Sanjiv Shah
OU ID: 113180542

DSA 5103
Fall 2016

Homework 5

Problem 1:

Please See attached R script for the function

Problem 2: Summary: A Data-Driven Approach to Predict the Success of Bank Telemarketing

(This is not a review of the paper as instructed. I don't believe I am qualified to review (like a peer reviewer would do, e.g., comment on developmental sequence for the study or reasoning for choosing a certain method for classification, etc.) the paper. However, this is a brief summary of the study detailed in the paper and lessons that I learned (takeaways) from it)

This paper, is about development of a specialized and intelligent decision support system (DSS) that can predict the result of a phone call to sell long term deposits (commonly known as CD's in the U.S.) using a Data Mining approach. The study compared four (4) data mining methods – namely linear regression, decision trees, neural networks and support vector machines using a realistic rolling window evaluation and two classification matrices.

The study utilized data collected by the client bank over a 6 year period (2008 – 2013) consisting of over 50,000 phone contacts made by the bank to its existing customers. The dataset is very unbalanced in that only approximately 12% of the records indicated "success" in selling the long term CD's to the bank's existing clients. The final dataset also included external macro-economic data from official sources resulting in a total of 150 attributes. As number of attributes directly affects complexity, performance (time to compute) and interpretability of the results, a semi-automatic two-step feature selection process was developed and tested.

Feature Selection: In the first step, domain expertise of the client bank was used to identify key criteria and its responses from the existing database that narrowed the attributes by approximately 54%. These reduced set was then processed using an adapted forward selection method. Four different threshold values were established. If from a group of similar attributes at least one individually tested attributes exceeded an AUC of first threshold value and all similar attributes exceeded an AUC of second threshold value it was considered that business hypothesis was confirmed. When a hypothesis was confirmed only the attribute that achieved AUC that exceeded the AUC for all similar attributes together or combination of attributes that exceeded fourth threshold.

For each group of attributes, they were tested by adding them to previous analysis (AUS measurement) and if it did not improve upon previous result, it was discarded. Otherwise, it was added to the final group of attributes. The order was kept the same as order of importance suggested by domain experts.

Modeling: Four binary classification models – logistic regression, decision trees, neural networks and support vector machines were tested for their suitability for the study. Each method was tested using part of the training dataset and compared to the remaining set of training dataset (training (2/3) and validation (1/3) datasets) AUC and ALIFT were the criteria used for comparing results of the each model. In order to assess validity of the feature selection process, each model was also tested using complete set of features and forward selection method.

Results: During modeling phase of the study, neural network produced the best outcomes and was chosen as the final model. However, it must be noted that that when this model was compared to the rolling window phase (test data – that was more recent than training data), its performance was

noticeably inferior (AUC from 0.929 vs 0.794 and ALIFT from 0.878 to 0.672). The authors attributed most of this performance degradation to the difference in time periods in the two datasets.

Takeaways: (What I conclude from the study)

- Feature selection is very important not only prior to modeling but also must be validated during modeling phase.
- Predictive models must be recalibrated when underlying assumptions (attributes and their influence) may have changed.
- Domain expertise must be utilized whenever possible.
- Knowledge extraction techniques applied after modeling and its implementation may produce additional insights that may be easier to interpret than model.

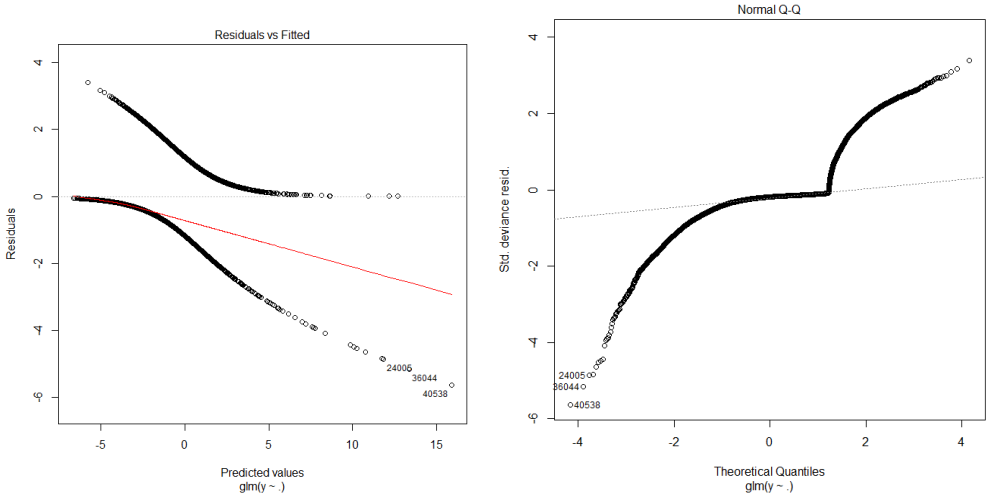
Q3. A.

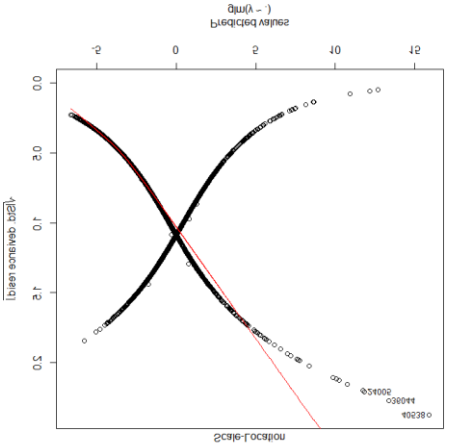
As the paper suggests, the data were collected between 2008 and 2013. As we are aware, these were very turbulent times in financial industry especially in Europe. Several countries in the Eurozone including Portugal were near financial collapse following the financial crisis that originated in the United States and affected European banks as well as some European countries significantly. People's financial perspectives also changed during that time. The data include the period during and following these turbulent times. However, the data are not provided as time-series data. So, it is not possible to analyze the data as time-series data. As the paper indicates, the results (success rate) after implementing the study's recommendations were better than predicted. This is a clear indication of customers' changed financial behavior a few years after the financial crisis.

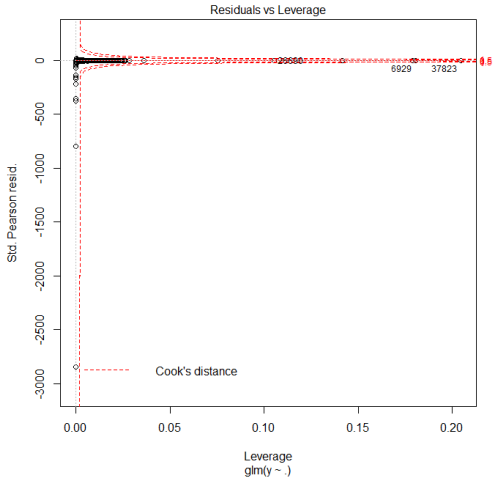
The data were divided in 70% training data and 30% testing data at random.

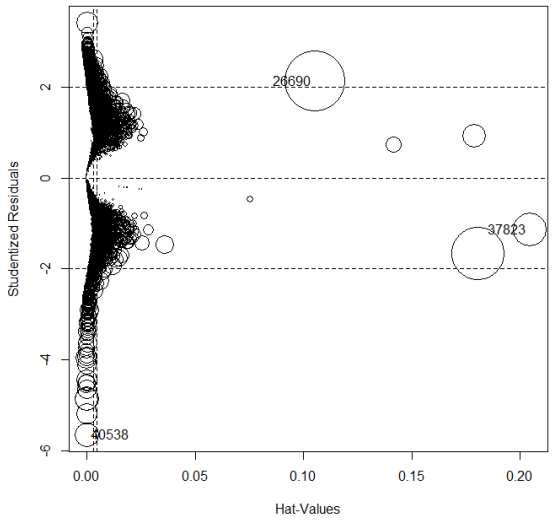
3(b): Logistic Regression:

Please see the attached R-script for complete procedure for logistic regression and evaluation metrics. Below are some of the charts for logistic regression.









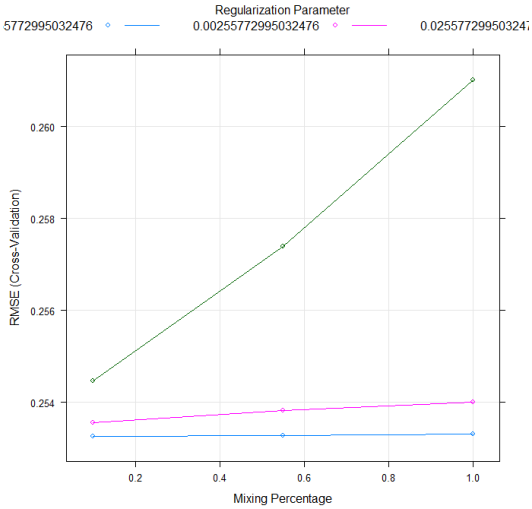
	GVIF	Df	GVIF^(1/(2*Df))
age	2.244816	1	1.498271
job	5.770588	10	1.091594
marital	1.438309	2	1.095123
education	3.220696	6	1.102374
housing	1.012485	1	1.006223
loan	1.005578	1	1.002785
contact	2.325759	1	1.525044
month	65.109125	9	1.261124
day_of_week	1.065165	4	1.007922
duration	1.233659	1	1.110702
campaign	1.055827	1	1.027534
pdays	11.079043	1	3.328520
previous	4.521459	1	2.126372
poutcome	24.753081	2	2.230526
emp.var.rate	142.316064	1	11.929630
cons.price.idx	67.309272	1	8.204223
cons.conf.idx	5.378113	1	2.319076
euribor3m	136.382762	1	11.678303

```
nr.employed    169.829416  1       13.031862
```

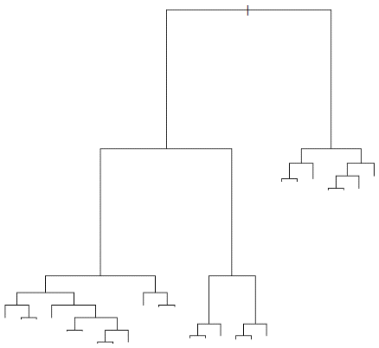
3(c): Elastic Net, Decision Tree, Random Forest and Boosted Tree methods. Please see accompanying R-script.

3(d): The C_P_E function (Classification Performance Evaluation) was applied to each of the method in 3(c).

Elastic Net:

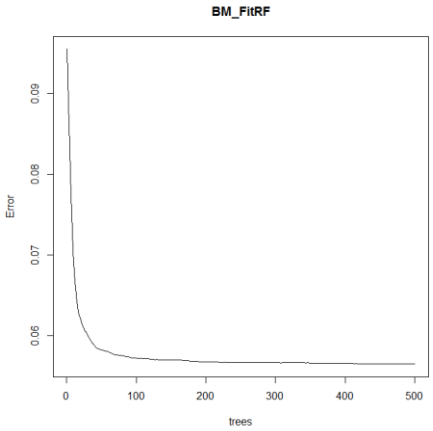


Decision Tree:

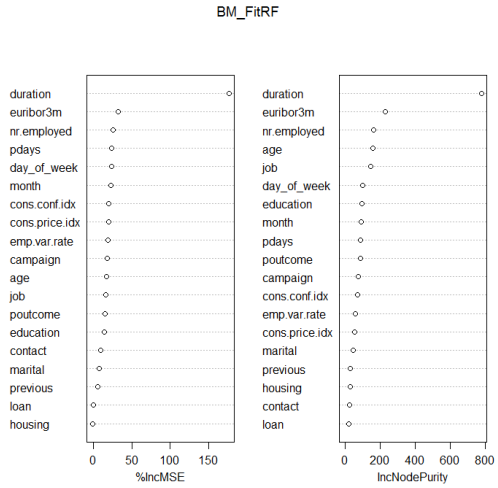


As shown in adjoining figure, only 1 pruning resulted in a manageable tree.

Random Forest



As shown in the adjoining figure, the error is minimized after about 200 trees. The final script was modified with ntree = 200 to run faster.



3(d): Comparison:

Based on the function call, all the evaluation metrics included in the function call can be compared. However for the sake of time and space only the AUC results are shown below in the summary table. Based on these results, the simplest and fastest method – logistic regression performed the best.

(This differs with the paper (the paper selected neural net method) primarily because the authors used more advanced package (rminer) that actually runs multiple times (20 in this case). I tried to use the same rminer package in my analysis but was unsuccessful in doing so. In my case, I was only able to run one method successfully out of 3 methods I tried. It was frustrating but, rminer being a fairly new package, not a lot of information is out there about its inner workings. However, I still think rminer is a better package for data mining tasks and I intend to educate myself in using it.)

Recommendation:

Due to limited testing, I can't recommend any other method than what authors found most advantageous. However, one suggestion I would make is that whichever model is chosen, it must be rerun periodically to keep it updated with most recent data. I recommend it be re-evaluated (in terms of its predictive accuracy and concurrency) every 6 months and old data from oldest 6 months be dropped off.

Summary

Method	AUC
Logistic Regression	0.9347
Elastic Net	0.9329
Decision Tree	0.9334
Support Vector Machine	0.8690

Problem 4: Support Vector Machine

The support vector machine script is included in accompanying R-script. Unfortunately, R will not generate a plot for me (no error but, no plot either).

Here is a histogram of predicted values