DATA 601 - HW05

Due date: November 27, 2022, 23:59 pm

Q1. (10 points)

Training and validation dataset: https://raw.githubusercontent.com/simsekergun/DATA601/main/2021Fall/datasets/HouseTraining.csv

- The last column of the HouseTraining.csv file lists price of 400 houses.
- There are 11 features
    - School rating (integer between 1 and 10)
    - House Area (sq ft)
    - Lot Area (sq ft)
    - Number of rooms
    - Number of bathrooms
    - Garage Yes:1, No: 0
    - Pool Yes:1, No: 0
    - Age of the House (years)
    - Walkability rating (something between 1 and 10)
    - Crime rate (something between 1 and 10)
    - Zipcode (Note that this is a fake data)
    - House price ($)

Here are questions

1.1 Calculate the average crime rate for each zip code determine the zipcode with highest average crime rate?

1.2 Calculate the average house price for each zip code determine the zipcode with lowest average house price? Do you see a pattern?

1.3 What feature has the strongest correlation with the "School_Rating"

1.4 Split your dataset into two (training 80%, validation (testing) %20, random_state=1). Build a multiple linear regression model to estimate the house price from all the other features we have and calculate the maximum relative error using $100 * \max|(y_i - \hat{y}_i)/y_i|$ and R^2, where y_i is the true value for the i^{th} case in your testing data set and $\hat{y}_i$ is the prediction.

1.5 Download the new test dataset (https://raw.githubusercontent.com/simsekergun/DATA601/main/2021Fall/datasets/HouseTest.csv), guess the prices of these 10 houses featured in this dataset and print your predictions. Note that this dataset doesn't include "House_Price" column which was given in the training dataset.

Q2. (10 points)

The Default data set of the ISLR2 package contains data about ten thousand customers. We know the balance of their bank account, their annual income and whether they are a student. You can download the dataset here: https://github.com/simsekergun/DATA601/blob/main/2021Fall/datasets/Default.xlsx?raw=true'

Let's replace yes' and no's with 1's and 0's using the factorize() function. Note that factorize() returns two objects: a label array and an array with the unique values. We are only interested in the first object, i.e.

```
df = pd.read_excel('https://github.com/simsekergun/DATA601/blob/main/2021Fall/datasets/Default.xlsx?raw=true',index_col=[0])
df['default'] = df.default.factorize()[0]
df['student'] = df.student.factorize()[0]
```

Here the steps/questions you need to follow

2.1 Plot the histograms of the features in this dataset. What kinds of distributions do you see?

2.2 Boxplot 'default vs balance' and 'default vs income'. Which one has outliers?

2.3 Split your dataset into two (training 80%, validation (testing) %20, random_state=1). Build a simple logistic regression model to predict default from balance feature only. Create the confusion matric and calculate accuracy, sensitivity, and specificity.

2.4 Repeat 2.3 but this time use 'balance','income', and 'student' features to predict the default with a multiple logistic regression model. Create the confusion matric and calculate accuracy, sensitivity, and specificity.

2.5 What does having a high sensitivity and a low specificity mean?