CPSC-6430 Machine Learning: Implementation & Evaluation

Project 2: Linear Regression

Dineshchandar Ravichandran
C19657741

Contents

Introduction **3**

1. Problem Statement **3**

2. Project Screen Shot: **3**

3. Project Input and Output **5**

3.1. Input Training Data: **5**

Details of training data 5

3.2. Input Testing: **5**

Details of testing data 6

3.3. Output: **6**

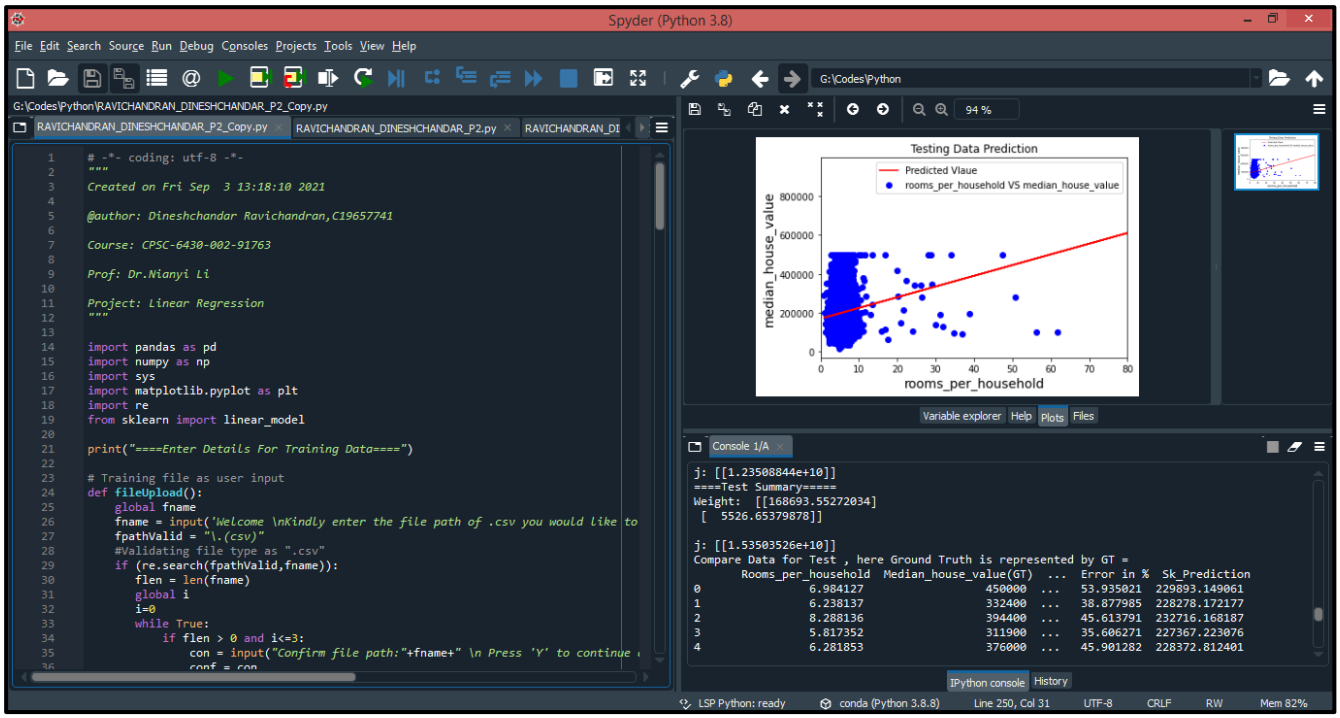
Introduction

Project 2 to implement the basic linear regression algorithm and predict value of Y (dependent value) for X (Independent Value).

1. Problem Statement

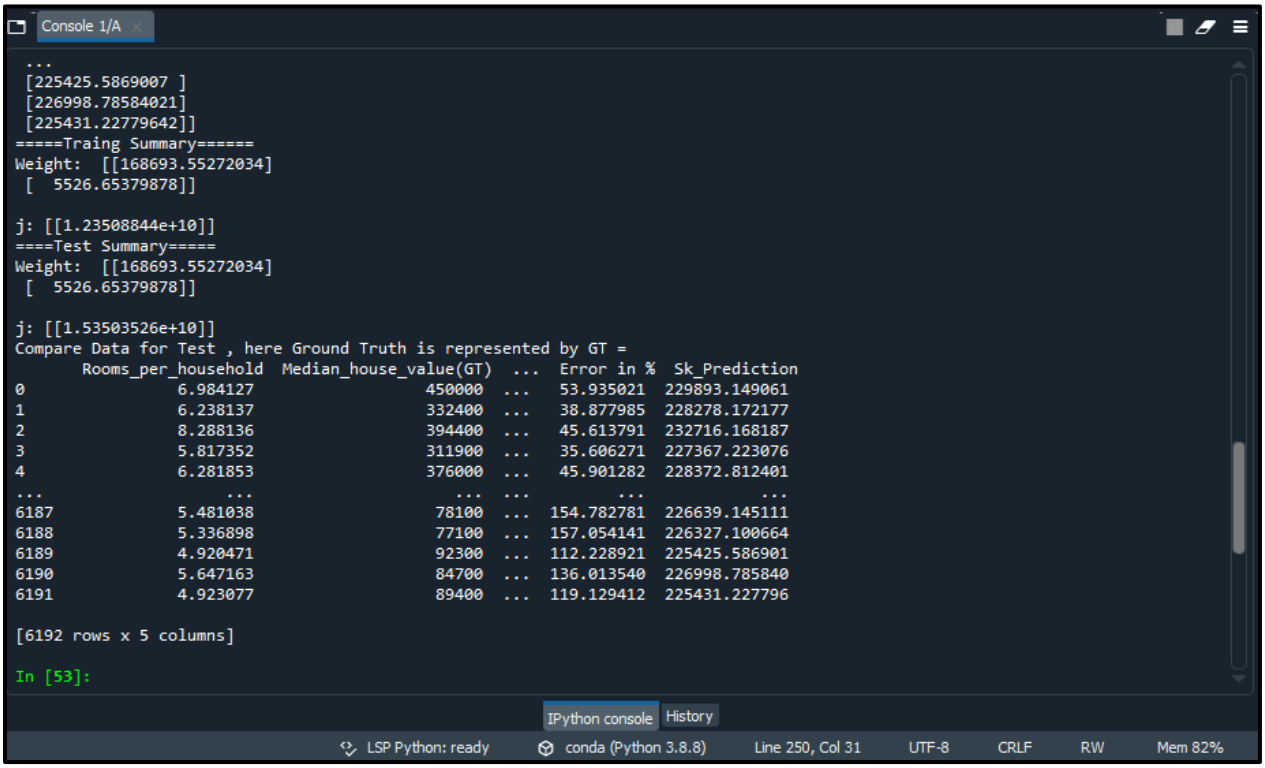
- To create a Python program to read the training housing data from a ".csv" file as per the user input and generate the weight.
- Based on the generated weight the program will predict the value of y which will be compared with the ground truth data from the data set.
- Post which the program should ask the user for a test file. Using the weights computed from the training file, it should then print out J for the test file.

2. Project Screen Shot:



- The above screen shots represents the code in the SPYDER IDE, along with Test Data Summary in console and the graphical representation of Test Data(blue) , Predicted values line(red).

- Console Screen Shot for the same:



Console Logs:

Console_log.txt

3. Project Input and Output

3.1. Input Training Data:

CSV file containing only 70% of original data present in "California Housing Data" ("housing.csv"):

housing_train.csv

Details of training data

- The training data set has the following data:
  Range Index: 14448 entries, 0 to 14447.
- Data columns (total 10 columns):

Sr.No.	Column	Non-Null Count	Dtype
0	longitude	14448 non-null	float64
1	latitude	14448 non-null	float64
2	housing_median_age	14448 non-null	int64
3	total_rooms	14448 non-null	int64
4	total_bedrooms	14310 non-null	float64
5	population	14448 non-null	int64
6	households	14448 non-null	int64
7	median_income	14448 non-null	float64
8	median_house_value	14448 non-null	int64
9	ocean_proximity	14448 non-null	object

- dtypes: float64(4), int64(5), object(1)

3.2. Input Testing:

CSV file containing 30% "California Housing Data" ("housing.csv"), which is not present in training data set:

housing_train.csv

Details of testing data

- Range Index: 6192 entries, 0 to 6191
- Data columns (total 12 columns):

Sr.No	Column	Non-Null Count	Dtype
0	longitude	6192 non-null	float64
1	latitude	6192 non-null	float64
2	housing_median_age	6192 non-null	int64
3	total_rooms	6192 non-null	int64
4	total_bedrooms	6123 non-null	float64
5	population	6192 non-null	int64
6	households	6192 non-null	int64
7	median_income	6192 non-null	float64
8	median_house_value	6192 non-null	int64
9	ocean_proximity	6192 non-null	object
10	rooms_per_household	6192 non-null	float64
11	area_rate	6192 non-null	float64

- dtypes: float64(6), int64(5), object(1)
- memory usage: 580.6+ KB

3.3. Output:

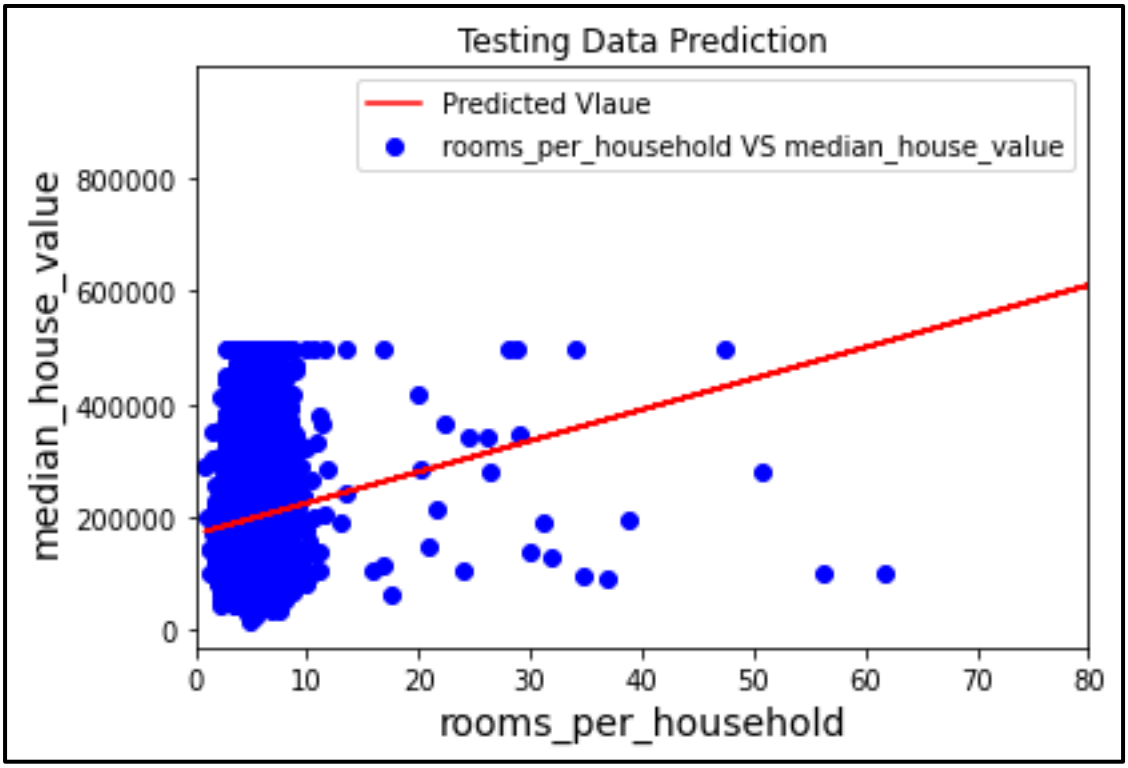
3.3. A Summary

```
=====Traing Summary======
Weight:  [[168693.55272034]
 [  5526.65379878]]

j: [[1.23508844e+10]]
====Test Summary=====
Weight:  [[168693.55272034]
 [  5526.65379878]]

j: [[1.53503526e+10]]
Compare Data for Test , here Ground Truth is represented by GT =
      Rooms_per_household  Median_house_value(GT)  ...  Error in %  Sk_Prediction
0                6.984127                  450000  ...   53.935021  229893.149061
1                6.238137                  332400  ...   38.877985  228278.172177
2                8.288136                  394400  ...   45.613791  232716.168187
3                5.817352                  311900  ...   35.606271  227367.223076
4                6.281853                  376000  ...   45.901282  228372.812401
...                   ...                     ...  ...         ...            ...
6187             5.481038                   78100  ...  154.782781  226639.145111
6188             5.336898                   77100  ...  157.054141  226327.100664
6189             4.920471                   92300  ...  112.228921  225425.586901
6190             5.647163                   84700  ...  136.013540  226998.785840
6191             4.923077                   89400  ...  119.129412  225431.227796
```

3.3. B. *Graph Representation*



3.4. *C. Training Comparison Table:*

Illustrating 5 of 14447:

	Rooms_per_household	Median_house_value (Ground Truth)	Predicted_Vlaues	Error in %	Sk_Prediction
0	6.984126984	452600	207292.4046	54.19964546	207292.4046
1	6.238137083	358500	203169.5767	43.3278726	203169.5767
2	8.288135593	352100	214499.2088	39.08003159	214499.2088
3	5.817351598	341300	200844.041	41.1532256	200844.041
4	6.281853282	342200	203411.181	40.557808	203411.181

CSV:

Traing.csv

3.4 D. Testing Comparison Table:

Illustrating 5 of 6191:

	Rooms_per_household	Median_house_value (Ground Truth)	Predicted_Vlaues	Error in %	Sk_Prediction
0	6.984127	450000	207292.4	53.93502	229893.1
1	6.238137	332400	203169.6	38.87799	228278.2
2	8.288136	394400	214499.2	45.61379	232716.2
3	5.817352	311900	200844	35.60627	227367.2
4	6.281853	376000	203411.2	45.90128	228372.8

CSV:

Testing.csv