Prediction Models of Default Payments of Credit Card Clients in Taiwan

Amir Omicevic

School of Computing Science and Digital Media

Robert Gordon University

Aberdeen, Scotland

a.omicevic@rgu.ac.uk

*Abstract*—This paper explores and evaluates binomial classification prediction models to predict the probability of the credit card customers defaulting on their payments. It is based on the dataset that contains 30000 credit card clients in Taiwan from April 2005 to September 2005. The dataset is characterised by 25 variables covering demographic, credit, history of payment and bill statements data. After an extensive process of data exploration and experiments including the addition of several new observations to mimic the real-world consumer behaviours, the final model presented an improved predictive power, while also indicating the potential for the future work.

Keywords—prediction; default payments; credit risk; credit card; consumer credit; machine learning; classification

# INTRODUCTION

Credit risk evaluation has been one of the core problems for the banking domains ever since they started providing loans to customers. The toughest challenge faced by the industry and researchers in the field of finance is uncertainty and risk. This is not helped by the ever-growing number of different and new types of loans that frequently emerge and target certain consumer types. They greatly increase risk and the management of it, and for any bank providing a large portfolio of varied loans, the risk may not be readily visible and can in time accumulate major losses [2]. The mere existence of risk makes financial-decision making very complicated and subject to constant change. But it also creates profitable opportunities for industry that can efficiently deal with the credit risk management. That is where computational finance as a division of applied computer science, or alternatively defined as the study of data and algorithms used in finance is applied to aid and automate the risk management. [1]

Credit risk problems are often formulated as classification techniques, where the predicted value is of a binary nature, i.e. it can be 0 or 1. It is dependent to multiple attributes/predictors being used by the function/machine learning algorithm that delivers the output (predicted value) based on these attributes. This document will do exactly that. It will explore the data in great depth, process it to make it suitable for training, and then evaluate a range of different machine learning algorithms. The results and further work will be critically evaluated.

# RELATED WORK

S. Chang, S. Kim and G. Kondo [4] worked on models to predict risk of default of Lending Club loans using feature expansion on demographic data and sentiment modelling on loan description. They evaluated Logistic regression, Naive Bayes and SVM and the results showed Naive Bayes with Gaussian distribution to perform the best, predicting default and achieving sensitivity of 80.1%, which in real terms showed progress over just simply predicting the majority class.

M. Charpignon, E. Horel and F. Tixier [5] investigated a large data set of 120000 records with 10 attributes. They evaluated Logistic regression, CART, Random Forests and Gradient Boosting Trees (GBT) with the best result of AUC of 0.86 (area under ROC curve) achieved by GBT. Their work comprised of using different classifiers, but without any meaningful data exploration beyond the basics.

A. Khandani, A. Kim and A. Law [6] worked on sample of customers from a major bank combining customer transactions and credit bureau data from January 2005 to April 2009. They achieve out-of-sample forecasts that significantly improve the classification rates of credit-card-holder defaults, with linear regression R2’s of forecasted/realized defaults of 85%.

A. Hamid and T. Ahmed [7] worked on a 1000 rows dataset with only 7 attributes, comprised of basic demographic data, credit history and job quality as nominal values. They evaluated J48 decision tree, BayesNet and NaiveBayes. Given the relatively low number of predictors, it was nonetheless a good start showing further potential, with J48 achieving a true prediction rate of 78%, but with a very low Kappa value. The issue here is they never focused on significant data exploration to improve the low predicting capability of the data, as exposed by the low Kappa value.

This work is inherently different since it focuses on two major areas. Evaluation of the performance of a multiple set of binary classification prediction and ensemble methods against the original dataset. And secondly, the introduction of a novel approach to data exploration techniques and addition of the new attributes to improve the quality of the prediction models. It then evaluates artificial neural network on the new transformed dataset to indicate improvements and the potential solution that may be applied to the real-life issues in the finance industry.

# DATA exploration AND VISUALISATION

The dataset is comprised of the following 25 attributes:

* ID: ID of each client
* LIMIT\_BAL: Amount of given credit in NT dollars
* SEX: Gender (1=male, 2=female)
* EDUCATION: (1=graduate school, 2=university, 3=high school, 4=others, 5=unknown, 6=unknown)
* MARRIAGE: Marital status (1=married, 2=single, 3=others)
* AGE: Age in years
* PAY\_0: Repayment status in September (-1=pay duly, 1=payment delay for one month, 2=payment delay for two months, and so on…)
* PAY\_2: Repayment status in August
* PAY\_3: Repayment status in July
* PAY\_4: Repayment status in June
* PAY\_5: Repayment status in May
* PAY\_6: Repayment status in April
* BILL\_AMT1: Bill statement in September (NT dollar)
* BILL\_AMT2: Bill statement in August (NT dollar)
* BILL\_AMT3: Bill statement in July (NT dollar)
* BILL\_AMT4: Bill statement in June (NT dollar)
* BILL\_AMT5: Bill statement in May (NT dollar)
* BILL\_AMT6: Bill statement in April (NT dollar)
* PAY\_AMT1: September payment amount (NT dollar)
* PAY\_AMT2: August payment amount (NT dollar)
* PAY\_AMT3: July payment amount (NT dollar)
* PAY\_AMT4: June payment amount (NT dollar)
* PAY\_AMT5: May payment amount (NT dollar)
* PAY\_AMT6: April payment amount (NT dollar)
* DEFAULT\_PAYMENT\_NEXT\_MONTH: (1=yes, 0=no)

Prior to using any tools, I carried out a **manual data quality analysis** which has revealed the following issues:

* no missing values, but a significant amount of inaccurate data with many rows containing nominal values outside their distinct categories
* class imbalance of 77% non-defaults and 23% defaults which is not unusual for datasets from this industry
* significant number of rows with patterns not reflecting real-life behaviour whatsoever, questioning the quality of the data sourcing.

The data quality at this point seems low and does not suggest the development of a high-quality classifier. We shall now **check data distribution, attributes correlation and strong predictors**. In general, data distribution (plots included in appendix) does not relate to Gaussian distribution and exposes 2 trends that continually repeat themselves, which are:

* exponential and bimodal distribution
* positively skewed distribution

**Attributes correlation** reveals a strong relationship between the monthly payments consumers paid over the last 6-month period (BILLAMT1-6). These inherently describe real-life behaviors, but given they are numerical values, this was expected, but irrelevant in terms of using attribute removal processes, such as principal component analysis. It also uncovers a strong relationship between the payment nominal categories showing whether consumers paid their bills on time, or in case they were late, by how many months (PAY0-6). Given these are payment history nominal values describing the same behaviour, but over time which is represented by 6 different attributes/columns in dataset, we can safely ignore the correlation.

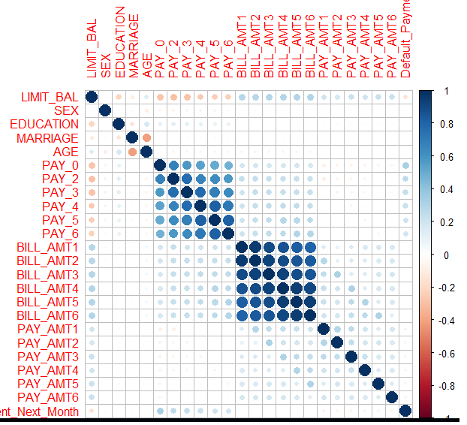
|  |  |  |
| --- | --- | --- |
| *Correlation* | *Attribute* | *Attribute* |
| 0.76 | PAY\_2 | PAY\_3 |
| 0.77 | PAY\_3 | PAY\_4 |
| 0.81 | PAY\_4 | PAY\_5 |
| 0.71 | PAY\_4 | PAY\_6 |
| 0.81 | PAY\_5 | PAY\_6 |
| 0.95 | BILL\_AMT1 | BILL\_AMT2 |
| 0.89 | BILL\_AMT1 | BILL\_AMT3 |
| 0.86 | BILL\_AMT1 | BILL\_AMT4 |
| 0.83 | BILL\_AMT1 | BILL\_AMT5 |
| 0.80 | BILL\_AMT1 | BILL\_AMT6 |
| 0.93 | BILL\_AMT2 | BILL\_AMT3 |
| 0.89 | BILL\_AMT2 | BILL\_AMT4 |
| 0.85 | BILL\_AMT2 | BILL\_AMT5 |
| 0.83 | BILL\_AMT2 | BILL\_AMT6 |
| 0.92 | BILL\_AMT3 | BILL\_AMT3 |
| 0.88 | BILL\_AMT3 | BILL\_AMT3 |
| 0.85 | BILL\_AMT3 | BILL\_AMT3 |
| 0.94 | BILL\_AMT4 | BILL\_AMT4 |
| 0.90 | BILL\_AMT4 | BILL\_AMT4 |
| 0.94 | BILL\_AMT5 | BILL\_AMT5 |

1. Attributes correlation (*significant correlation*)

The process also reveals **strong predictors**, which are the payment nominal categories (PAY0-6). We should not be confused that they are also correlated amongst themselves, because they inherently mimic the real-life events being the strongest indicators of the customer’s credit worthiness. For example, a consumer with a perfect 6-month history will have a nominal value of -1 set for each month, giving a false correlation of 100%. **It is very important to note that these predictors will play a major role when novel approach is required to deliver a much-improved classifier later.**

|  |  |  |
| --- | --- | --- |
| *Correlation* | *Attribute (output)* | *Attribute (input)* |
| -0.15 | Default payment | LIMIT\_BAL |
| 0.33 | Default payment | PAY\_0 |
| 0.27 | Default payment | PAY\_2 |
| 0.24 | Default payment | PAY\_3 |
| 0.22 | Default payment | PAY\_4 |
| 0.20 | Default payment | PAY\_5 |
| 0.19 | Default payment | PAY\_6 |

1. Example of a figure caption. *(figure caption)*



1. Attributes correlation diagram

# BASELINE EXPERIMENTS

Even though we now know that data quality leaves a lot to be desired, we shall run some experiments to arrive to a baseline model, which we can hopefully improve later by employing further data exploration. The experiment will use 10-fold cross validation with 3 repeats as it represents the best data coverage for testing and validation. It will also factor all nominal values to enable the following 7 binary classification algorithms in R to work properly:

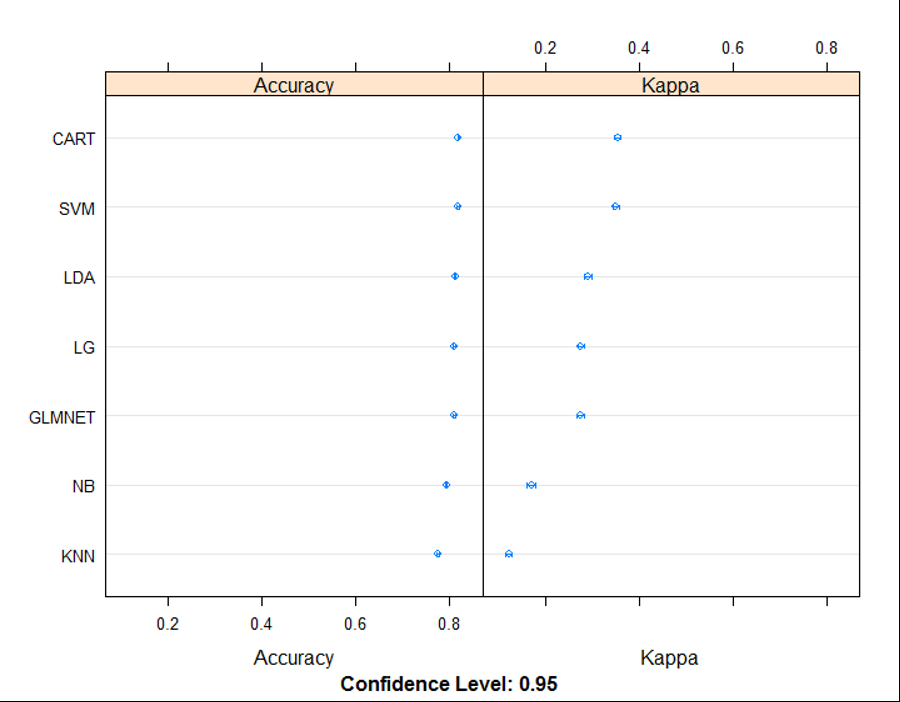
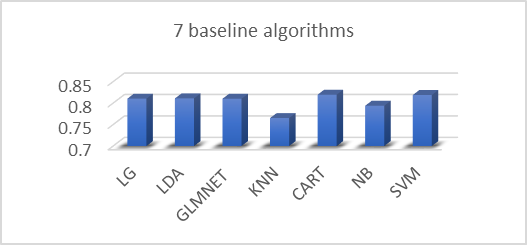
* LG Logistic Regression
* LDA Linear Discriminate Analysis
* GLMNET Regularized Logistic Regression
* KNN k-Nearest Neighbours
* CART Classification and Regression Trees
* NB Naive Bayes
* SVM Support Vector Machines with Radial Basis Functions

The algorithms were chosen because they represent different methods [8]. The data is randomly partitioned into two sets, a training set consisting of 80% and a validation data set consisting of 20% records. **The first training run** produced the following results:



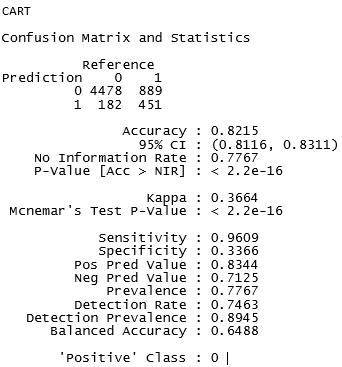
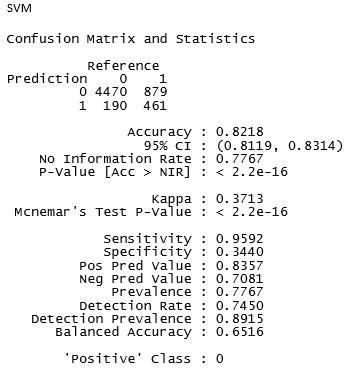
1. First run results

The results reveal that all classifiers have very similar performance, with KNN slightly falling behind. It is interesting to note that LG, GLMNET and LDA have almost identical mean accuracy with CART and SVM being the best performers, but ever so slightly in statistical terms. **As the current base prediction is 77% non-default outputs, and the accuracy mean is 80%, it is reasonable to conclude that the current models do not explain the variability of the dataset and as such do not provide much value [3].**



1. First run results (*prediction accuracy and Kappa*)

Furthermore, **Kappa values are very low, which indicates how well the classifier performed as compared to how well it would have performed simply by chance.** In other words, a **model will a low Kappa score indicates that there is not a big difference between the accuracy and the null error rate.** Null error rate shows how often predictor would be wrong if it always predicted the majority class. After performing predictions against the validation data set using the CART and SVM trained classifiers to find out how well they perform, and also what the null error rate and Kappa values are, the following results are achieved:



1. SVM and CART confusion matrix

The following conclusions can be drawn:

* Validation accuracy of around 80% is very similar to training accuracy which shows the training worked well with 10-fold cross validation with 3 repeats.
* Kappa values while statistically improved, still very low, so not much better than predicting the non-default all the time.
* Negative prediction value that indicates how well the classifiers predict that consumer will default are very low, around 70%.
* Specificity that shows how well the classifiers predict default when it is a default is very low around 34%, which along with the low negative prediction rate confirms they could not be used professionally and attempts must be made to improve them.

# FURTHER BASELINE EXPERIMENTS

The following experiments will be performed to try to improve the models:

* Box-Cox data transformations to normalise data [8]
* Resample dataset to get a class balance ratio of 1:1
  + Take all data with output class set to 1 and put it into a temporary dataset Default. Divide the rest of the data with output class set to 0 into 5 equal sub-sets, and randomly take an equal amount of data from each and put it into a temporary dataset NonDefault. Merge the two sets into a final dataset ResampledTrainingDataset totalling 13271 rows.
* Manual resampling of ResampledTrainingDataset
  + Indexing of the 13271 rows was randomly changed resulting in a better mix of class output throughout the dataset
* Boosting and Bagging on ResampledTrainingDataset
  + Bagged CART (BAG) and Random Forest (RF).
  + Stochastic Gradient Boosting (GBM) and C5.0 (C50).

Numerous and mostly time-consuming and long-running tests were performed with the above transformations in hope of achieving improvements, but the results were disappointing, particularly running the boosting and bagging algorithms against a perfectly balanced dataset ResampledTrainingDataset.

|  |  |  |
| --- | --- | --- |
| *Algorithm* | *Kappa* | *Negative prediction* |
| Box-Cox original data SVM | 0.37 | 0.70 |
| Box-Cox original data CART | 0.36 | 0.71 |
| GBM resampled data | 0.39 | 0.69 |
| C50 resampled data | 0.39 | 0.70 |
| SVM resampled data | 0.38 | 0.48 |
| CART resampled data | 0.37 | 0.48 |

1. Further baseline experiments (*results overview*)

It showed that **rebalancing output class does not guarantee to yield any improvements**. The SVM (which took 9 hours to complete with Box-Cox method) and CART performed much worse on a smaller and rebalanced dataset. The only very slight statistical improvements in Kappa values was delivered by GBM and C50, but still unusable for any real-life scenarios.

# FURTHER DATA EXPLORATION – TRANSFORMED DATASET

It is obvious that further data exploration should be carried out in search of a good classifier. The objective here is clear: the **negative prediction rate (default) and Kappa value must be improved.** The following actions were performed:

* Remove all rows where PAY1-6 values are outside -1, 1, 1+. values. They are inaccurate and their meaning is not defined.
* Update all rows where Education values are outside 1-4 range to be 4 (unknown).
* Update all rows where Marriage values are outside 1-3 range to be 3 (others).
* Add a penalty points columnPAY\_PENPOINTS to add a point to each missed payment: if any PAY (0-6) is >=1 then PAY\_PENPOINTS=(SUM(PAY\_0 + PAY\_2 + PAY\_3 + PAY\_4 + PAY\_5 + PAY\_6)) \* 2 Otherwise, PAY\_PENPOINTS=0. **Data exploration earlier has shown under ‘Attributes correlation’ that the association exists between the PAY(0-6) and output, so this should help classifier towards predicting default when it is a default.**
* Add a column PAYAMT\_ZERO: if there are at least two columns with 0 for PAY\_AMT (1-6), and the corresponding BILL\_AMT are >0, then PAYAMT\_ZERO = 1. Otherwise, set it to 0. **This should also help classifier to better predict default output.**
* Add a column AVG\_BILLAMT to capture a MEAN of all BILL\_AMT values to **help identify large spending patterns**.
* Add a column SD\_BILLAMT to capture a SD of all BILL\_AMT values to **put more weight on large spending patterns**.
* Add a column LAST\_BILL\_AVG\_RATIOto capture the ratio between the last bill BILL\_AMT6 and AVG\_BILLAMT to **show the tendency of increased spending**.
* Add a column SD\_AVG\_RATIO **(coefficient of variance)** to capture the ratio between the SD\_BILLAMT and AVG\_BILLAMT. This will really **accentuate large peaks in spending**.
* Remove all rows where SD\_BILLAMT = 0 as the bill and payment data for that consumer appears to be creating no sensible context or association to the real-life behaviours, and as such these rows are treated as outliers.
* Remove all rows where LAST\_BILL\_AVG\_RATIOor SD\_AVG\_RATIOare negative. These rows represent imbalanced bill statements where some statements are negative suggesting massive overpayments, but the consumer payment for that month does not reflect that.

**There were 11 transformations in total** applied to ResampledTrainingDataset. **A lot of care was taken to ensure that the transformations are applicable in the real-life scenarios.** Otherwise, the model would be false. The transformations were designed to **help the classifier predict a default when it is a default.** Then the decision was made to attempt an interesting experiment to use an artificial neural network against the transformed dataset in hope to obtain a much better fit.

# ARTIFICIAL NEURAL NETWORK

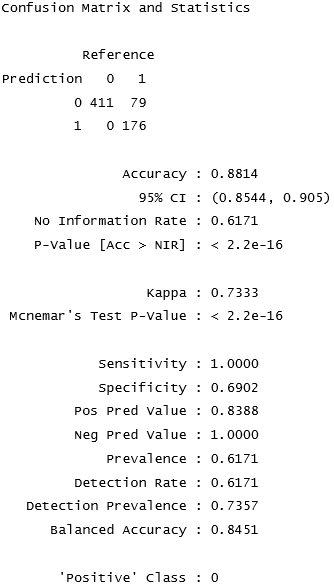
The original idea was to use two different R libraries, **nnet** and **neuralnet,** but getting neuralnet to work proved too complex. The feed-forward nnet library exposes a method with the same name, as illustrated below:

ann = nnet(observationsFormula, data=trainingDataset, rang = 0.001, size=10, maxit=10000, decay=0.001)

The parameters are:

* observationsFormula is a vector of class output and attributes used to calculate weights
* data is the pointer to the training data set
* rang sets initial random weights, normally starts at 0.5. Given the usage of LIMIT\_BAL and PAY\_PENPOINTS, the lowered value of 0.001 performed best.
* size represents a number of cells in the hidden layer (nnet library only allows a single hidden layer, which is totally adequate for our experiment)
* maxit represents a maximum number of iterations
* decay is a parameter for weight decay, normally set to 0. The best results were achieved by a much lower value of 0.001.

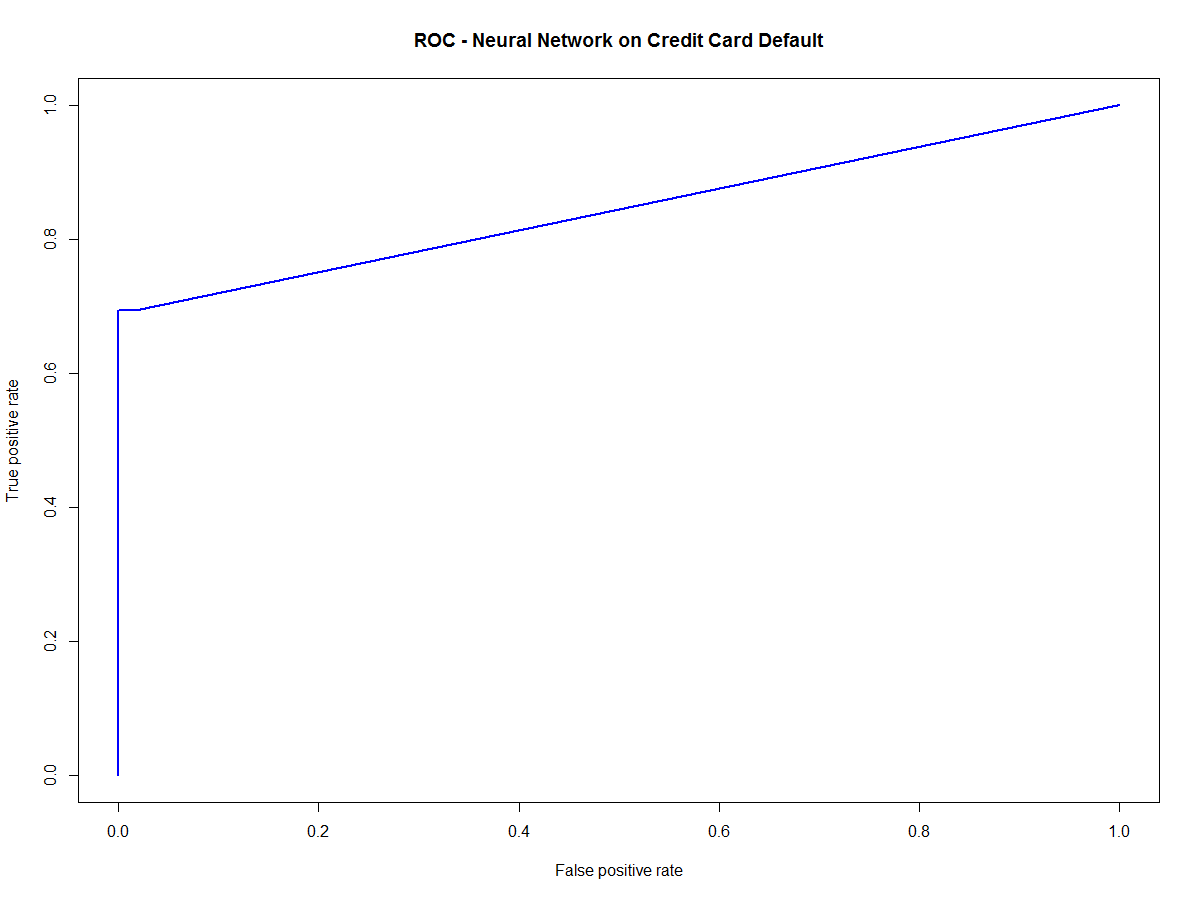
Numerous experiments were run with different data set attributes and different values for rang, size and decay parameters. The best fitting model emerged to be as follows:



1. Artificial neural network confusion matrix *(best results)*

A noticeable improvement has been achieved with the following results:

* Sensitivity on non-defaults of 100%. Negative prediction value of 100% concluding that **network predicted a default when it was a default**. That was the main objective.
* **Kappa value significantly increased from 0.39 to 0.73** concluding that the **model prediction power moves up from the null error rate** that simply predicts the majority class.
* Specificity of 0.69 also improved, but still misclassified 79 true defaults as non-defaults out of 255. Far from perfect, but a definite improvement and a **move in the right direction.**
* The ROC curve shows the **potential of the model**, but also exposes the large area for improvement.

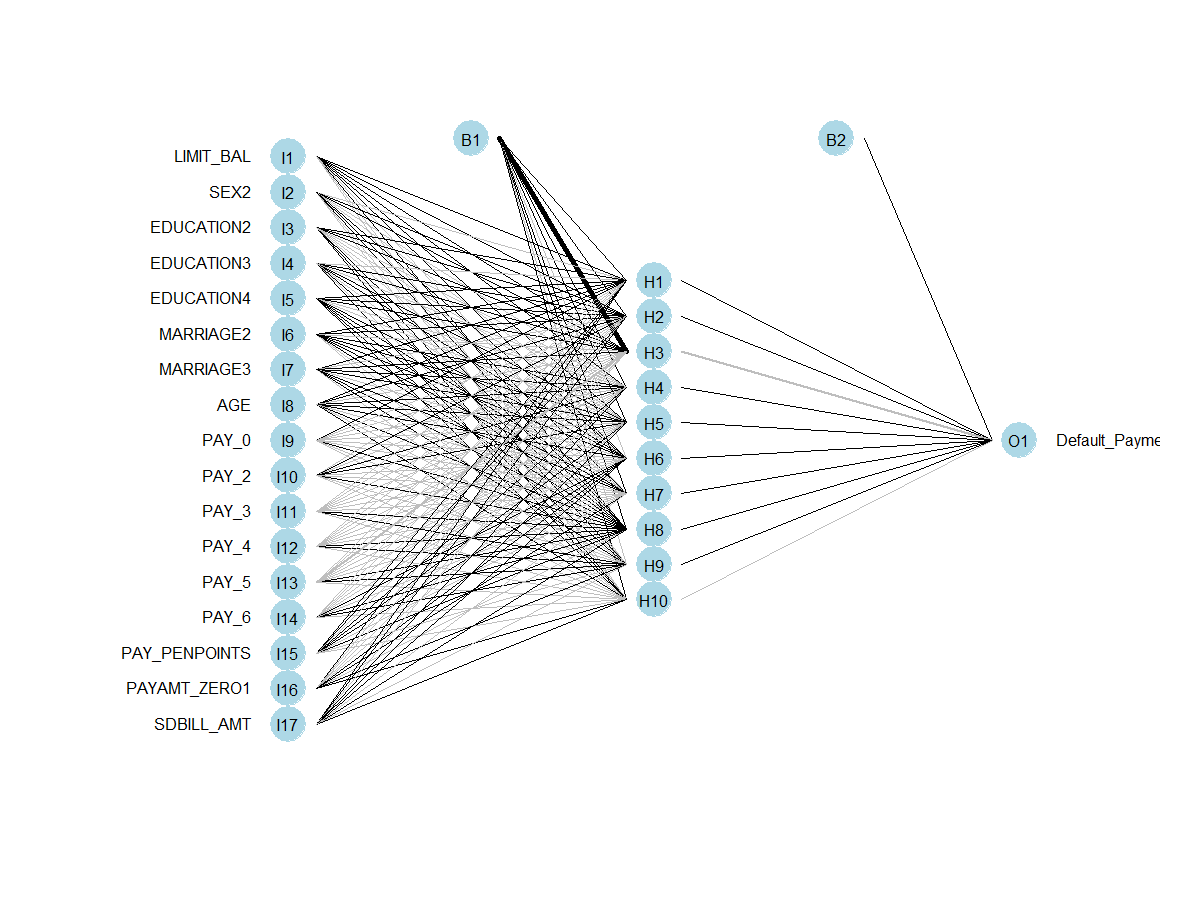


1. ROC curve (*based on neural network confusion matrix*)

The resulting neural network has the following properties:

17-10-1 network with 191 weights

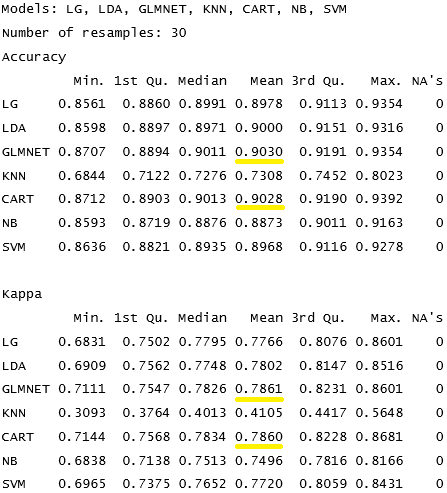
inputs: LIMIT\_BAL SEX2 EDUCATION2 EDUCATION3 EDUCATION4 MARRIAGE2 MARRIAGE3 AGE PAY\_0 PAY\_2 PAY\_3 PAY\_4 PAY\_5 PAY\_6 PAY\_PENPOINTS PAYAMT\_ZERO1 SDBILL\_AMT



1. Plot of artificial neural network. *(191 weights)*

# ORIGINAL CLASSIFIERS WITH TRANSFORMED DATASET

Encouraged by the visible improvements and keeping in mind that we have not tried any of the classifiers we used on the original data set, we shall conduct the last experiment to find out if the classifiers will respond well to the transformed data set.



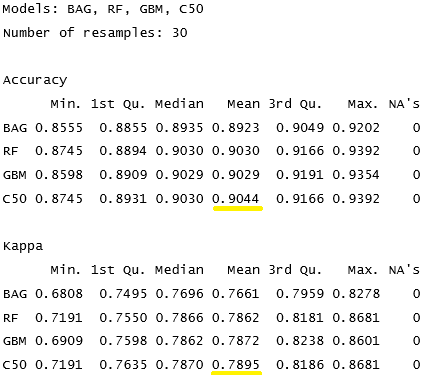
1. Original classifiers with transformed data set results

If we exclude KNN, **the results are very encouraging** for all classifiers, particularly for GLMNET Regularized Logistic Regression and CART Classification and Regression Trees showing **an improvement in Kappa value from 0.73 to 0.79**.

This shows that the transformed dataset can be explained well with **regression models resulting in a slightly better, but similar output to the neural network**. The new attributes have especially helped the CART to improve performance, and **for the ease of interpretation, this model should be chosen to visualize the decision rules [9]**.

We shall do one more test, and that will exhaust this exercise within the current scope. We have already used some ensemble algorithms with the original data set, so we shall run them against the transformed data set to see how they perform.

The last test with the ensemble methods shows a further, ever so slight improvement using C50 algorithm.



1. Results of ensemble methods on transformed dataset

# CONCLUSION AND FUTURE WORK

* Data analysis, exploration and clean-up was the largest and most involved part of the project.
* Dependable results very difficult to achieve with the highly inaccurate data.
* Artificial neural network indicated a good way forward based on **transformed data** with new attributes providing the new direction. This was further confirmed using other binary and ensemble classifiers.
* For a real-world professional model, new attributes and accurate data would be required to capture more aspects of consumer credit behaviour. Potential attributes could be:
  + country-dependent financial indicators, such as level of debt per household, distribution of debt per household based on social and financial sub-categories.
  + longer historical view of the consumers past monthly payments, i.e. 12, 24 or 36 months. Research by Kennedy [10] has shown that the 12 month window is the most optimal in terms of accuracy.
  + consumer credit rating with some ordinal values from credit reference agencies.
  + consumer overall debt level as a continuous value.
  + consumer income as a continuous value.
  + consumer address with some ordinal categories (bad, good, excellent).
* The new attributes could uncover the new strong predictors, and in combination with the existing ones sway the classifier towards a model that predicts very well.
* Combination of separate ANN, C50, CART, GLMNET models with a final most frequent class output derived from the 4 separate models could potentially provide a solid model with the good statistical fit to consumer behaviours.

The **neural model** is included as it **provides adaptability, robustness and predictive accuracy** for the evaluation of bank conditions and credit risk. From the other angle, it may be true to say that the **other machine learning algorithms** could be used to **compliment and further support the classification output achieved by the neural model**. Particularly when a quantified measure of a decision is close to 0.5 and an additional evaluation is required to confirm the final output.

The **binomial prediction could be upgraded into a multi-class classifier** with the precisely defined categories that have a quantified association to the real risk of consumers defaulting on a payment. For example, there could be five categories, highly likely, likely, middle, unlikely and highly unlikely to default. This approach would provide good interpretation of the default likelihood to the bank management.

Another approach might involve the use of **self-organising maps** as one of the most flexible clustering techniques used for failure prediction and visualisation. [1]

##### References

1. C. Dunis, P. Middleton, A. Karathanasopoulos, K. Theofilatos, “Artificial Intelligence in Financial Markets (Cutting-Edge Applications for Risk Management, Portfolio Optimisation and Economics)”, pp. 3, pp. 179-186
2. E. Siegel, Predictive Analytics, pp.148-149.
3. G. James, D. Witten, T. Hastie, R. Tibshirani “An Introduction to Statistical Learning”, pp. 144-145.
4. http://cs229.stanford.edu/proj2015/199\_report.pdf, “Predicting Default Risk of Lending Club Loans,”, Accessed 07/04/2017.
5. http://cs229.stanford.edu/proj2014/Marie-Laure%20Charpignon,%20Enguerrand%20Horel,%20Flora%20Tixier,%20Prediction%20of%20consumer%20credit%20risk.pdf, “Prediction of consumer credit risk,” Accessed 07/04/2017.
6. http://mitsloan.mit.edu/media/Lo\_ConsumerCreditRiskModels.pdf, “Consumer Credit Risk Models via Machine-Learning Algorithms,” Accessed 12/04/2017.
7. http://aircconline.com/mlaij/V3N1/3116mlaij01.pdf, “Developing prediction model of loan risk in banks using data mining”, Accessed 14/04/2017.
8. J. Brownlee, “MACHINE LEARNING MASTERY WITH R, Binary Classiﬁcation Machine Learning Case Study Project”, pp. 190-221
9. J. Brownlee, “MASTER MACHINE LEARNING ALGORITHMS”, pp. 77-82
10. K. Kennedy, N. Mac, N.B., Delany, S.J., O’Sullivan, M., N. Watson “A window of opportunity: Assessing behavioural scoring”, Expert Systems with Applications, pp. 1372-1380

##### APPENDIX

All R scripts and data distribution plots are contained in the

appendix.

# Acknowledgements

I am grateful to Professor Andrew Ng at Stanford University who has been my first connection with Machine Learning through his Coursera course before I started my MSc studies here at RGU. The course was a valuable experience that created a hunger for more.

I would also like to say a ‘big thank you’ to Professor Chrisina Jayne at RGU for keeping this newly found inspiration alive with her passion for the subject and excellent lectures which I have immensely enjoyed.

And finally, my thanks to my family for being supportive and

understanding during the long hours away from home studying

at RGU library.