# Final Exam AMS 597

# Name:\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_SBU ID:\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

# The final exam is due by 1:45pm. Please email your completed exam to me at: [wei.zhu@stonybrook.edu](mailto:wei.zhu@stonybrook.edu)

# Please include (1) R code

# (2) Output from R

# (3) Answers to all the questions asked

# Please keep yourself on Zoom video until you have emailed your solutions.

# Please keep your power plugged in.

# Please work on the exam independently on your own.

# There are two parts to this exam: Part I & Part II – all in this single document.

#### Part I. General Linear Model with the Math Performance Data

The accompanying **math.csv** file contains data including student math scores in three school periods, demographic, social and school related features. Each case is a student, and the variables are:

1 school - student's school (binary: each student is from one of the two schools, 'GP' or 'MS')

2 sex - student's sex (binary: 'F' - female or 'M' - male)

3 age - student's age (numeric: from 15 to 22)

4 address - student's home address type (binary: 'U' - urban or 'R' - rural)

5 famsize - family size (binary: 'LE3' - less or equal to 3 or 'GT3' - greater than 3)

6 Pstatus - parent's cohabitation status (binary: 'T' - living together or 'A' - apart)

7 Medu - mother's education (numeric: 0 - none, 1 - primary education (4th grade), 2 - 5th to 9th grade, 3 - secondary education or 4 - higher education)

8 Fedu - father's education (numeric: 0 - none, 1 - primary education (4th grade), 2 - 5th to 9th grade, 3 - secondary education or 4 - higher education)

9 Mjob - mother's job (nominal: 'teacher', 'health' care related, civil 'services' (e.g. administrative or police), 'at\_home' or 'other')

10 Fjob - father's job (nominal: 'teacher', 'health' care related, civil 'services' (e.g. administrative or police), 'at\_home' or 'other')

11 reason - reason to choose this school (nominal: close to 'home', school 'reputation', 'course' preference or 'other')

12 guardian - student's guardian (nominal: 'mother', 'father' or 'other')

13 traveltime - home to school travel time (numeric: 1 - <15 min., 2 - 15 to 30 min., 3 - 30 min. to 1 hour, or 4 - >1 hour)

14 studytime - weekly study time (numeric: 1 - <2 hours, 2 - 2 to 5 hours, 3 - 5 to 10 hours, or 4 - >10 hours)

15 failures - number of past class failures (numeric: n if 1<=n<3, else 4)

16 schoolsup - extra educational support (binary: yes or no)

17 famsup - family educational support (binary: yes or no)

18 paid - extra paid classes within the course subject (Math or Portuguese) (binary: yes or no)

19 activities - extra-curricular activities (binary: yes or no)

20 nursery - attended nursery school (binary: yes or no)

21 higher - wants to take higher education (binary: yes or no)

22 internet - Internet access at home (binary: yes or no)

23 romantic - with a romantic relationship (binary: yes or no)

24 famrel - quality of family relationships (numeric: from 1 - very bad to 5 - excellent)

25 freetime - free time after school (numeric: from 1 - very low to 5 - very high)

26 goout - going out with friends (numeric: from 1 - very low to 5 - very high)

27 Dalc - workday alcohol consumption (numeric: from 1 - very low to 5 - very high)

28 Walc - weekend alcohol consumption (numeric: from 1 - very low to 5 - very high)

29 health - current health status (numeric: from 1 - very bad to 5 - very good)

30 absences - number of school absences (numeric: from 0 to 93)

# each student’s Math exam scores in three periods:

31 G1 - first period grade (numeric: from 0 to 20)

32 G2 - second period grade (numeric: from 0 to 20)

33 G3 – third period grade (numeric: from 0 to 20)

Our goal is to use the first 30 variables as predictors for the last variable, **G3** (3rd period math performance).

Please note that we will **not** include G1 and G2 in our analysis.

1. Please use the random seed 123 to divide the data into 75% training and 25% testing.

1. Please first find the best Ridge Regression model using the training data. Please (a) find the best **λ** valuethrough cross-validation and display this value; (b) display the coefficients of the fitted model; and (c) make prediction on the testing data, and report the RMSE and the Coefficient of Determination .
2. Please first find the best LASSO model using the training data. Please (a) find the best **λ** valuethrough cross-validation and display this value; (b) display the coefficients of the fitted model; and (c) make prediction on the testing data, and report the RMSE and the Coefficient of Determination .
3. Please first find the best Elastic Net model using the training data. Please (a) find the best **tuning parameter** valuesthrough cross-validation and display these values; (b) display the coefficients of the fitted model; and (c) make prediction on the testing data, and report the RMSE and the Coefficient of Determination .
4. Please find the best model using the stepwise variable selection method using the training data and the BIC criterion. Please (a) display the coefficients of the fitted model; and (b) make prediction on the testing data, and report the RMSE and the Coefficient of Determination .
5. Please find the best model using the best subset variable selection method using the training data. Please (a) display the coefficients of the fitted model; and (b) make prediction on the testing data, and report the RMSE and the Coefficient of Determination .
6. Which model selection method among the 5 we have used above is the best? Please discuss any modifications you can do to further improve your model(s).

#### Part II. Classification Task with the Titanic Data

The **Titanic3.csv** data has 1309 passengers and 11 variables:

* *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, a significant portion of values are missing
* *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 (75%) and testing (25%). Then you shall first build various classification models to predict passenger survival using the training data, and subsequently use that model to predict whether each passenger in the testing data survived or not.

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 75% training and 25% testing.

1. (a) Please first build the best random forest to predict passenger survival using the training data. Please compute the Confusion matrix and report the sensitivity, specificity and the overall accuracy using the out of bag (OOB) samples. (b) Next please use this random forest 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.
2. Now we will build the CART model. (a) Please first build a fully grown tree using the training data, and draw the tree plot using *rattle*. (b) To make the tree more robust, we will prune the fully grown tree using the training data with 10-fold cross-validation. Please draw the pruned tree using *rattle*. (c) Finally, please use this optimal pruned tree to predict survival for each passenger in the testing data. Please compute the Confusion matrix and report the sensitivity, specificity and the overall accuracy for the testing data.
3. The third and last classification model we will build is the logistic regression model. (a) For the training data, please find a logistic regression model that best predicts passenger survival using the stepwise variable selection method, and using the BIC. Please report the final model and the associated BIC value. (b) Please use this model to predict passenger survival in the testing data. Please compute the Confusion matrix using the threshold of 0.5 and report the sensitivity, specificity and the overall accuracy for the testing data.
4. (a) Please obtain the predicted class results for each test subject from each classifier [using the Predict () function in R]; (b) Please then combine these data sets from different classifiers, using the majority vote method -- and then compute the sensitivity, specificity and overall accuracy.
5. limitation
6. fully grown tree