Practice Final Exam

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.

Problem 1

Use the question1 data, fit the regression model on Y by using all the variables (X6 and X7 are categorical variables). Create development sample (70% of the data) and hold-out sample (30% of the data). Perform statistical tests, use graphs or calculate the measures (e.g. VIF, Leverage Points, Cook's Distance) for questions below. Use the development sample for part a to d. Use the hold-out sample for part e.

a-) Is the model significant? Is there a Multicollinearity in the data? Are the errors Normally distributed with constant variance?

b-) Are there any influential or outlier observations?

c-) Can X5, X6, and X7 be dropped from the model? Perform the statistical test and state your final model.

d-) Develop an alternative model by using the Regression Tree and compare the performance against the regression model built in part a.

f-) Score the model on hold-out sample, and recalibrate the model on the holdout sample. Compare the results against the final model derived part c.

Problem 2

Use the question2 data set to answer this question. We are interested in predicting (Y) the number of customers who complained about the service.

a-) Build a model to predict the number of complaints, perform the statistical tests that shows that model is significant

b-) Find the predicted number complaints given the independent variables below and predict 95% confidence interval

X1=606 X2=41393 X3=3 X4=3.04 X5=6.32

Problem 3

Use question 3 data sets. Monthly data on amount of billings (Y) and on number of hours of staff time (X) for the 20 most recent months are recorded.

a-) Build a model to predict Y based on the independent variables and test if there is an autocorrelation persists in the data. If autocorrelation persists, remediate the autocorrelation.

b-) X (Staff time) in month 21 is expected to be 3.625 thousand hours. Predict the amount of billings in constant dollars for month 21, using a 99 percent prediction intervaL Interpret your interval.

Problem 4

Use question 4 data set, Create development sample (70% of the data) and hold-out sample (30% of the data) use set.seed(1023) before creating the samples.

a-) Use the development sample , fit a linear regression model, regression tree and Neural Network Model, and calculate the SSE for each model, which method has the lowest SSE?

b-) test the models performances on teh hold out sample, which model would you choose?

Problem 5

Use Question 5 dataset, Y is a dichotomous response variable and X2, X3, and X4 are categorical variables.

a-) Fit a regression model containing the predictor variables in first-order terms and interaction terms for all pairs of predictor variables on development sample.

b-)Use the likelihood ratio test to determine whether all interaction terms can be dropped from the regression model; State the alternatives, full and reduced models, decision rule, and conclusion.

c-)For logistic regression model in part (a), use backward elimination to decide which predictor variables can be dropped from the regression model. Which variables are retained in the regression model?

d-) 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.

e-) Make the prediction for the following two cases and calculate 95% confidence interval

X1=(33,6) X2=(1,1) X3=(1,1) X4=(0,0)