Home Work # 7. AMS 380

Name:_____SBU ID:_____

Dear all, the homework is due on Thursday, Nov 4, 2021, at 11:59 PM. Please submit your homework to the Blackboard in a pdf or word (.doc) document. Rmarkdown is highly recommended.

Please include (1) R code; (2) Output from R; (3) Answers to all the questions asked.

CART with the Titanic Data – Classification Task

The Titanic.csv data we will use for our homework is taken from the Kaggle competition site (https://www.kaggle.com/c/titanic) where it was called the train.csv. **We will treat this dataset as our entire data** because we do not know the survival status in the Kaggle test.csv data. Our Titanic data has 891 passengers and 12 variables:

- *PassengerId*: Passenger ID: 1- 891
- *Survived*: A binary variable indicating whether the passenger survived or not (0 = No; 1 = Yes); this is our response variable
- *Pclass*: Passenger class (1 = 1st; 2 = 2nd; 3 = 3rd)
- *Name*: A field rich in information as it contains title and family names
- *Sex*: male/female
- *Age*: Age, asignificant portion of values aremissing
- *SibSp*: Number of siblings/spouses aboard
- *Parch*: Number of parents/children aboard
- *Ticket*: Ticket number.
- *Fare*: Passenger fare (British Pound).
- *Cabin*: Cabin number
- *Embarked*: Port of embarkation (C = *Cherbourg*; Q = *Queenstown*; S = *Southampton*)

First, one must clean the data and decide which variables to exclude from our analysis. My recommendation is that we exclude *Name*, *Ticket*, and *Cabin* in the ensuing analysis. Next, please note that *Age* has many missing values - my suggestion is to delete those with missing values. Now after the data cleaning step, your task is to split the data randomly into training (80%) and testing (20%), first build the full tree and then establish an optimal model with pruning and 10-fold cross validation using the training data, and then use that model to predict whether each passenger in the testing data survived or not. Please use *rpart* to build the classification tree and *rattle* to plot the tree.

Please review the following websites for related methods and concepts:
http://www.sthda.com/english/articles/35-statistical-machine-learning-essentials/141-cart-model-decision-tree-essentials/
https://trevorstephens.com/kaggle-titanic-tutorial/r-part-3-decision-trees/

1. For the entire dataset, please perform the data cleaning as instructed before; namely, exclude

the variables *Name*, *Ticket*, and *Cabin* and delete missing values in the variable *Age*. Please report how many passengers are left after this step. Then please use the random seed 123 to divide the cleaned data into 80% training and 20% testing.

2. Please first build a fully grown tree using the training data, and draw the tree plot using *rattle*. Next please use this tree to predict the survival of passengers in the testing data. Please compute the Confusion matrix and report the sensitivity, specificity and the overall accuracy for the testing data.

3. To make the tree more robust, we will prune the fully grown tree using the training data with 10-fold cross-validation. Please (1) show the complexity plot, (2) report the best CP value, and (3) draw the pruned tree using *rattle*.

4. Please use this optimal pruned tree to predict the survival of passengers in the testing data. Please compute the Confusion matrix and report the sensitivity, specificity and the overall accuracy for the testing data.