Comparison of methods of data mining techniques for the predictive accuracy.

Pyzhov, Vladislav and Pyzhov, Stanislav

23 May 2017

On the comparison of data mining techniques for the predictive accuracy: The case of credit card payment default

Pyzhov Stanislav, Pyzhov Vladislav

May 23, 2017

Introduction

Our project aims to assess the performance of the different data mining models on the prediction of the credit card payment default. we used only three data mining techniques from original paper: k-nearest neighbors, Logistic regression, and Neural Networks. Moreover, there is another widely-used technique - Random Forest - that we decided to include in our paper. In order to evaluate the performance of each model, we used 2 criteria: Accuracy (by error rate) and area under the curve (AUC).

This project is based on the work of Yeh, Lien (2009) (*6*). In the paper, authors used the payment data set from the important bank in Taiwan. To build a model, the whole sample was divided in two subsets - training and testing sets - so each model could be trained on the first one and then be evaluated on the second. Our motivation was to see whether the same result could be obtained if we repeatedly apply the models to the different data sets. To do so, Monte Carlo simulation was implemented to generate these sets.

The rest of the paper is organizes as following:

- Section 1: we describe all chosen data mining techniques, two measures of accuracy, and

methods of estimating this measures.

- Section 2: we describe the dataset and explain the empirical methodologies.
- Section 3: we train 4 models on the data set with 5000 thousand observations using two different estimating methods (cross-validation and repeated cross-validation).
- Section 4: we implement Monte-Carlo simulation on the smaller data set and compare the results.
- Section 5: conclusion.

Section 1: Theory

Theoretical models

Data mining is the computing process to discover patterns in large data set, which is very useful in the implication of the credit card payment systems. For example, the decision making on the issues such like whether or how much to extend credits, when to collect delinquent credits and how to evaluate the quality of an applicant. In this project, we focus on the credit scoring, which means to classify applicants by their probability of the default. By comparing the predictions of our different models and the real data on the default, we can find out which model performs the best.

Logistic regression

Logistic regression (Logit) is a linear regression with categorical dependent variable. This regression can be used to analyze the probability of the binary outcome based on the effect of several independent variables. As a result, it is possible to define the risk associated with increase in the level of particular factor. This is especially crucial for our analysis due to the

nature of our dependent variable - ”default” and ”not default” and independent variables which define the characteristics of the applicant. The main advantage of this method is that it provides probabilities about certain outcomes while keeping variance at the low level, and disadvantage is such that estimates could be biased if there are nonlinear relation between the predictors.

K nearest neighbors (KNN) model

K-nearest neighbors is non-parametric method which can be used for classification tasks. Its main idea is that the object is assigned a class Which is the most common among the K neighbors of this object. Neighbors are taken from the data set with predefined classes, and according to the value of K, the algorithm showed what class is mostly distributed around the object. This method is particularly efficient when the training data set is large, in multi-class cases, and when there are anomalies. The disadvantages are such that the data must be representative, algorithm relies sufficiently on the data due to the fact that in order to classify an additional object it needs to use all objects, and computations of the algorithm are costly. (*4*)

Neural Networks Model

This model is based on the principal of the biological neural networks. The distinctive feature of this model is such that it is defined as a system of interconnected ”artificial neurons”. Each of them is connecting with each other through weights. As a result, their intercommunication can be used in order to understand the effect of specific input variables on the output through the unknown neurons (hidden units) which are represented as a hidden layer. The simplest case of the artificial neural network with one hidden layer is presented on the figure 1. Algorithm is based on the repetition of a two phase cycles: propagation and assignment of weights. This method is useful in the case of unknown causalities due to the hidden layers, sustainable to noises in the input variables and it is highly adaptive to the changing environment. The main disadvantages are such that it is not easily interpretable in a form of probabilistic formula and

computationally intensive to train.

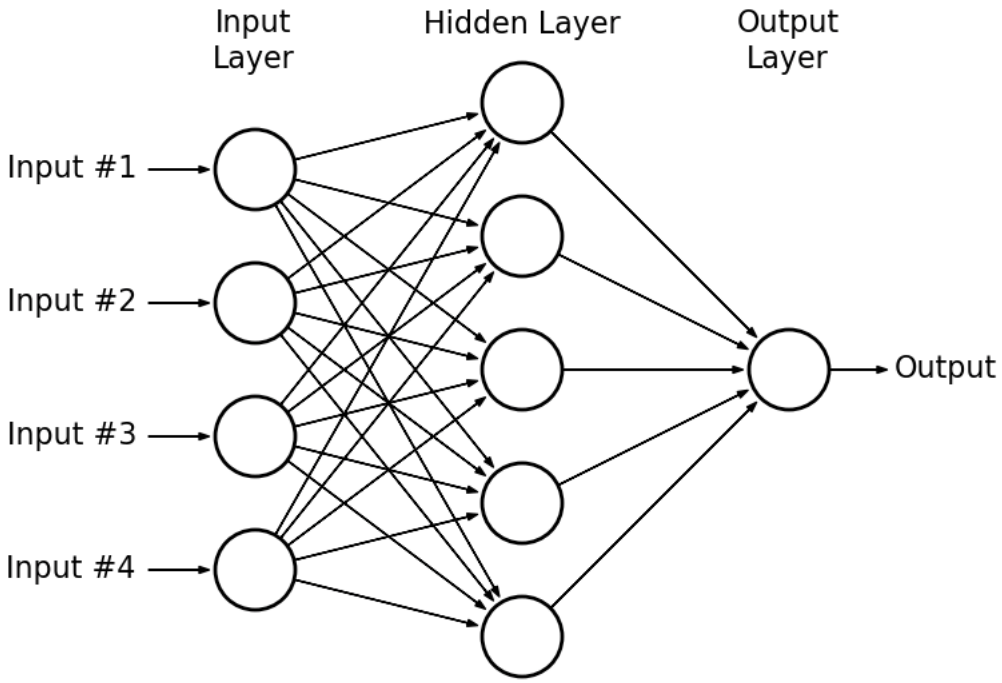

Figure 1: Artificial Neural Network algorithm

Random Forest

Random forest is a classification method which constructs a set of decision trees in order to define the class that is the mode of the classes output by individual trees. The main idea is that each tree constructs nodes which are represented as a test of specific attributes of the objects which then split in branches according the presence of this attribute. Figure 2 illustrates the conceptual diagram of the random forest algorithm. This method is one of the most accurate algorithms which is efficiently work both with large datasets and large number of variables. However, it overfits training datasets, the outcome is highly difficult to be interpreted by human, and its results cannot be implemented for other contexts. (*5*)

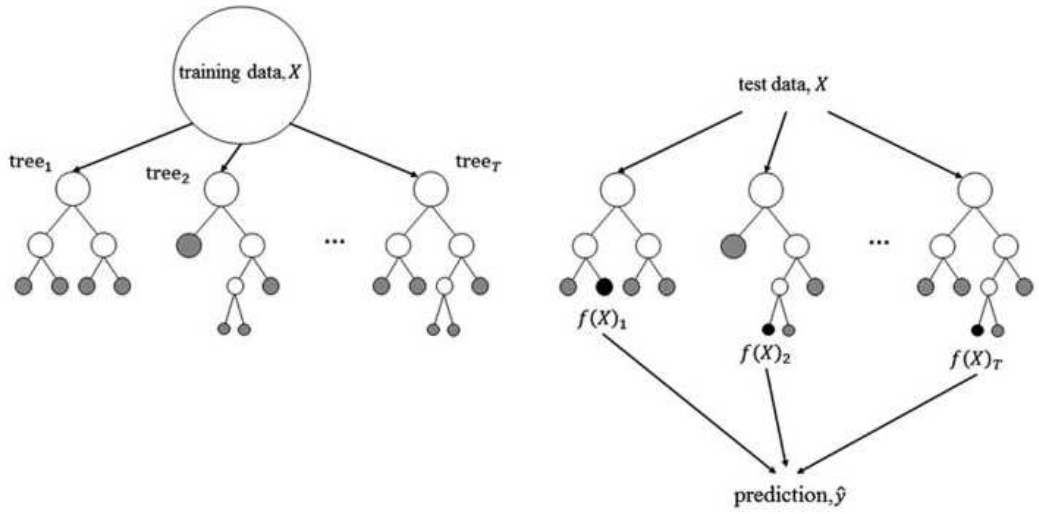

Figure 2: Random Forest algorithm

Methods of estimating model accuracy

In order to evaluate model's accuracy and to receive more or less unbiased result, it is necessary to use a specified method. Taking into account the limitations of our data, decision about allocation of the data set to the training and test sets is crucial. Therefore, we divide our sample of 1000 observations in two parts by half each. When operating with all data set of 30 000 observations this share could be reconsidered. As a result, in our case choosing the right resampling technique (bootstrap, cross-validation, etc.) can significantly affect the effectiveness of the performance. (*3*)

k-fold cross-validation (CV)

The idea is to split the training set into k-folds, then in rotation to build a model on k-1 subset and to evaluate on the remaining one. After that, when the model is built, we can estimate the error rate on the total testing set. The most popular size of k is 10, due to the fact that it is relatively less computationally expensive, and as it was shown by Kohavi (1995) 10-folds is the

most balanced decision in terms of bias and variance. (*2*) The main disadvantage of CV is that even producing little biased results, it could be highly variable according to the sample size.

Bootstrapping

Bootstrapping involves producing a number of random samples from the training data (with replacement), which are then used to evaluate the performance of the model. The method is more computationally heavier than CV, but the variance of an estimator is shown to be much smaller. However, the main disadvantage of bootstrapping is that it produces more biased estimations, especially on large data samples.

Repeated k-fold cross-validation (RCV)

The methodology is the same as in CV, but the whole process is repeated multiple times. CV method suffers from high internal variance due to the random division into k-folds, and RCV is used in order to overcome this problem. Ji-Hyun Kim (2014) has argued that RCV should be used as the preferred method due to its high evaluation characteristics. (*1*)

Section 2: Data and methodology

Data description The original data set is downloaded from the archive of University of California, Irvine. It is quite extensive with 30,000 observations and 24 variables. The responsive variable is default payment. The other 23 variables are explanatory ones. There are 5 demographic variables. The next 6 variables are about the history of payment. Remaining variables are about past amount of bill statement and paid amounts.

For section 3, we used data set of 5000 randomly selected observations, estimated measures using 10-fold cross-validation and 5 times repeated 10-fold cross validation.

For section 4, due to limited computers capacity, we randomly selected 1000 observations

with no replacement from the original data set. We performed Monte-Carlo simulation with 100 number of repeats, and also used simple 10-fold cross-validation for the estimation.

Pre-processing In some observation, there were missing values for all variables related to the past bill statements and paid amounts. These variables are crucial in identifying a class of card-holder, thus we excluded them from the original data set, so we got nearly 28500 instead of 30,000 samples.

Furthermore, another important aspect of the performance of the machine learning methods is a choice of data pre-processing techniques. Prerequisite for it was the fact that all models have its distinctive features and sensitivity to the various types of predictors in the model. These techniques are basically mean transformation, addition or deletion of the data set used for training the models. In our analysis, we used transformation technique, in particular, "centering and scaling" of the training data sample. As a result, predictors will have a zero mean and standard deviations equal to one, these as a consequence will improve estimates of the models. Data pre-processing is common and very important aspect of data-mining process, and it is especially important for the artificial neural networks algorithm, because otherwise it would lead to bias estimates.

Methodology for Monte-Carlo On the data set with 1000 observations, we did the following:

- Randomly selected samples from the reduced data with replacement, so in each iteration we had different data set.
- For each iteration, we divided the data equally into two groups: the training set and the testing set.
- We trained 4 data mining techniques on the training set.

- Then we tested the performance of each model by two measures.
- This procedure was repeated 100 times, and we calculated the number of times when each model was the best according to the value of Accuracy and AUC.

Section 3: Estimation of the models

We randomly selected 5000 observations from original data without replacement, and then divided into two equal subsets for training and testing. In order to evaluate the accuracy (error rate) of each model, we used 10-fold cross-validation technique.

When the models are built on training set, we estimated all of them on the testing set, and compared in terms of accuracy and AUC.

Accuracy is calculated as the number of all correct predictions divided by the total number of the data set. The best accuracy is 1.0, whereas the worst is 0.0.

ROC curves for all 4 models are presented on the graph 1. Each curve shows the classification ability of the model used. In particular, we plot True Positive Rate (TPR) against False Positive Rate (FPR).

TPR is also known as the sensitivity measure, and shows the rate when the person with positive class has been recognized correctly.

On the other hand, FPR is known as "the probability of false alarm" and can be calculated as (1 - Specificity). Specificity in its turn shows the rate when the person with negative class has been associated with it correctly. Thus, having 100 $\%$ Specificity yields that all people with negative class will be recognized correctly.

According to this measure, the best model that could be obtained is the one with highest TPR and the lowest FPR, that is represented on graph with the upper left corner point. Thus, the larger the area under the ROC curve, the greater is the classification ability of the model.

From the figure 3, it can be seen that Random forest perform greater than all other models, kNN seems to have the lowest result, while it is not clear between Neural Networks and Logistic regression.

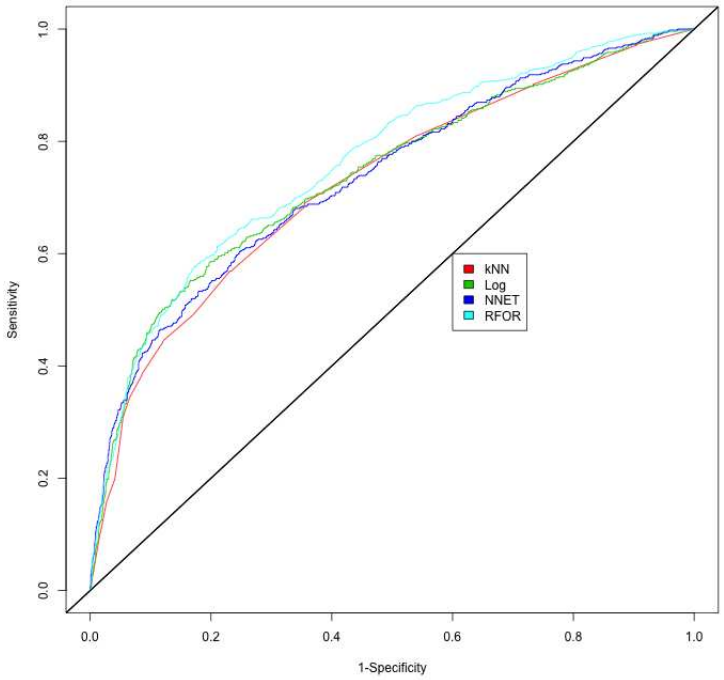

Figure 3: ROC curves with CV

On the table 1, we can see the precise measures of accuracy and AUC obtained for all 4 models. kNN-model shows the worst result with both measures. The greatest value according to accuracy measure receives Neural Networks method. However, it is not significantly higher than the ones obtained by Logistic regression and Random forest methods. Interestingly, when comparing the AUC results, Random forest receives the highest value of 0.762 which is significantly higher than all other. Moreover, it is possible to see that logistic regression performs better than Neural networks, having the value of AUC equal to 0.739.

In Yeh, Lien (2009) researchers argue that due to the fact that data contains greater number of non-risky card-holder, error rate (Accuracy) is insensitive to classification accuracy of the

Table 1: Accuracy and AUC values with CV

	kNN	Logistic	Neural Networks	Random Forest
Accuracy	0.811	0.819	0.82	0.819
AUC	0.725	0.739	0.735	0.762

model, and that it is more precise to use AUC as the measure. Thus, we can conclude that Random forest shows the best result in terms of AUC. It worth noting that in the corresponding paper, it was found that logistic regression performed worser than Neural networks method. The explanation for that change could be that (1) we sampled only part of the whole available data set, (2) we used cross-validation in order to compute the error rate (researcher did not specify in the paper which method they used or did they use one at all).

We wanted to see whether the result will be changed if we use another estimating method - 5 times repeated 10-fold cross-validation. Table 2 below shows the obtained results from this estimations.

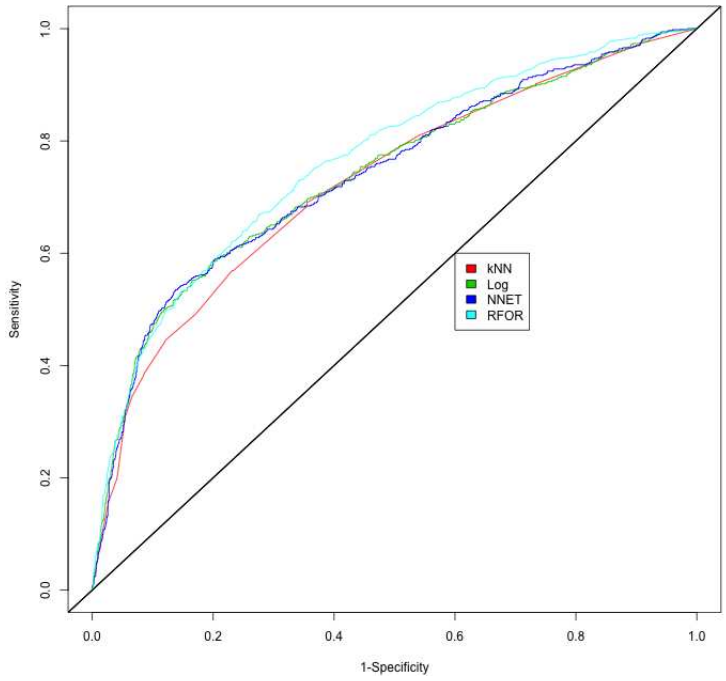

Figure 4: ROC curves with repeated CV

Table 2: Accuracy and AUC values with repeated CV

	kNN	Logistic	Neural Networks	Random Forest
Accuracy	0.811	0.819	0.816	0.816
AUC	0.725	0.739	0.739	0.763

It can be seen that now, according to the accuracy measure, Logistic regression performs greater than all other models with the value of 0.819. However, when comparing by AUC-measure, random forest shows the best result with 0.763 value, while logistic regression and Neural networks received only 0.739.

Section 4: Monte-Carlo

Performance on the training set

Even though the performance on the training set does not give any reliable estimations of accuracy, we show here the results in order to see how each model fits. On the figures 5 and 6, we can see the dynamics of Accuracy and AUC measures according to Monte-Carlo iterations.

The graph shows that for K-nearest neighbors and Logistic model, both ACC and AUC relatively fluctuate in the same range. However with Neural Network there are major parts that have 100% accurate. For Random Forest, it particularly performed well with every time it reaches 100% accuracy. Thus, when we counted the times when each model was the best, we got that Random Forest model was the best 100 times out of 100. In its turn, Neural Networks method was having 100% accuracy in 66 iterations, and 100% AUC 75 times. Both Logistic regression and k-nearest neighbors have never been received the highest value on both measurement. The result above seems suspicious. However, in machine learning, model error on the training data is meaningless. It is neither a limitation nor overfitting. It is merely a consequence of the fact that methods are built just to perform well on training data.In the case of RF, ”train set error” is expected to be near zero because RF uses not-pruned decision trees, hence naturally it performs

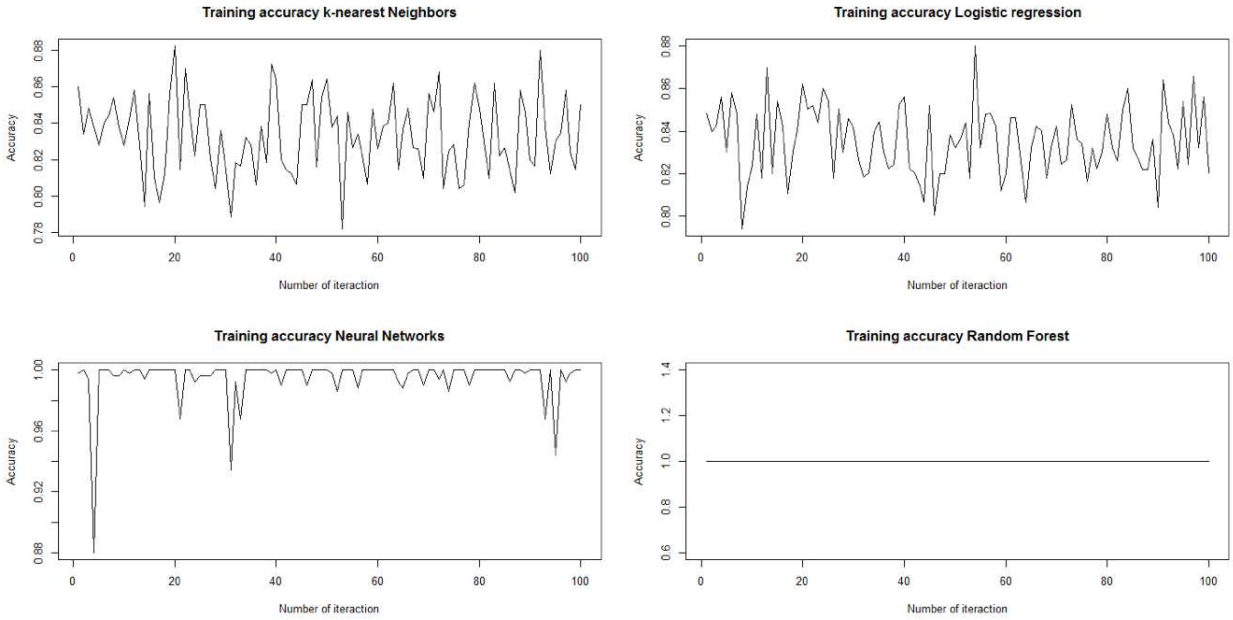

Figure 5: Dynamics of Accuracy on the training set

perfectly. In the case of NN it is the same since it uses complicated algorithm to get the best fit model.

Performance on the testing set

Figures 7 and 8 exhibit the dynamics of Accuracy and AUC measures for each model.

For ACC, Logistic, KNN and NN has similar range from 0.76 to 0.86. Whereas the lower bound of RF is 0.86 and its best prediction rate on test set is 0.94. It appears that performance on training set is much better than on test set. This is actually normal in Machine learning and agrees with literature.

When we counted for the times when each model was the best, it has showed that with Accuracy measure RF performed better in 99 times, while Neural Networks were the best only once. The situation with Logistic regression and kNN is the same as on the training set - these two models were never the best.

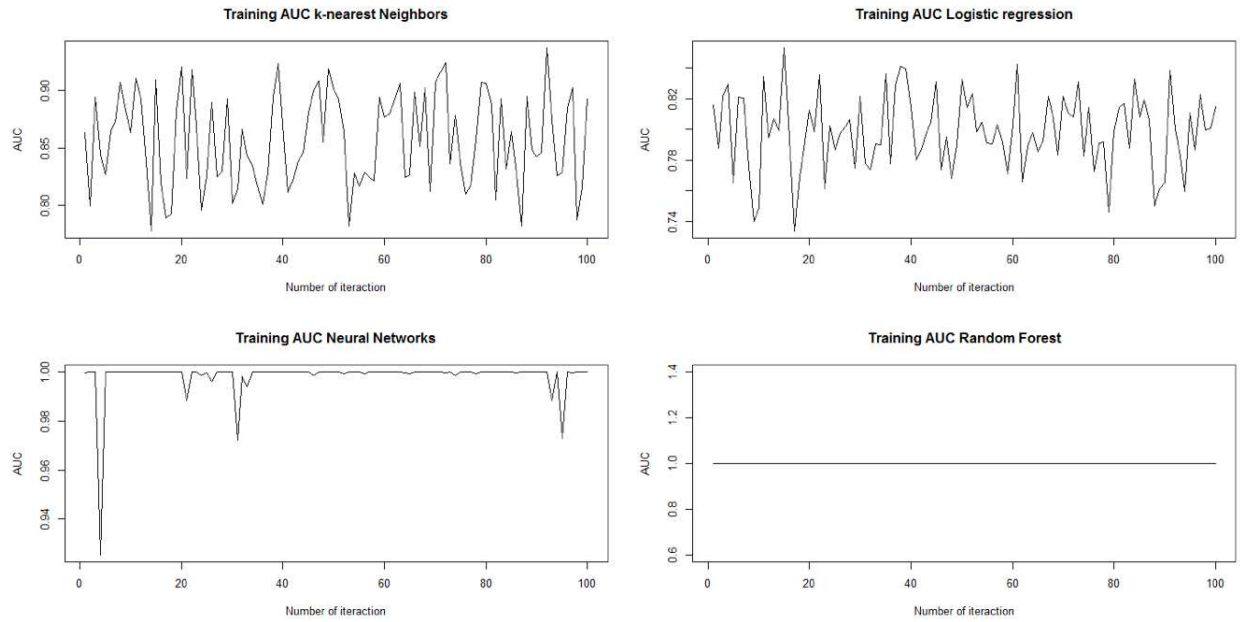

Figure 6: Dynamics of AUC on the training set

Thus, we can conclude that the RF did a very good job on classification. These results is not coming from one time estimation but from 100 times so it is not a matter of chance.

Choosing a better model by mearly selecting the higher ACC and/or AUC seems to be not very convincing if the numbers are relatively close to each other. It is more reliable if we use t-test for checking what if the difference in ACC and AUC of RF compared to other model is significant.

Since the accuracy in training set does not matter, we do the t-test on the accuracy criteria (ACC, AUC) of estimated model on test set. For each test, we receive a P-value much lower than 0.05. Therefore we are confident to conclude that the ACC and AUC difference between RF and other model is significant. Hence, RF does perform better than other techniques.

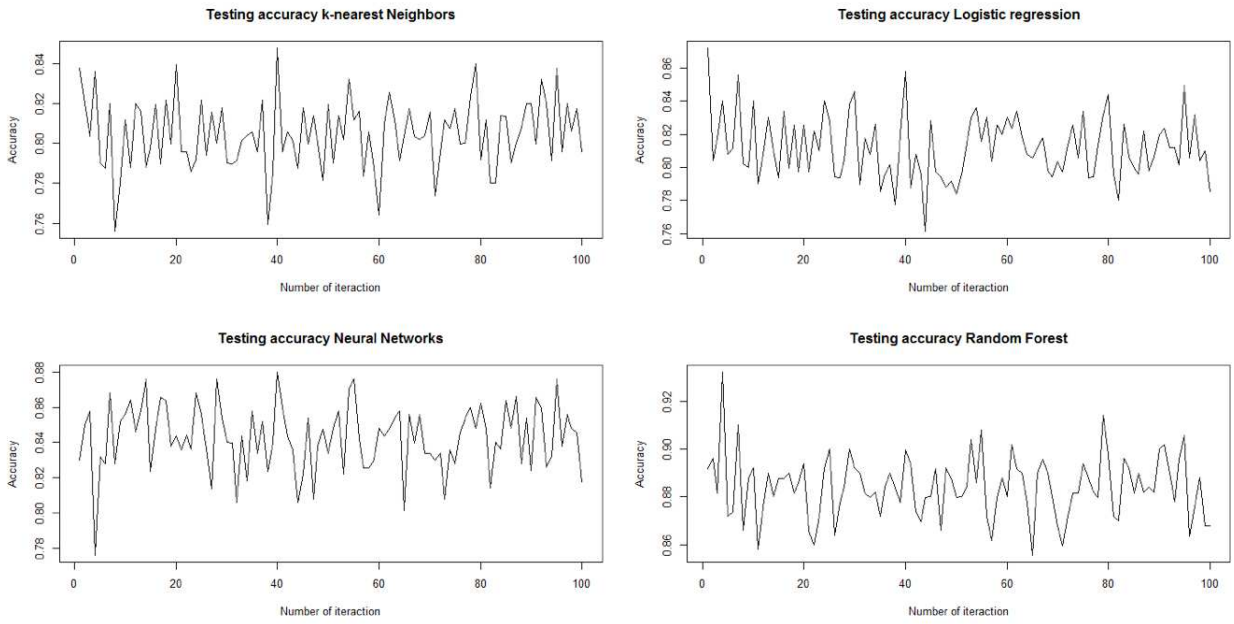

Figure 7: Dynamics of Accuracy on the testing set

Conclusion

In this project, we chose 4 data mining models - k-nearest neighbors, logistic regression, artificial neural networks, and random forest - and assessed their performance on the payment data about card-holders from important bank of Taiwan.

Firstly, we evaluated the performance of all 4 models on the subset with 5000 observation. It was found that depending on the estimating method, we received different results. Using 10-fold cross-validation, we found that Neural Networks received the highest value of Accuracy measure, while Random forest had the highest AUC value. When we changed the method to 5 times repeated 10-fold cross-validation, we found that Logistic regression performed better in terms of Accuracy, while, nevertheless, Random forest was still the best with the highest AUC.

Secondly, we used Monte Carlo simulation to assess the performance of several classification methods. We found that in the case of credit card holder profile, RF performed better in

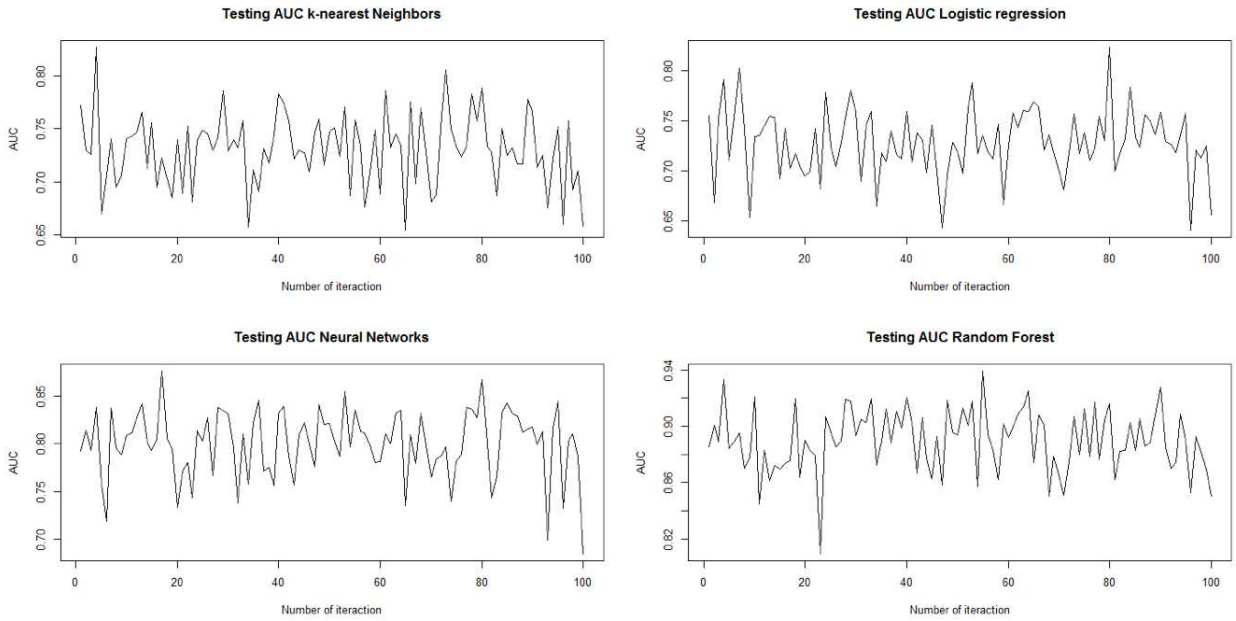

Figure 8: Dynamics of AUC on the testing set

predicting the overall rate of bad and good ones. Thus, it is more relevant for banks to use RF instead of other methods. RF is not commonly used since the method is very costly on computation. However with todays advanced computers, it is possible for any bank to implement RF to get better prediction rate.

However, We performed Monte-Carlo only 100 times and on the set of just 1000 observations. Thus, it would be better if we would take the bigger share from the original data (or even used the whole set) and then randomly take samples from it. We suspect that by boostraping the sample from a small pool of 1000 observations, there might be little difference in the performance of each model. If it is the case, it would explain the outpeformance of RF.

It is more important for bank to predict Good profile than predict bad profile since good ones give them benefits. In the particular case of banking, it would be better if we use Sensitivity than overall accuracy rate (ACC). We suspect that for a bank, a model could have the best ACC since it has sophisticated algorithm. Yet using it would bring less profits to bank compared

to simpler method since its Sensitivy is not large. The law of parsimony appears here when choosing analysis model: a complicated model should not be better than a simple one.

References

1. Ji-Hyun Kim. Estimating classification error rate: Repeated cross-validation, repeated hold-out and bootstrap. *Computational Statistics & Data Analysis*, 53(11):3735–3745, 2009.

2. Ron Kohavi et al. A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Ijcai*, volume 14, pages 1137–1145. Stanford, CA, 1995.

3. Max Kuhn and Kjell Johnson. *Applied predictive modeling*, volume 26. Springer, 2013.

4. Gordon S Linoff and Michael JA Berry. *Data mining techniques: for marketing, sales, and customer relationship management*. John Wiley & Sons, 2011.

5. Daniel Mennitt, Kirk Sherrill, and Kurt Fristrup. A geospatial model of ambient sound pressure levels in the contiguous united states. *The Journal of the Acoustical Society of America*, 135(5):2746–2764, 2014.

6. I-Cheng Yeh and Che-hui Lien. 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, 2009.