Improving Credit Scoring Accuracy by Population Segmentation

Shubhamoy Dey

Abstract

In the research work presented in this paper the possibility of improving scoring accuracy by pre-segmenting the population and then building subpopulation scorecards for each segment, is examined. A novel subspace selection technique together with a two-step cluster detection technique is used to identify naturally occurring segments (represented by clusters) in the data. Empirical results obtained from a real credit card data set demonstrate the strength of the approach. The segmentation method is able to separate the population into parts with widely different risk profiles (measured by the subpopulation odds). In the following step, different scorecards are built for each of these population segments. Significant improvement of overall accuracy of the scoring system is shown to be achieved by these segment-specific scorecards. Further, the scoring system resulting from the proposed approach is easy to implement in credit card account management systems that are currently used in the credit card industry.

Keywords: Credit risk, consumer credit, cluster detection, dimensionality reduction

1. Introduction

Credit scoring is an application of techniques from areas of statistics, operations research, and related disciplines like machine learning and data mining, that attempts to forecast financial risk associated with lending. It is essentially a way of distinguishing groups with different credit risk in a population, based on observed characteristics.

Although the concept of recognizing groups in a population was introduced in statistics by Fisher (1936), Durand (1941) was the first to propose that the technique could be used to distinguish between good and bad loans. A good account of the early successes can be found in Myers and Forgy (1963), and Churchill et al. (1977). Discussion on the basics of credit scoring has been excluded from this paper, as such details can be found in text books such as Thomas, Crook and Edelman (1992), Mays (1998), and Mays (2001). A number of previous surveys have described recent developments from different perspectives (e.g. Rosenberg and Gleit, 1994; Hand and Henley, 1997; Thomas, 2000; Thomas et al., 2005; and Crook et al, 2007).

Traditionally, linear discriminant analysis and logistic regression have been the core techniques used to construct scoring models. Since credit decisions often involve several billions of dollars, even an improvement in accuracy of a fraction of a percentage point can lead to significant gains. This has prompted both researchers and practitioners to investigate every possibility of improving scoring model accuracy. The use of logistic regression models, nonparametric models such as k-nearest neighbor, classification trees, and neural network models have been examined by Henley (1995), Henley and Hand (1996), Makowski (1985) and Angelini et al. (2008) respectively, in the quest for higher scoring accuracy.

Practitioners have been able to achieve higher scoring accuracy by segmenting the population based on implicitly different risk profiles. For example, in behavioural scoring, the population is usually split into 'up-to-date payers', '1-payment-defaulters' and '2-payments-defaulters' sub-populations, and building separate scorecards for each sub-population. The research described in this article focuses on improving the accuracy of credit scoring models by pre-segmenting the population and then building scorecards for each segment. The segmentation is done using a combination of a novel subspace selection technique proposed by Dey and Roberts (2002), and a two-step cluster detection technique. Empirical results obtained from real data demonstrate the strength of the two-stage approach. The segmentation stage is able to separate the population into parts with widely different risk profiles (measured by the population odds). In the second stage, different scorecards are built for each of these population segments, thereby significantly improving the overall accuracy of the scoring system.

2. Related Work

Classification trees essentially reduce heterogeneity by segmenting and re-segmenting the population. Frydman et al. (1985), in their classification tree approach, were among the first to suggest that splitting the population (into 'movers' and 'stayers' as described in their paper) would improve accuracy. However, Banasik et al. (1996) opine that it may not always result in improved scoring systems in practice unless the segments are distinctive enough. According to Thomas (2000), how many segments of the population should have different scorecards remains an important research question.

Various two-stage approaches have been proposed and reported in research literature (e.g. Hsieh, 2004; Huang et al., 2006; Laha, 2007; Chuang and Lin, 2009; Sustersic et al, 2009). Lee et al. (2002) use a combination of discriminant analysis and neural networks and report that the neural network training time is reduced and that scoring accuracy improves outperforming traditional discriminant analysis and logistic regression approaches. Hsieh (2005) uses clustering to segregate unrepresentative samples into 'isolated and inconsistent' clusters, and neural networks to construct the credit scoring model out of the samples in the 'consistent' clusters. A self-organizing map clustering algorithm is used to determine the number of clusters and k-means to do the clustering. Lee and Chen (2005) use multivariate adaptive regression splines to build a credit scoring model and use the significant variables as input to a neural network model. Huang et al. (2007) report promising results by combining genetic algorithms (GA) with support vector machine (SVM) classifier. The proposed hybrid GA-SVM strategy can simultaneously perform feature selection task and model parameter optimization. Chen et al. (2009) use classification and regression trees and multivariate adaptive regression splines to perform feature selection and a SVM classifier resulting in reduced type-II errors, while Xu et al. (2009) propose a combination of link analysis and SVM. Similarly Lin (2009) examines a two-stage model of logistic regression and neural networks, and Huang (2009) proposes the use of a kernel graph embedding scheme for dimensionality reduction to improve the performance of SVM classifiers.

It is evident from the results of these recent works that population segmentation is a promising approach but determining the number of 'optimum' segments and high dimensional data encountered in practical credit scoring situations pose challenges. One of the most popular time tested methods of unsupervised population segmentation is clustering. As discussed later, segments obtained by this method have the additional advantage of being easily implementable in the (usually mainframe based) software environments used by most lending institutions.

2.1 Dimensionality and Subspace Selection: In Dey and Roberts (2002), and Dey (2002) the problem of high dimensionality has been examined in some detail and a novel approach for subspace selection has been proposed. As discussed in Dey and Roberts (2002), data mining applications, and applications such as credit scoring, usually encounter high-dimensional data spaces. Most of these dimensions contain 'uninteresting' data, which would not only be of little value in terms of discovery of any rules or patterns, but have been shown to mislead some classification algorithms. Since, the computational effort increases very significantly in the presence of a large number of attributes, it is highly desirable that all irrelevant attributes be discarded.

Often, patterns of interest are embedded in lower dimensional subspaces of data. If the data space S has k attributes $\in \{a_1,a_2 \ldots a_k\}$, then an n-dimensional subspace sn of the data space S can be formed by selecting a combination of n attributes from the set $\{a_1,a_2 \ldots a_k\}$, where $n \leq k$. Subspaces of feature set are all the more important for credit scoring applications because many of the features in the data are usually correlated. Correlation effects impact the model weights during scorecard construction using techniques such as logistic regression and discriminant analysis (Coffman, 1986). Thus, only certain subsets of features (i.e. subspaces) can be used to build robust scorecards.

It is usual to tackle this problem by getting some attributes and subspaces identified by the user (or domain experts). For even moderately large number of attributes, the number of possible subspaces is so large (2^k-1) that it is quite unlikely that the 'experts' would be able to identify all the 'interesting' subspaces. Although there are a number of techniques such as principal component analysis and multi-dimensional scaling available for dimensionality reduction and feature selection, they cannot be used to identify 'interesting' subspaces.

Agarwal et al. (1998) argued that the clustering tendency of a feature space (or subspace) was a good indicator of its 'interestingness'. Based on those ideas and a direct measure of clustering tendency known as Hopkins Statistic (after Hopkins, 1954), a *'Measure of Merit'* for subspaces was proposed by Dey and Roberts (2002). The *measure of merit* (M) proposed was a product of two components coverage and concentration of 'dense' areas (in other words clusters) within the feature space. The paper also reported success in identifying interesting subspaces using a highly scalable algorithm based on the rank ordering of subspaces using the *Measure of Merit*. The 'harshness' with which 'uninteresting' subspaces are filtered out by the algorithm is controlled by a threshold value parameter (τ) and a cut-off value for the *Measure of Merit* (C_M).

2.2 Two-step Clustering: Apart from being one of the most important data mining tasks (Agrawal et al., 1993), cluster detection is one of the few data mining techniques that can be classed as undirected data mining or unsupervised learning. It is the task of discovering structure in data by segmenting the heterogeneous population (represented by the complete dataset) into a number of more homogeneous subgroups called clusters. The process is based on similarity between 'objects', and the number and nature of the subgroups (or clusters) is generally not known in advance. It has been the subject of intense research for many years and a large number of algorithms have been proposed. A good comparison of several algorithms can be found in the study by Michaud (1997). Most algorithms require the number of clusters to be specified in advance, and the few that do not require this (e.g. Nakamura and Kehtarnavaz, 1998; Kothari and Pitts, 1999) are computationally intensive.

Based on Balanced Iterative Reducing and Clustering using Hierarchies (BIRCH) algorithm presented in the papers of Zhang et al. (1996) and Chiu et al. (2001), a two-step clustering method has been developed by SPSS Inc. (for SPSS Versions 11.5 and higher). It can handle both continuous and categorical variables and requires only one data pass. It has two steps: (1) pre-cluster the records into many small sub-clusters; (2) cluster the sub-clusters resulting from pre-cluster step into the desired number of clusters. It can also automatically detect the number of clusters. Euclidean distance and log-likelihood criterion are used as distance measures for purely continuous variables and mixture of categorical and continuous variables respectively. The problem of determining the optimal number of clusters is overcome using the Schwarz Bayesian Criterion. Although the log-likelihood distance measure is based on the assumption that the continuous variables follow normal distribution and that the categorical variables have multinomial distribution, the algorithm behaves reasonably well even when the assumption is not met. A study regarding its efficacy has been reported by Kayri (2007).

3. Two-stage Segmented Credit Scoring

The proposed method consists of a pre-processing step and a scoring step. The scoring step itself is a two-stage process. These process steps are depicted in the schematic diagram in Figure-1. The pre-processing step is an implementation of the *Measure of Merit* based algorithm of Dey and Roberts (2002). The output of this step is a set of interesting subspaces. Next, SPPS (Version 15.0) two-step clustering method is used to find the clusters in these subspaces. Subsequently, credit risk scorecards are built for each of the population segments (formed from the clusters) using logistic regression. The output of the 2-stage process is a set of 'conventional' credit risk scorecards (one for each population segment). As shown in Figure-1, the pre-processing step is performed by a set of C-language programs and both stages of the 2-stage credit scoring process is performed using SPSS. All the steps (pre-processing, segmentation and scorecard building) are performed using the same data set, the two new steps introduced by our proposed approach do not require any additional data.

As set out in Section 2.1 above, even a data set with a moderate number of attributes has an exponentially large number of subspaces. The pre-processing step uses inherent clustering tendency of the data to identify a sub-set of 'interesting' subspaces small enough

to be closely examined for the existence of useful population segments. This population segmentation is achieved by a cluster detection technique capable of utilizing both continuous and categorical variables (as discussed in Section 2.2) - an appropriate one to use, since both these types of variables are generally present in credit card data sets. Finally, in the second stage of the 2-stage process, we build conventional scorecards using the traditional logistic regression technique due to a number of reasons. Firstly, scorecards are used to score millions of credit card accounts through very specialized automated scoring software, and the output of our process being a set of conventional scorecards keeps the interface to those scoring software unchanged (apart for an additional step to determine the appropriate population segment - described fully in Section 4). Secondly, the dependent / target variable in credit scoring data sets is binary (good or bad account) in nature, and logistic regression is one of the strongest techniques applicable to this type of data. Thirdly, scorecards and logistic regression are well understood and trusted by financial regulators (such as the Financial Services Authority in UK), and are therefore more readily acceptable as 'sound' scoring systems to implement.

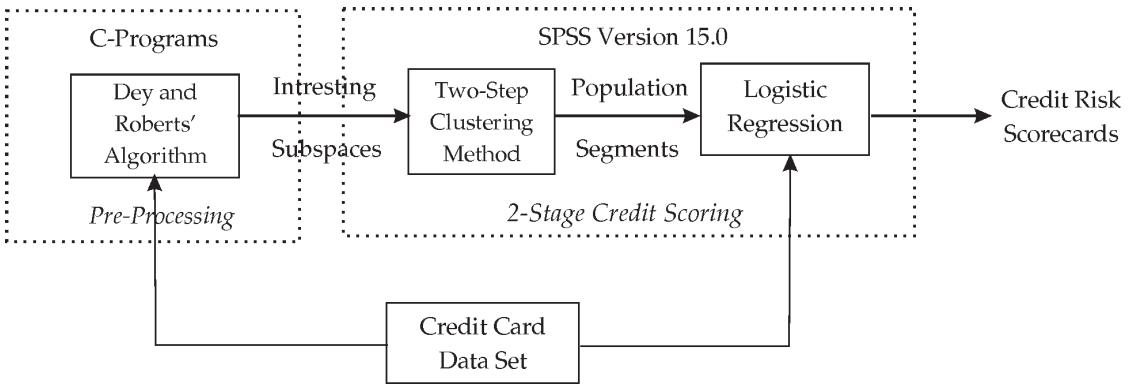


Figure-1 : Proposed 2-stage Segmented Credit Scoring Approach

3.1 Empirical Study Data Set: A credit card dataset containing behavioural data from a major credit card company has been used to demonstrate the feasibility and effectiveness of credit scoring using the proposed two-stage approach. There were totally 62599 credit card accounts in the data set and the ratio of good to bad accounts (the good-bad odds) of the population was 2.742. Out of them, the data of 21477 randomly selected accounts (approximately ?) were kept aside for validating the model, and the rest 41122 were used for building the model. The good-bad odds of the validation and build subsets were 2.740 and 2.743 respectively, confirming the absence of bias in the selection of accounts for the validation and build subsets. Each account record in the data set contained 142 independent variables: 125 internal variables (e.g. payment regularity measures, purchases and cash advances, changes in balances and credit-line utilization) and 17 credit bureau variables (e.g. total credit limit of all credit cards, total outstanding debt). The dependent variable was the credit status of the customer-good or bad. In keeping with the practice in the credit card industry, the good / bad definition was arrived at after a rigorous study of the good-bad evolution characteristics of the portfolio of accounts. A 'bad' account was defined as '3+ in 9 months', that is, one with 3 or more missed payments or declared bankrupt during a 9-month 'observation' period, and a 'good' account was defined as one with at most 1 missed payment during the lifetime of the account[1].

Although most of the independent variables were continuous, many were categorical. Some of the independent variables had strong correlations amongst themselves[2].

3.2 Identification of subspaces: The number of subspaces that can be constructed from 142 independent variables is quite large: 2^{142} - 1 (or 5.5×10^{42} approx.), making it almost impossible to exhaustively search for clusters in all of them. Yet, only a few of them are likely to have strong clustering tendency or interestingness. Using the *Measure of Merit* based ranking of subspaces, the Dey and Roberts' algorithm avoids searching through all the possible subspaces but locates those with strong clustering tendency. The algorithm identified five 9-dimensional, three 10-dimensional, three 11-dimensional and one 12-dimensional subspaces for consideration (see Table-1 below). The threshold value parameter (t) and the cut-off value for the *Measure of Merit* (C_M) were intentionally set to a high value to restrict the number of identified subspaces to a handful. The small number of subspaces found by this preprocessing step can be exhaustively examined for population segmentation, as described in the next section.

No. of Dimns	Sub -space No.	Subspace Variables[3]	Scaled Measure of Merit
9	1	DGT0, MSD0, MXDL, MCD0, MCIB, AP6M, L59, B61, B71	3.22
	2	DGT0, MSD0, MXDL, MCD0, MCIB, AP6M, W1, B61, B71	3.03
	3	DGT0, MSD0, MXDL, MCD0, MCDB, AP6M, W1, B61, B71	2.89
	4	DGT0, MSD0, MXDL, MCD0, B1B2, AP6M, W1, B61, B71	2.73
10	5	DGT0, MSD0, MXDL, MCD0, B1B2, AP6M, L59, B61, B71	2.70
	6	DGT0, MSD0, MXDL, MCD0, MCIB, AP6M, W1, L59, B61, B71	2.91
	7	DGT0, MSD0, MXDL, MCD0, MCIB, AP6M, W1, S63, B61, B71	2.86
	8	DGT0, MSD0, MXDL, MCD0, MCDB, AP6M, W1, L59, B61, B71	2.74
	9	DGT0, MSD0, MXDL, MCD0, MCIB, AP6M, W1, L59, B61, B71, PR	2.88
11	10	DGT0, MSD0, MXDL, MCD0, MCIB, AP6M, W1, L59, B61, B71, FC1_3	2.87
	11	DGT0, MSD0, MXDL, MCD0, MCDB, AP6M, W1, L59, B61, B71, FC1_6	2.77
	12	DGT0, MSD0, MXDL, MCD0, MCIB, AP6M, W1, L59, B61, B71, PR, FC1_6	2.63

Table-1: Results of preprocessing: Interesting Subspaces

A number of lower-dimensional subspaces that had lower (dimensionality scaled[4]) values of *Measure of Merit* were also identified by the algorithm but have not been shown in the table above for the sake of brevity.

3.3 Detecting Clusters in the Subspaces: All the 12 subspaces identified above were examined for existence of 'useful' clusters (i,e. clusters that could represent 'natural' population segments) using the two-step clustering method. A summary of the results is displayed in Table-2 below. The two-step clustering method automatically determined the number of clusters in each subspace. It found three distinct clusters in each subspace[5] (as can be observed from the table below).

Subspace Number	Cluster 1		Cluster 2		Cluster 3	
	% of Accounts in Cluster	Good-Bad Odds	% of Accounts in Cluster	Good-Bad Odds	% of Accounts in Cluster	Good-Bad Odds
1	18.47	1.49	34.68	1.67	46.85	6.39
2	18.44	1.49	35.05	1.69	46.51	6.39
3	38.63	1.47	11.47	2.74	49.91	5.24
4	54.40	1.73	6.67	4.19	38.94	6.08
5	37.89	1.49	13.62	2.34	48.48	5.56
6	18.50	1.49	36.16	1.75	45.34	6.41
7	18.45	1.49	39.90	1.93	41.65	6.31
8	44.95	1.69	13.09	2.39	41.96	5.81
9	18.83	1.53	34.56	1.67	46.61	6.38
10	18.47	1.49	35.03	1.69	46.49	6.40
11	18.08	1.53	43.68	2.07	38.24	6.15
12	18.43	1.49	34.70	1.67	46.87	6.39

Table-2: Results of the Clustering Step: Distinct Population Segments

It can also be observed from Table-2 that the good-bad odds of the accounts in each cluster within a given subspace were distinctly different. For example, the three clusters in Subspace-6, had good-bad odds 1.49, 1.75 and 6.41 respectively. This itself is a significant result. It implies that the clusters represent population segments with different credit risk profiles, and that these segments can be identified by the two-step clustering algorithm, using a small subset of the independent variables.

3.4 Building Scorecards for the Clusters: Once the population segmentation has been achieved, either by the usual 'intuitive' means such as the ones described in Section 1, or by more definitive empirical methods such as clustering described in Section 3.3, it is useful to build scorecards for the population segments. Scorecards along with a set of 'cut-off scores' represent credit strategies that are used to manage the risk in credit card portfolios and control losses due to payment default. In theory, each identified population segment should have a dedicated scorecard built. However, building, monitoring and rebuilding scorecards, and devising credit strategies based on them involve practical difficulties and costs. Some segments that are small in size (compared to others) and/or have similar risk profiles or other similar characteristics are often merged together.

Building scorecards involves a number of interrelated steps. To begin with, the number and choice of independent variables included in scorecards, and how continuous variables are classed is determined by a number of factors including the inter-correlation between the variables and the variation of good-bad odds across classes. Details of the 'art' of building good scorecards can be found in Chapters 5 and 7 of Mays (2001). Ideally, scorecards should be built (and validated using the validation subset) for as many of the subspaces as possible, to find the best possible set of scorecards. For illustration, a set of three scorecards (one for each population segment/cluster) was built for the population segments (i.e. clusters)

identified in Subspace-1. Some of the performance measures of the scorecards are summarized in Table-3 behind.

Segment	No. of Accounts	Population G-B Odds	No. of Indep. Variables	Scorecard Odds Range	K-S Statistic	Gini Coeff.
1	7596	1.494	4	0.51 to 30.7	34	41.6%
2	14261	1.674	7	0.03 to 160	42	54.8%
3	19265	6.393	6	0.002 to 130	51	67.4%
Combined Effect of the three Segmented-Population Scorecards						
1,2&3	41122	2.743	N/A	0.002 to 160	49	63.4%
Scorecard for the Whole (Un-segmented) Population						
Single	41122	2.743	6	0.002 to 59.7	44	59.0%

Table-3: Results of the Logistic Regression Step: Credit Risk Scorecards

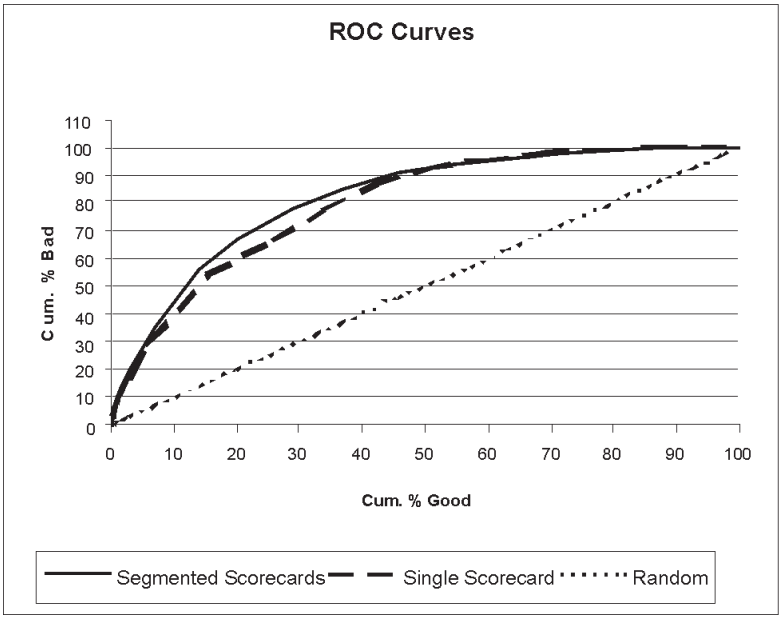
3.5 Results: It can be observed from the entries in Table-3 that the use of three segment-specific scorecards instead of a single scorecard for the whole population led to an increase of 4.4 percentage points in the Gini coefficient and an increase of 5 percentage points in the K-S Statistic. The Receiver Operating Characteristic (ROC) and Kolmogorov-Smirnov (K-S) curves shown in Figure-2 below further illustrate the fact that the segment-specific scorecards (together) better differentiate the bad accounts from the good, compared to a single whole-population scorecard. The curves shown in Figure-2 have been drawn using the validation data set, demonstrating that the increase in K-S Statistic and Gini Coefficient observed in the development sample holds for the validation sample as well.

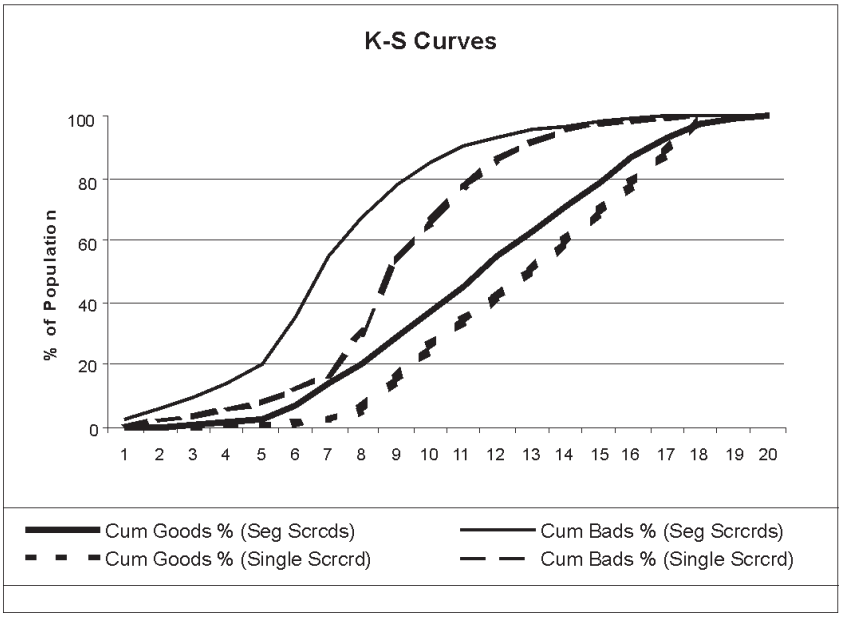
Thus, not only does the set of three segment-specific scorecards gives a better separation of good and bad accounts (evident from the higher K-S Statistic), but also ranks risk more uniformly (evident from the higher Gini Coefficient). Also, qualitatively, the ROC curves and the K-S curves pertaining to the set of segmented scorecards are 'smooth' (i.e. without abrupt changes in slope / inflection) implying that the risk-ranking and the good-bad separation achieved by the new scorecards are uniform throughout the scored population.

3.5 Results: It can be observed from the entries in Table-3 that the use of three segment-specific scorecards instead of a single scorecard for the whole population led to an increase of 4.4 percentage points in the Gini coefficient and an increase of 5 percentage points in the K-S Statistic. The Receiver Operating Characteristic (ROC) and Kolmogorov-Smirnov (K-S) curves shown in Figure-2 below further illustrate the fact that the segment-specific scorecards (together) better differentiate the bad accounts from the good, compared to a single whole-population scorecard. The curves shown in Figure-2 have been drawn using the validation data set, demonstrating that the increase in K-S Statistic and Gini Coefficient observed in the development sample holds for the validation sample as well.

Thus, not only does the set of three segment-specific scorecards gives a better separation of good and bad accounts (evident from the higher K-S Statistic), but also ranks risk more uniformly (evident from the higher Gini Coefficient). Also, qualitatively, the ROC curves and the K-S curves pertaining to the set of segmented scorecards are 'smooth' (i.e. without abrupt changes in slope / inflection) implying that the risk-ranking and the good-bad separation achieved by the new scorecards are uniform throughout the scored population.

Figure-2: Scorecard Performance Characteristics based on the Validation Data Set





4.Implementation Considerations

Since the first scorecard was developed by Fair Isaac and Corp. (FICO) in the late 1950s the raw regression weights, obtained from running logistic regression on the model build data, are scaled to bring them to a range of positive integral values[6]. For example, the traditional FICO scores (still used widely in US and many other countries) range between 300 and 900. A plot of the log-odds against the scaled scores for the combined segmented scorecards is depicted in Figure-3.

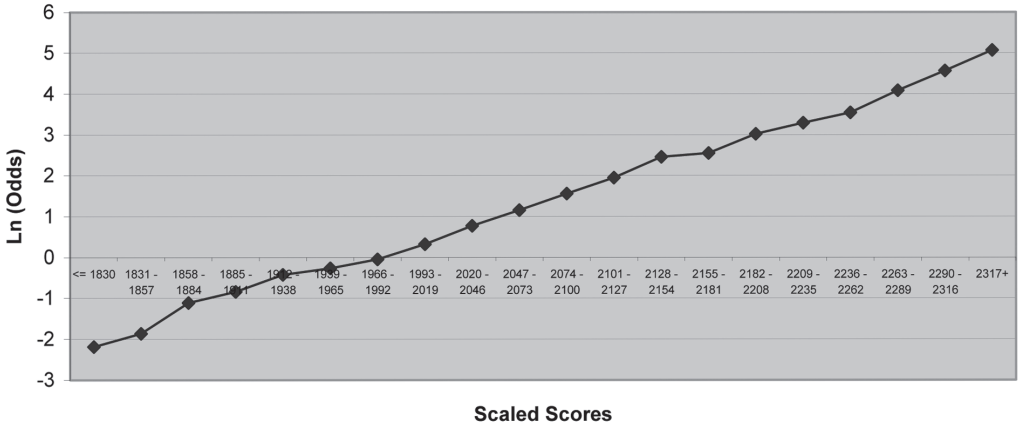


Figure-3: Log-Odds against Scaled Scores

The scaling parameters are not standardized across the industry. Different parameters are used by different lenders and credit bureaus.

Once the empirical model represented by a scorecard (or a set of segmented scorecards) have been built using data from existing accounts with known outcomes (i.e. whether good or bad account), the probability of default of any account (new or existing) with (hitherto unknown outcome) can be estimated by scoring the account using the scorecard and independent variables' data, and translating the score to a log-odds value using the (scaled) log-odds curve. Credit strategies can then be determined through cut-off scores for various credit management actions. As mentioned in Section 3, the interface to this part of the scoring and credit management system will remain unaffected by the implementation of the proposed 2-stage scoring process. The parts of the scoring system that would need to be changed to implement the segmentation scheme of the 2-stage process are discussed in the following part of this section.

In practice, credit card issuers service millions of credit card accounts using automated systems such as TRIAD Adaptive Control System (Fair Isaac Corp., 2008). Credit scoring models formulated as 'scorecards' are widely used due to the ease with which they can be implemented in those automated credit card account management systems. Although implementing a set of scorecards is fairly straight forward, complicated segmentation schemes may not be implementable. For example, if a neural network model is required to determine the population segment to which a particular account belongs, implementation could be very difficult or impossible.

The approach proposed here requires 'non-standard' steps such as Dey and Roberts' algorithm and two-step clustering method for the model building part of the credit scoring process. However, the model building is a one-time activity carried out by analysts outside the automated scoring systems. Once the population segmentation has been done (using the 'non-standard' steps necessary), determining the segment to which an account is to be assigned (which in turn will decide the scorecard that will be used for scoring that account) reduces to the task of selecting one out of a small number of clusters. Since, the centers and boundaries of the clusters used to segment the population during the model building, are fixed, finding the appropriate cluster amounts to calculating the distances to a few (known) clusters. This kind of simple logic can be accommodated in TRIAD-like automated systems.

5.Conclusions and Further work

In this paper, a practical two-stage approach for consumer credit scoring has been proposed. It has been shown that clustering is an effective and practical way of achieving good population segmentation. The problem of high dimensionality has been tackled using an algorithm published by Dey and Roberts (2002). The reduction in the number of subspaces achieved by the algorithm makes it possible to exhaustively examine all of them for clusters. In this work, the BIRCH inspired two-step clustering method has been used to address the problem of finding an unknown number of clusters, but a few other methods are also available. In the second stage of the two-stage approach, standard logistic regression-based risk scorecards have been deliberately proposed for the ease of implementation, and because most practitioners are familiar with scorecards and cut-off score-based credit strategies. Furthermore, scorecards are well understood and accepted as 'sound' credit control instruments by regulators in the credit industry. The efficacy of the approach has been demonstrated using a real data set. The results show that scorecards with better performance characteristics can be built using the approach proposed. Although, the data set pertains to the behavioural scoring area there is no reason why the approach would not work with application scoring data. However, this needs to be investigated further.

Notes

1 All other accounts were classed as 'indeterminate' and excluded from the data set.

2 This is quite usual with most large practical data sets and is routinely dealt with during the scorecard build process.

3 Descriptive long names of variables have been substituted by abbreviated names to prevent identification of variables proprietary to the data provider.

4 For details of the 'Dimensionality Scaling Factor' please refer to Dey and Roberts (2002).

5 However, for the purpose of our approach, the number of clusters in each subspace need not be the same.

6 The scaling parameters are not standardized across the industry. Different parameters are used by different lenders and credit bureaus.

References

Agrawal, R., Imielinski, T., & Swami, A. (1993). Database mining: a performance perspective. I*EEE Transactions on Knowledge and Data Engineering,* 5 (6), 914-925

Angelini, E., Tollo, G., & Roli, A. (2008). A neural network approach for credit risk evaluation. *The Quarterly Review of Economics and Finance,* 48, 733-755.

Banasik, J., Crook, J. N., & Thomas, L. C. (1996). Does scoring a subpopulation make a difference? *International Review of Retail, Distribution and Consumer Research,* 6, 180-195.

Chen, W., Ma, C., & Ma, L. (2009). Mining the customer credit using hybrid support vector machine technique. *Expert Systems with Applications,* 36 (4), 7611-7616.

Chiu, T., Fang, D., Chen, J., Wang, Y., & Jeris, C. (2001). A robust and scalable clustering algorithm for mixed type attributes in large database environment. *Proceedings of the seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.*

Chuang, C., & Lin, R. (2009). Constructing a reassigning credit scoring model. *Expert Systems with Applications*, 36 (2), 1685-1694.

Churchill, G. A., Nevin, J. R., & Watson, R. R. (1977). The role of credit scoring in the loan decision. *Credit World,* 3(March), 6-10.

Coffman, J. Y. (1986). The proper role of tree analysis in forecasting the risk behaviour of borrowers. *Management Decision Systems,* Atlanta, MDS Reports 3,4,7, and 9.

Crook, J. N., Edelman, D. B., & Thomas, L. C. (2007). Recent developments in consumer credit risk assessment. *European Journal of Operational Research*, 183, 1447-1465.

Dey, S. (2002). *Application of knowledge discovery techniques to a combination of georeferenced, Tuple-oriented data and spatial data.* PhD thesis, University of Leeds: UK.

Dey, S., & Roberts, S. A. (2002). On high dimensional data spaces, Data mining III (ISBN: 1-85312-925-9): *Proceedings of the Third International Conference on Data Mining,* Bologna, Italy.

Durand, D. (1941). *Risk elements in consumer installment financing.* National Bureau of Economic Research: New York.

Fair Isaac Corporation, http://www.fico.com/en/Products/DMApps/Pages/FICO-TRIAD-Customer-Manager.aspx, 2008.

Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems. *Annals of Eugenics* 7, 179-188.

Frydman, H., Kallberg, J. G., & Kao, D. L. (1985). Testing the adequacy of Markov chains and Mover-Stayer models as representations of credit behavior. *Operations Research,* 33, 1203-1214.

Hand, D. J., & Henley, W. E. (1997). Statistical classification methods in consumer credit. *Journal of the Royal Statistical Society,* 160 (Series A), 523-541.

Henley, W.E. (1995). *Statistical aspects of credit scoring.* PhD thesis, Open University: UK.

Henley, W. E., & Hand, D. J. (1996). A k-NN classifier for assessing consumer credit risk. *The Statistician,* 65, 77- 95.

Hopkins, B. (1954). A new method for determining the type of distribution of plant individuals. *Annals of Botany,* 18, 213-227.

Hsieh, N. (2004). An integrated data mining and behavioral scoring model for analyzing bank customers. *Expert Systems with Applications,* 27 (4), 623-633.

Hsieh, N. (2005) Hybrid mining approach in the design of credit scoring models. *Expert Systems with Applications,* 28 (4), 655-665.

Huang, C., Chen, M., & Wang, C. (2007). Credit scoring with a data mining approach based on support vector machines. *Expert Systems with Applications,* 33(4), 847-856.

Huang, J., Tzeng, G., & Ong, C. (2006). Two-stage genetic programming (2SGP) for the credit scoring model. *Applied Mathematics and Computation,* 174 (2), 1039-1053.

Huang, S-C. (2009). Integrating nonlinear graph based dimensionality reduction schemes with SVMs for credit rating forecasting. *Expert Systems with Applications,* 36(4), 7515-7518.

Kayri, M. (2007). Two-step clustering analysis in researches: A case study. *Eurasian Journal of Educational Research,* 28, 89-99.

Kothari, R., & Pitts, D. (1999). On finding the number of clusters. *Pattern Recognition Letters,* 20(4), 405-416.

Laha, A. (2007). Building contextual classifiers by integrating fuzzy rule based classification technique and k-nn method for credit scoring. *Advanced Engineering Informatics,* 21, 281-291.

Lee, T., Chiu, C., Lu, C., & Chen, I. (2002). Credit scoring using the hybrid neural discriminant technique, *Expert Systems with Applications,* 23(3), 245-254.

Lee, T., & Chen, I. (2005). A two-stage hybrid credit scoring model using artificial neural networks and multivariate adaptive regression splines. *Expert Systems with Applications,* 28 (4), 743-752.

Lin, S. (2009). A new two-stage hybrid approach of credit risk in banking industry. *Expert Systems with Applications,* 36(4), 8333-8341.

Makowski, P. (1985). Credit scoring branches out. *The Credit World,* 75, 30-37.Mays, E. (1998). Credit risk modeling, Glenlake Publishing: Chicago.

Mays, E. (1998). *Credit risk modeling,* Glenlake Publishing: Chicago.

Mays, E. (2001). *Handbook of credit scoring.* American Management Association: New York.

Michaud, P. (1997). Clustering techniques. *Future Generation Computer Systems,* 13, 135-147.

Myers, J. H., & Forgy, E. W. (1963). The development of numerical credit evaluation systems, Journal of American Statistics Association 58 (September), pages 799-806, 1963.

Nakamura, E., & Kehtarnavaz, N. (1998). Determining number of clusters and prototype locations via multi-scale clustering. *Pattern Recognition Letters,* 19, 1265-1283.

Rosenberg, E., Gleit, A. (1994). Quantitative methods in credit management: A survey. *Operations Research,* 42, 589-613.

Sustersic, M., Mramor, D., & Zupan, J. (2009). Consumer credit scoring models with limited data. *Expert Systems with Applications,* 36 (1), 4736-4744.

Thomas, L. C., Crook, J. N., & Edelman, D. B. (1992). *Credit scoring and credit control,* Oxford University Press: Oxford.

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

Thomas, L. C., Oliver, R. W., & Hand D. J. (2005). A survey of the issues in consumer credit modelling research. *Journal of the Operational Research Society,* 56, 1006-1015.

Xu, X., Zhou, C., & Wang, Z. (2009). Credit scoring algorithm based on link analysis ranking with support vector machine. *Expert Systems with Applications,* 36 (2), 2625-2632.

Zhang, T., Ramakrishnon, R., & Livny, M. (1996). BIRCH: An efficient data clustering method for very large databases. *Proceedings of the ACM SIGMOD Conference on Management of Data,* Montreal, Canada, 103-114.

Author's Profile

Dr. Shubhamoy Dey is a faculty of Information Systems at Indian Institute of Mangement Indore since 2002. He has worked and served as an independent consultant for 15 years in the information technology industry in India, UK and USA. He has obtained his Ph.D. from the School of computing, University of Leeds, and holds B.E. and M.Tech degrees from Jadavpur University, and Indian Institute of Technology Kharagpur respectively. His research interests include data mining, spatial databases, text mining, data warehousing, empirical modelling and computational finance. He has published numerous research articles in national and international journals and conferences.