COMP4211 PA1

20316792

Shayan Imran

2019

# Empirical Study on Linear Regression

## Regression Metrics

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | MSE Training Set | MSE Test Set | R2 Training Set | R2 Test Set |
| Fifa | 3.373700e-03 | 3.305544e-03 | 0.838373 | 0.842251 |
| Finance | 9.557832e-12 | 9.462171e-12 | 1.000000 | 1.000000 |
| Orbits | 5.115185e-03 | 4.999975e-03 | 0.686264 | 0.695042 |

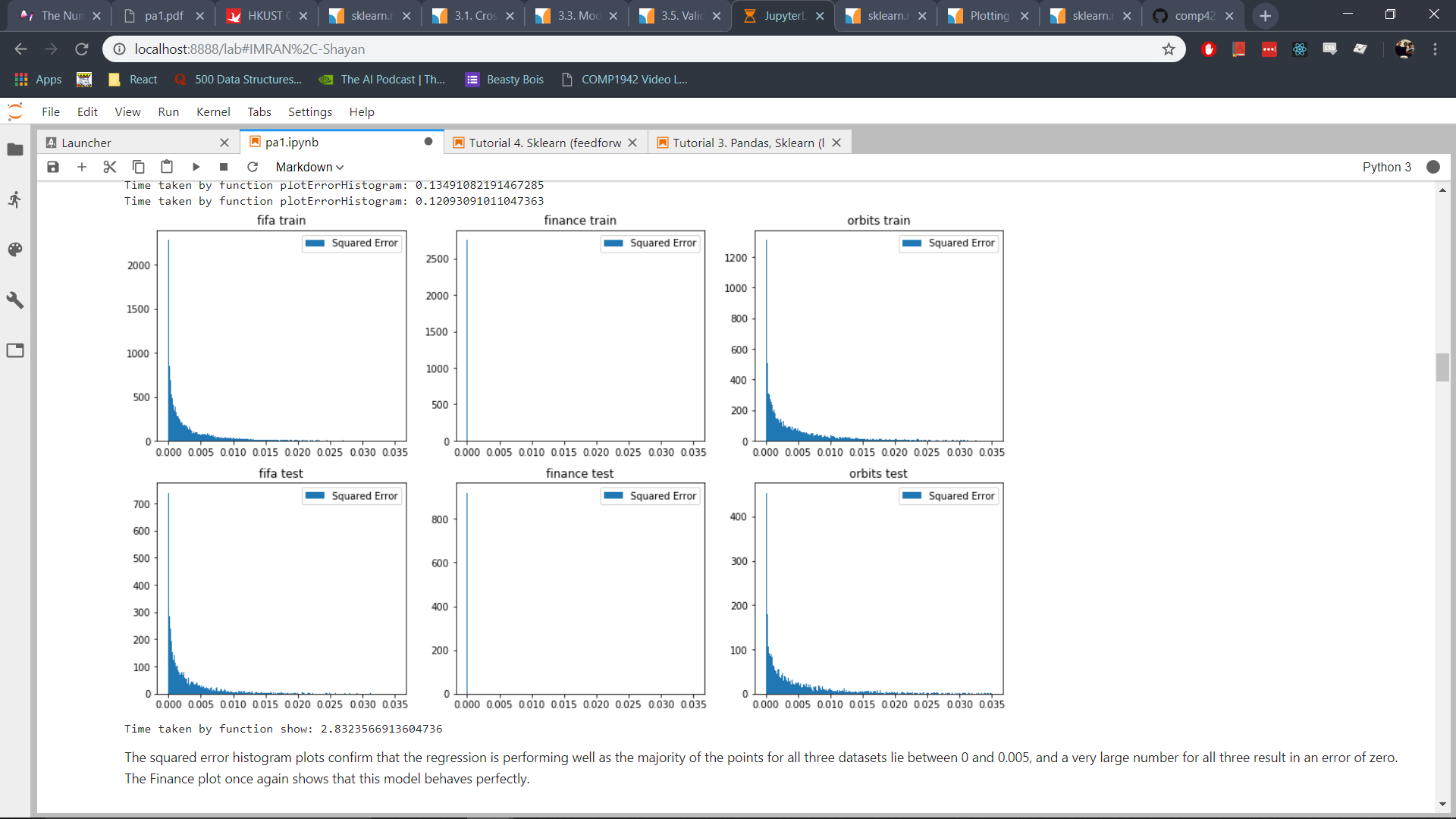
As can be observed from the table, linear regression results in a very low mean squared error for the Finance training and test data, and a perfect r2 score suggesting the regression perfectly fits the data. The Fifa dataset fits the data well with a low mean squared error and r2 scores that are higher than 80% on both the training set and the test set. The Orbits dataset has a mean squared error and an r2 score that is less than the other two datasets indicating lower performance. The r2 scores on their own may not be enough, however, to check the performance of the model. For this reason, residual vs fitted value scatter plots for the regression are provided below for further insight.

## Residual vs Fits



The Finance dataset is perfectly modelled so all the points lie on the horizontal axis meaning that the predicted and actual values are all the same within a very small margin of error. This further proves that this dataset works extremely well with linear regression. The residual vs fits plots for the Fifa datasets shows that the residues are somewhat randomly distributed across the horizontal axis, which means the data is appropriate for modelling with linear regression. The plot for Orbits, however, does not show a very randomly distributed pattern, with y values showing an increasing trend from 0 to 0.5, and then a decreasing trend from 0.5 to 1. This indicates that linear regression may not be appropriate to model this data, a non linear model may have better performance.

## Squared Error Histograms



The squared error histogram plots confirm that the regression is performing well as the majority of the points for all three datasets lie between 0 and 0.005, and a very large number for all three result in an error of zero. The Finance plot once again shows that this model behaves perfectly.

# Empirical Study on Logistic Regression

## Model Settings

The regression was performed using sklearn.linear\_model.SGDClassifier as this performs gradient descent for optimization. The following parameters are set for the model:

1. loss = ‘log’ *(This is set to perform logistic regression using SGDClassifier)*
2. max\_iter = 10,000
3. tol = 1e-5
4. random\_state = 1 *(To ensure reproducibility)*
5. learning rate = ‘constant’ *(Learning rate does not change during optimization)*
6. eta0=0.0011 *(This is the step size parameter, this value produces stable convergence in all models)*