Journal of
Applied & Computational Mathematics

Wijewardhana, J Appl Computat Math 2019, 8:1

Research Article Open Access

Mathematical Model for Predicting Debt Repayment

Wijewardhana PDUA*

Department of Mathematics, University of Colombo, Kumaratunga Munidasa Mawatha, Sri Lanka

Abstract

Debt collection is a massive industry, with in 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. It has been shown that all three models could classify the debt repayments with a considerable accuracy, if the assumptions of the models are satisfied.

Keywords: Modeling; Regression; Bankruptcy; Debt

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 [1]. The debt collection industry is enormous in countries such as USA. In the US, third-party debt collection agencies employ more than 140,000 people and recover more than $50 billion each year, mostly from consumers, and nearly 30% of American consumers had an account in a collection agency as at 2011 ((Federal Reserve Bank of New York 2011) [2].

Creditors possess legal and informational advantages, and the information available is limited [3]. Therefore predicting whether a particular customer is likely to repay a debt is a complicated and inherently tedious exercise [4]. 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 [5]. 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 [6]. 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 [7].

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 [8]. Knowledge discovery is defined as the process of identifying valid, novel, and potentially useful patterns, rules, relationships, rare events, correlations, and deviations in data [9]. 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 [10]. Then the main results in each of the approaches are compared and results are analyzed [11].

The format of this report is as follows, in the next section literature regarding classifying problems of debt recovery and other related problems are analyzed [12]. In addition the instances of data mining and statistical methods used in literature are also reviewed. The next section is the body of the report comprising of the methodology and results and conclusions [13].

Literature review

Debt recovery modeling and related classification problems such as credit risk modeling, bankruptcy modeling and bad debt modeling have enjoyed a significant interest in research of late. It is evident from literature that in such classification problems involving complex relationships were modeled better using data mining methods than conventional statistical models such as discriminant analysis.

Logistic regression and neural networks were used to model credit scoring, and also compared with traditional methods such as discriminant analysis. Using data from three credit unions found that when classifying accepted loans into good and bad neural networks correctly classified a greater percentage of both the total samples, and of the goods, than either linear discriminant analysis or logistic regression. It shared such conclusions that discriminant analysis performed with less accuracy than neural networks in a similar problem.

It has been designed several neural network models and also employed genetic algorithms to classify the financial performance of Finnish companies [7].

The use of neural networks and regression analyses in classifying farmers defaulting on farmers Home Administration Loans. Barney concluded that neural networks outperform logistic regression in correctly classifying farmers into those who made timely payments and those who did not [6].

It compared performances of neural networks, logistic regression, memory-based reasoning and a combined model to predict debt recovery in health care industry. They have used a typically unbalanced dataset of a healthcare company and found that logistic regression,

***Corresponding author:** Wijewardhana PDUA, Department of Mathematics, University of Colombo, Kumaratunga Munidasa Mawatha, Sri Lanka, Tel: (9411) 2501731; E-mail: udaniwijewardhana@gmail.com

Received February 01, 2019; **Accepted** February 10, 2019; **Published** February 15, 2019

Citation: Wijewardhana PDUA (2019) Mathematical Model for Predicting Debt Repayment. J Appl Computat Math 8: 429.

neural networks performed better in terms of the overall accuracy but the decision tree performed better when classifying the 'good' cases [4].

It have examined a range of classification techniques such as Logit and Probit models, mixed logit models and neural networks in Advances in Credit Risk Modeling and Corporate Bankruptcy Prediction [3]. It has been used Cox's hazard model in addition to neural networks in predicting credit card debt recovery [1].

Methodology

Variable derivation

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 two hundred thousand odd 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 same procedure was applied to take it back to the original 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. One reason to use these derived quantities is the fact that this data set has only six continuous variables and it's not sufficient to do a good analysis.

Logistic regression model

Logistic regression is a regression method used when the response variable is dichotomous. Regression methods are an integral component of any data analysis concerned with describing the relationship between a response variable and one or more explanatory variables. The purpose of a logit model is to derive a mathematical equation that predicts the membership of a given instance and also attempts to predict the probability that a given instance belongs to a particular group using the logistic function.

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.

The logistic function: The probabilities for a particular instance are modeled, as a function of the explanatory (predictor) variables, using a logistic function.

Let the logistic function be denoted by f, then it is defined as Figure 1: $(x)=\frac{1}{1+e^{-x}}$ For each real number *x*.

Therefore it is immediately clear that <(x)<1 for each real number *x*. Also observe that $\underset{x\to\infty}{i}(x)=1$ *and* $_{x\to-\infty}f(x)=\cdot$

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 (x) is defined as a linear combination of the independent variables.

$(x)=+_1X_1+_2X_2+\cdots{}_nX_n$

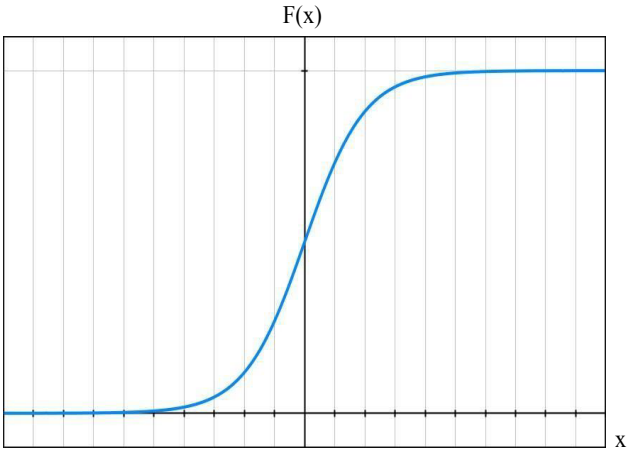

Figure 1: Logistic function.

where the parameters are to be estimated.

The independent variables can be predictor variables, or interaction terms among them whenever it is appropriate. When the predictor variables are categorical dummy variables are used accordingly.

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 ($i=1,2\ldots n$) are the independent variables. Then the probability that a particular instance with X=($X_1, X_2 \ldots X_n$) belonging to the group labeled as 1 is defined to be

$$(D=1\,|\,X=X_1,X_2\ldots X_n)=\frac{1}{1+^{-\left(+\sum^{n}{}_{i=1}{}_iX_i\right)}}=\frac{1}{1-^{-(x)}}$$

Observe that due to the properties of the above mentioned Logistic function this probability is always between 0 and 1, and , s are parameters to be estimated using the maximum likelihood method. Therefore using a predetermined cutoff probability (usually 0.5) one could classify the group a particular instance belongs to.

Logistic regression also avoids many of the unrealistic assumptions which govern other statistical methods assumptions for instance of normality, linearity, and homogeneity of variance for the independent variables. Because it does not impose these requirements, it is preferred to discriminant analysis when the data does not satisfy these assumptions.

The main assumptions of logistic regression are the following which pose no problem in using this model for modeling debt recovery in this setting.

- Response variable is a dichotomous categorical variable.
- Logistic regression assumes independent variables are linearly related to log odds
- The error terms are independent.

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. The training data set was then used to build the model and once several models are built each was tested on the validation data set.

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.

Due to missing values several variables such as the age of the debtor, prior insurance payments were not considered. Then categorical variables with more than 10 categories such as client class and the state were not involved in order to simplify the model and also to increase accuracy as estimating a high number of coefficients for dummy variables can adversely affect the model accuracy. In order to minimize the effect of outliers the continuous variable 'original balance' was categorized into 4 categorizes using its quartiles as shown below.

Opening Balance<1st Quartile-1

1st Quartile <=Opening balance<Median - 2

Median <=Opening Balance<3rd Quartile - 3

Opening Balance => 3rd Quartile 4.

Another rational for selecting variables was statistical significance of the variable in the estimated model. However variables were not omitted purely on the basis of statistical significance. Some variables have been included in the final model despite being deemed insignificant if they improved the results considerably. It is possible since although a variable may have low predictive power individually, but when taken collectively, considerable confounding can be present in the data.

Other accepted variable selection strategies such as purposeful selection were also used to further improve variable selection.

Finally an initial model was built using statistical software, and further improvements were made to conclude at the final model which will be illustrated in the next section.

Artificial neural network model

Theory behind ANN: Artificial Neural Network is a standard machine learning procedure which is commonly uses for classification in data science. ANN can learn patterns in data by mimicking the structure and learning process of neurons in the brain. Artificial neurons are linked together according to specific network architecture and transform the inputs into meaningful outputs [8]. It is typically composed of layers of neurons or processing elements, which are interconnected by set of correlation weights. It is recognized as a complex nonlinear mathematical function that converts input data to a desired output.

An artificial neural network consists of interconnected neurons. The neurons are usually assembled in layers. Feed forward network is a biologically inspired classification algorithm. It consist of large number of simple neuron-like processing units, organized in layers. 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. The weights on these connections encode the knowledge of a network. Often the units in a neural network are also called nodes. 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 [10]. Therefore here we used feed forward networks, 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. Each layer consists of multiple neurons that are connected to neurons in adjacent layers [8]. Here we used three hidden layers. Since multiple hidden layers make the neural network computationally quite complex. But the reason for our use is 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 [9]. Therefore while considering our training sample size we used three hidden layers.

Each neuron receives a signal from the neurons in the previous layer. Then each of those signals are multiplied by a separate weighted value. The weighted inputs are summed. The result is non-linearly scaled between 0 and +1. And then passed through a limiting function called an activation function. It scales the output to a fixed range of values. The output of the limiter is then broadcast to all of the neurons in the next layer [11].

Here we used sigmoid function which is also known as tan-sigmoid function. The reason why you would use a sigmoid as opposed to something else 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) (Figure 2).

Back propagation is a method which adjusts weights to build output with minimum error. BP algorithm could be broken down to four main steps. After choosing the weights of the network randomly, the back propagation algorithm is used to compute the necessary corrections. The algorithm can be decomposed in the following four steps which are Feed-forward computation, Back propagation to the output layer, Back propagation to the hidden layer and Weight updates. The algorithm is stopped when the value of the error function has become sufficiently small [12].

Clustering: Clustering is a task of assigning a set of objects into groups (clusters) that the objects in the same cluster are more similar to each other than to those in other clusters.

K-means clustering:K-means is one of the simplest unsupervised learning algorithms that solve the well-known clustering problem. The procedure follows a simple and easy way to classify a given data set through a certain number of clusters (assume k clusters). The main idea is to define k centers, one for each cluster. These centers should be placed in a cunning way because of different location causes different result. So, the better choice is to place them as much as possible far away from each other. The next step is to take each point belonging to a given data set and associate it to the nearest center. When no point

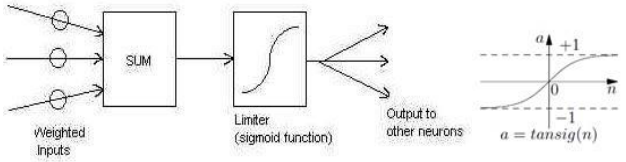

Figure 2: Inside the neural network and sigmoid function.

is pending, the first step is completed and an early group age is done. At this point we need to re-calculate k new centroids as barycenter of the clusters resulting from the previous step. After we have these k new centroids, a new binding has to be done between the same data set points and the nearest new center. A loop has been generated. As a result of this loop we may notice that the k centers change their location step by step until no more changes are done or in other words centers do not move any more [13].

Finally, this algorithm aims at minimizing an objective function known as squared error function given by:

$$J(v) = \sum_{=1}^{c} \sum_{=1}^{c_i} \left(\|x - v\| \right)^2$$

Where '$||x - v||$' is the Euclidean distance between x and , 'c' is the number of data points in ith cluster and 'c' is the number of cluster centers [13].

Principle component analysis: Principal component analysis (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.

The first axis (Z1) is chosen to be lying in the direction of maximum variability. Z1 can be written as, $Z1 = {}_1 x_1 + {}_2 x_2$ where $_1$ and $_2$ are coefficients for the principal components and are constants. And the second axis (Z2) to be perpendicular to first one.

Total variability on the direction of first axis alone is almost same as the total variability of the data. This has one dimension which is Z1 instead of two dimensions X1 and X2.

Coefficients of the principal components are obtained by calculating the Eigen vectors of the covariance matrix corresponding to each Eigen value of covariance matrix. And variance explained by each principal component is the Eigen values of the covariance matrix. The largest Eigen value explains a larger amount of variability.

It was asserted in that the relaxed solution of k-means clustering, specified by the cluster indicators, is given by the PCA principal components, and the PCA subspace spanned by the principal directions is identical to the cluster centroid subspace. However, that PCA is a useful relaxation of k-means clustering and it is straightforward to uncover counterexamples to the statement that the cluster centroid subspace is spanned by the principal directions.

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.

Input data and data separation: 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:(RuleSupport)/(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.

Results and Discussion

Lift curve

A lift curve will be used to evaluate the performance of the model. Lift curves basically have 3 curves. They are random curve in maroon, best curve in blue and actual curve in green in the below graph.

It can be explained from an example, assume there are 58438 debtors of whom 9245 have been paid. The intersection points in the below graph to explain each of the curves. Y-Axis and x-axis represents paid % and total % respectively. Intersection point in the x-axis suggests that 20% of debtors have been captured. That is 58438*20%=11866 debtors have been captured. That is showed by the random curve. If we take 20% of total debtors, it should include 20% of the paying debtors. That is 9245*20%=debtors should be inside the 11866. What the best curve says is, if we take 100 debtors, all those 100 should pay. In this example, green curve or the actual curve suggests that it has capture

75% of the paying debtors. The more the green curve moves towards the blue curve, better the results would be in Figure 3.

Results of logistic regression model

Two logit models were built. The first model consisted of the variables,' *noofcontacts*' which is the number of times a debtor was contact by the agency, the account category, existence of SSN number and insurance claims the opening balance category and '*acpif*' whether or not the debtor had any previous accounts which were paid in full. The model had high predictive power although it depended highly on the '*acpif*' variable due to this reason a second model was derived excluding that variable. The table below shows the parameters which were estimated by the maximum likelihood method. Observe that in case of categorical variables it is the relevant number of dummy variables that are presented (Figure 4).

The second table illustrates the results of the model, where selected and unselected cases indicate the training and validation datasets respectively (Figure 5).

The second model was built deliberately excluding the '*acpif*' variable for the reasons stated earlier (Figure 6). In addition to some of the variables used in the previous model, this model uses the binary variables indicating existence of work and home phone numbers, recency of the service date in terms as described in the data section and the debt as a percentage of total. The cutoff probability was lowered to .4 to adjust for the unbalanced nature of the dataset (Figure 7).

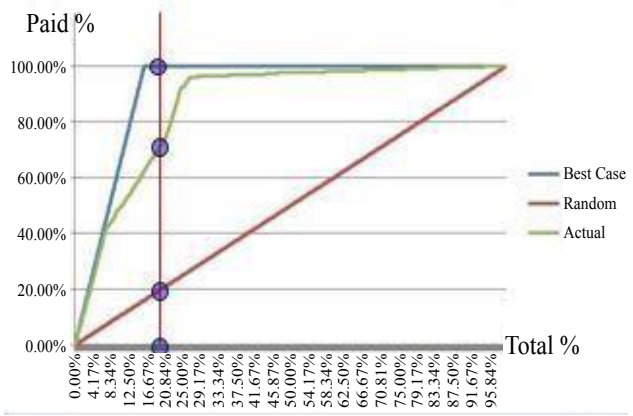

Figure 3: Example of a lift curve.

Results of ANN model

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 (Figure 8).

From ANN results for continuous with and without binary gave a lift curve as below.

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%.

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 in Table 1.

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.

Figure 4: Logistic Regression Model 1.

			Predicted					
			Selected Cases			Unselected Cases		
			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

Figure 5: Logistic Regression results of model 1.

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	nocontacts	.200	.029	47.937	1	.000	1.221
	accountcat			3246.372	2	.000	
	accountcat(1)	-1.235	.505	5.986	1	.014	.291
	accountcat(2)	.615	.505	1.486	1	.223	1.850
	insurance(1)	-.257	.027	93.197	1	.000	.774
	SSN(1)	.593	.036	267.428	1	.000	1.810
	OB			2372.552	3	.000	
	OB(1)	-1.695	.043	1555.844	1	.000	.184
	OB(2)	-.337	.036	87.326	1	.000	.714
	OB(3)	-.035	.035	1.002	1	.317	.966
	recency_sd	-.001	.000	32.997	1	.000	.999
	workphone_D(1)	-.276	.035	61.065	1	.000	.759
	homephone_D(1)	-.372	.026	212.413	1	.000	.689
	debt_as_per_total	-6.181	.046	17930.103	1	.000	.002
	Constant	4.272	.513	69.468	1	.000	71.696

a. Variable(s) entered on step 1: nocontacts, accountcat, insurance, SSN, OB, recency_sd, workphone_D, homephone_D, debt_as_per_total.

Figure 6: Logistic Regression Model 2.

Classification Table

			Predicted					
			Selected Cases			Unselected Cases		
			Paid		Percentage Correct	Paid		Percentage Correct
	Observed		0	1		0	1	
Step 1	Paid	0	81581	7198	91.9	31212	2813	91.7
		1	6056	10818	64.1	2346	4051	63.3
	Overall Percentage				87.5			87.2

Figure 7: Logistic Regression results of model 2.

No. of clusters	Largest cluster size	Sum of squares within the cluster	Order	Sum of squares between clusters	Order
3	72723	5854801665	1	41477936087	4
4	66664	4189731884	2	45171913885	3
5	55779	1892361200	3	47456316724	2
6	57160	1107091110	4	48491270565	1

Table 1: Best cluster size.

Then to find out the best number of clusters, check within sums of squares and between sums of squares for size 3, 4, 5 and 6 would be more practical for this dataset.

Therefore it is identified that the most suitable cluster size is 6. Then the dataset is clustered, selected the largest cluster and ANN is formulated on that.

The ANN results for non-clustered and clustered data provides lift curves as below respectively (Figure 9).

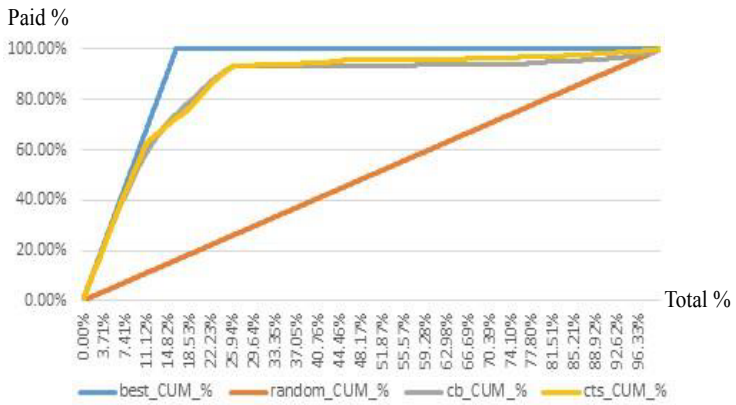

Figure 8: Continuous with and without binary variables.

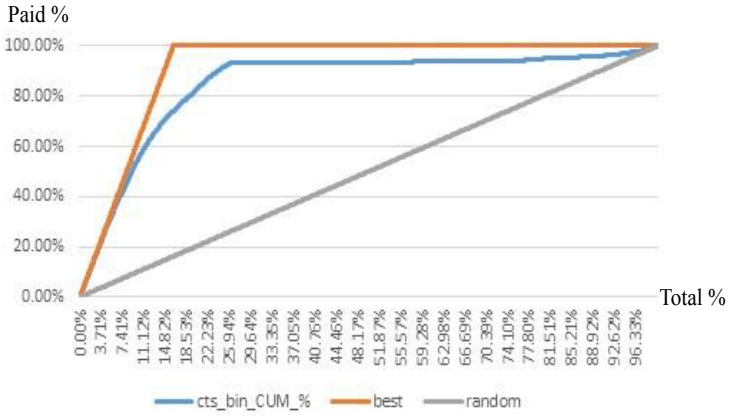

Figure 9: Non-clustered continuous with binary variables.

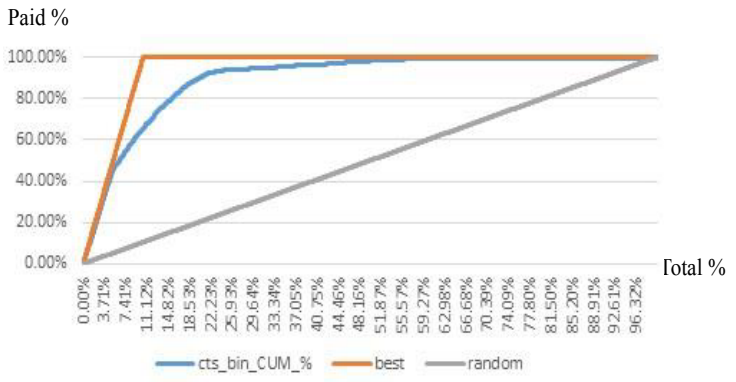

Figure 10: Clustered continuous with binary variables.

When compared to non- cluster results (54.64%) of top 10% the clustered data (64.4%) covered 9.76% more than that. That implies clustering has a significant effect for our data. Therefore clustered data are used for further analysis (Figure 10).

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 PCA on that and finally do ANN. After PCA we found that 98.73% is covered by 4 PCs. Therefore using those 4 PCs we trained and tested the ANN. Those PCs are;

Y=−0.04204 original bal −0.05999 age_of_debt−0.00822age_at_dp-0.9934recency_sd

−0.04484currentbal_at_dp−0.07562TotNetBal_dp+0.00011debt_as_per_total

−0.0001homephone_D+0.00006workphone_D−0.00009SSN +0.00008insurance

Y=0.49759originalbal−0.00682age_of_debt+0.01212age_at_dp+0.09747recency_sd

−0.48786currentbal_at_dp−0.71042TotNetBal_dp+0.00049debt_as_per_total

+0.00015homephone_D+0.00015workphone_D+0.00003SSN +0.00111insurance

Y=0.51689originalbal−0.03082age_of_debt−0.05146age_at_dp+0.0114recency_sd

+0.49124currentbal_at_dp−0.69839TotNetBal_dp+0.00459debt_as_per_total

−0.00111homephone_D−0.00059workphone_D−0.00038SSN−0.00087insurance

Y=0.01236originalbal+0.99759age_of_debt+0.01236age_at_dp−0.05883recency_sd

+0.00766currentbal_at_dp−0.03135TotNetBal_dp+0.00002debt_as_per_total

−0.00007homephone_D+0.00016workphone_D−0.0002SSN −0.00104insurance.

For these three steps the lift curves are as below (Figure 11).

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.

Then it is tested whether PCA has any effect before clustering. First PCA is done. Then it is found that 99.24% covered by 4PCs. Then those 4PCs are clustered and after taking the largest cluster and ANN is done on that (Figure 12).

However it is top 10% also captured only 56% which is less than previous best.

Therefore for this data set it is highly recommended to keep both continuous and binary variables cluster and to do ANN.

Results of affinity analysis model

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 13 shows the predicting power of this model.

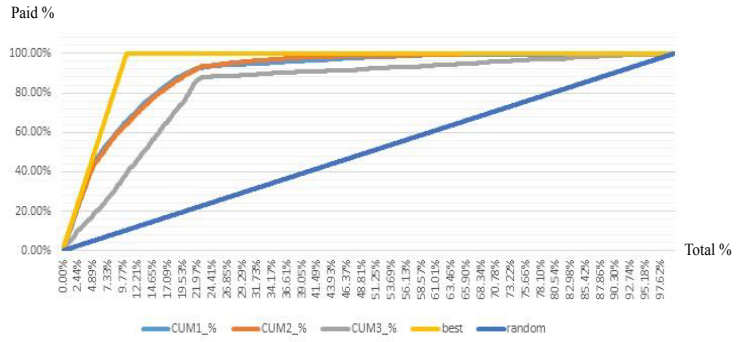

Figure 11: Lift Curve Comparison.

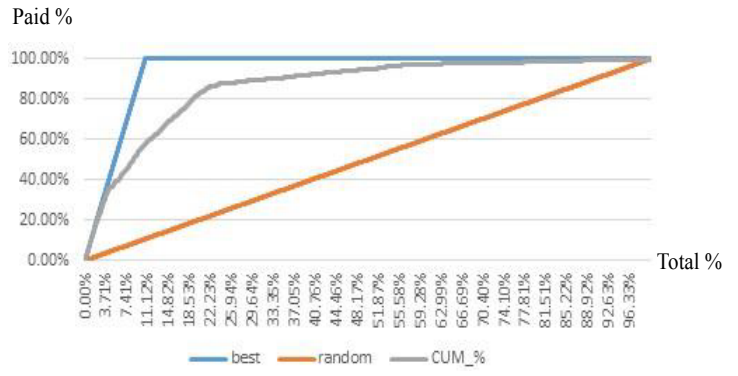

Figure 12: Lift Curve PCA before clustering

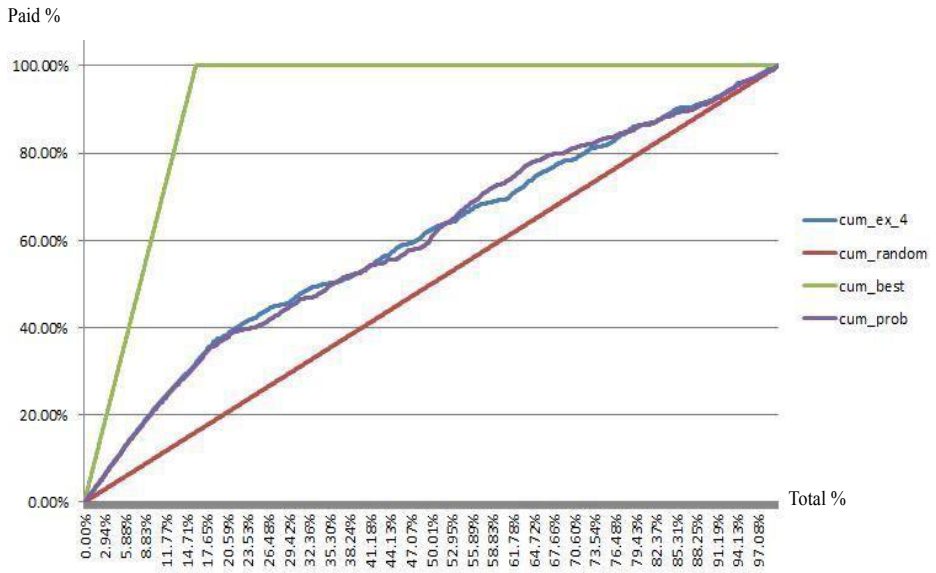

Figure 13: Predictability of affinity analysis.

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 2: Top 10% results of the models.

The analyses of the results are based on the top ten percent of the debtors. Figure 13 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. The results of the models are given in the Table 2.

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.

It 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 2 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.

References

1. Ho Ha S, Krishnan R (2012) Predicting repayment of the credit card debt. Comput Oper Res 39: 765-773.
2. Hosmer D, Lemeshow S (1989) Applied logistic regression. New York: Wiley.
3. Jones S, Hensher D (2008) Advances in credit risk modelling and corporate bankruptcy prediction. Cambridge, UK: Cambridge University Press.
4. Zurada J, Lonial S (2005) Comparison Of The Performance of Several Data Mining Methods For Bad Debt Recovery In The Healthcare Industry. Applied Business Research.
5. Desai V, Crook J, Overstreet G (1996) A comparison of neural networks and linear scoring models in the credit union environment. Eur J Oper Res 95: 24-37.
6. Barney D, Finley Graves O, Johnson J (1999) The farmers home administration and farm debt failure prediction. Journal of Accounting and Public Policy 18: 99-139.
7. Back B, Laitinen T, Sere K (1996) Neural networks and genetic algorithms for bankruptcy predictions. Expert Systems with Applications 11: 407-413.
8. Tseng FM, Hsiao-Cheng Y, Gwo-Hsiung T (2002) Combining neural network model with seasonal time series ARIMA model. Technol Forecast Soc Change 69: 71-87.
9. Patterson DW (1998) Artificial Neural Networks: Theory and Applications
10. http://www.fon.hum.uva.nl/praat/manual/Feedforward_neural_networks_1What_is_a_feedforward_ne.html
11. McCollum P (2018) An Introduction to Back-Propagation Neural Networks.
12. Cilimkovic M (2008) Neural Networks and Back Propagation Algorithm.
13. Data Clustering Algorthims.