Machine Leraning Project: Credit Card default Classification

Rudrani Angira and Srivatsan Iyer

April 26, 2016

1 Data and Literature Review

The dataset has the financial indicators for the customers from Taiwan and the goal is to predict the probability of default based on the given financial and demographic indicators. The common data mining methods studied for credit scoring in literature have been derived from the classification techniques which employs a classifier to divide the data into categories based on some variables. In this report we analyze the three classification mechanisms namely decision trees, K-nearest neighbours and Artificial neural networks as we need to classify the data into two classes. In Analysis section we compare the results obtained from the three classifiers.

2 Dataset Description

The dataset has continuous and categorical variables.There are 23 features in total. Below is the description of the variables[1]:

1. Limit Balance: Amount of the given credit (NT dollar): it includes both the individual consumer credit and his/her family (supplementary) credit.
2. Gender: (1 = male; 2 = female).
3. Education: (1 = graduate school; 2 = university; 3 = high school; 4 = others).
4. Marital status: (1 = married; 2 = single; 3 = others).
5. Age: Age (year).
6. PAY_0 - PAY_6: History of past payment. We tracked the past monthly payment records (from April to September, 2005) as follows: PAY_0 = the repayment status in September, 2005; PAY_2 = the repayment status in August, 2005; . . . ;PAY_6 = the repayment status in April, 2005. The measurement scale for the repayment status is: -1 = pay duly; 1 = payment delay for one month; 2 = payment delay for two months;. . . ; 8 = payment delay for eight months; 9 = payment delay for nine months and above.
7. BILL_AMT1-BILL_AMT6: Amount of bill statement (NT dollar). BILL_AMT1 = amount of bill statement in September, 2005; BILL_AMT2 = amount of bill statement in August, 2005;. . . ; BILL_AMT6 = amount of bill statement in April, 2005.
8. PAY_AMT0-PAY_AMT6: Amount of previous payment (NT dollar).PAY_AMT0= amount paid in September, 2005,PAY_AMT1 = amount paid in August, 2005;. . . ;PAY_AMT6= amount paid in April, 2005.
9. default.payment.next.month: 1 if the person defaulted in the next month. 0 otherwise.

The original data contains around 30,000 observations with around 24,000 samples of no-default and 6000 samples of default. We have sub-sampled this data to be around 12,000 samples with 6000 samples of each category.

3 Exploratory Analysis

The correlation matrix has been plotted between the continuous variables to see if any of the variables can be eliminated. The matrix is plotted in the R markdown file Code.Rmd consisting of the code and the plots. It can be observed that the continuous variables are not correlated with each other. Monthly bill payments show some correlation with other monthly bill payments and monthly bill amounts show some correlation with other monthly bill amounts which is an expected result.

But there is no correlation between monthly bill amounts and monthly bill payments and hence all the continuous features can be retained for further analysis.
For classification algorithms the normality assumption of the features is not mandatory hence we do not plot the Quantile-Quantile plots of the features. After some initial experiments, we decided to bucket the continuous features to make them categorical. We experimented with various number of bucket before we settled with 3 buckets for decision tree and 15 buckets for neural networks.

4 Implementation

In this section we briefly explain the algorithms used for the analysis.

1. K-nearest Neighbours: It does not require any predictive model prior classifying the observations. The algorithm searches the training data space for nearest neighbours and classifies the new observation in the most frequent class of its K closest observations which is calculated based on a distance metric. In our implementation we have used Euclidean distance.

2. Decision Trees: From the given dataset, we tried to find if a sequential set of decisions can help us predict if a person would default. ID3 Decision Tree classification algorithm fits in perfectly. In the algorithm, our main goal is to create the tree of decisions. Each node in the tree corresponds to a feature vector. Every outward edge represents one unique value of the feature. Once the tree is generated, we simply follow along the tree, taking decisions at every node and following the edge that represents the unique value from our test data. The process create the decision tree relies on calculating minimum entropy or maximum information gain.

3. Neural Network: It is a non-linear statistical method which can work on any kind of dataset. It basically establishes the mathematical relationship between the inputs (feature vectors) and outputs (response variable) and minimizes the error through the back propagation. The weights are updated at every iteration and the errors are recalculated till the error rate converges. The training process builds a probabilistic mapping between the input features and classes and based on this model the neural network predicts the probability of the new observation belonging to each class. In our implementation there are four hidden layers. Below is the equation model for the neural network.

$$net_{hi} = \sum_{i=1}^{n} w_i * input_i$$

where, net_{hi} is the value on a neuron in hidden layer. The output of every neuron out_{hi} in the hidden layer will be

$$out_{hi} = \frac{1}{1 + e^{-net_{hi}}}$$

The input for the neuron in the next layer o_i

$$net_{oj} = \sum_{j=1}^{m} w_j * net_{hj}$$

and the output will be

$$out_{oj} = \frac{1}{1 + e^{-net_{oj}}}$$

, the total error is calclated E_{total} = back propagation is done to update the weights

$$w_k^{'} = w_k - \eta * \frac{\partial E_{total}}{\partial w_k}$$

5 Fixing Parameters

As in any other prediction process, we have had our share of experiments to arrive at a set of parameters that perform the best. The section below relates to experiments we have performed on tweaking the parameters of classifiers we have used.

1. Decision Tree
   The ID3 algorithm requires the data to be categorical. Some of the features are real valued data. In order to bucket them as categorical data, we have to define the number of buckets. We fixed cut-off size to be 5, varied the number of buckets on the data and evaluated the accuracy. We got a plot that looks like below:

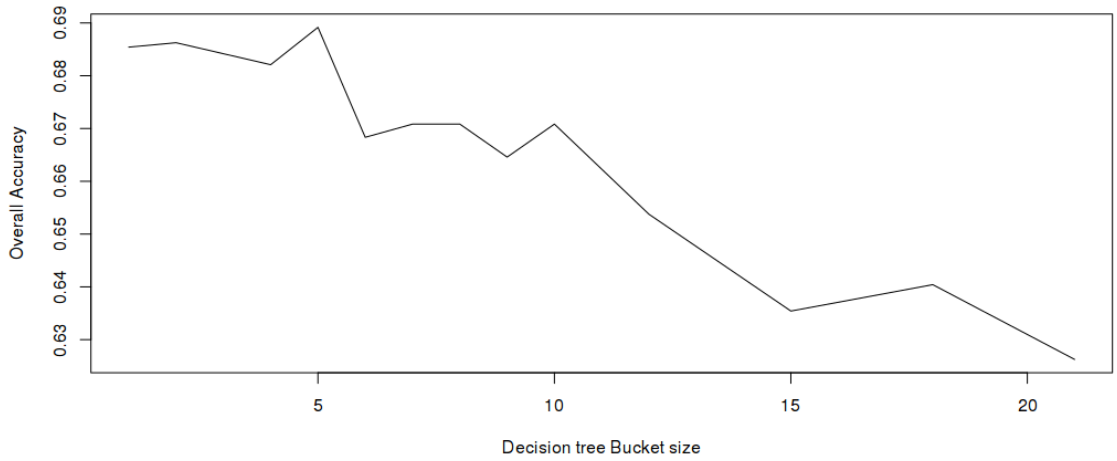

Figure 1: Decision Tree Bucket Size

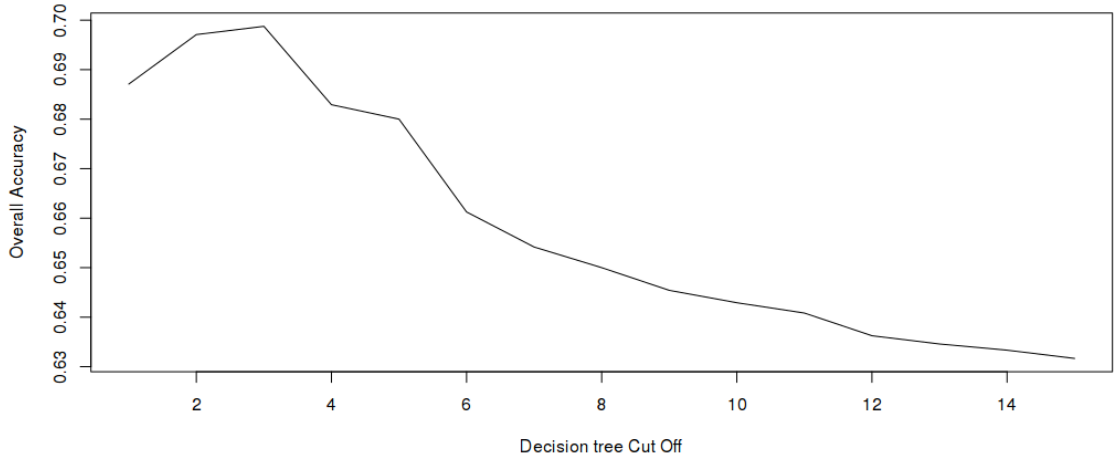

Figure 2: Decision Tree Cut Off

From **Figure 1**, we observe that the accuracy peaks when there are 5 buckets. We have used this for further experiments.Next, we conducted another set of experiments to arrive at an optimal cut-off. This number represents the maximum depth of the decision tree. We cut off the tree and do not let the algorithm fully create the tree to prevent over-fitting. Below is the plot of overall accuracy v/s the cut-off. Clearly, the graph in **Figure 2** exhibits the typical characteristics of overfitting. As we let the ID3 algorithm classify deeper nodes, the overall accuracy falls down. We have settled on cutoff=3

2. Nearest Neighbors
   In this classifier, we need to fix the value of K ,the number of nearest neighbors to consider. We present below the plot of values of accuracy v/s K.This experiment leads us to using the value of 70 for K.

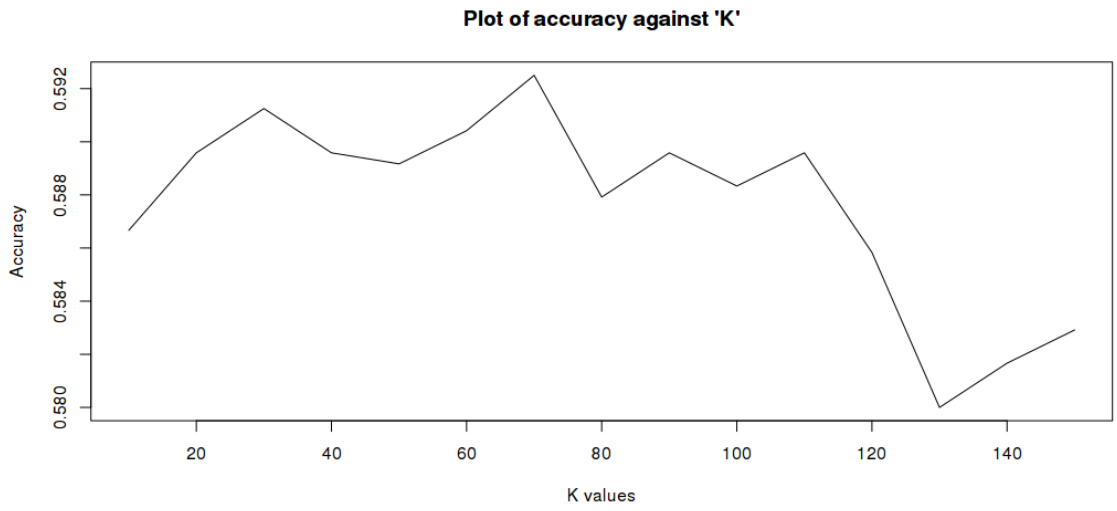

Figure 3: Knn best

6 Result

The accuracy for Decision Tree is 70.00%

Decision Tree Confusion Matrix		
	0	1
0	995	208
1	512	685

The accuracy for 70-Nearest Neighbour is 60.5%

KNN Confusion Matrix		
	0	1
0	668	413
1	535	784

The accuracy for Neural Network is 70.04% (with 10 nodes in hidden layer)

Neural Network Confusion Matrix		
	0	1
0	922	280
1	439	759

7 Conclusion

1. The classifiers incline towards using the feature PAY_0 because it closely relates to the final output. In our dataset description, this makes sense because PAY_0 relates to the most recent payment.
2. Neural Network and Decision Tree give the highest accuracy.
3. The customers who default in past payments are more probable to default in future payments.

8 References

1. Dataset from: http://archive.ics.uci.edu/ml/datasets/default+of+credit+card+clients.
2. Yeh, I. C., Lien, C. H. (2009). The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients. Expert Systems with Applications, 36(2), 2473-2480.
3. Lee, C. C., Lin, T. T., Chen, Y. T. (2011). An Empirical Analysis of Credit Card Customers Overdue Risks for Medium-and Small-Sized Commercial Bank in Taiwan. Journal of Service Science and Management, 4(02), 234.