DSA/ISE 5103 Intelligent Data Analytics
Assignment #5
Name: Jiwon Jeon
Date: 11/20/2016

1 Classication Performance Evaluation

Function name: myModel

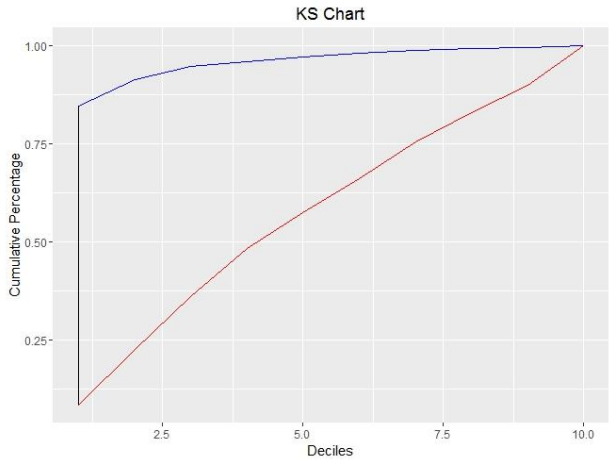
Input variables:

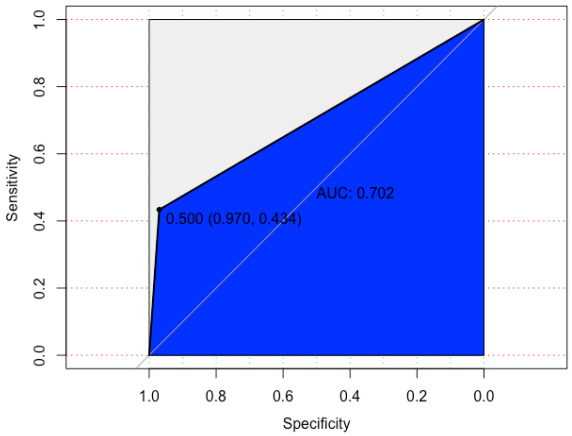
i. trueVal: actual outcome values of test data; represented as 0 or 1 to assign binary variables.
ii. predClass: predicted outcome values of the model using test data; this binary values are represented either 0 or 1 determined by predProb
iii. predProb: prediction probability; for this function, the binary predicted outcome (predClass) is denoted as 1 if predProb is greater than 0.5, and denoted as 0 if predProb is less than 0.5

Output values: the following information is generated from the function syntax.

i. Confusion matrix: a cross-tabulation of true (reference) value and predicted values. For binary data, the classes are displayed as 0 or 1.
ii. KS chart: KS indicates the degree of separation between the positive and negative distributions for the target values.
iii. ROC curve and AUC: ROC curve provides the accuracy level of prediction in terms of sensitivity and specificity.
iv. Distribution of predicted probabilities values: distribution of TP (true positive) and TN (true negative) values are displayed.
v. Concordant pairs: percent concordant, discordant and tied are calculated in the function.
vi. D statistic: the difference in average probabilities between the two predicted classes (of binary data).
vii. Lift chart: a 'lift plot' is created showing the fitness of model for the chosen samples of one class.

The following plots are sample KS chart and ROC curve, respectively:





Return values: the function also returns *accuracy* and *AUC* values for the purpose of predictability comparison between classification models.

2 A data-driven approach to predict the success of bank telemarketing: Part 1 (Review)

This paper aims to develop a decision support system by data mining approach to help bank managers strategize their telemarketing campaigns to sell long-term deposits for new customers. The effort is to predict the customer responses of subscribing before the marketers make calls by classifying them into success or not, therefore, the model will be a binary classifier.

The subject bank dataset contains 52,944 records from 2008 to 2013, which shows the influences of the global financial crisis in 2008. Data prior to July 2012 are selected as training data to build the prediction models, and the data after July 2012 were used as test data to assess the final models. The training data were further split into training and validation sets, by a random split with 2/3 and 1/3, respectively.

To build a model, the researchers performed feature engineering. 150 features were manually selected considering traditional selection criteria such as bank client, plus business oriented criteria including product and social-economic attributes. The selected features were further reduced into 22 variables in modeling phase. A semi-automatic feature selection technique was applied for this stage:

First, 150 features were grouped into 14 question sets by business intuitive knowledge of bank managers, which were later identified as 14 factors.

Then, a forward selection method was applied to each factor in sequence. A prediction model with training data was run with the first factor, then AUC was calculated over the validation data for this factor and model. The second trial included the first factor and the second factor, then AUC was computed again to compare the value with that of the first factor. If the AUC value increases, the features in the current factors were all chosen for the modeling.

To build a model, the researchers performed feature engineering. 150 features were manually selected

Four types of prediction models were employed in this research: logistic regression (LR), decision trees (DT), neural network (NN) and support vector machine (SVM). As shortly introduced in feature engineering description, the four models were evaluated using AUC and ALIFT metrics to find the best predictive performance. Results show that neural network (NN) was concluded to be the best model in terms of AUC (0.80) and ALIFT (0.67) evaluations.

A sensitivity analysis and a decision trees method also depicted key attributes which had the most impact on the model responses: the three month Euribor turned out to be the most relevant feature for the models, followed by the direction class, the bank agent experience, etc. This information would enhance the efficiency of telemarketing by focusing on several factors rather than considering all possible options.

This research placed sufficient effort on feature engineering part, however, the validation was only done by a simple holdout method. Validation process should be improved by random sampling methods for the final model to be applied to bigger datasets in future analysis. Moreover, since 2008 was under extraordinary economic situation, the records of the year may not be suitable for regular long-term deposit responses. Feature engineering process needs to treat this abnormality in the raw dataset to generalize the model application.

3 A data-driven approach to predict the success of bank telemarketing: Part 2

(a) i. Data split

The given dataset is first split into training and test datasets by stratified random sampling: 80% for training data, 20% for test data. Hence, the training set includes 32,951 data, while test set has 8,237 data. All 21 features were all used for modeling.

ii. Cross-validation strategy

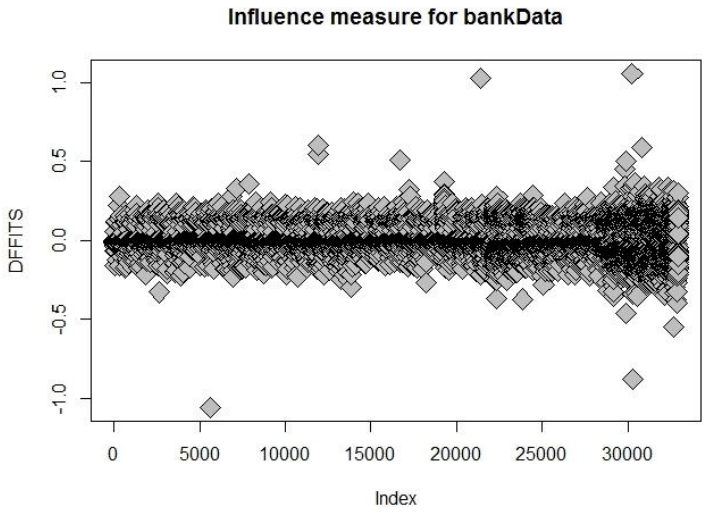
Repeated k-fold cross-validation was selected for resampling. The other validation methods such as (single) k-fold cross-validation, leave-one-out cross-validation and bootstrap method may be chosen to employ.

(b) The actual classification of test data shows 7,309 non-target customers and 928 target customers. **Logistic regression** model validated by repeated k-fold cross-validation method provides 7,632 non-target and 605 target customer classification. Sample classifications are:

```
      trueVal predClass  predProb
41167       1         1 0.5841175
41172       1         0 0.1572695
41175       1         1 0.6163378
41178       0         0 0.2525710
41183       0         1 0.5351589
41186       0         0 0.3415322
```

i. Influence

From the influence chart using diffits command shown below, it is noticed that there are several data points which highly influence the logistic regression model. These data points show significant deviations from the zero unity line and the majority is found around 30,000-th values.

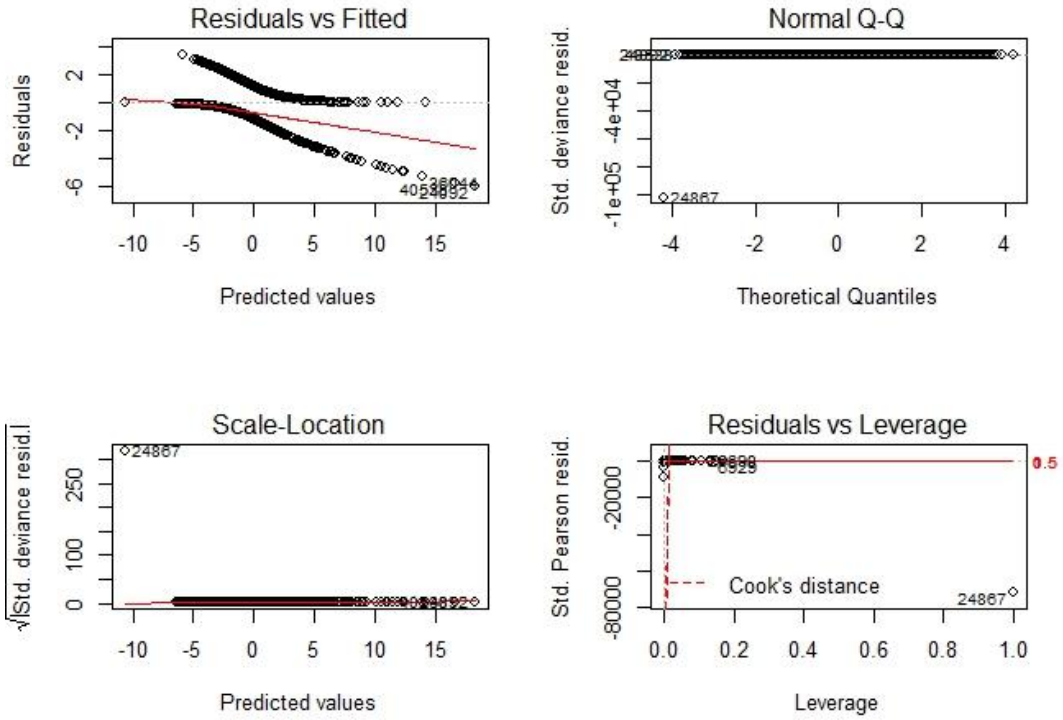


ii. Variance inflation

vif and alias functions show that there exist multicollinearity in the logistic regression model. By implementing stepwise AIC method, the formula of 21 features used in the model can be rearranged with reduced number of features. The proposed formula is:

```
formula = target ~ default + contact + month + duration + pdays + poutcome +
                   emp.var.rate + cons.price.idx + euribor3m
```

iii. Residuals

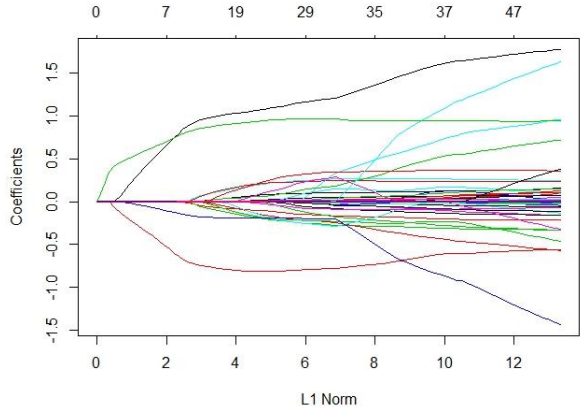


The residuals of predicted values from logistic regression shows a little considerable deviation from zero line. There must exist some significant outliers to be carefully reviewed and removed from dataset if necessary.

(c) i. **Elastic net** model provides 7,637 non-target customer and 600 target customer classification. Sample classifications are:

	trueVal	predClass	predProb
41167	1	1	0.5981238
41172	1	0	0.1653751
41175	1	1	0.6181254
41178	0	0	0.2484586
41183	0	1	0.5407045
41186	0	0	0.3415690

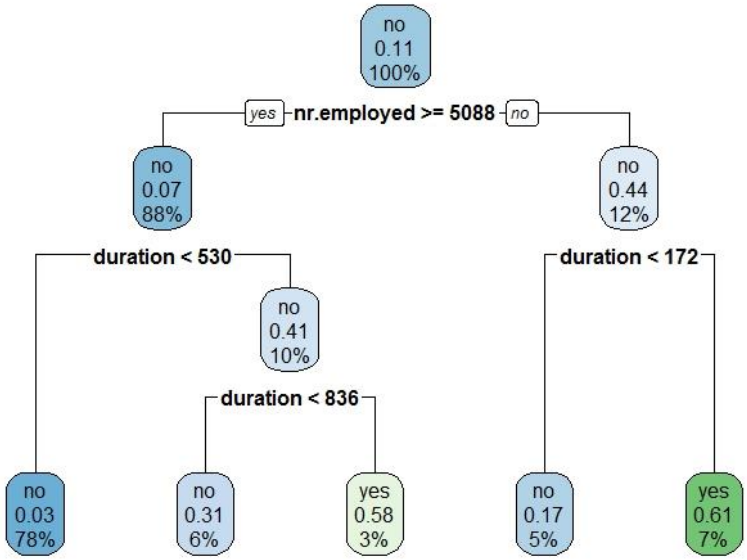
The elastic net plot has the following converging shape:



ii. **Decision tree** model provides 7,368 non-target and 869 target customer classification. Sample classifications are:

```
trueVal predClass  predProb
41167       1        1 0.6068826
41172       1        0 0.1671105
41175       1        1 0.6068826
41178       0        0 0.1671105
41183       0        0 0.1671105
41186       0        1 0.6068826
```

Classification scheme is shown as below:



iii. **Random forest** model provides 7,551 non-target and 686 target customer classification. Sample classifications are:

```
     trueVal predClass
8232       1         0
8233       1         0
8234       1         1
8235       0         0
8236       0         0
8237       0         0
```

(d) The function 'myModel' was used to evaluate the predictability of each model: logistic regression, elastic net, decision tree and random forest. The accuracy and AUC values for each model are as below:

```
                Model   accuracy       auc
Logistic Regression 0.9095544  0.6978502
Elastic Net         0.9094330  0.6959003
Decision Tree       0.9117397 0.7654058
Random Forest       0.9125895 0.7428358
```

It is generally expected that random forest method offers the best accuracy among the suggested modeling techniques for classification and/or prediction, and it is confirmed for the bank data using 'myModel' evaluation. However, AUC value of random forest method is slightly lower than that of decision tree, which requires additional investigation to find better fitness between the two models.

Moro et al. (2014) employed four modeling strategies including logistic regression, decision tree, random forest and neural network, and concluded that neural network method provides the best prediction in terms of AUC. In order to compare the result from 'myModel' and that of Moro et al. neural network classification as well as SVM will be conducted in Problem 4.

(e) From the four classification method, random forest techniques is most recommended as it shows the highest accuracy. However, this method has lower value of AUC compared to decision tree, hence, other techniques including SVM and neural network will be conducted in Problem 4 for further comparison.

4 Extra-Credit (SVM and Neural Network)

(a) **SVM model** provides 7,652 non-target customer and 585 target customer classification. Sample classifications are:

	trueVal	predClass
8232	1	1
8233	1	0
8234	1	1
8235	0	0
8236	0	0
8237	0	0

(b) **Neural network** model provides 7,919 non-target customer and 318 target customer classification. Sample classifications are:

	trueVal	predClass
8232	1	0
8233	1	0
8234	1	1
8235	0	0
8236	0	1
8237	0	0

The overall evaluation results are compared as below:

	accuracy	auc
Logistic Regression	0.9095544	0.6978502
Elastic Net	0.9094330	0.6959003
Decision Tree	0.9117397	0.7654058
Random Forest	0.9125895	0.7428358
SVM	0.9141678	0.7187947
Neural Net	0.9095544	0.7486517

Interestingly, the accuracy of neural network (NN) model does not have the highest value among the six tested methods. However, AUC of NN shows the best value of approx. 0.75, hence, NN is most recommended to use for the subject bank data. Moro et al. (2014) obtained 0.8 of AUC from NN modeling. The difference between the result of 'myModel' and that of Moro et al. can be explained in terms of feature engineering. In this assignment, the 21 features of training data were all used for model construction for simplicity in coding although a formula with reduced features were suggested during data process. Moreover, the selection of training/test datasets as well as the validation strategy may have affected the result.