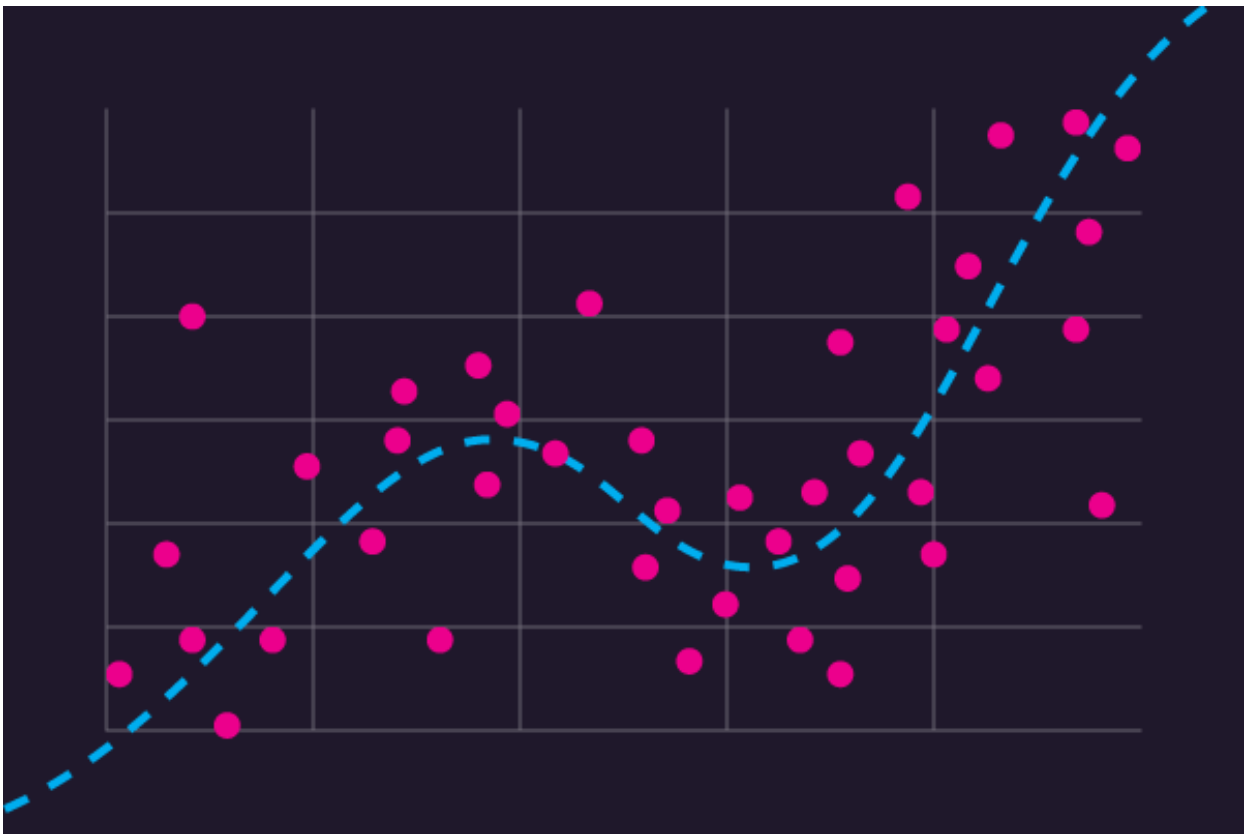

Fig 1

Lasso Regression

30.11.2022

Harpreet Singh

2020MCB1237

Mathematics and Computing 2020

Foundation of Data Science MA515

Project

Do exploratory data analysis on the data. Use regression techniques

to predict the salary of baseball players. Use the lasso technique also

and compare the results

Goals

1. Performing regression using both ridge regression and lasso regression and comparing two models.
2. Performing ridge and lasso for different alpha values to get optimal alpha for both models.
3. Studying the relationship between data points using correlation heatmap and lasso regression coefficients

Lasso vs Ridge Regression

Ridge Regression

Ridge regression is the method of estimating coefficients with a high correlation between two or more data points (dependent on each other). This method performs L2 regularization; this method penalizes the coefficients with the summation of squared coefficients. $H_{ridge} = X(X'X + \lambda I)^{-1}X$, based on minimizing the ordinary least square (OLS) technique.

Lasso Regression

Lasso Regression is the same as ridge regression. The difference is that it does not add squared terms of coefficients but instead adds the absolute value of coefficients. This is the reason it is called L1 regularization. This equation is solved using various numerical techniques. Lasso regression not only finds the model's coefficients but acts as a **Variable**

selection model since it drops the coefficients to 0 directly, eliminating them from the model.

Observations

I. Alpha values for Lasso Regression vs R^2 value

Following is the graph produced by lasso regression using different alpha's.

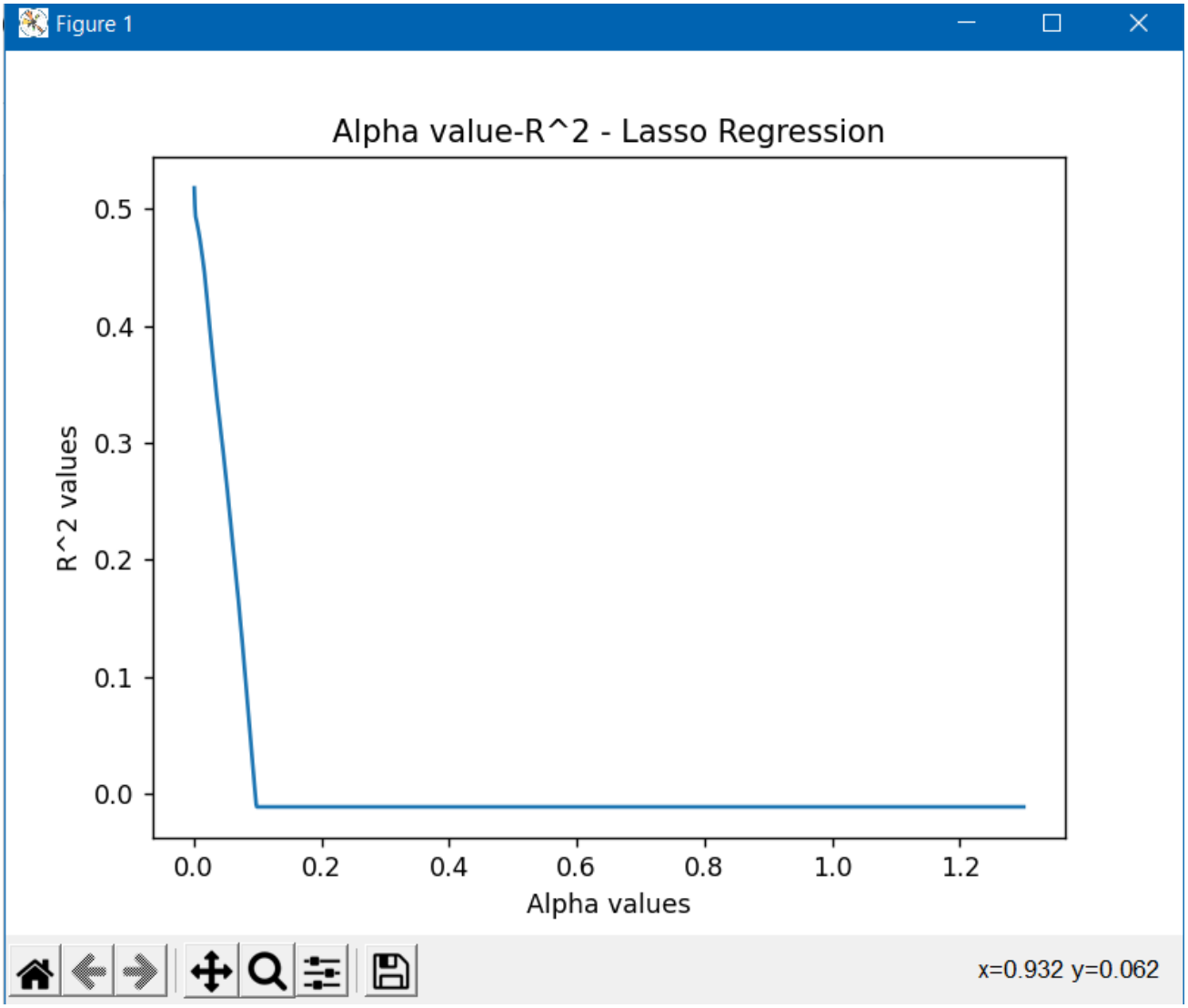

Fig 2

II. Alpha values for Ridge Regression vs R^2 value

Following is the graph produced by lasso regression using different alphas.

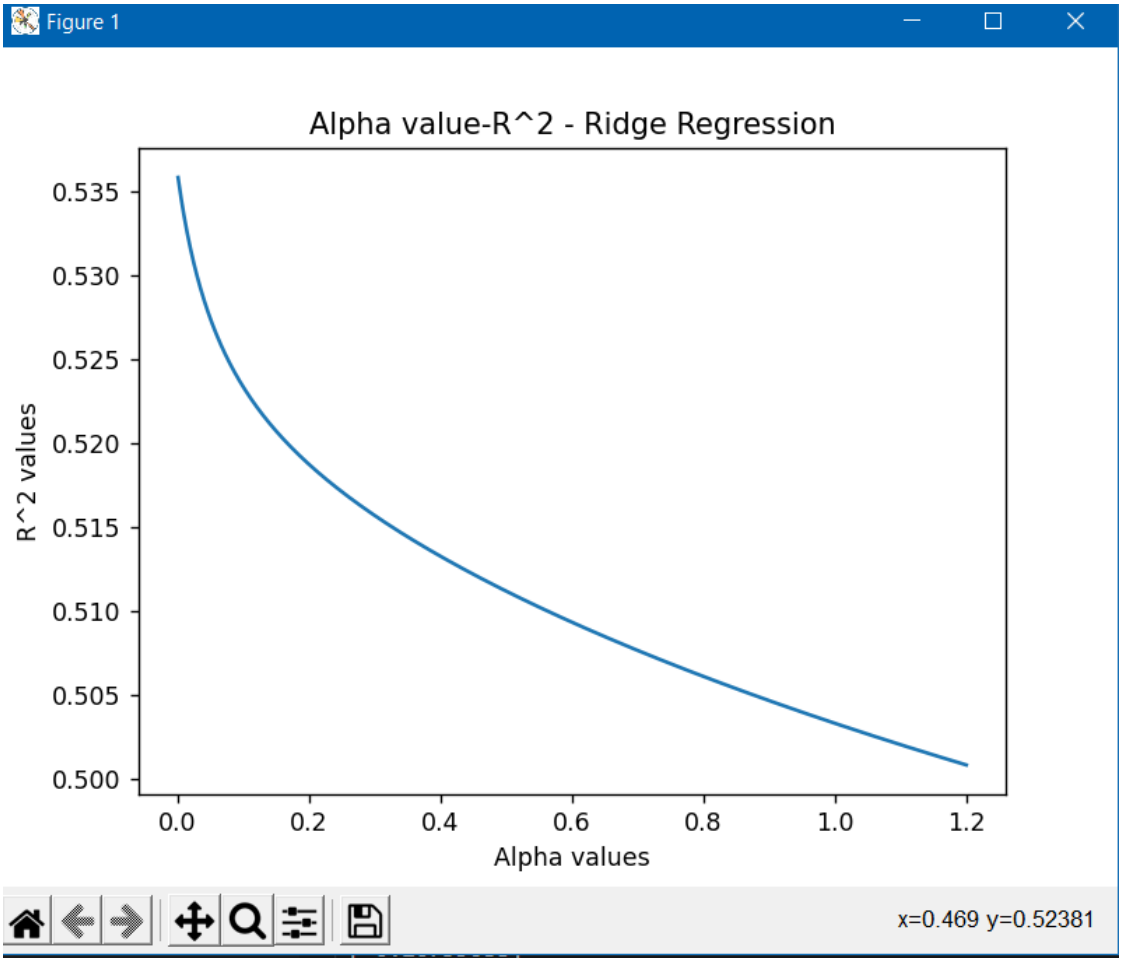

Fig 3

III. Coefficients given by Lasso Model, "Variable Selection"

```
This is optimal lambda for lasso regression : 0.001 for getting best score of 0.5178865209879437
[-1.59815423e-01  6.76432302e-18  3.98160802e-01 -1.58717593e-17
 -0.00000000e+00  0.00000000e+00 -1.43378615e+00  1.70910655e+00
  3.75762732e-01  0.00000000e+00  1.69789191e-02  9.68186282e-01
 -1.86096705e-01 -0.00000000e+00  5.27266252e-01  0.00000000e+00
  2.28508117e+00  7.63330420e-01 -9.63783932e-01  6.24008539e-01
  3.06368646e-01 -3.02127237e-01]
```

Fig 4

As we can observe, some of the coefficients have become 0, hence will not affect the results anymore; this is the specificity of the Lasso technique as compared to Ridge regression, where coefficients do not become 0 and instead get some non-zero value.

IV. Dependency of data attributes suggested by Correlation heatmap.

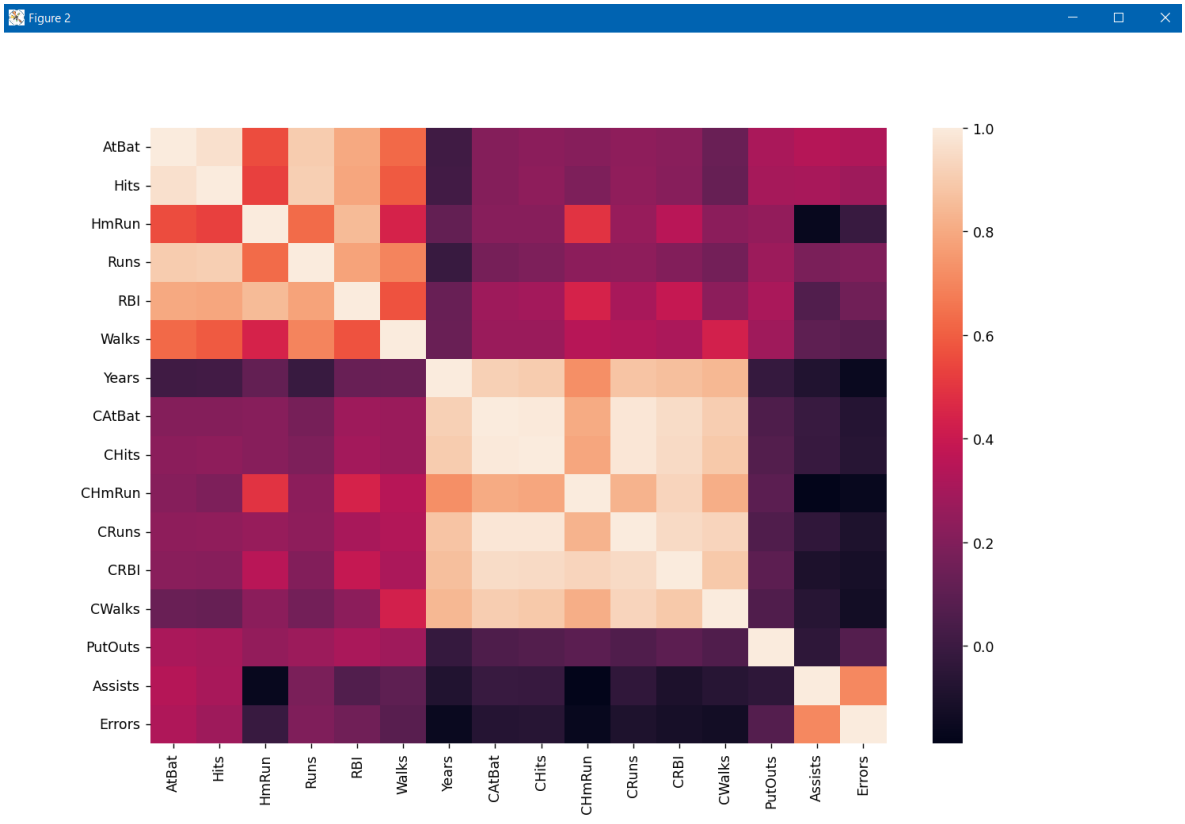

Fig 5

Similar to the heatmap suggestion, some variables will be dropped automatically while performing regression (coefficient = 0) with high correlation, as suggested by the above Correlation heatmap.

V. Data summary

Data summary is obtained by using data.describe() command that provided statistical variables like standard deviation, mean, 25,50,75th percentiles, max, min, etc

VI. Predictions obtained so far.

```
These are the predicted salaries
[ 573.46407965  946.32772995    8.70251583  797.84088606  533.89008437
  282.10171036 1032.3647321   435.69468392 1158.85314111  928.12677739
  539.03857517  879.80953872  429.98796656  825.17638817   89.73810954
  632.11451295  140.08297959  873.44754264  899.03148342  198.16064264
  132.27387418  384.90311781  760.98843942   92.52516658  148.8421107
  410.82700696  756.53631855  264.71538291  249.89871112  567.60646313
  363.14926246  747.88897037   95.30468283  183.34012906  563.77444236
  416.4833679   606.86414186  493.47877698 1119.22068683  360.89286217
 1093.89167976  193.42005886  181.84739455  303.86906662  687.94753121
 1063.44235408  302.15318677  619.82710556  980.62824868 1178.73364891
  479.54228524  749.18153679  201.50274331]
```

Fig 6

These salaries were obtained in lasso regression as X_test as input. After multiplying with standard deviation and adding the mean (obtained earlier in data manipulation), we obtained these salaries.

References:

- Fig 1: https://www.imsl.com/blog/what-is-regression-model (image source)
- ISLR book for definitions
- Matplotlib library for producing Fig 2, Fig 3, Fig 4, Fig 5