A Mathematical Model for Predicting Debt Repayment: A Technical Note

Udani Wijewardhana[1], Chinthaka Bandara[2] & Thesath Nanayakkara[3]

Abstract

Debt collection is a massive industry, within the USA alone more than $50 billion recovered each year. However the information available is often limited and incomplete, and predicting whether a given debtor would repay is inherently a challenging task. This has amplified research on debt recovery classification and prediction models of late. This report considers three main mathematical, data mining and statistical models in debt recovery classification, in logistic regression, artificial neural networks and affinity analysis. It also compares the effectiveness of the above-mentioned tools in evaluating whether a debt is likely to be repaid. The construction and analysis of the models were based on a fairly large unbalanced data sample provided by a debt collection agency. We have shown that all three models could classify the debt repayments with a considerable accuracy, if the assumptions of the models are satisfied.

JEL Classification: C45, C53.

Keywords: Debt collection, Artificial Neural Network, Affinity analysis.

[1] University of Colombo. Sri Lanka.
[2] University of Colombo. Sri Lanka.
[3] University of Colombo. Sri Lanka.

1. Introduction

Many industries around the globe have been plagued with bad debt, the cost of debt collection has been ever increasing which has led many companies to outsource debt collection to a collection agency. The debt collection industry is enormous in countries such as USA. In the U.S, third-party debt collection agencies employ more than 140,000 people and recover more than $50 billion each year, mostly from consumers (Fedaseyeu and Hunt 2015), and nearly 14% of American consumers had an account in a collection agency as at 2011 (2014 Annual Report & Form 10-K).

Creditors possess legal and informational advantages, but the information available is limited. Therefore predicting whether a particular customer is likely to repay a debt is a complicated and inherently tedious exercise. This difficulty is amplified because many accounts are forwarded to a collection agency from the healthcare sector, and due to the nature of the industry the information is incomplete and lacks financial information. Hence if it could be accurately predicted if a debt could be repaid is hugely beneficial to a collection agency.

This research project is on predicting debt repayments using historical data of a US based debt collection agency. We use the data to develop mathematical and data mining models to classify and predict if a debt could be recovered or not. The ultimate objective of the study is to build a model which could accurately classify new data using the training data set.

This report mainly focuses on data mining and knowledge discovery tools in logistic regression artificial neural networks and market basket analysis to classify and predict. Knowledge discovery is defined as the process of identifying valid, novel, and potentially useful patterns, rules, relationships, rare events, correlations, and deviations in data (Fayyad 1996). Mathematical and data mining tools are an integral part of the knowledge discovery process, as they can be used to identify hidden patterns and underlying structures in the otherwise unstructured data.

There has been ample evidence in literature of instances where data mining methods were used in classifying problems such as debt scoring, credit scoring and bankruptcy predictions. Logistic regression and neural networks were used to model credit scoring (Desai, Crook and Overstreet, 1996). Zurada and Lonial compared performances of neural networks, logistic regression memory-based reasoning and a combined model to predict debt recovery in health care industry (Z. L and L. S 2005). Hensher and Jones have examined a range of classification techniques such as logit and probit models and neural networks in Advances in Credit Risk Modeling and Corporate Bankruptcy Prediction (UK: Cambridge Press 2008). Ho Ha and Krishnan have used Cox's hazard model in addition to neural networks in predicting credit card debt recovery (Ho Ha and Krishnan 2012).

2. Methodology

The debt collection data set consist of thirty seven variables which includes six continuous variables, seven binary variables, five date variables, thirteen categorical variables and four identifiers. The data set consisted with over two hundred thousand transactions. The input data set was a worked data set and hence adjustments were done to take the data set back to its original phase. The main adjustment done was to take the current balance back to the original phase (i.e. current balance before the payment). The important change was done to the total net balance. The original variables itself is not sufficient to do the analysis. To overcome this issue, several variables were derived using the original data set.

2.1 Logistic Regression Model

Logistic regression is a regression method used when the response variable is dichotomous. The purpose of a logit model is to derive a mathematical equation that predicts the membership of a given case. There are ample examples in literature in cases where Logistic regression has been used in classifying problems such as debt recovery, credit scoring and bad debt modelling.

Let the logistic function be denoted by f, then it is defined as $f(x) = \frac{1}{1+e^{-x}}$ for each real number x.
Therefore it is immediately clear that$0 < f(\mathrm{x}) < 1$ for each real number x.
Also observe that $\lim_{x\to\infty} f(x) = 1$ and $\lim_{x\to-\infty} f(x) = 0$.

It is these two properties which causes the logistic function so popular in classification problems. In logistic regression the logistic function is used to obtain a probability that the response variable belongs to a particular group.

In order obtain the Logistic regression model from the logistic function the logit $g(x)$ is defined as a linear combination of the independent variables. Then $g(x) = \alpha + \beta_1 X_1 + \beta_2 X_2 + \cdots \beta_n X_n$ where the parameters are to be estimated.

To illustrate a Logistic Regression model suppose we want to model a dichotomous response variable D, with the two groups represented as D=1 and D=0 and X_i (i=1,2…n) are the independent variables. Then the classification probability of a group is defined to be;

$$\mathrm{P(D = 1|X = X_1, X_2 \ldots X_n)} = \frac{1}{1 + \mathrm{e}^{-(\alpha+\sum_{i=1}^{n}\beta_i X_i)}} = \frac{1}{1 - \mathrm{e}^{-g(x)}}$$

α, β_is are parameters to be estimated using the maximum likelihood method. The β_i s can be interpreted as the amount in which the log odds of the response variable belonging to the group classified as 1 will change, when a unit increase is made to the variable $\mathrm{X_i}$ all others held constant.

Modelling debt recovery data using Logistic Regression

The objective of using Logistic regression was to determine if a given debtor will repay even a portion of his due debt or not. Therefore a new the response variable named 'Paid' was created using the payments data. "Paid' was code as 1 if any payments have been made, or else as 0.

The first step towards modeling was to randomly partition the final cleansed data set into 60-40 training and validation data sets. Next phase was to select covariates. The traditional approach in statistical model building was to select the most parsimonious model which accurately describes the data. And also it was important to minimize the number of variables involved in the model not only for simplification but the more variables included in a model, the greater the estimated standard errors become, and the more dependent the model becomes on the observed data.

Results of Logistic Regression Model

The final logistic regression model and its results are shown by the two tables below. The first table shows the variables included in the model. The variables used in this model were, 'nocontacts', the number of contacts the agency has made, account category, opening balance category, existence of previous accounts paid in full (acpif) and insurance. Dummy variables were used for categorical variables. The 2nd column illustrates the estimated parameters for each coefficient. The next columns represent the standard error of each estimate, its Wald statistic, degrees of freedom and the significance compared to a chi squared test respectively.

Table 2.1.1 Variables in equation

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	nocontacts	1.754	.036	2381.256	1	.000	5.779
	accountcat			185.263	2	.000	
	accountcat(1)	-1.127	.734	2.359	1	.125	.324
	accountcat(2)	-.459	.734	.392	1	.532	.632
	insurance(1)	-.211	.050	17.637	1	.000	.810
	SSN(1)	.022	.071	.098	1	.754	1.023
	OB			162.708	3	.000	
	OB(1)	-.940	.080	138.581	1	.000	.391
	OB(2)	-.211	.061	11.817	1	.001	.810
	OB(3)	.018	.055	.107	1	.744	1.018
	acpif(1)	-8.898	.095	8793.218	1	.000	.000
	Constant	5.967	.741	64.855	1	.000	390.148

a. Variable(s) entered on step 1: nocontacts, accountcat, insurance, SSN, OB, acpif.

With this model we could correctly classify around 88% of the debtors, from the debtors who have paid in both the training and validation data sets. As expected due to the unbalanced nature of the data set, predicting that a debtor who hasn't paid correctly is of a very high accuracy. However it must be said that the model depends heavily on the variable 'acpif'.

Table 2.1.2 Results of logistic regression model

			Predicted					
			Selected Cases[b]			Unselected Cases[c,d]		
			Paid		Percentage Correct	Paid		Percentage Correct
	Observed		0	1		0	1	
Step 1	Paid	0	88567	212	99.8	33925	100	99.7
		1	2069	14805	87.7	757	5640	88.2
	Overall Percentage				97.8			97.9

2.2 Artificial Neural Network Model

Artificial Neural Network (ANN) is a standard machine learning procedure which is commonly uses for classification in data science. Artificial neurons are linked together according to specific network architecture and transform the inputs into meaningful outputs.

An ANN consists of interconnected neurons. The neurons are usually assembled in layers. Feed forward network is a biologically inspired classification algorithm. Every unit in a layer is connected with all the units in the previous layer. These connections are not all equal, each connection may have a different strength or weight. Data enters at the inputs and passes through the network, layer by layer, until it arrives at the outputs. During normal operation, that is when it acts as a classifier, there is no feedback between layers (Asadi, Roya, Kareem & Sameem 2014). Therefore here feed forward networks is used, because they are relatively simple.

Feed forward network mainly has an input layer at the start and an output layer at the end and a single or multiple hidden layers in the middle. The hidden layers can capture the nonlinear relationship between variables (Tseng, Yu and Tzeng 2002). Here three hidden layers are used, because too small or too many hidden layers make the system inconsistent. Generally, the number of hidden neurons primarily depends on the number of training samples (Neural Networks 2018).

Here sigmoid function which is also known as tan-sigmoid function is used as the activation function. The reason is that it is continuous and differentiable and its derivative is very fast to compute and has a limited range (from 0 to 1, exclusive).

After transforming and cleansing of data there are 146093 records. Since ANN can be used only for numerical data the seven continuous variables (originalbal, age of debt, age at dateplaced, recency of service date, currentbal at dateplaced, totalnetbal at dateplaced and debt as percentage of total) with four binary variables (homephone, workphone, SSN and insurance) are trained.

To check the most accurate way from with and without binary variables trained an ANN separately. For each ANN 60% is used as the training set and the remaining 40% is used as the predicting set. And also when comparing the results using lift curve check the prediction accuracy of the top 10% of cumulative percentage amount of ANN results and draw conclusions. Because we can assume that if a model satisfied for the top 10% which is the easiest part to collect debt can most probably predict future patterns.

For the top 10% only continuous is predicted 57.22% when both continuous with binary is predicted only 54.64%. This means only continuous make a good sense. But to represent our dataset without any biasness we cannot neglect binary variables, since it has supported to predict more than 50% of the data with continuous variables. Since our dataset has only few continuous variables without derived variables, most prominent way to do further is with binary variables.

Then whether the clustering has any effect or not is found. Therefore for the dataset with both binary and continuous variables trained an ANN and compared the results for the top 10%. When compared non - cluster results (54.64%) of top 10% with the clustered data (64.4%) we found that clustered data covered 9.76% more than non – clustered data. That implies clustering has a significant effect for our data. Therefore clustered data are used for further analysis.

For clustering k-means technique is used. It is a primitive algorithm. However it is used because it is easy to handle big data with low time complexity. The purpose of clustering is to find out the best cluster which is lowest within cluster sums of squares and largest between cluster sums of squares. Therefore Hartigan and Wong is used since it minimizes the sum of squares within-clusters.

Then to find out the best number of clusters, checked cluster within sums of squares and between sums of squares for size 3, 4, 5 and 6 which would be more practical for this dataset. Then identified that the ideal cluster size is 6 which has the lowest cluster within sums of squares and largest cluster between sums of squares. Therefore the dataset is clustered, selected the largest cluster and ANN is formulated on that.

After clustering using k-means with 6 as the no. of clusters the largest cluster size is identified as 57160. Then sub setting by that cluster and ANN is done on that.

Similarly to enhance this method two advanced methods are followed as well. First method is to normalize suitable variables after clustering and the ANN. For normalization, only continuous variables are used and the flow is checked before and after normalization. Normalization has been done to get the variables for the same scale. Then it is noticed that originalbal, recency of service date, currentbal at dateplaced and totalnetbal at dateplaced gave interpretable results. Therefore only normalize those variables while keeping others as the same and do ANN on that.

The second methods is, after getting the largest cluster and do Principal component analysis (PCA) on that and finally do ANN. PCA is a mathematical procedure that transforms a set of correlated response variables into a smaller set of uncorrelated variables called principal components. It check whether the total variability of the data can be explained through a few artificially created linear combinations. After PCA we found that 98.73% is covered by 4 PCs.

For these three steps the lift curves are as below.

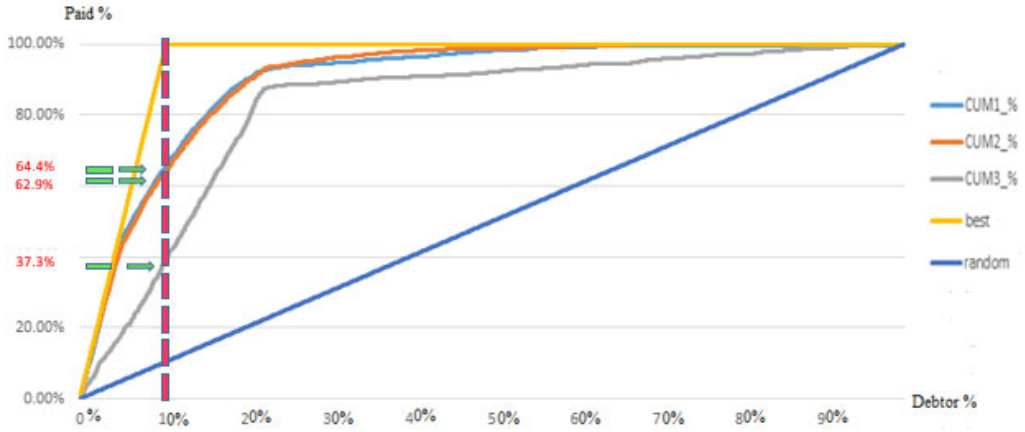

Figure1.2.1 Lift Curve Comparison

When comparing the top 10% captured cumulative percentages clustering and ANN predicted 64.4% while other two steps capturing 62.9% and 37.23% respectively. That means for our data set only do ANN after clustering is much better than standardizing or doing PCA. Therefore for this data set it is highly recommended to keep both continuous and binary variables cluster and to do ANN.

2.3 Affinity Analysis Model

Item based collaborative filtering is an approach used to measure the similarity of a certain item {i} to a set of items {i1, i2, i3… in}. Cosine-based similarity and correlation-based similarity are few examples for algorithms which can use to find the measure of similarity. The general usage of item-based collaborative filtering is for the purpose of recommendation. The measure of similarity is used as an input to recommendation algorithm.

In this project, we use item-based collaborative filtering as a model to predict the ability of a debt repayment for a given debtor. Attributes of the debtors are used as items while the payment status has been used as the base item and the similarities are measured according to this base item. R studio is the software used for collaborative filtering and Apriori algorithm was used to determine the association rules between items in the database.

Only the categorical and the binary data were used as inputs to the algorithm. Binary variables were converted into categorical variables by adding a prefix to each binary variable (i.e. insurance 0 was converted as ins - 0). User id is the primary key in the data set.

Data set was divided into 2 parts as the train data and the test data. Sixty percent of the data were taken to the train set and forty percent were taken for the test set. The alternative way to divide the data set is to take seventy percent for training and thirty percent for testing.

Association rule mining Finds items that tend to co-occur in the data and specifies the rules that govern their co-occurrence. Apriori algorithm was used to generate the rules of interest. Apriori performs market basket analysis by identifying co-occurring items (frequent item sets) within a set. Apriori finds rules with support greater than a specified minimum support and confidence greater than a specified minimum confidence.

- Support - how frequently the items in a rule occur together
- Confidence - conditional probability of the consequent given the antecedent
- Lift -(Rule Support) /(Support(Antecedent) * Support(Consequent))

Support and confidence are inputs to the Apriori algorithm and user can define values for those variables. For this project we persisted with a support of 0.05 and a confidence of a 0.1. One of the reasons to use low values for these variables is the fact that we desired a rule to have a minimum of six values. When the values of those two variables are high, it's hard to generate variables with the length of six or above.

The qualities of the rules are reflected through the value of the output variable "lift". The rules which have a lift value more than one is considered to be a quality rule and only rules which has a lift value more than one is considered for the analysis purposes. In this project we considered only the top twenty five rules to do the analysis.

The training set is used to generate the rules. These generated rules are then matched with the transactions in the test data set. Two approaches are used to match the rules. They are,

1. Exact match - Check for the transaction which matches exactly with the rules.
2. Average match - Check how many instances matches per rule.

Affinity analysis for predicting debt repayment was done using the R studio software and Apriori algorithm was used to generate the rules. Support of 0.05 and confidence of 0.1 was given as input variable while minimum length of a rule was given as six. Figure 3.7 shows the predicting power of this model.

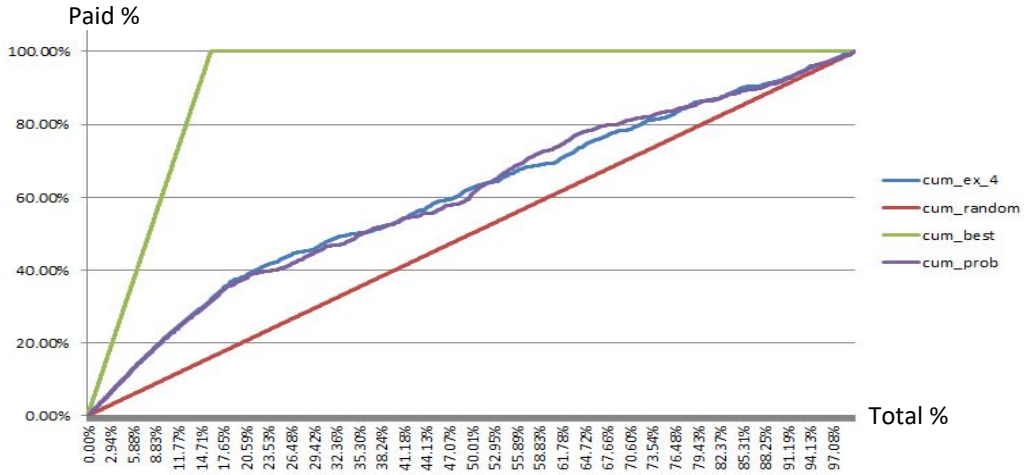

Figure 1.3.1 Predictability of affinity analysis

The analyses of the results are based on the top ten percent of the debtors. Figure 3.7 depicts that in the best case scenario around 70% of the debtors are captured as paying debtors. This is depicted from the green line. The random scenario is the case where when we choose ten percent of the debtors, ten percent of the paying debtors should be included. This is shown by the maroon line.

The prediction results are depicted from the blue and the purple lines. Blue line represents the model with exact matching criteria while the purple shows the average matching criteria. Both the criteria show similar results when considering the top ten percent of the debtors. This method captures around twenty two percent of the paying debtors inside the top ten percent.

3. Conclusion

Artificial neural network is a powerful tool to do predictions but the main limitation of a neural network is that it only works with numerical variables. Since this data set contains many categorical variables, a fully representative data set might not be trained to predict the debt repayment when using only a neural network. The affinity analysis provides a way to use the categorical variables to do the prediction of debt repayment. The results of the models are given in the table 3.1.

Table 3.1 Top 10% results of the models

Model	Paid percentage in the top 10%
ANN only	57.22%
Standardizing and ANN	62.9%
PCA and ANN	37.23%
Clustering and ANN	64.4%
Affinity analysis	22.0%

Table 3.1 shows that artificial neural network alone has a higher predicting power than the affinity analysis model. It also depicts that clustering improves the quality of the artificial neural network. To bring more credibility into the model, a mix model of a neural network and an affinity analysis can be experimented as a future work. The current results in the table 3.1 clearly suggests that categorical variables do have a predicting power and hence using them along with the numerical variables will increase the predicting power of the model.

After comparing all the models with their predicting ability, we find that the model which includes clustering and then running a neural network on the data set provides the model with the best predicting ability. It should be noted that the accuracy of the prediction has a direct influence from the data set which is used for prediction and hence we can only say that this model fits to this type of a data set. The research can be extended further by using various data sets which falls into various categories and then evaluate whether there is a best fit model for majority of the data sets or the models to be used for prediction depends on the data sets feeds into the model.

4. References

Asadi, Roya & Abdul Kareem, Sameem 2014, Review of feed forward neural network classification preprocessing techniques. AIP Conference Proceedings. 56718411146. 10.1063/1.4882541.

Desai, V., Crook, J. and G. Overstreet 1996, 'A comparison of neural networks and linear scoring models in the credit union environment', European Journal of Operational Research, vol. 95, no. 1, pp. 24-37.

Fayyad 1996, 'From Data Mining to Knowledge Discovery in Database', American Association for Artificial Inteligence.

Fedaseyeu, V. and R. Hunt, 'The Economics of Debt Collection: Enforcement of Consumer Credit Contracts', SSRN Journal.

Ho Ha, S. and Krishnan, R. 2012, 'Predicting repayment of the credit card debt', Computers & Operations Research, vol. 39, no. 4, pp. 765-773.

Investors. 2014 Annual Report & Form 10-K. [Online].Available at: https://Investors.nytco.com/Investors/Financials/Annual-Reports/Default.aspx, 2014, s1.q4cdn.com/156149269/files/doc_financials/annual/2014/2014-Annual-Report-(FINAL).pdf.

Neural Networks. 2018. Neural Networks. [ONLINE] Available at: https://www.doc.ic.ac.uk/~nd/surprise_96/journal/vol4/cs11/report.html.

Tseng, F., Yu, H. and Tzeng, G. 2002, 'Combining neural network model with seasonal time series ARIMA model', Technological Forecasting and Social Change, vol. 69, no. 1, pp. 71-87.

Zurada, J., and Lonial, S. 2005, 'Comparison of the performance of several Data Mining Methods for Bad Debt Recovery in the Healthcare Industry', Applied Business Research, Spring. UK:

Cambridge Press 2008, 'Advances in credit risk modeling and corporate bankruptcy prediction',.