Final Assignment

MSDS 598 Spring 2022

Directions

- Submit a .ipynb notebook to Canvas.
- The Notebook should begin with a Markdown cell with your **Name** and the title of the Assignment, **Final Assignment**. If you are working in a group you must put your group members names and state that everyone worked equally on this project, or indicate any issues with your group. Failure to do so will result in points lost.
- Use Markdown Cells to **clearly** indicate which code answers which question and to answer short answer questions. Failure to do so will result in points lost.
- The filename for your notebook should be formatted like

  $$FirstName\_LastName\_AssignmentName.ipynb.$$

  Failure to do so will result in points lost.
- This is due on March 10th by Midnight PM Pacific time. There will be **NO** late submissions allowed.

In this Final Assignment you will explore Melbourne Housing data. The goal will be to explore relationships between the variables and then construct a model that can predict housing price from different features of a house. The data can be found here:

https://www.kaggle.com/dansbecker/melbourne-housing-snapshot

along with a description of each of the variables in the dataset.

1. Do the following to explore the data.

(a) Print the number of missing values for each variable. Are there any variables that you think might have issues from this?

(b) Create a bar chart that shows the median housing price for each `Regionname` Do you think that `Regionname` will affect housing price?

(c) Use seaborn's `lmplot` function to create a scatter plot with `Lattitude` on the x-axis and `Longtitude` on the y-axis colored by `Distance` (Hint: set `legend` and`fit_reg` to False). What is this?

2. Assume that the mean of Price in this dataset is the population mean for housing prices in Melbourne. Choose a particular Region, take a random sample of 25 houses from `Regionname` and conduct a z-test with level of significance $\alpha = 0.5$ to decide if housing prices in this Region differ from the population mean significantly. Use the standard error of the sample as the variance for the distribution of sample means.

(a) What are the Null and Alternative Hypothesis?

(b) Use the p-value to decide whether to reject or fail to reject.

3. Find two variables that are linearly correlated and provide the three pieces of evidence below:

(a) Scattor plot of the two variables

(b) Pearson Correlation Coefficient between the two variables

(c) Scatter plot of the fitted values of the line of best fit against the residuals

4.

(a) Pick three variables that might be linearly correlated with `Price` (anything with Correlation Coefficient ρ satisfying $|\rho| > 0.1$) that satisfy a *weak* assumption of non-Multicolinearity (use a Correlation Coefficient ρ satisfying $|\rho| > 0.75$ to determine strongly colinear)

(b) Perform OLS regression and report the R^2.

5.

(a) Create a validation set using ten percent of the data.

(b) Engineer one new feature.

(c) Use this validation set and the R^2 metric to decide if this new feature improves the model.

6.

(a) Create a new variable called `above_median` which indicates whether a house's price is above the median price for houses in this dataset.

(b) Report the accuracy of your model from Q5 created from the training set at predicting this new variable on the validation set.

(c) Perform logistic regression using the same training set and report the accuracy from this model on the same validation set.

(d) Compute the Precision, Recall, and AUC for both models on the validation set.

7. Perform a linear regression using the training set from Q5 with all numerical variables to predict `Price` and use either Lasso or Ridge regression. Report the R^2 on the validation set.