UC San Diego Extension
Cloud Services for Machine Learning

Summer 2020
Homework#2

Date Given: July 6, 2020 Due Date: July 12, 2020

==========

Analyze the data source in 'kc-house-data.csv' file. This data source is a part of databases available in the public domain. This file contains 21,613 observations of real-estate properties of King county in Washington state. The data for the following 21 variables are provided.

1. id
2. date
3. price
4. bedrooms
5. bathrooms
6. sqft_living
7. sqft_lot
8. floors
9. waterfront
10. view
11. condition
12. grade
13. sqft_above
14. sqft_basement
15. yr_built
16. yr_renovated
17. zipcode
18. latitude
19. longitude
20. sqft_living15
21. sqft_lot15

Read the raw data source file 'kc-house-data.csv'. Build a linear regression model (as described in Problem#1 and Problem#2 on the next page) using the following variables.

Response Variable:

- price (numerical)

Predictor Variables:

- sqft_living (numerical)
- bedrooms (numerical)
- waterfront (categorical):
  - Levels of waterfront: 0 = no waterfront, 1 = waterfront
- condition(categorical)
  - Levels of condition: 1,2,3,4,5

Problem#1

Build a regression model using the following characteristics.

- Programming Language: Python
- Cloud Platform: Colab
- Package: Scikit-Learn

Verify that your regression equation is as follows.

$$price = 66{,}581.53 + 305.72 * sqftliving - 52{,}704.36 * bedrooms + 783{,}090.68 * waterfront$$
$$-25{,}698.33 * condition2 - 8{,}811.75 * condition3 + 28{,}198.78 * condition4 + 100{,}565.72 * condition5$$

Predict the price of a home with following characteristics:

- sqft_living = 3,000
- bedrooms = 4
- waterfront = No (0)
- condition = 4

Verify that the $Predicted\ price = \$801{,}112.20$

==========

Problem#2

Build a regression model using the following characteristics.

- Programming Language: None
- Cloud Platform: AutoML GCP

The procedure to build a regression model on GCP is as follows.

1. GCP/Storage
    a. Create a Bucket in GCP
    b. Region: us-central1(Iowa)
    c. Upload Data file in that bucket
2. GCP/Table/Dataset
    a. Import data in a GCP Dataset from the bucket: Takes time
3. GCP/Table/Model
    a. Train the Model
        i. Select Target Variable + Budget: Takes time
    b. Evaluate the Model
    c. Test & Use: Deploy the Model: Takes time
        i. Prediction

Predict the price of a home with following characteristics:

- sqft_living = 3,000
- bedrooms = 4
- waterfront = No (0)
- condition = 4

The predicted value of the 'price' variable using GCP should be approximately equal to $801,112.20. Also compute the 95% prediction interval of the response variable 'price'.

Building a regression model on GCP will cost a certain amount. Please check the GCP charges on your account before and after you complete this assignment. Make sure you do not deplete the $300 credit you have on your account.

How to handle 'condition" categorical variable:

The 'condition' variable is categorical with 5 levels. The values of this variable are 1,2,3,4,5. This does NOT mean that 5 > 4 > 3 > 2 > 1. Since there are 5 levels of this variable, we need to replace the 'condition' variable with 4 (k-1) dummy (indicator) variables.

We must convert the 'condition' variable into 4 separate dummy variables using one-hot-encoding. The logic used for prediction is shown in the table below. The 'condition=1" will be our base condition. All values will be computed relative to 'condition=1'.

	Var: **condition2**	Var: **condition3**	Var: **condition4**	Var: **condition5**
Condition1 (Base)	0	0	0	0
Condition2	1	0	0	0
Condition3	0	1	0	0
Condition4	0	0	1	0
Condition5	0	0	0	1

Regression equation is as follows.

$$price = 66{,}581.53 + 305.72 * sqftliving - 52{,}704.36 * bedrooms + 783{,}090.68 * waterfront$$
$$-25{,}698.33 * condition2 - 8{,}811.75 * condition3 + 28{,}198.78 * condition4 + 100{,}565.72 * condition5$$

- This means that the price of a house with '**condition=2**' is $25,698.33 **less** compared with the house with condition=0.

- This means that the price of a house with '**condition=5**' is $100,565.72 **more** compared with the house with condition=0.

Now let us predict the price of the house using different value of the 'condition' categorical variable.

==========

Predict the price of a home with following characteristics:

- sqft_living = 3,000
- bedrooms = 4
- waterfront = No (0)
- **condition = 1**

$$price = 66{,}581.53 + 305.72 * sqftliving(3{,}000) - 52{,}704.36 * bedrooms(4) + 783{,}090.68 * waterfront(0)$$

$$price = 772{,}913.4$$

==========

Predict the price of a home with following characteristics:

- sqft_living = 3,000
- bedrooms = 4
- waterfront = No (0)
- **condition = 2**

$$price = 66{,}581.53 + 305.72 * sqftliving(3{,}000) - 52{,}704.36 * bedrooms(4) + 783{,}090.68 * waterfront(0) - 25{,}698.33 * condition2(1)$$

$$price = 747{,}215.1$$

==========

Predict the price of a home with following characteristics:

- sqft_living = 3,000
- bedrooms = 4
- waterfront = No (0)
- **condition = 3**

$$price = 66{,}581.53 + 305.72 * sqftliving(3{,}000) - 52{,}704.36 * bedrooms(4) + 783{,}090.68 * waterfront(0) - 8{,}811.75 * condition3(1)$$

$$price = \texttt{764,101.7}$$

==========

Predict the price of a home with following characteristics:

- sqft_living = 3,000
- bedrooms = 4
- waterfront = No (0)
- **condition = 5**
- $price = 66{,}581.53 + 305.72 * sqftliving(3{,}000) - 52{,}704.36 * bedrooms(4) + 783{,}090.68 * waterfront(0) + 100{,}565.72 * condition5(1)$
-

$$price = \texttt{873,479.2}$$

==========