

Empirical evaluation of sampling and algorithm selection for predictive modeling for default risk

Satchidanand S.S.

HP India

sssatchidananda@gmail.com

and

Jay B.Simha

Abiba Systems, India

Jay.b.simha@abibasystems.com

Abstract— Default of credit in credit card business is a major challenge to be tackled. The objective of minimizing the loss due to default can be reduced by early detection of defaulters. Predictive modeling using data mining has been successful in tackling the defaulter modeling. Since the data set is imbalanced and there are number of algorithms for classification, it becomes difficult to select a suitable method for data preprocessing and selection of the algorithm. In this paper an attempt has been made to evaluate the suitable sampling plan and algorithm for default risk modeling on a real data set. Results indicate that a radial basis function with segmentation and logistic regression has been found to be quite effective in identifying the true positives.

Index Terms— predictive modeling, credit risk, data mining, algorithms

I. INTRODUCTION

Predictive modeling is the most frequently used data mining application for building business intelligence and decision support solutions. Essentially, it provides a robust and automated mechanism for building classification systems in data rich environments. There are three major steps in this data mining process. Historical data is first mined to train patterns/models for predicting future behavior. These behaviors can include typical business goals such as predicting response to direct mail, defects in manufactured parts, declining activity, credit risk, delinquency, likelihood to buy specific products, profitability, etc. These patterns/models are then used to score new transactions to determine their likelihood to exhibit the modeled behavior. These scores are then used to act upon for optimizing a business objective.

Credit cards are one of the popular non-cash instruments around the world. They are convenient tools for consumers to make purchase first and pay back the debt later. However,

credit card lending is risky for the card issuers because the loans are usually not secured by any assets. Also, unlike the traditional loans that are discrete, generally involve an individual analysis of credit risk, and have a specific maturity date, credit cards invite a continuous flow of borrowing with limited subsequent checks of financial status after the initial issuance of the card. This is also a market of information asymmetry in the sense that the borrowers know better of their own ability and willingness to repay the debt than do the card issuers. Given the risk associated with credit lending, it is important for card issuers to identify consumer risk type as early as possible in order to prevent risky consumers from borrowing too much before the default occurs. This paper aims to provide a model for new consumers' repayment behavior with which card issuers can determine the default risk of consumers from their initial card usage data. In order to clarify the goal of the paper, we outline the decision problems facing credit card issuers and customers.

When card holders default, card issuers undertake collection procedures. These procedures are expensive relative to the size of most loans. They are often futile when the card holder has gone bankrupt. Therefore, it is important for the card issuers to identify card holder type at early stage in order to minimize lending to risky customers before the default occurs. It means that it becomes necessary to maximize the 'True positives (TP)'.

Predictive modeling defaulter risk is one of the important problems in credit risk management. There are quite a few aggregate models and data driven models available in literature. The main contribution of this paper is empirical evaluation of the effectiveness of the various components of the predictive modeling process. In particular sampling selection and algorithm used for learning classification are evaluated on a real world dataset.

The rest of the paper is organized as follows. The available literature for credit and default risk is briefly discussed in section two. In section three the different sampling schemes for selecting the examples are discussed. The classification algorithms used in the research are briefly discussed in section

four. The experiments and results are given in section five followed by conclusion.

II. LITERATURE SURVEY

There has been much work in the field of classification and risk management. Most work is based on Bayesian networks, decision trees, neural networks and logistic regression. However, it appears that very little published literature is available on default modeling in finance.

Many real-world data sets exhibit skewed class distributions in which almost all cases are allotted to one or larger classes and far fewer cases allotted for a smaller, usually more interesting class. Sampling the data for balancing the classes is a tested method for mitigating imbalance risks. Off the sampling schemes three of the interesting works are from [5], [8], [12]. Japkowicz [5] discussed the effect of imbalance in a dataset. The author evaluated three strategies: under-sampling, resampling and a recognition-based induction scheme. The author noted that both the sampling approaches were effective, and also observed that using the sophisticated sampling techniques did not give any clear advantage in the domain considered.

Another approach which is interesting is by Ling and Li [8]. They combined over-sampling of the minority class with under-sampling of the majority class. They used lift analysis instead of accuracy to measure a classifier's performance. They proposed that the test examples be ranked by a confidence measure and then lift be used as the evaluation criteria. A lift curve is similar to an ROC curve, but is more tailored for the marketing analysis problem. In one experiment, they under-sampled the majority class and noted that the best lift index is obtained when the classes are equally represented. A similar result has been observed in this research also.

Satchidanand et.al.[12], have used in addition to random sampling, a scheme based on cluster prototype sampling for modeling credit risk. The idea was to minimize the possible information loss due to random sampling. The majority class instances were segmented using a standard clustering algorithm like k-means and the number of clusters will be set to cardinality of the minority classes.

Bharatheesh et.al[1] have used modified naïve Bayesian classifier to model the delinquency in credit card business. They have used under sampling and over-sampling schemes, with chi-squared discretization. They have found that under sampling of majority classes yield better results compared to over sampling of majority classes.

The approach utilized in the research described in this paper was to evaluate standard machine learning algorithms applied to classify default behavior. All previous research cited in this paper use contexts, classes, features, and machine learning methods that are different from the research described herein, and therefore, a direct comparison of the results with the previous research work is beyond the scope of this paper.

III. ALGORITHMS

Eleven classifiers of five different types are used in the experiment in this research. They are:

- Bayesian – simple Bayes, complete Bayes and NBTree
- Neural networks – FBP, RBF
- Statistical – Logistic regression, k nearest neighbors
- Machine learning – decision trees, decision rules and decision tables
- Kernel based - Support Vector Machines, RBF

Simple Bayesian classifier: is based on the Bayesian rule of estimating posterior probability with an independence assumption of attributes [3]. It is given by the equation,

$$\text{classify}(f_1, \dots, f_n) = \text{argmax}_c P(C = c) \prod_{i=1}^n p(F_i = f_i | C = c)$$

Where f_x is the feature set and c is the class.

Naïve Bayesian classifiers are very robust to irrelevant features. In practice naïve Bayesian classifier has been found to perform surprisingly well in spite it's sever limitation of attribute independence.

Complete Bayesian classifier: A Bayesian network (or a belief network) is a probabilistic graphical model that represents a set of variables and their probabilistic dependencies. Formally, Bayesian networks are directed acyclic graphs whose nodes represent variables, and whose arcs encode the conditional dependencies between the variables. Mathematically it is given by,

$$P(X_1, \dots, X_n) = \prod_{i=1}^n P(X_i | \text{parents}(X_i)).$$

Using a space of features defined on X . The inference on fully built network is done using several algorithms available in literature. Bayesian classifier has found to work well on known data sets for credit risk[10].

NBTree: simple Bayesian classifiers have shown to suffer from scalability and complete Bayesian networks suffer from learning complexities. A hybrid between decision trees, which have shown to be scalable and simple Bayes, which are simple to learn and infer has been developed by Kohavi [xx]. NBTree is both scalable and has the classification performance comparable to that of state-of-art classification algorithms[7].

FBP artificial neural networks: An artificial neural network (ANN), often just called a "neural network" (NN), is an interconnected group of artificial neurons that uses a mathematical model or computational model for information processing based on a connectionist approach to computation. ANNs are essentially simple mathematical models defining a function $f : X \rightarrow Y$. A widely used type of composition is the *nonlinear weighted sum*, where

$$f(x) = K \left(\sum_i w_i g_i(x) \right)$$

where K is some predefined function, such as the hyperbolic tangent. In a FBP network, the errors are propagated back as feedback for convergence correction [13].

RBF neural networks: Radial Basis Functions are powerful techniques for interpolation in multidimensional space. A RBF is a function which has built into a distance criterion with respect to a centre. They are similar to three layered feed forward network with a radial basis function in the hidden layer. RBF networks have the advantage of not suffering from local minima in the same way as multi-layer perceptrons. This is because the only parameters that are adjusted in the learning process are the linear mapping from hidden layer to output layer. Linearity ensures that the error surface is quadratic and therefore has a single easily found minimum [9]. However, RBF networks have the disadvantage of requiring good coverage of the input space by radial basis functions. In this research this problem is overcome with kernel with Gaussian distribution, which uses the k-means clustering algorithm to provide the basis functions and learns logistic regression classifier for each cluster[14].

Logistic regression: Logistic regression is a variation of ordinary regression which is used when the dependent variable is a binary variable (i. e., it takes only two values, which usually represent the occurrence or non-occurrence of some outcome event) and the independent (input) variables are continuous, categorical, or both. Unlike ordinary linear regression, logistic regression does not assume that the relationship between the independent variables and the dependent variable is a linear one. Nor does it assume that the dependent variable or the error terms are distributed normally. The form of the model is

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_k X_k$$

where p is the probability that $Y=1$ and X_i are the independent variables, β_i are known as the regression coefficients, which have to be estimated from the data. Logistic regression estimates the probability of a certain event occurring. Logistic regression, thus, forms a predictor variable ($\log(p/(1-p))$) which is a linear combination of the explanatory variables. The values of this predictor variable are then transformed into probabilities by a logistic function. This has been widely used in credit scoring applications due to its simplicity and explainability [4].

KNN classifier: is a method for classifying objects based on closest training examples in the feature space. k -NN is a type of instance-based learning, or lazy learning where the function is only approximated locally and all computation is deferred until classification. In KNN result of new instance query is classified based on majority of K -nearest neighbor category. The purpose of this algorithm is to classify a new object based on attributes and training samples. The classifiers do not use any model to fit and only based on memory. Given a query point, K number of objects (training points) closest to the query point is found. The classification is using majority vote among the classification of the K objects. Any ties can be broken at random. K Nearest neighbor algorithm used neighborhood classification as the prediction value of the new query instance[xx]

Decision trees: A decision tree or a rule based classifier is a predictive model; that is, a mapping of observations about an

item to conclusions about the item's target value. Each interior node corresponds to a variable; an arc to a child represents a possible value of that variable. A leaf represents the predicted value of target variable given the values of the variables represented by the path from the root. In decision tree learning, a decision tree describes a tree structure wherein leaves represent classifications and branches represent conjunctions of features that lead to those classifications [12]. A decision tree can be learned by splitting the source set into subsets based on an attribute value test. This process is repeated on each derived subset in a recursive manner. The recursion is completed when splitting is non-feasible or when a singular classification can be applied to each element of the derived subset. C4.5 is an improved version of an earlier decision tree based learner called ID3, which it self was based on Hunt's [12] concept learning system. The attributes are tested for selection on several criteria like minimum entropy, information gain etc., this has become a popular classifier in data mining due to both flexibility of learning and comprehension of results.

Decision rules: Decision trees use greedy global search methods for tree induction and later flattened to get decision rules. On the other hand, Witten et.al. proposed a simple local search method for induction rules directly without flattening step. Though they use same concept as that of decision trees for induction, the results will be simpler rules than decision trees [14].

Decision tables: were originally conceived as a tool for representing managerial knowledge and needs. It is a detailed, logically organized tabular structure with two major sections; conditions and actions. The condition may be thought of as 'If' statements, while the actions may be thought of as 'Then' statements. For each combination of conditions, we will take the actions specified. Induction of decision tables is done in a similar way as that of decision trees and rules, but the resulting knowledge will be parsimonious compared the other two[6].

Support Vector Machines: Support vector machines map input vectors to a higher dimensional space where a maximal separating hyperplane is constructed. Two parallel hyperplanes are constructed on each side of the hyperplane that separates the data. The separating hyperplane is the hyperplane that maximises the distance between the two parallel hyperplanes. An assumption is made that the larger the margin or distance between these parallel hyperplanes the better the generalisation error of the classifier will be. Formally the representation of SVM in its dual form is,

$$\max \sum_{i=1}^n \alpha_i - \sum_{i,j} \alpha_i \alpha_j c_i c_j \mathbf{x}_i^T \mathbf{x}_j$$

Where X_i is the training set with labeled examples and α terms constitute a dual representation for the weight vector in terms of the training set.

IV. EXPERIMENT AND RESULTS

The data for this research is from a real world for credit card business. The data contained about 20,000 cases with defaulters as the minority class. It has been observed that the

defaulter class is about 10% of the population making it a skewed or imbalanced data set. In this research we have used three different schemes for balancing using samples. They are over sampling of the minority class by resampling, under sampling of the majority class by random sampling the majority class and cluster proto type sampling of the majority class using k-means clustering. The different classifiers discussed earlier have been run on all the samples and a 10 fold cross validation has been used for extracting the performance statistics. In general classification accuracy (the generally used model validation criteria), accuracy of true positive classification (in this case classification accuracy of defaulters) and true negative classification accuracy (in this case the classification accuracy of non-defaulters) have been compared. The main objective of the research was to find the best sampling scheme and best classifier for maximizing the true positives i.e. defaulters. The results are tabulated in table 1, table 2 and table 3.

Table 1. Results for over-sampling the minority class for balancing

Sample plan →	Over sampling		
	CA	TP	TN
Algorithm ↓			
Simple Bayes	57.3	36.2	79.0
Complete Bayes	56.7	49.1	68.7
NBTree	57.0	43.4	70.5
Neural networks (FBP)	57.8	35.3	78.7
Neural networks (RBN)	59.2	51.3	61.2
Logistic regression	59.2	40.2	74.5
Decision trees	58.7	46.3	68.6
Decision rules	58.5	38.5	73.8
Decision tables	57.7	41.2	74.5
Support vector machines	56.9	38.1	68.5
Knn	56.5	39.2	67.0

Table 2. Results for under-sampling the majority class for balancing

Sample plan →	Under sampling		
	CA	TP	TN
Algorithm ↓			
Simple Bayes	60.5	37.5	81.5
Complete Bayes	59.2	50.1	68.2
NBTree	60.0	46.1	73.3
Neural networks (FBP)	59.8	36.2	81.9
Neural networks (RBN)	60.3	55.4	64.9
Logistic regression	61.1	44.2	75.1
Decision trees	60.2	48.7	71.2
Decision rules	59.0	40.5	77.9
Decision tables	60.8	43.4	77.4
Support vector machines	58.2	41.2	70.2
Knn	55.7	42.0	65.0

Table 3. Results for under-sampling the majority class for balancing with cluster prototypes

Sample plan →	Over sampling		
	CA	TP	TN
Algorithm ↓			
Simple Bayes	57.4	37.1	80.1
Complete Bayes	55.9	48.2	70.3
NBTree	57.6	43.5	71.7
Neural networks (FBP)	56.8	36.7	79.5
Neural networks (RBN)	60.3	53.5	60.5
Logistic regression	58.9	42.3	75.2
Decision trees	58.6	47.6	69.7
Decision rules	59.1	40.5	75.9
Decision tables	57.0	41.5	73.3
Support vector machines	57.2	39.7	69.4
Knn	57.1	40.6	62.6

* CA –Classification Accuracy, TP – True Positive accuracy, TN – True negative accuracy

The results indicate that sampling has an effect on classification accuracy of the learned model on the test data. It can be observed that the classification accuracy is quite better on balanced samples with under sampled scheme. However the under sampling based on cluster prototypes is not effective in the domain under examination.

It can be observed that the original objective of maximizing the true positives has been achieved by very few algorithms, even though the overall classification accuracy of most of the algorithms is statistically same.

For the business objective in this research i.e. in maximizing the true positives, the radial basis function (RBF) neural network outperforms all other classifiers, even though it has worst true negative classification performance. This leads to a hypothesis that, when a specific business objective like what is set in this research is the criteria for selection of the classifier, it may be necessary not to select the classifier based on classification accuracy or lift alone.

An interesting point to be observed is that the true negative accuracy of simple Bayesian classifier, which has the second best accuracy. This indicates the features are independent in the true negative space. This characteristic of the results lead us to believe that accuracy of classification accuracy for true positive and true negative for different algorithms do not vary proportionately. It brings us to an important point that algorithms are better at identifying true positives and true negatives as separate spaces. It may be a good idea to combine the classifiers for different spaces. Further research has to be carried out to profile the data characteristics for such a behavior.

V. CONCLUSIONS

Defaulter modeling is an important function in credit risk management. The imbalance in data set and idiosyncrasies of learning methods pose a challenge of ‘no best classifier for all’. In this research different balancing methods are evaluated and it has been observed that under sampling of majority classes works best for the data set on hand. Similarly a kernel

based RBF neural network has performed better than other classifiers, in identifying the true positives, which is the objective of the study. The work is under progress to understand the performance of mixed classifiers for classifying true positives and true negatives separately.

REFERENCES

- [1] Bharatheesh T.L, Iyengar S.S, "Predictive Data Mining for Delinquency Modeling", . ESA/VLSI 2004: 99-105
- [2] Dmitriy Fradkin and Ilya Muchnik "Support Vector Machines for Classification" in J. Abello and G. Carmode (Eds) "Discrete Methods in Epidemiology", DIMACS Series in Discrete Mathematics and Theoretical Computer Science, volume 70, pp. 13-20, 2006.
- [3] Hand, DJ, & Yu, K., "Idiot's Bayes - not so stupid after all?" *International Statistical Review*. Vol 69 part 3, pages 385-399, 2001
- [4] Hosmer, David W.; Stanley Lemeshow, *Applied Logistic Regression, 2nd ed.*. New York; Chichester, Wiley, 2000
- [5] Japkowicz, "The Class Imbalance Problem: Significance and Strategies," in *Proceedings of the 2000 International Conference on Artificial Intelligence (IC-AI'2000): Special Track on Inductive Learning*, (Las Vegas, Nevada), 2000. (18)
- [6] Kohavi R, "The Power of Decision Tables", *Proceedings of European Conference on Machine Learning, 1995*.
- [7] Kohavi R, "Scaling Up the Accuracy of Naïve-Bayes Classifiers: a Decision-Tree Hybrid", *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining, 1996*.
- [8] Ling, C., & Li, C., "Data mining for direct marketing: Problems and solutions", *Proceedings of the Fourth International Conference on Knowledge Discovery and Data Mining (KDD '98)* (pp. 73-79). Menlo Park, CA: AAAI Press, 1998
- [9] Martin D. Buhmann, M. J. Ablowitz, *Radial Basis Functions: Theory and Implementations*, Cambridge University., 2003
- [10] Neil M, Fenton N, Tailor M, "Using Bayesian Networks to model Expected and Unexpected Operational Losses", *Risk Analysis*, Vol 25(4), 963-972, 2005
- [11] *Nearest-Neighbor Methods in Learning and Vision*, edited by Shakhnarovich, Darrell, and Indyk, The MIT Press, 2005
- [12] Satchidananda S.S and Jay B.Simha, "Comparing the efficacy of the decision trees with logistic regression for credit risk analysis ", SUGI Asia conference, Mumbai, India, 2006
- [13] Wasserman, P.D. , *Neural computing theory and practice*. Van Nostrand Reinhold, 1989
- [14] Witten, I.H., Frank, E., Trigg, L., Hall, M., Holmes, G., Cunningham, S.J.: Weka: Practical machine learning tools and techniques with java implementations, *Proceedings of ICONIP/ANZIIS/ANNES'99 Int. Workshop: Emerging Knowledge Engineering and Connectionist-Based Info. Systems*, 1999