Account
Dashboard
Courses
Calendar
297
Inbox
History
Commons
10
Help

2023 May/Summer (04/22...
Home
Announcements
Assignments
Discussions
Media Gallery
Grades 4
People
Files
Syllabus
Modules
Collaborations
Chat
Google Drive
Student Rating of Teaching
NameCoach Roster
Gradescope

Lab 6

Start Assignment

Due Jul 17 by 11:59pm **Points** 30 **Submitting** a text entry box or a file upload
Available Jul 4 at 12am - Jul 31 at 11:59pm

EE5373: Data Modeling Using R

Summer, 2023

Department of Electrical and Computer Engineering

University of Minnesota

Lab 6: House price predictions.

-

Due date: See the due date shown on the class web page.

Goal: This lab gives you practice developing regression models using a new data set, and introduces you to data segmentation using the dplyr package.

What to do:

In this lab, you will develop linear regression models to predict house prices using the data set available here: https://www.kaggle.com/harlfoxem/housesalesprediction. (also in the lab 6 module)

This public domain data set contains information about sales of houses in the Seattle area from May, 2014 - May, 2015. The definitions of the columns in the data set are here: https://www.kaggle.com/harlfoxem/housesalesprediction/discussion/207885

You are to develop the following linear regression models to predict house prices using the above data set to develop, train, and test your models. Specifically, do the following tasks:

1. Use the backward elimination process to develop a linear regression model using the training and testing process discussed in class with f=0.6. You should exclude potential predictors that are obviously not useful, such as the property ID number, before beginning the backward elimination process. After partitioning the data into appropriate training and testing sets, show how well this model predicts the house prices in the testing set using the root-mean-square error (RMSE) as the overall measure of the quality of your predictions. Repeat the computations five times to produce a total of five different training-testing samples all with the same value of f. Report the RMSE values for each of these five repetitions.

2. Segment the data by zip code and train a model for each zip code that uses these predictors: bedrooms, bathrooms, sqft_living, sqft_lot, grade, and yr_built. Use this model to predict the prices of each house in your testing set within each zip code. Compute the corresponding RMSE value for each zip code. How do these predictions compare across zip codes? For instance, are the RMSE values somewhat similar across all zip codes, or do some zip codes tend to have much higher or lower errors? If there is a significant difference, speculate why you think some zip codes might have significantly higher or lower errors than other zip codes. How do these predictions compare to those you made in Problem 1 based on these RMSE values?

What to turn in for grading:

Write a short lab report that shows how you developed the model for Problem 1 and explains and answers the questions above for both problems. Include a listing of your R code for Problem 2.

Additional information:

As discussed in lab, I recommend using the dplyr package (https://dplyr.tidyverse.org/) for segmenting the data into subsets. See the slides presented in lab for some example R code that uses the dplyr package to perform this type of segmentation.