UCLA Extension Data Science Intensive

Instructor: William Yu

Project 3

Submit your project in R script through Canvas.

A. Analyze Breast Cancer Predictor

- In class (D02c_funct), we learn how to download data directly from UCI machine learning library. In particular, we imported and dealt with the breast cancer dataset. Use that script to import the dataset again. And take a look at the dataset from the UCI website.
- Note that there is some irregularity in one variable (shown as "?"). We want to use some simple missing value method (D03b_na) to replace those. Before you do that, you might need to do the something like the following:

  Ex: bcancer$nuclei=as.numeric(gsub("\\?","NA",bcancer$nuclei))

- Run some imputation (check D03b_na.R) to replace NA.
- Use a simple logistic function (include all explanatory variables) to find which variables are statistically significant to predict that a patient is benign or malignant (**variable "class": 2 is benign; 4 is malignant**) for the breast cancer. No need to do model/variable selection here.

B. Get the Data for the Ride-Sharing Industry

- In the gig economy project folder, you will find the report I wrote about the gig economy. In the folder, I provided part of the data and R script (nonemployer_us.R) I used for my research. Read through the R script carefully so you can understand how I collected and managed the data. The data output is shown as us.nemp.csv.
- Based on that R script and make some changes of codes to get the data of employment and income for only the ride-sharing industry (NAICS code: 4853. Taxi and Limousine Service) from 2005 to 2018. Hint: For example: la18.d2 = la18 %>% subset(NAICS= =4853) %>% rename…
- Export the data in the excel file and plot the employment and income for ride-sharing workers from 2005 to 2018. Hint: I provided the outcome (ride_share.xlsx) for your reference.

C. Analyze Zillow Prize Project

- Run D03c_zillow and enjoy its amazing EDA (exploratory data analysis) and download the related dataset in Project 3 folder into your laptop and read the following link for this project: https://www.kaggle.com/c/zillow-prize-1/kernels.
- Now let's find out what variables/predictors will be able to predict the Zestimate's forecast errors:

logerror = log(Zestimate) - log(SalePrice)

- Keep in mind that if Zestimate is a good model, its forecast error (logerror) should be like an independent/uncorrelated noise. It means it will be difficult to find additional variables to explain logerror. But, after all, Zestimate is not a perfect model. So there might be a chance to find some statistically significant predictor. Since the dependent variable (y) in this case is logerror, you don't need to be surprised to see a very low R^2.

- To begin with your project, you can use the following code to source the script D03c. That is to run the whole script. Make sure you are in the working directory of where D03c_zillow is.
  source("D03c_zillow.R", echo = TRUE)
- Next, we want to change the standard of being good_features from with missing_pct <0.75 to < **0.25**. Note: by doing so, the number of good feature variable will be reduced to 20 plus.
- In D03d, it shows how to select good feature variables in the big dataset (cor_tmp) with dependent variables (logerror and abs_logerror). In case you cannot understand it, here is my way to do it:
  ```
  good_feature = filter(missing_values, missing_pct<0.25)
  good_feature
  gfeature = as.vector(good_feature[,1])
  ## For logerror
  zdata = cor_tmp %>% select(logerror, gfeature)
  ## For abs_logerror
  zdata3 = cor_tmp %>% select(abs_logerror, gfeature)
  ```
- Before running the regression analysis, let's remove these variables because they are (1) geographic information and ID, (2) one value, or (3) pure linear combination of other variables.
  (1) id_parcel, fips, latitude, longitude, zoning_landuse_county, zoning_property, rawcensustractandblock, region_city, region_zip, censustractandblock.
  (2) tax_year
  (3) tax_building and tax_land (note that tax_building + tax_land = tax_total)
- Now you should have less than 20 variables in the data frame.
- Use cor and corrplot functions to check the correlations among these variables. There are some variables which are extremely correlated (correlation>0.95). Remove those highly correlated variables (only keep one).
  Hint: num_bathroom_calc, num_bathroom, num_bath; area_live_finsihed, area_total_calc; tax_total, tax_property.
- Use str to see the structure of this data frame. There are two variables that are integer. Convert them to factor.
- Now we are ready to run the linear regression for the dependent variable: logerror. Use lm to run regression including all these variables. And then use regsubsets to find the best model.
- Change the dependent variable from logerror to abs_logerror and do the regression.

- Briefly explain the results.