Intro to Data Science: Homework 2 — Submitted By: Mayanka Jha

Ques 2: Predict 2013 home prices using state information only.

a) What is the intercept? What does it correspond to?
The intercept value is 281730. The intercept value generally indicates the average home price.

b) How do you get this information from your regression?
After creating linear regression model, we print the summary of our regressor, it gives us the value of the intercept.

c) Based on your regression coefficients, what states have the most and least expensive average homes?
DC has the most expensive average homes and West Virginia has the least expensive average homes.

d) How do you get this information from your regression?
To get this information, we see the regression coefficient values, the highest regression coefficient will correspond to most expensive average homes and the lowest regression coefficient will correspond to the least expensive average homes.

e) What is the average price of homes in those states?
Average Price of home in DC = 232558.9+281730 = 514288.9
Average Price of home in WV = 281730 - 183306.9 = 98423.1

f) How do you get this information from your regression?
We add the observed intercept value and regression coefficient value from our regressor to obtain this information.

Ques 3: Predict 2013 home prices from state and county information.

a) What US counties have the highest and lowest regression coefficients? Why?
County Pitkin has the highest regression coefficients and county Calaveras has the lowest regression coefficient. The lowest regression coefficient value indicates the county with lowest average home price and the highest regression coefficient value indicates the county with highest average home price.

Ques 4: Build a regressor that best predicts average home values in this dataset.

a) Describe what you did to build the best predictor possible
The house_test data set does not contain the price2013 value, so if I directly create a model using house_train dataset and try to predict the price2013 for house_test dataset, I wouldn't be able to validate it. So, in order to build the best regressor, I followed below steps:
    1. Split the train dataset into two dataset, training and validation (70, 30 partition).
    2. Created a model with price2007, state and poverty as the predictors and training dataset
    3. Predicted the values for validation dataset
    4. Calculated the RMSE value
    5. This model appears to be good, but I wanted to see what value the feature poverty adds in my model
    6. Created a new model using only state and price2007 and repeated steps 3 and 4
    7. The RMSE value increases and we want to have lower RMSE value. **Hence, I have trained my final model using state, price2007 and poverty as predictors.**

b) Give your best Kaggle score -> **58855.600**

c) Give your Kaggle username -> **Mayanka Jha**

Ques 5: Suppose you have 2 bags. Bag #1 has 1 black ball and 2 white balls. Bag #2 has 1 black ball and 3 white balls. Suppose you pick a bag at random and select a ball from that bag. What is the probability of selecting a white ball?

Suppose A= Event that Bag#1 is chosen
and B= Event that Bag#2 is chosen
and C= Event that white ball is selected
Also, $P(A) = ½$, $P(B) = ½$, $P(C|A) = 2/3$, $P(C|B) = ¾$

Hence using, P(C)= P(C|A) * P(A) + P(C|B) * P(B)
= (1/2 * 2/3) + (1/2 * 3/4)
= 17/24= 0.71

Ques 6: A soccer team wins 60% of its games when it scores the first goal, and 10% of its games when the opposing team scores first. If the team scores the first goal about 30% of the time, what fraction of the games does it win?

Suppose A = The soccer team scores the first goal
B = The opposing team scores the first goal
C = The soccer team wins the game

Also, P(A) = 0.3, P(B) = 0.7, P(C|A) = 0.6, P(C|B) = 0.1

Hence using, P(C) = P(C|A) * P(A) + P(C|B) * P(B)
= (0.6* 0.3) + (0.7 * 0.1)
= 0.25