CSCI E-106: Final Exam - Fall 2020

Hakan Gogtas

12/5/2020

Instructions

Open book and open notes exam (textbooks (print or pdf), lecture slides, notes, practice exam, homework solutions, and TA slides, including all Rmd's*).

You are allowed to use RStudio Cloud (https://rstudio.cloud.) , Microsoft Word, Power Point and PDF reader, and canvas on your laptop.

Proctorio is required to start this exam. If you are prompted for an access code, you must Configure Proctorio on your machine.

Please read the list of recording and restrictions provided by Proctorio carefully before taking the exam

Please pay attention any timing and technical warnings that popped up your screen

The exam will be available from Monday December 14th at 8 am EST through Tuesday December 15th at 8:00pm EST. Once you start the exam, you have to complete the exam in 3 hours or by Tuesday December 15th at 8:00pm EST, whichever comes first.

In order to receive full credit, please provide full explanations and calculations for each questions

Make sure that you are familiar with the procedures for troubleshooting exam issues Preview the document Make sure you submit both .Rmd and (knitted) pdf or html files.

You need to have a camera on your laptop.

Question 1

Use the "Final Exam Fall 2020 Question 1.csv" data set. Company executives want to be able to predict market share of their product (Y) based on merchandise price (X_1), the gross Nielsen rating points (X_2), an index of the amount of advertising exposure that the product received; the presence or absence of a wholesale pricing discount ($X_3 = 1$ if discount present: otherwise $X_3 = 0$); the presence or absence of a package promotion during the period ($X_4 = 1$ if promotion present: otherwise $X_4 = 0$). (10 points, 5 points each)

a-) Use X_1, X_2, X_3, and X_4 to predict Y. Develop a best subset model for predicting Y. Justify your choice of model. Ensure that all variables are significant in your final model, use $\alpha = 0.05$.

b-) Check all the model assumptions (using the residual plots). And test auto correlation, use $\alpha = 0.01$. If auto correlation is present, revise the model to eliminate the correlation.

Question 2

Use the "Final Exam Fall 2020 Question 2.csv" data set. Residential sales that occurred during the year 2002 were available from a city in the midwest. Data on 522 arms-length transactions include following variables:

sales price,

finished square feet,

number of bedrooms,

number of bathrooms,

air conditioning (1 if yes; 0 otherwise)

garage size

pool (1 if yes; 0 otherwise),

year built,

quality (3 different qualities),

style (there are 7 different styles, 1 to 7),

lot size,

year built.

(50 points)

a-) Use "set.seed(300)" to create development sample (70% of the data) and hold-out sample (30% of the data). (5 points)

b-) Use all variables to predict the sales price on the development sample. Transform the sales price and refit the model. Explain why a transformation would be necessary in this case.(5 points)

c-) Use stepwise (both ways) model selection to select the best model for predicting transformed sales price on the development sample.**Ensure that all variables are significant, use $\alpha = 0.05$. Justify your choice of model. Check the appropriate model assumptions visually from the graphs.** (5 points)

d-) Use Regression Tree approach to predict the sales price on the development sample (use 3 digits). (5 points)

e-) Use Neural Network approach to predict the sales price on the development sample. (10 points)

f-) Use elastic net approach to predict the sales price on the development sample. (10 points)

g-) Score all models on hold-out sample. Compare the SSEs, R^2 and select the best model. (10 points)

Question 3

Use the "Final Exam Fall 2020 Question 2.csv" data set in Question 2. Create a binary response variable Y, called high quality, by letting Y=1 if quality variable equals to 1 otherwise 0. (20 points)

a-) Fit a model to predict Y, ensure that all variables are significant by using the backward elimination to build your model. Use $\alpha = 0.05$ and ensure that all variables are significant.(10 points)

b-) Conduct the Hosmer-Lemeshow goodness of fit test for the appropriateness of the logistic regression function by forming five groups. State the alternatives, decision rule, and conclusion. (5points)

c-) What is the estimated probability of each house from the data given below having good quality? (Output should have 3 probabilities). Copy and paste the code below on r (5points)

test.dat<-data.frame(matrix(c(559000,2791,3,4,1,3,0,1992,1,30595,0,535000,3381,5,4,1,3,0,1988,7,23172,
0,525000,3459,5,4,1,2,0,1978,5,35351,0),byrow=T,nrow=3,ncol=11))

dimnames(test.dat)[[2]]<-c("Sales.price","Finished.square.feet","Number.of.bedrooms","Number.of.bathroom",
"Air.conditioning","Garage.size","Pool","Year.built","Style","Lot.size","Adjacent.to.highway")

Question 4

Use ships data sets in the MASS package. Copy and paste and following code “library(MASS);data(ships,package =”MASS“)”. (10 points, 5 points each)

Data contains the number of wave damage incidents and aggregate months of service for different types of ships broken down by year of construction and period of operation.

a-) Fit linear models with the number of damage incidents as the response and all other variables are predictors. Is the model significant?

b-) Predict the number of incidents for the following data point. ("copy and paste onto R)

test.data<-data.frame(type=“B”,year=60,period=60,service=44882)

Question 5

Refer to question 2-C and your final model . There could be potential outliers in the model. Build a robust regression model (use the same variables) and compare your regression model and outputs with the model you built in question 2-C. (10 Points)