

Ramya Tekumalla

Assignment 2 due 10/07/2019

rtekumalla1@student.gsu.edu

The following are the details of the submitted files

- hw2.xlsx – contains the details of all the teams from Serie A, Italy League of 18/19 season.
- hw2a1.py – code for A question part 1.
- hw2a2.py – code for A question part 2.
- hw2b1.py – code for B question part 1.
- hw2b2.py – code for B question part 2.

Question A Part 1

Since the question asked was to use top 10 teams as a training set, I have sorted the hw2.xlsx based on their points and for this question; I have not shuffled the data. The first 10 rows are training set and the second 10 rows are the test set. The following picture depicts the teams and the statistics of the teams.

club	Squad	Age	Foreigners	Total Market Value	Average Market Value	Matches	win-loose	Pts
Juventus FC	59	24,0	25	996,10	16,88	38	40	90
SSC Napoli	54	23,3	27	593,54	10,99	38	38	79
Atalanta BC	88	21,7	29	287,99	3,27	38	31	69
Inter Milan	60	23,0	30	667,90	11,13	38	24	69
AC Milan	45	23,5	22	599,61	13,32	38	19	68
AS Roma	57	22,8	27	495,76	8,70	38	18	66
Torino	62	22,8	29	223,92	3,61	38	15	63
Lazio	57	23,4	34	395,64	6,94	38	10	59
Sampdoria	59	22,5	30	194,51	3,30	38	9	53
Bologna	48	24,0	27	115,08	2,40	38	-8	44
Sassuolo	59	22,8	15	186,02	3,15	38	-7	43
Udinese Calcio	56	23,7	41	181,37	3,24	38	-14	43
SPAL	51	24,3	19	83,65	1,64	38	-12	42
Parma	70	24,6	19	73,82	1,05	38	-20	41
Cagliari Calcio	49	24,9	19	139,94	2,86	38	-18	41
Fiorentina	56	22,0	36	281,58	5,03	38	2	41
Genoa	69	22,8	34	122,52	1,78	38	-18	38
FC Empoli	59	23,1	24	66,46	1,13	38	-19	38
Frosinone	44	24,8	10	68,54	1,56	38	40	25
Chievo Verona	68	23,7	32	70,02	1,03	38	-50	17

Fig1: 20 teams that participated in the Serie A, Italy League of 18/19 season

For this model, we only use one feature i.e Average Market value and predict the points Since the regression model is a linear, the relationship between the target variable (y) and the feature variable (x) is defined as

$$y = \text{weight} * x + \text{intercept}.$$

By running, the code hw2a1.py, we get the following weight coefficients and Y axis intercept

```
(base) [ramyat@deepml 2]$ python hw2a1.py
Weight coefficients: [0.02078956]
y-axis intercept: 49.25608738948132
```

Plugging in the min and max values into the equation, we can plot the regression fit to our training data:

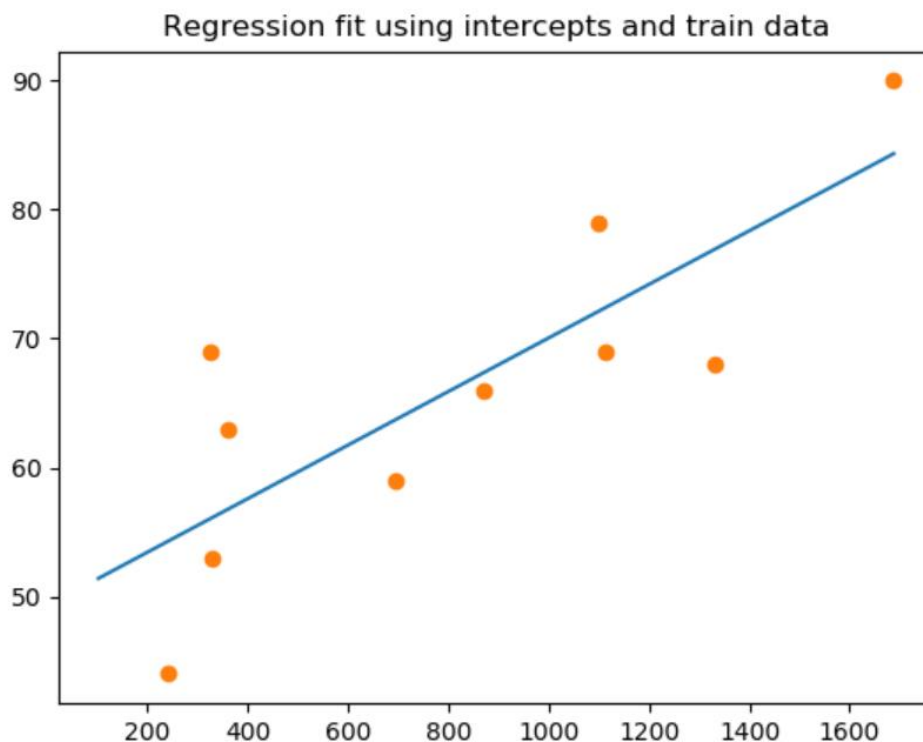


Fig2 : Regression fir using intercepts and train data

The predict method is used to predict the target variable. We expect these predicted values to fall onto the line that we plotted in the Figure2. The following Figure 3 depicts the predictions on train data.

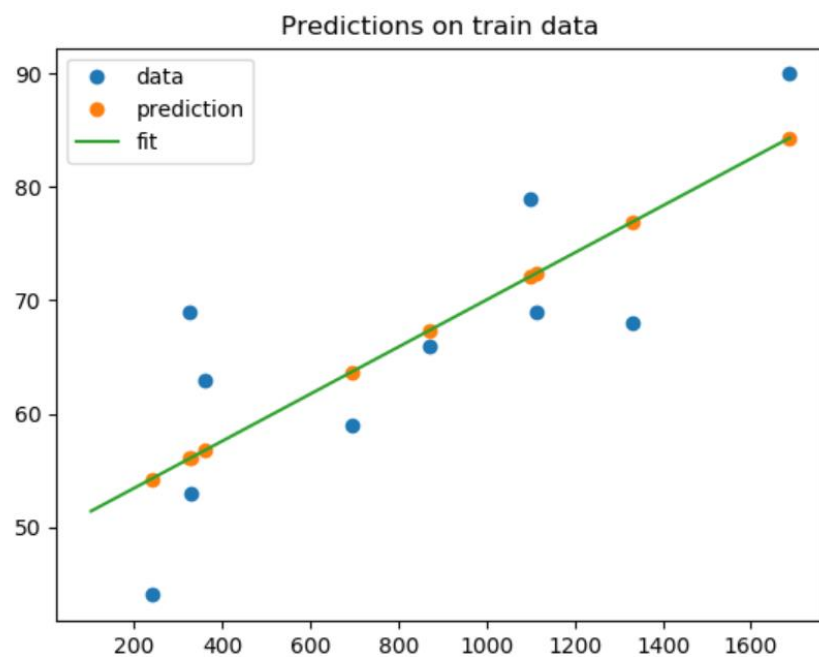


Fig 3: Predictions on train data

As we can see in the plot above, the line is able to capture the general slope of the data, but not many details. Using the predict on the test set, we get the following result,

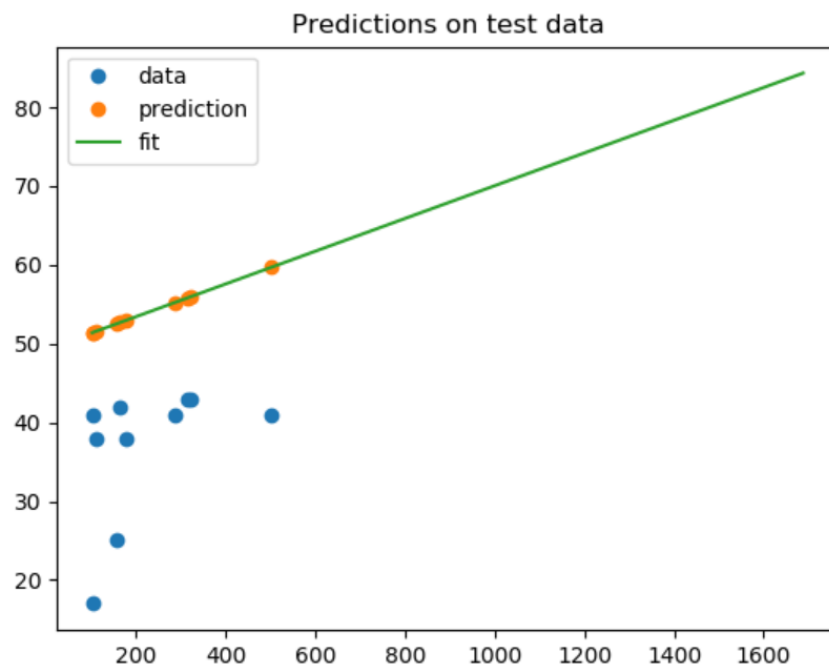


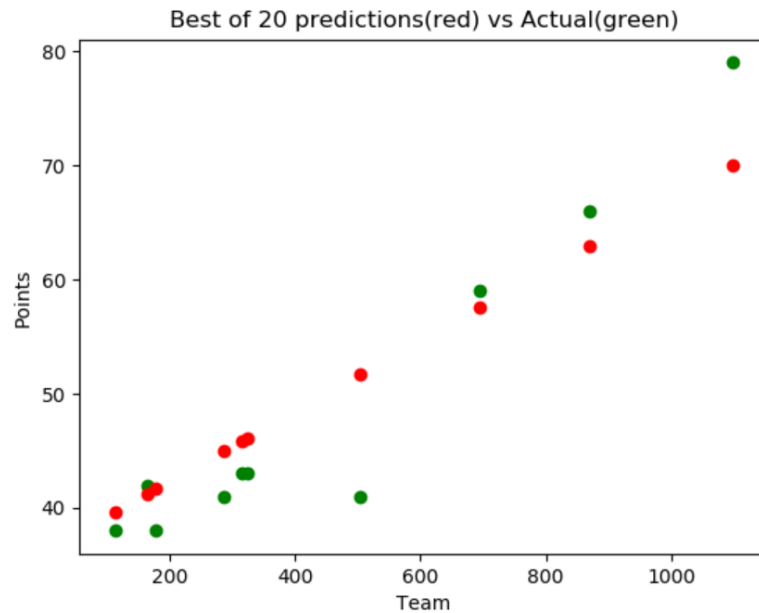
Fig 4: Predictions on test data

```
regressor.score(X_test, y_test) - -4.001527189726083
```

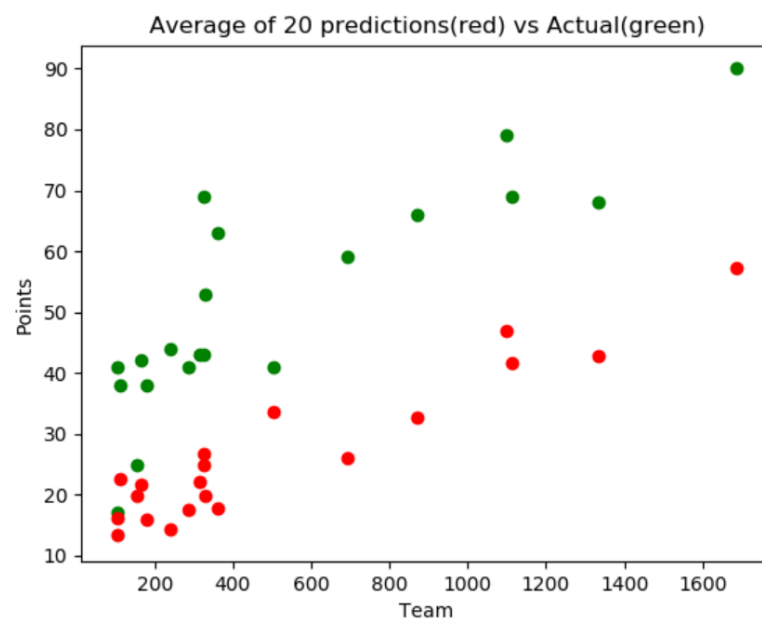
```
((y_pred_test - y_test) ** 2).mean() - 345.555513538175
```

Question A part 2

For this part, we enable the shuffle mode. We randomly generate the train & test set and repeat the experiment for 20 times. The following figure depicts the best of 20 predictions.



Next, we determine the average of all predictions. The following figure depicts the average of the 20 predictions.



`((df['Pts'] - df['Avg_Problem1']) ** 2).mean() - 733.1362764475546`

Conclusion for Question 1

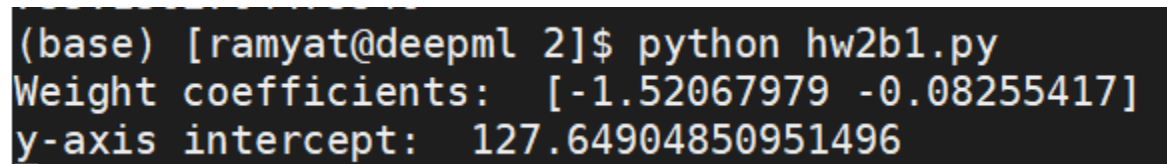
Model A2 is better when compared to Model A1 because of the following reasons

- 1) Model A1 is only trained on top 10 rows of the training data and hence the model hasn't seen all the variations in the data.
- 2) Model A2 shuffled data, the training data is uniformly distributed, and the model fitted well. The MSE value is higher when compared to A1

Question B Part 1

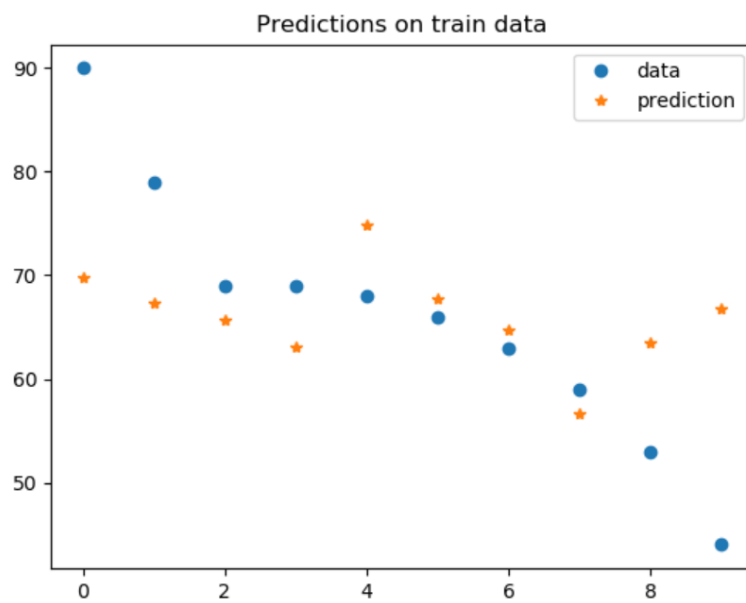
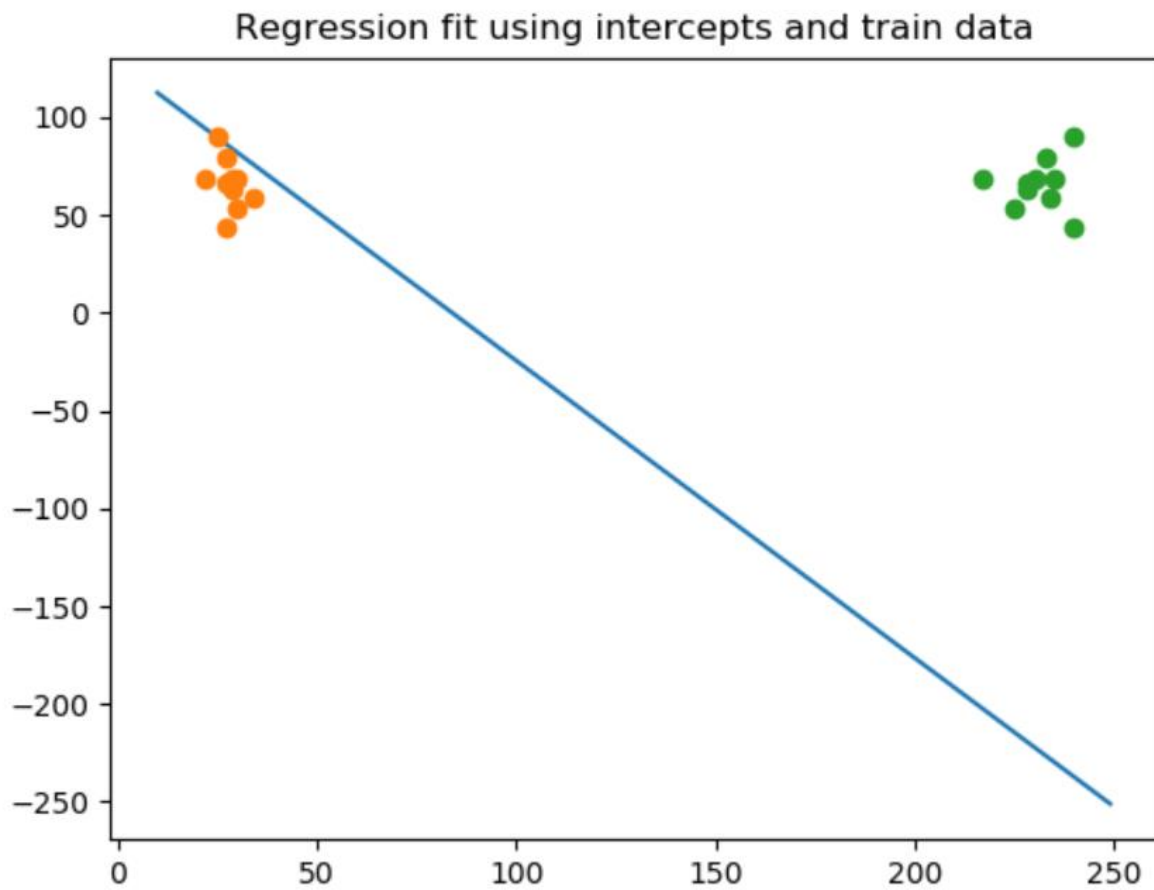
For this question, the code is similar to Question A Part 1 with an exception that, two different features are used. Age and Foreigners are the two different features that are in this model.

The weighted coefficients and the y intercepts are calculated. The following image is the answer generated from the running the code hw2b1.py



```
(base) [ramyat@deepml 2]$ python hw2b1.py
Weight coefficients: [-1.52067979 -0.08255417]
y-axis intercept: 127.64904850951496
```

The following figure depicts the regression fit using the weight coefficients and y axis intercepts and the train data.



The following figure depicts the predictions using the test data.

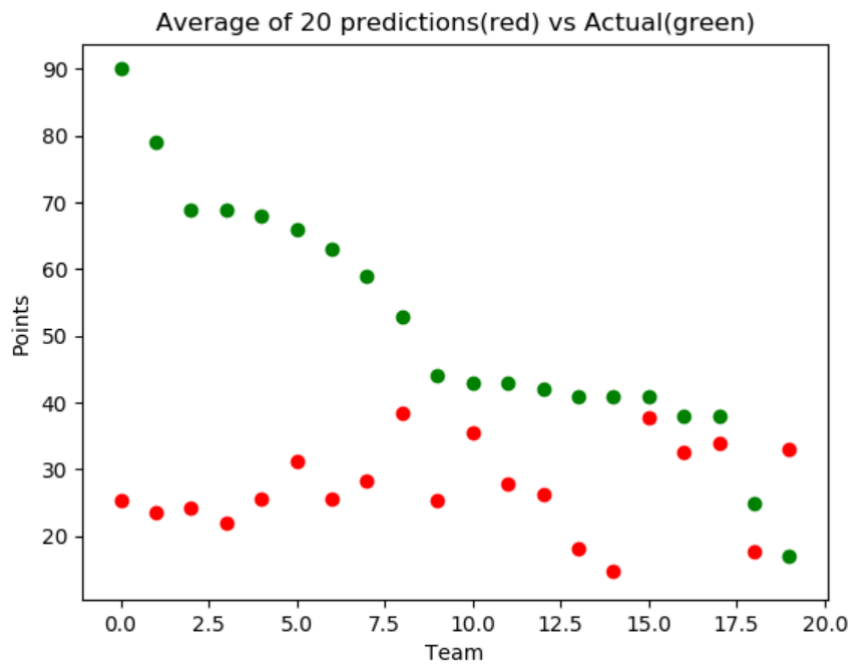
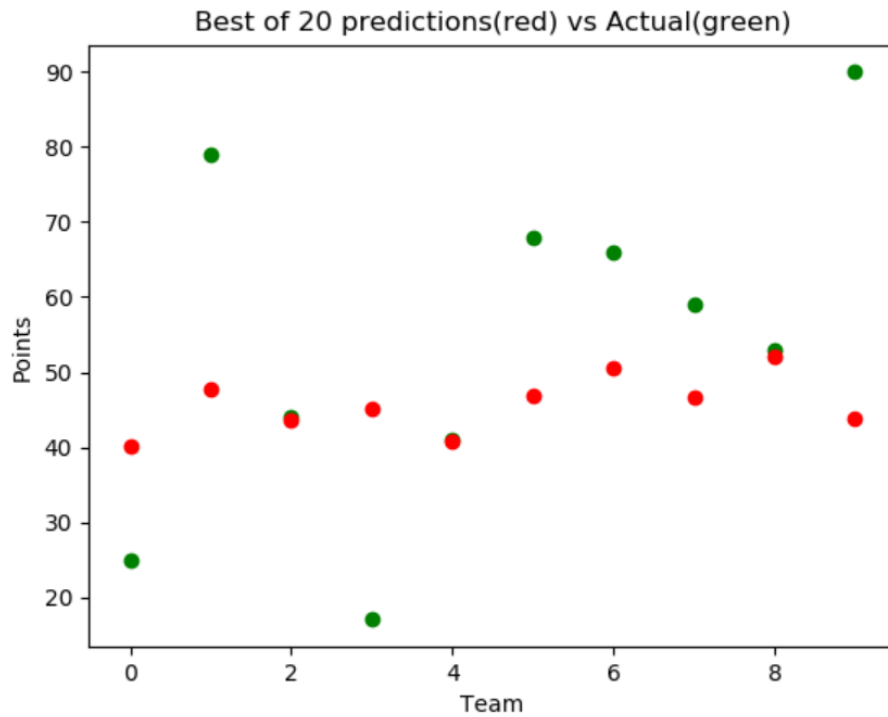


`regressor.score(X_test, y_test)` - **-19.250830241905398**

`((y_pred_test - y_test) ** 2).mean()` - **1399.1298614132438**

Question B Part 2

This is similar to Question A Part 2 except that we are using two different features, i.e age and foreigners. In addition, we shuffle the data for this part.



$((\text{np.array}(\text{best_y_true}[\text{best}]) - \text{np.array}(\text{best_y_pred}[\text{best}])) ** 2).mean() - 497.76914536265014$

Conclusion for Question 2

Though the Value of MSE for Part B1 is better than Part B2, Model B2 is better model when compared to B1. It is the same reason as A1. The model B2 uses shuffled data and the experiment is repeated for 20 times. Whereas for Part B1, the model does not use shuffled data and uses only top 10 rows.

Comparison between Question A and Question B

The models of B are better when compared to Model A because they use more number of features when compared to model A.