Who Will Subscribe A Term Deposit?

Jiong Chen (jc4133), Yucen Han (yh2645), Zhao Hu (zh2210),
Yicheng Lu (yl3071), Mengni Sun (ms4783)

December 7, 2014

Abstract

In this project, we aim to increase campaign efficiency by identifying the main factors that affect the success of a campaign and predicting whether the campaign will be successful to a certain client, namely, whether the client will subscribe a term deposit. As our data are imbalanced, we use resampling methods before building models. After preprocessing the data, we build four models: logistic regression, feedforward neural network, random forest and k-NN. The optimal model we get is the one using neural network algorithm.

Keywords: logistic regression, neural network, random forest, imbalanced data, bank marketing campaign

1. Background

The increasing number of marketing campaigns over time has reduced their effects on the general public. First, due to competition, positive response rate to mass campaigns are typically very low, according to a recent study, less than 1% of the contacts will subscribe a term deposit. Second, direct marketing has drawbacks, such as causing negative attitude towards banks due to intrusion of privacy. In order to save costs and time, it is important to filter the contacts but keep a certain success rate.

2. Objective

Our objective is to build a classifier to predict whether or not a client will subscribe a term deposit. If the classifier has high accuracy, the banks can arrange a better management of available resources by focusing on the potential customers "picked" by the classifier, which will improve their efficiency a lot. Besides, we plan to find out which factors are influential to customers' decision, so that a more efficient and precise campaign strategy can be designed to help to reduce the costs and improve the profits.

3. Data Source

Our data were collected from a Portuguese marketing campaign related with bank deposit subscription for 45211 clients and 20 features, and the response is whether the client has subscribed a term deposit. Our data set is downloaded from
http://archive.ics.uci.edu/ml/datasets/Bank+Marketing

The marketing campaigns were based on phone calls. Sometimes more than one contact to the same client was required.

The following are attribute descriptions:

Table 1: Description of Variables

Varaible	Type	Varaible	Type	Varaible	Type	Varaible	Type
age	Num	contact	Char	emp_var_rate	Num	poutcome	Char
compaign	Num	day_of_week	Char	euribor3m	Num	job	Char
cons_conf_idx	Num	default	Char	nr_employed	Num	loan	Char
cons_price_idx	Num	education	Char	pdays	Num	marital	Char
duration	Num	housing	Char	previous	Num	month	Char

4. Preliminary Data Analysis

There are two main issues with the dataset:

4.1 Missing Data

Since our data were collected from phone call interviews, many clients refused to provide their personal information due to the privacy issue. The existence of missing data may blur the real pattern hidden in the data thus making it more difficult to extract information. Therefore, we chose three methods to deal with those missing data for different attributes.

The first method is to do imputation, which we used on the "education" and "housing" attribute. For example, for the education attribute, there are 1731 missing lines marked as "unknown" in this attribute. We built a two-by-two way contingency table(Table 2) compared our target response and the known/unknown status of education attribute. A chi-square test is applied to check the independence and the p-value of the result is less than 0.01. In this case, the unknown status is obviously related to our target response and we cannot simply ignore those missing values. Therefore, we use the rest known data to impute the missing terms.

Table 2: Contingency table of missing value

	No	Yes	Proportion of Yes
Education: known	35068	4389	0.125
Education: unknown	1480	251	0.170

The second method is partial deletion, which we used on the "marital" and "job" attributes. For those attributes that do not show a strong relationship between the known/unknown status and target response, we partially delete those lines with missing value because we believe that those value are missed at random and ignore them will not affect our model.

The third method is a method for a special case. We find that in the "default" attribute, the total amount of "yes" response is very small, only have three clients. However, the number "unknown" status is quite large which is 8,598 in total. After chi-square test, there is also a strong evidence shows that there is a relationship between this unknown status and our target response. In this case, we cannot make imputation because of the rare population of "yes" response. Therefore, we decided to keep the "unknown" status as a new category and use it in our algorithm.

4.2 Imbalanced Data

The presence of imbalanced data may distort the algorithms and its predicting performance. This problem often happens in real world dataset, since people with some certain behaviors account for relatively smaller part. In this case, the responses in the training data are 90% "no" and 10% "yes", which is surly a significantly imbalanced dataset. How to deal with this problem can be divided into two parts.

First, change the way of measuring algorithm's performance. As a traditional and common measurement of performance, the test accuracy rate can not be simply used here because the model will tend to fit the majority class better to improve the overall accuracy. However, we prefer to be more successful in identifying people who will subscribe a term deposit than the overall power of prediction. Therefore, we will use ROC (Receiver Operating Characteristic) curve and AUC (Area Under Curve) as the performance measurement.

Second, change the dataset using resampling method or apply different weights to the observations in objective function. The most commonly used resampling method is oversampling (sample with replacement from the group with less data until the number equals to the larger group), undersampling (sample with replacement from the group with more data until the number equals to the lesser group) and mix sampling (mixture of oversampling and undersampling). The method with reweighting can vary a lot by applying different weights. Therefore, how to deal with the imbalanced data can vary according to the algorithms and the real situations.

5. Methodology

Since this is a classification problem with binary response, the method we attempt to try includes logistic regression, neural network, random Forest and k-nearest neighbors algorithms.

5.1 Logistic Regression

Logistic Regression, or logit regression, is a kind of probabilistic statistical classification model. It is also used to predict a binary response from a binary predictor. What's better, logistics model doesn't suffer a lot from severe class imbalance.

Logistic Regression models the log odds of the event as a linear function:

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x_1 + \cdots + \beta_P x_P$$

$$p = \frac{1}{1 + \exp\left[-(\beta_0 + \beta_1 x_1 + \cdots + \beta_P x_P)\right]}$$

This nonlinear function is a sigmoidal function of the model terms and constraints the probability estimates to between 0 and 1. Also, this model produces linear class boundaries, unless the predictors used in the model are nonlinear transformations of the original features.

Many of the categorical predictors in our project have sparse and unbalanced distributions. Because of this, we would expect that a model using the full set of predictors would perform worse than the set that has near-zero variance predictors removed.

5.2 Neural Network

The multilayer feedforward neural network model consists of multilayer of computational units. Each units in one layer known as a perceptron can compute a continuous output using the logistic function by using the outputs of last layer as its inputs. Easily speaking, each perceptron acts similarly to a logistic regression classifier and all these linear classifiers are interconnected in a feed-forward way. Therefore, this model can compute a probabilistic output like logistic classifier and also give a more powerful non-linear decision boundary.

However, there are still some limitations of neural network. The biggest problem is that neural networks are too much of a black box. It is hard to understand how the network is solving the problem and therefore we may never feel confident that it will generalize well to data not included in the training set.

5.3 Random Forest

The classification method we will try here is random forest. This method is good for prediction but a little bit difficult to interpret. Since we are facing the binary category, Random Forest is a good classification method to try.

Random Forest will grow a big tree without trimming, then, take majority vote of the results of all the trees. The process of this method is:

1. Take a sample of size n from the training dataset;
2. Randomly choose p variables from all the variables available;
3. Train a single big tree on the sample dataset and using p variables;
4. Repeat the step above B times;
5. Take a majority vote of the results for all of the B trees.

5.4 k-NN

Regular linear regression makes assumptions about the structure of the data (high bias), but its predictions are stable (low variance). We need a more flexible model that makes fewer assumptions. In contrast to linear regression methods, the k-nearest neighbor methods implement non-linear boundaries to our training and test data.

The k-NN method uses the average outcome value of its k nearest neighbors based on Euclidian distance.

$$\hat{Y}(x) = \frac{1}{k} \sum_{x_i \in N_k(x)} y_i$$

As we are facing a classification problem and the outcome variable is binary, our k-NN approach will assign a point to a group based on the majority vote of its k nearest points.

As our data is imbalanced, misclassification rate may not be a propriate way to explain our data. We should use other models to fit our data.

6. Results

6.1 Logistics Regression

We practiced 10-fold cross validation to tune parameters for logistic classifier. After we got a tuned model, we tested its performance on the testing set and the accuracy achieved 0.9132. Because our data is imbalanced, we practiced some remedial method such as changing the prediction threshold and oversampling, and got 0.859, 0.8597 for test accuracy, respectively. We also tried a "raw" version of logistic model without cross validation, and its accuracy is 0.9106.

According to these results, the best logistic model is the original tuned model.
Since the best logistic model is determined, we will display its performance.

Resampling results:

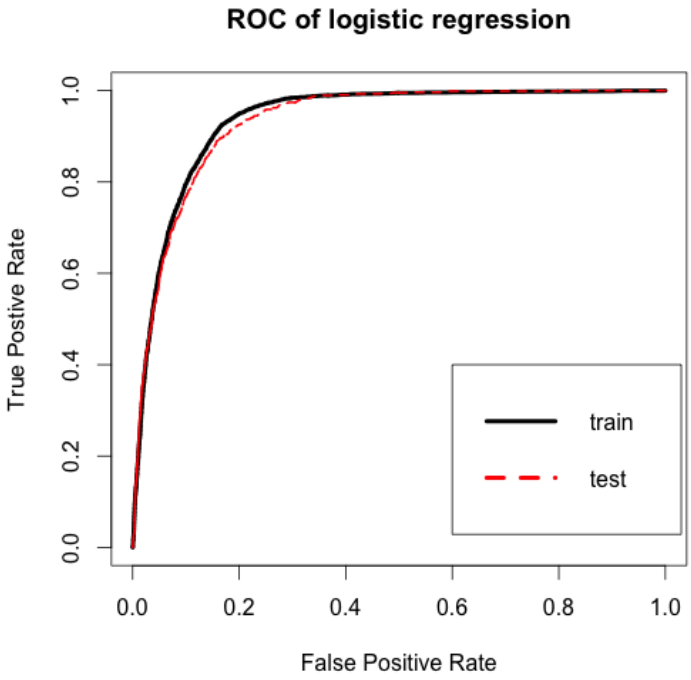
AUC	Sens	Spec	Accuracy	Kappa
0.936	0.973	0.424	0.912	0.471

Confusion matrix:

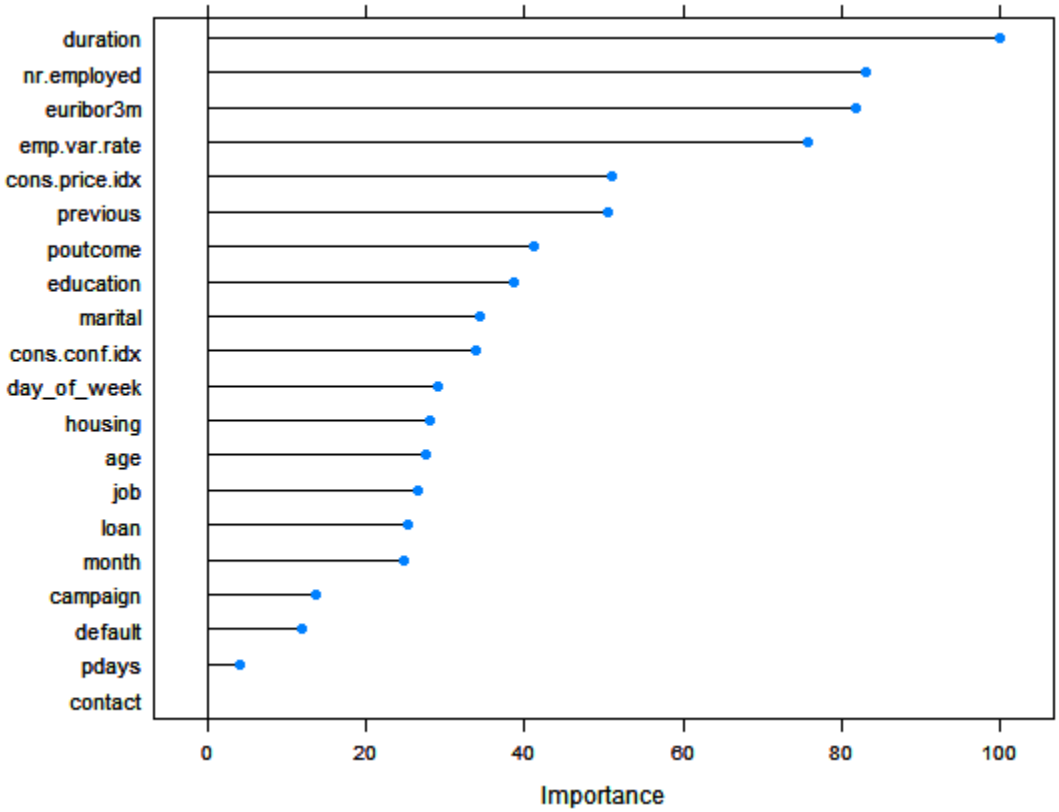
Table 3: Confusion Matrix for Logistic Regression

Logistic Model		Reference	
		no	yes
Prediction	no	6625	492
	yes	172	359

ROC Curve(the area under curve is 0.936):



Importance Ranking:



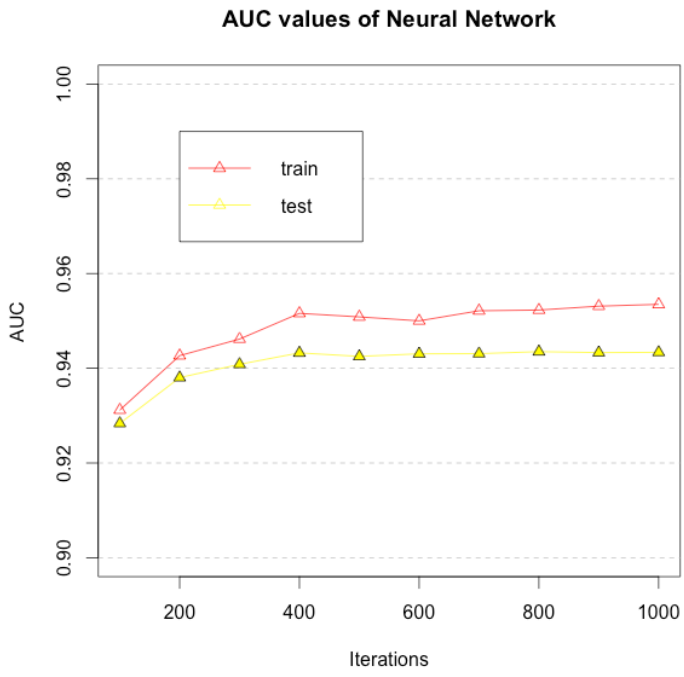
Estimated Coefficient of the four most important predictors:

duration(last contact duration, in seconds): 4.71e-03
nr.employed(number of employees): 6.26e-03
euribor3m(euribor 3 month rate): 4.16e-01
emp.var.rate(employment variation rate): -1.92

Hosmer and Lemeshow goodness of fit (GOF) test:
X-squared = 7648, df = 8, p-value < 2.2e-16
According to the test result, logistic regression model doesn't fit the data well.

6.2 Neural Network

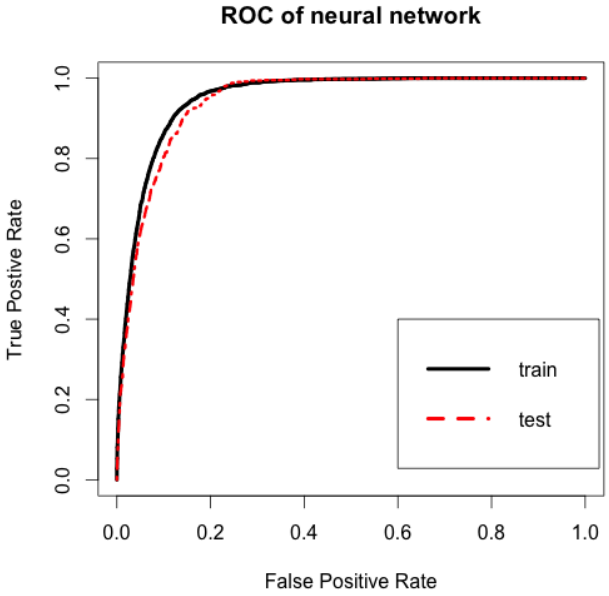
For the training procedure, the back propagation is chosen as the numerical optimal method to fitting the model. A 5-fold cross-validation is also designed to select best parameters for neural network, which use mean AUC value on validation set as model selection measurement. We have tried 275 different kind of parameters combination for grid search cross-validation. In each search we set the maximum iterations equal to 500. And the tuning parameters includes the number of layers, converge rate and the range of initial random weight.



After the searching, we find our best mean AUC value equal to 0.9436 on the validation set. And then, we fit the model on the entire training set and the figure shows the evaluation of our model on the training set and test set varies by the different maximum iterations. For each iteration, we fit the model three times and use the mean AUC value as it performance measurement.

We can observe that the model fits the training sets perfectly after 400 iterations, where the mean AUC values are 0.9516 and 0.9423 for training set and test set . And the initial

weights are randomly initialized within the range [-0.1,0.1] for this case, where we did not use any regularization of our features. This shows that the model indeed fit well for the loan response, at least for our data set.



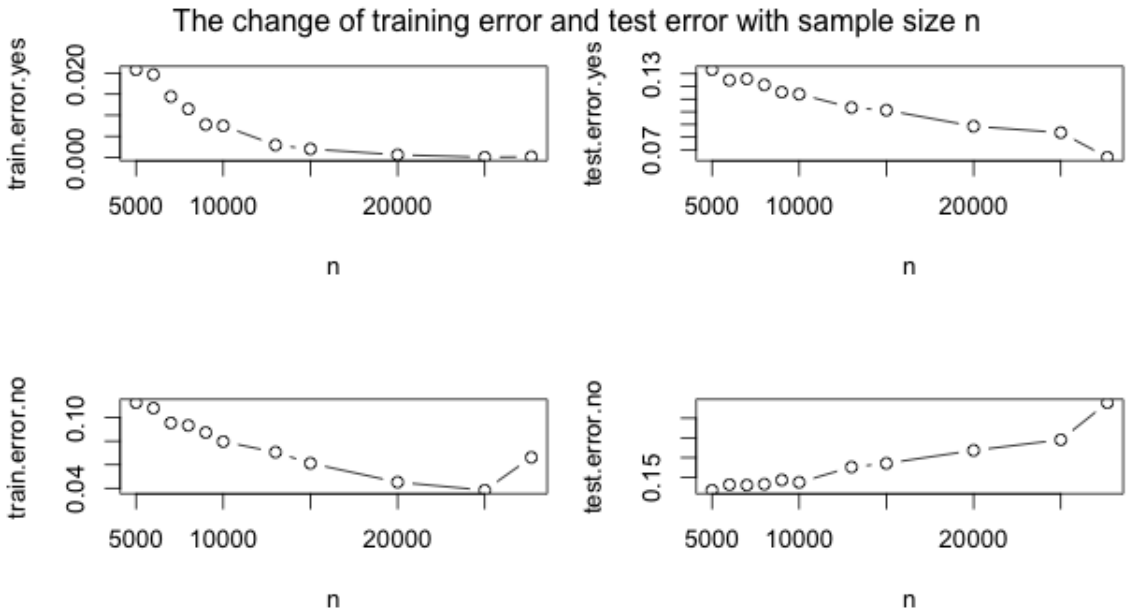
The figure shows the ROC curve of the model based on iterations equal to 400. The optimal points closest to the [0,1] is shown below as a confusion matrix.

Table 4: Confusion Matrix for Neural Network

Neural Network		Reference	
		no	yes
Prediction	no	5831	82
	yes	966	769

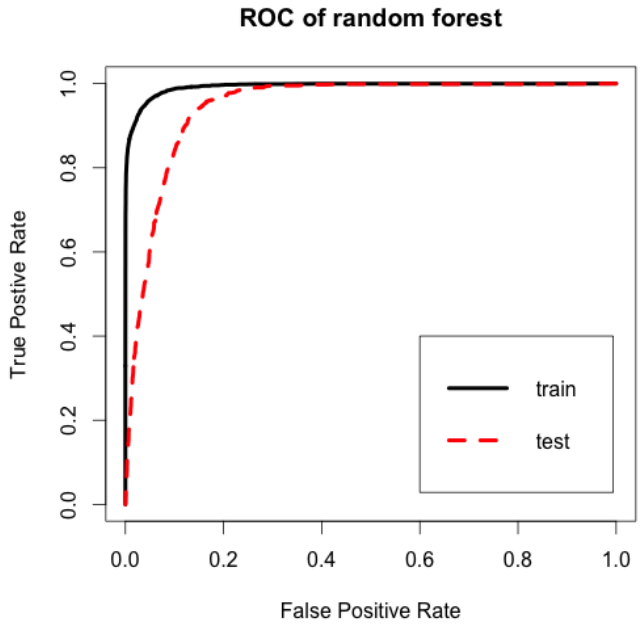
6.3 Random Forest

Since random forest classifier tends to be biased towards the majority class, we will use resampling to fix this problem. The commonly used methods are under sampling, oversampling and mix sampling. In case of the data observes the "yes" class too hard, we will use mix sampling to protect over fitting. The size of the sample should be chosen carefully, we would plot the training error rate in each class and test error rate in each class to choose n. The plots are shown below:



We can see from the plot that for the "no" class, the training error and test error increase after n achieves 20000. It is because we observe the "yes" class too hard, we ignore some information in class "no". By examining the test error in "yes" class, we can see that from 20000 to 25000, the test error does not decrease a lot, therefore, we believe the most appropriate sample size is 20000.

Next, we can plot ROC Curve and use AUC to evaluate the performance of this algorithm. The ROC and AUC are shown below:



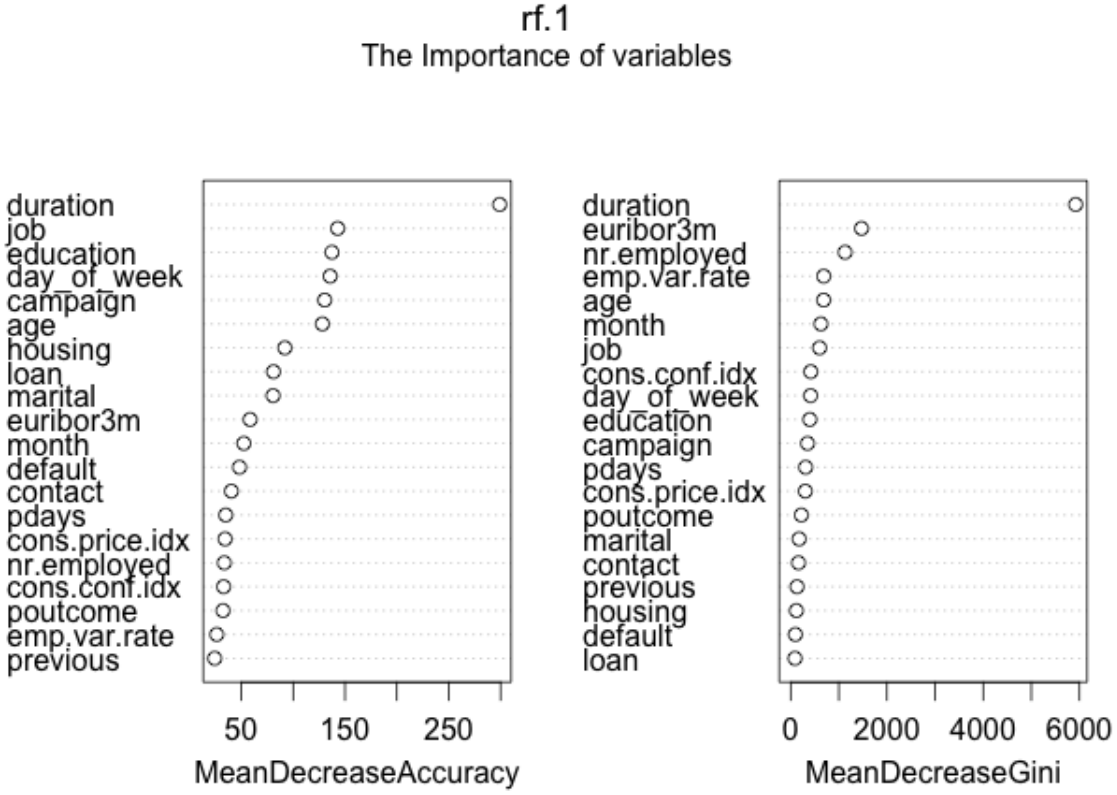
We can see that AUC for training data is 0.991, for test data is it is 0.9427.

We can also create a table to show the accuracy rate in each group.

Table 5: Confusion Matrix for Random Forest

Random Forest		Reference	
		no	yes
Prediction	no	6176	192
	yes	621	660

In addition, we have the importance of variables plot, as shown below:



Since we want to know which variable includes the largest amount of information, we should focus on the mean decrease of Gini Index, which is a measurement of statistical dispersion. We can see that the most important variable is duration, which is the last contact duration; the second is euribor3m, which is euribor 3 month rate; the third is nr.employed, which is the number of employees in the bank; the fourth is emp.var.rate, which is the employment variation rate.

6.4 k-NN

As k increases, bias increases but variance decreases. To counteract this, we use 10-fold

cross-validation. The best k is the one that minimizes the misclassification rate for the validation data set.

The IBk function in RWeka package can do classification with k-NN. This function can turn categorical variables into dummy variables and also use normalized terms so that the different scales have the same effect on the distance function.

Although k-NN gives us a more flexible decision boundary and it is easy to implement, but it also has some disadvantages. For example, it is not interpretable as we cannot interpret the effect of different predictors on our dependent variable.

When k=50, it approximately has the lowest misclassification rate for validation set.

Based on the cross-validation, the test error of our best k-NN model is 0.0982.

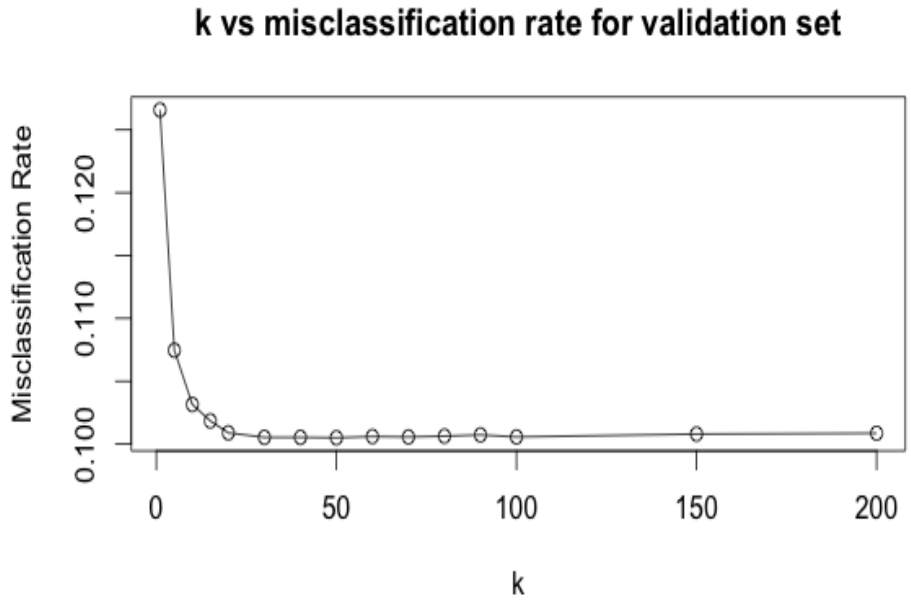
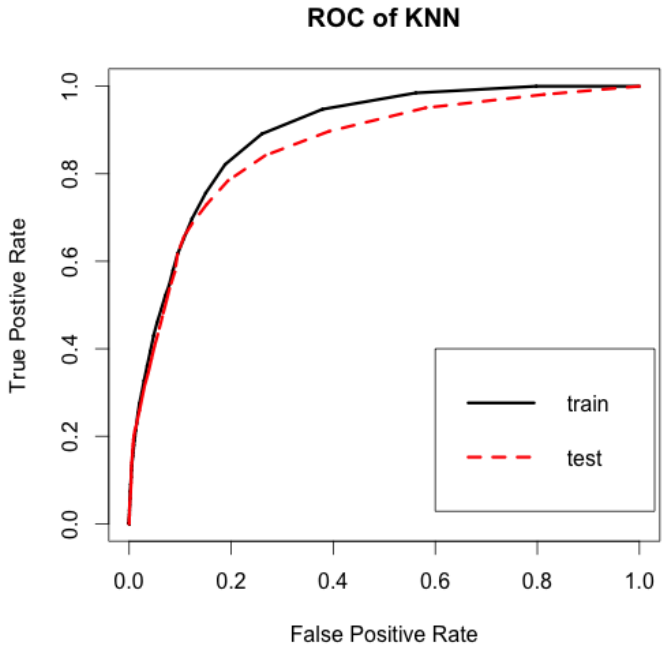


Table 6: Misclassification Rate based on k in k-NN

k	misclassification rate	k	misclassification
1	0.126585	60	0.100601
5	0.107465	70	0.100569
10	0.103151	80	0.100634
15	0.101843	90	0.100732

20	0.100896	100	0.100569
30	0.100536	150	0.100797
40	0.100536	200	0.100863
50	0.100503		



7. Conclusion

7.1 Model Comparison

In this session, we list performance measurements including AUC, ROC, test accuracy and FPR at TPR=0.99 for all the four models.

AUC and ROC Curve are designed to measure the discrimination, that is, the ability of the model to correctly classify those would and wouldn't subscribe a term deposit. Test accuracy is simply a measurement of the accuracy of the models on a new contact.

As for FPR at TPR=0.99, one of our objectives of the project is to build a model, which is able to identify almost all (i.e. 99%) of the contacts who will eventually subscribe a term deposit, while keeping a high overall prediction accuracy. With this model, we may significantly reduce the workload and costs, by researching over a much smaller group in which people are predicted to subscribe a term deposit instead of the massive population. In this case, we introduced as measurement of the power of a model the false positive rate

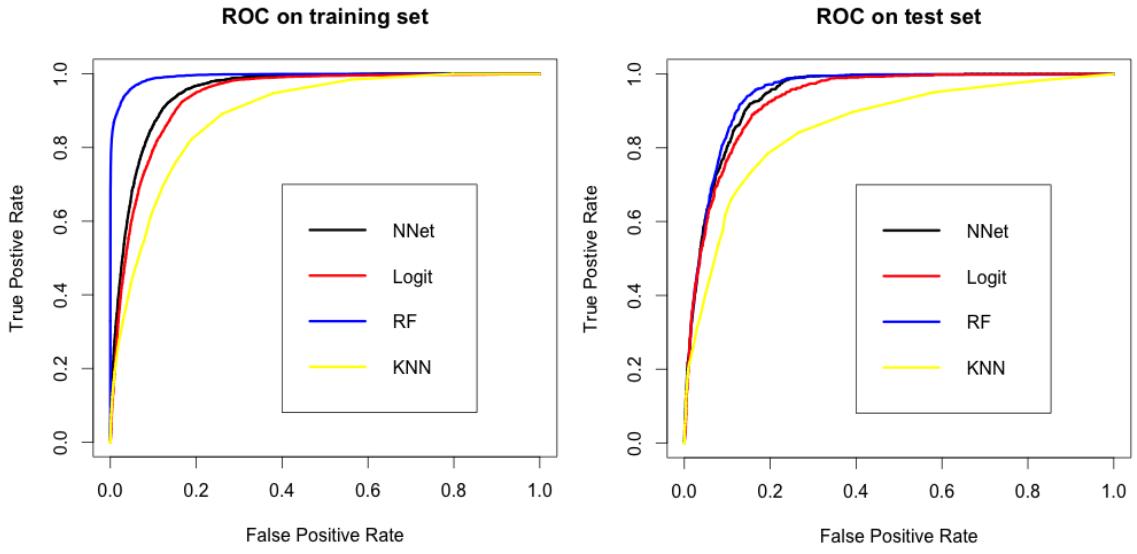
at true positive rate=0.99. That is to compare those models which are able to identify almost all of the potential customers by how much do they mistakenly classify people who won't subscribe a term deposit into class "yes".

Table 7: Comparison of the Models

	Random Forest	Neural Network	k-NN	Logistic Regression
AUC	0.9427*	0.9423	0.8531	0.9364
Test accuracy	0.9136	0.9145*	0.9018	0.9137
FPR at TPR=0.99	0.2642	0.2566*	0.3054	0.3809

(* means this algorithm is the best in the given measurement)

Neural network dominates in two measurements and ranked 2nd in AUC, so it's the most powerful model. Random forest has a similar result to neural network with slightly worse performance. Logistic regression has an acceptable performance. Despite awful result in FPR at TPR=0.99, it provides a practical way to make inferences. k-NN, as a baseline model, has the worst performance.



7.2 Recommendation Based On Model Performance

In the light of overall test accuracy and AUC, and FPR at TPR=0.99, the best model is neural network. It has the most powerful prediction ability. Next, we need to find out which factors are most important and how these factors influence customers' decision.

According to the plot for both logistic regression and random forest, we can tell that the most influential variables are duration, nr.employed, euribor3m, and emp.var.rate.

Based on signs of coefficients of variables in logistic regression, “duration” has positive effect on people saying “yes”. This is because the longer the conversations on the phone, the higher interest the customer will show to the term deposit. “nr.employed”, which is the number of employees in the bank, has positive effect for turning people to subscribe the term deposit. This can be due to the fact that the more employees the bank have, the more influential and prestigious this bank is. “euribor3m” is another important variable, which denotes the euribor 3 month rate. This indicator is based on the average interbank interest rates in Eurozone. It also has positive effect since the higher the interest rate the more willingly customer will spend their money on financial tools. Employment variation rate (emp.var.rate) has negative influence, which means the change of the employment rate will make customers less likely to subscribe a term deposit. This makes sense because the employment rate is an indicator of the macroeconomy. A stable employment rate denotes a stable economic environment in which people are more confident to make their investment.

Therefore, if banks want to improve their lead generation, what they should do is to hire more people to work for them, improve the quality of conversation on the phone and run their campaigns when interest rates are high and macroeconomic environment is stable.