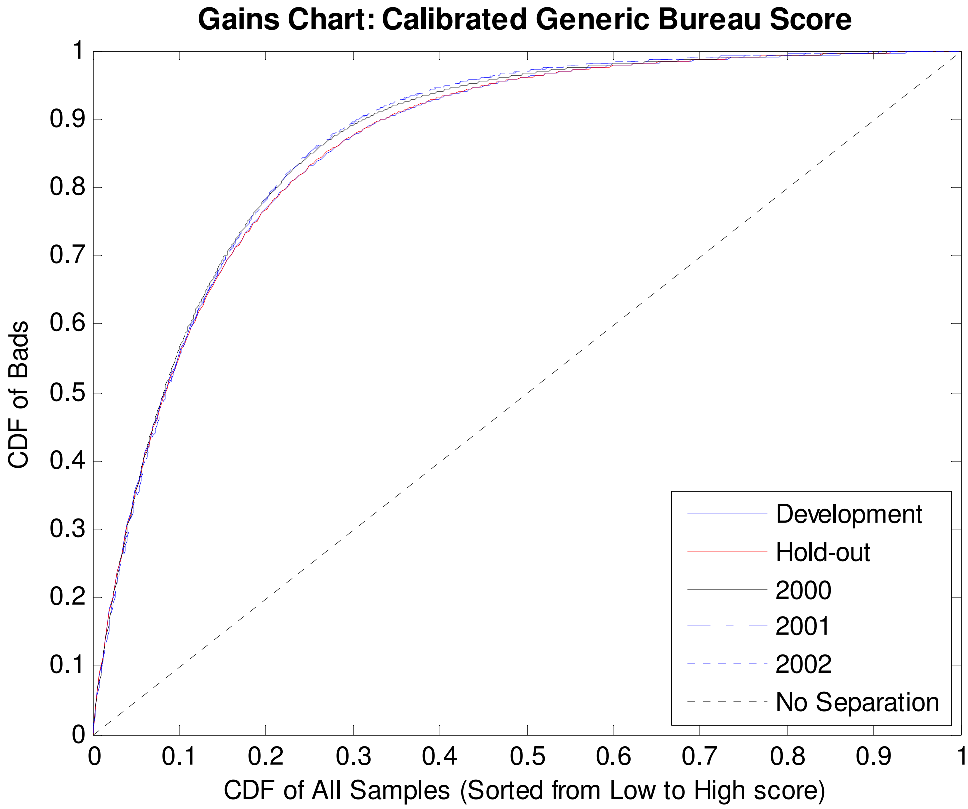

Gains Chart: NonParametric (CHAID) Model
CDF of Bads
CDF of All Samples (Sorted from Low to High score)
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
Development
Hold-out
2000
2001
2002
No Separation

Figure 17:


Gains Chart: Calibrated Generic Bureau Score
CDF of Bads
CDF of All Samples (Sorted from Low to High score)
Development
Hold-out
2000
2001
2002
No Separation
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1