**Homework Assignment #2 (CS 156 Spring)**

**Due: Wednesday February 24 midnight**

**(Regression – Simple Linear, Multiple Linear, Decision Tree, Random Forest)**

The objective of this homework assignment is to predict house prices by deploying various predictive models that accept as inputs, variables that significantly influence the price. We will use 4 different models and compare their performance with respect to their predictive accuracy. Here are the models we will use:

1. Simple Linear Regression
2. Multiple Linear Regression
3. Decision Tree Regression
4. Random Forest Regression

The dataset for this project contains house sale prices. There are 16 column headers:

1. Waterfront Dummy variable indicating if the house was overlooking a waterfront
2. Renovated If the house was renovated
3. View An index from 0 to 4 indicating how good the view was. Higher is better
4. Condition An index from 1 to 5 on the condition of the apartment. Higher is better
5. Grade An index from 1 to 4. Higher the better
6. Bedrooms Number of Bedrooms
7. Bathrooms Number of Bathrooms (can have 0.5 to indicate half bathroom)
8. Sqft\_living Square footage of Interior living space
9. Sqft\_lot Square footage of Interior land space
10. Floors Number of floors
11. Sqft\_above Square footage of the interior living space that is above ground level
12. Sqft\_basement Square footage of the interior living space that is below ground level
13. Yr\_built The year the house was initially built
14. Sqft\_living15 Square footage of the living area of the nearest 15 neighbors
15. Sqft\_lot15 Square footage of the land lots of the nearest 15 neighbors
16. Price Price of sale

**Part (A):** Data Import, Data Pre-processing

* 1. Read the file Housing-Data-one-zip-3.csv
  2. Convert categorical data: Waterfront, Renovated, View, Condition, Grade
  3. Transform some data. For example, you may transform the column Yr\_built to reflect the age of the building by subtracting Yr\_vuilt from 2020.
  4. Divide the data set into Training set and Test set be

We will use the same data set for all 4 prediction algorithms in this assignment. Here are the assumptions for the first 5 fields of the data set and the inputs for your program to do the prediction of house prices. Predict the house price for the following cases (Note: **Age = 2020-Yr\_built**)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Assume**  waterfront | renovated | view | condition | grade |
|  |  |  |  |  |
| 0 | 0 | 0 | 3 | 3 |

[Bedroom, Bathhrooms, Sqft\_living, Sqft\_lot, Floors, Sqft\_above, Sqft\_basement, Age, Sqft\_living15, Sqft\_lot15]

* + 1. [3, 0.75, 2510, 20000, 2.0, 2510, 0, 59, 2130, 20000]
    2. [4, 2.25, 1500, 5393, 2.0, 1500, 0, 21, 1500, 5952]
    3. [4, 2.25, 2870, 5393, 2.0, 2870, 0, 21, 1500, 5952]
    4. [4, 3.50, 4083, 68377, 2.0, 4083, 0, 15, 2430, 41382]
    5. [4, 3.50, 4500, 68377, 2.0, 4500, 0, 15, 2430, 41382]
    6. [4, 3.50, 2870, 68377, 2.0, 2870, 0, 15, 2430, 41382]
    7. [4, 3.50, 750, 68377, 2.0, 750, 0, 15, 2430, 41382

**Part (B):** Use Simple Linear Regression to predict the house price using **Sqft\_living** as the independent variable

1. Print Rsquare
2. Plot the linear regression line for the Training Data Set
3. Plot the linear regression line for the Test Data Set
4. Predict the house prices for the test data set given above.

**Part (C):** Use Multiple Linear Regression using all variables to predict the house price

1. Print Rsquare
2. Predict the house prices for the test data set given above.

**Part (D):** Use Decision Tree Regression model to predict the house price

1. Print Rsquare
2. Predict the house prices for the test data set given above.

**Part (E):** Use Random Forest Regression model (use 10 Random Trees) to predict house price

1. Print Rsquare
2. Predict the house prices for the test data set given above.

**Summarize your observations:**

1. Tabulate the result as follows:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Test Data Point | Simple Linear Regression | Multiple Linear Regression | Decision Tree Regression | Random Forest Regression |
| (i) | 356363.12752274 | 322853.75707537 | 363000 | 405900 |
| (ii) | 232887.30257624 | 223043.61780165 | 215000 | 218050 |
| (iii) | 400374.31265219 | 402892.93768066 | 299000 | 317590 |
| (iv) | 548667.55588001 | 557862.19128196 | 359000 | 474178.8 |
| (v) | 599647.17865495 | 588308.23996124 | 359000 | 474178.8 |
| (vi) | 400374.31265219 | 469298.50531561 | 359000 | 422128.8 |
| (vii) | 141197.33355656 | 314512.83816915 | 194820 | 294180 |
| R-Square | 0.6682006794899293 | 0.8072554741507528 | 0.9952504116289396 | 0.9503025303839485 |

1. Which predictive model performed the best and why do you think so?
   1. although the r-squared value for decision tree method is the highest, there are repeated values in the table when the 7th parameter is the same, which shows that it isn’t the best estimator. random forest performs somewhat worse than the decision tree in terms of r-squared, but the predictions seem to line up better with the actual data.
2. Which variables are most important for prediction? Use Multiple Linear Regression Model to justify your answer. Hint: use **print(regressor.coef\_)** to print out the coefficients for the independent variables and focus on the last 10 coefficients.
   1. we get the following coefficients, where those in boldface and purple are the coefficients of focus:

8.90689220e+04

4.81272955e+04

1.59078621e+04

7.76704263e+03

3.35765190e+04

**-6.92876340e+03**

**4.32242173e+03**

**4.66709071e+01**

**6.49099103e-01**

**-4.18199657e+03**

**2.63412000e+01**

**2.03297072e+01**

**-6.92693042e+02**

**1.71650799e+01**

**2.25297017e+00**

* 1. now, we can see the scores of each of the coefficients: we can conclude the most important feature is number of bedrooms, floors, and bathrooms. the others are in boldface below:

1. Feature: 5, Score: -6928.76340
2. **Feature: 6, Score: 4322.42173**
3. **Feature: 7, Score: 46.67091**
4. **Feature: 8, Score: 0.64910**
5. Feature: 9, Score: -4181.99657
6. **Feature: 10, Score: 26.34120**
7. **Feature: 11, Score: 20.32971**
8. Feature: 12, Score: -692.69304
9. **Feature: 13, Score: 17.16508**
10. **Feature: 14, Score: 2.25297**