Roy Chan
chanroyc@sfu.ca
301202770

CMPT 459 Fall 2017
DataMining
Martin Ester
TA: Zhilin Zhang

Programming Assignment 2

1. Check pa2.r

2. Table of Missing values per attribute in the training and test dataset.

	Training Dataset	Test Dataset
pclass	0	0
survived	0	0
name	0	0
sex	0	0
age	217	46
sibsp	0	0
parch	0	0
ticket	0	0
fare	1	0
cabin	0	0
embarked	0	0
boat	0	0
body	945	243
home.dest	0	0

Roy Chan
chanroyc@sfu.ca
301202770

3. Using only past data to predict the future with the assumption that we want to predict the survival of a passenger at the time of the accident, i.e. when the Titanic hit the iceberg, the attributes that we would use as features are as follows:

   pclass: Passenger class - Irrelevant
   survived: Survival - Relevant
   name: Name - Irrelevant
   sex: Sex - Relevant
   age: Age - Relevant
   sibsp: Number of Siblings/Spouses Aboard - Relevant
   parch: Number of Parents/Children Abaord - Relevant
   ticket: Ticket Number - Irrelvant
   fare: Passenger Fare - Irrelevant
   cabin: Cabin - Relevant
   embarked: Point of Embarkation - Relevant
   boat: Lifeboat the passenger took - Relevant
   body: Body Identification Number - Irrelevant
   home.dest : Home/Destination - Irrelevant

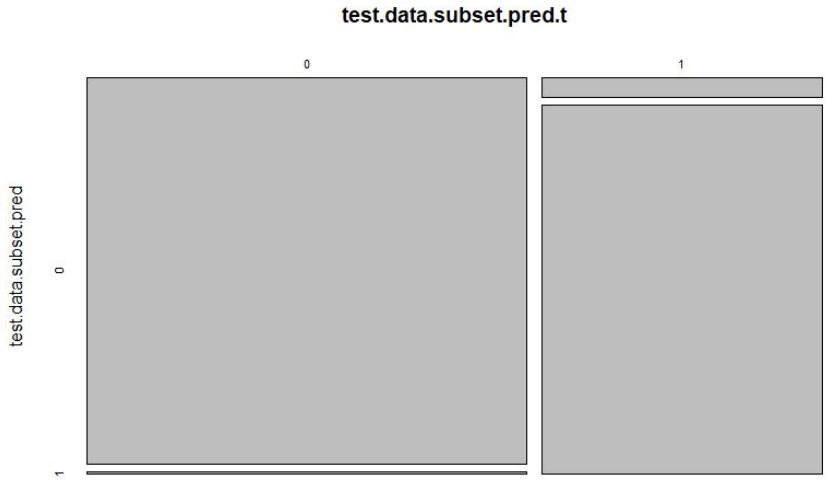
   Passenger class, name, ticket number, passenger fare, body identification number, home/destination are all attributes which are irrelevant to the accident. The rest of the attributes are chosen as features (survival, sex, age, sibsp, parch, cabin, embarked, boat).

4. The typical approach is the replace the missing values with the average, the median, or the mode of the existing one. In my plan, I will be using the average to replace the missing values. Check pa2.r for preprocessing procedures.

5. After learning a logistic regression model from the training data in pa2.r, the significance of the attributes are as follows (most significant to least significant): boat, sex male, parch, sibsp, and age. Sex male and several boats being the the only attributes with p-value < 0.05

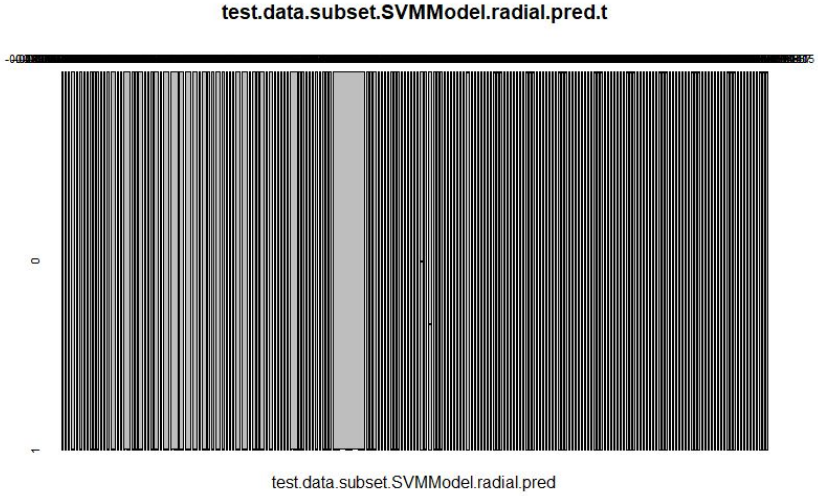
Roy Chan
chanroyc@sfu.ca
301202770

6. After applying the logistic regression model to predict the class labels of the test data, the following is a plot of the confusion matrix. The accuracy is 0.9770992



7. I was unable to figure out how to plot ROC and find the AUC.
8. After learning the SVM models from the training data using linear and radial kernels, the tune() function was used to obtain the best parameters for linear and radial kernels. The best parameters for the linear kernel is cost = 32 and gamma = 0.5. The best parameters for the radial kernel is cost = 2 and gamma = 0.5.

9. The radial kernel from my results is the best SVM model to predict the class labels of the test data, based on the results from training data. The following is the confusion matrix with the best SVM model applied to predict the class labels of the test data. I believe that there are some mistakes in my code that lead to the following confusion matrix. It does not seem to be correct.



Roy Chan
chanroyc@sfu.ca
301202770

10. I was unable to figure out how to plot ROC and find the AUC.