COMP4211 PA1

20316792

Shayan Imran

2019

# Empirical Study on Linear Regression

## Regression Model

The regression was performed using sklearn.linear\_model.LinearRegression with the following parameters:

Fit\_intercept = True

The time taken by the regression is:

|  |  |
| --- | --- |
| Dataset | Time taken(seconds) |
| Fifa | 0.019989490509033203 |
| Finance | 0.0029990673065185547 |
| Orbits | 0.003998994827270508 |

## 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 |

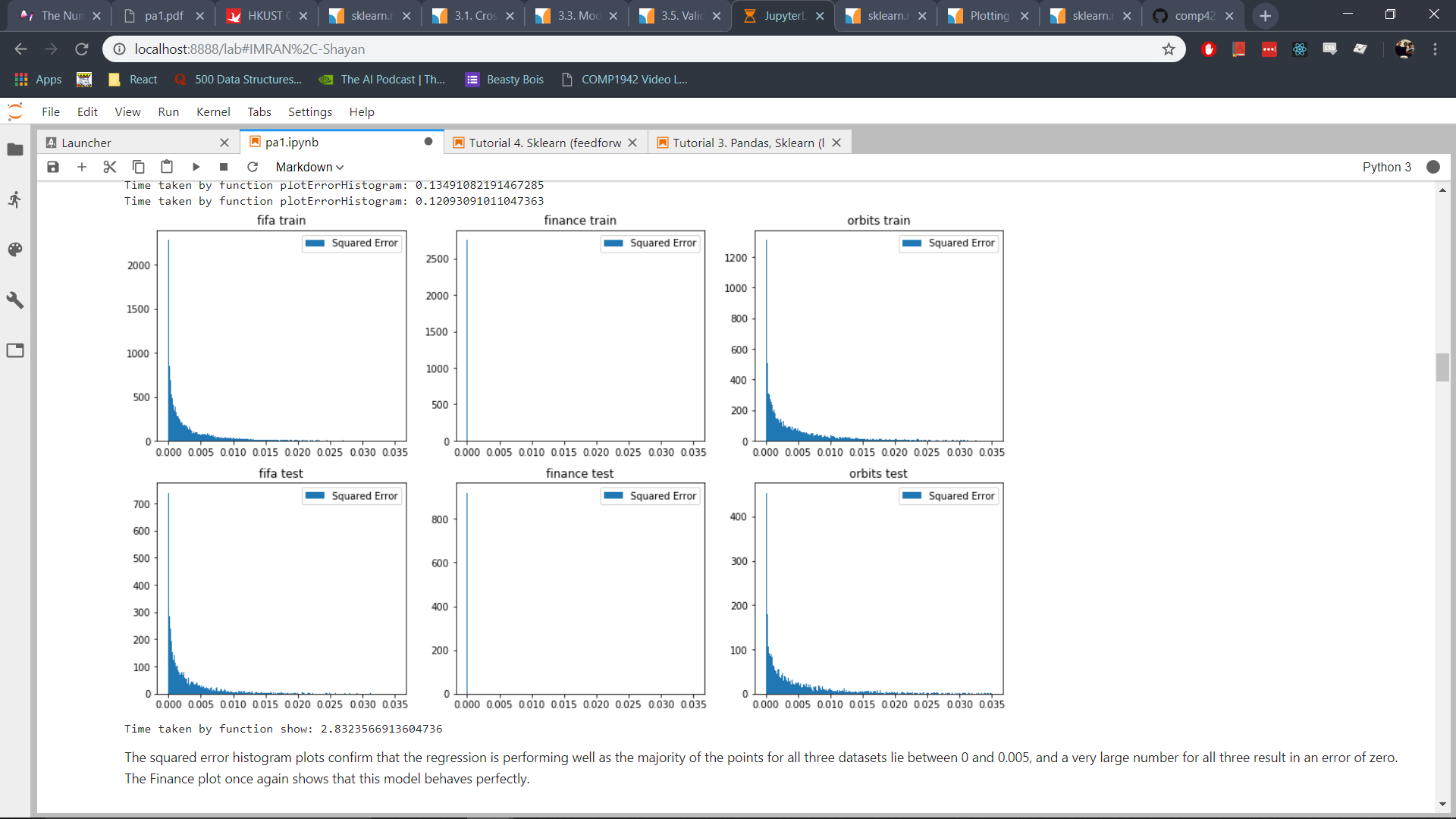
As can be observed from the table, linear regression results in a very low mean squared error for the Finance training and test data, and a perfect r2 score suggesting the regression perfectly fits the data. The Fifa dataset fits the data well with a low mean squared error and r2 scores that are higher 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.

Time taken to plot the histograms:

|  |  |
| --- | --- |
| Function | Time taken(seconds) |
| Plot Fifa Test | 0.11893081665039062 |
| Plot Finance Test | 0.11691975593566895 |
| Plot Orbits Test | 0.11493349075317383 |
| plt.show() | 2.750332832336426 |

# 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)*

The time taken for logistic regression is:

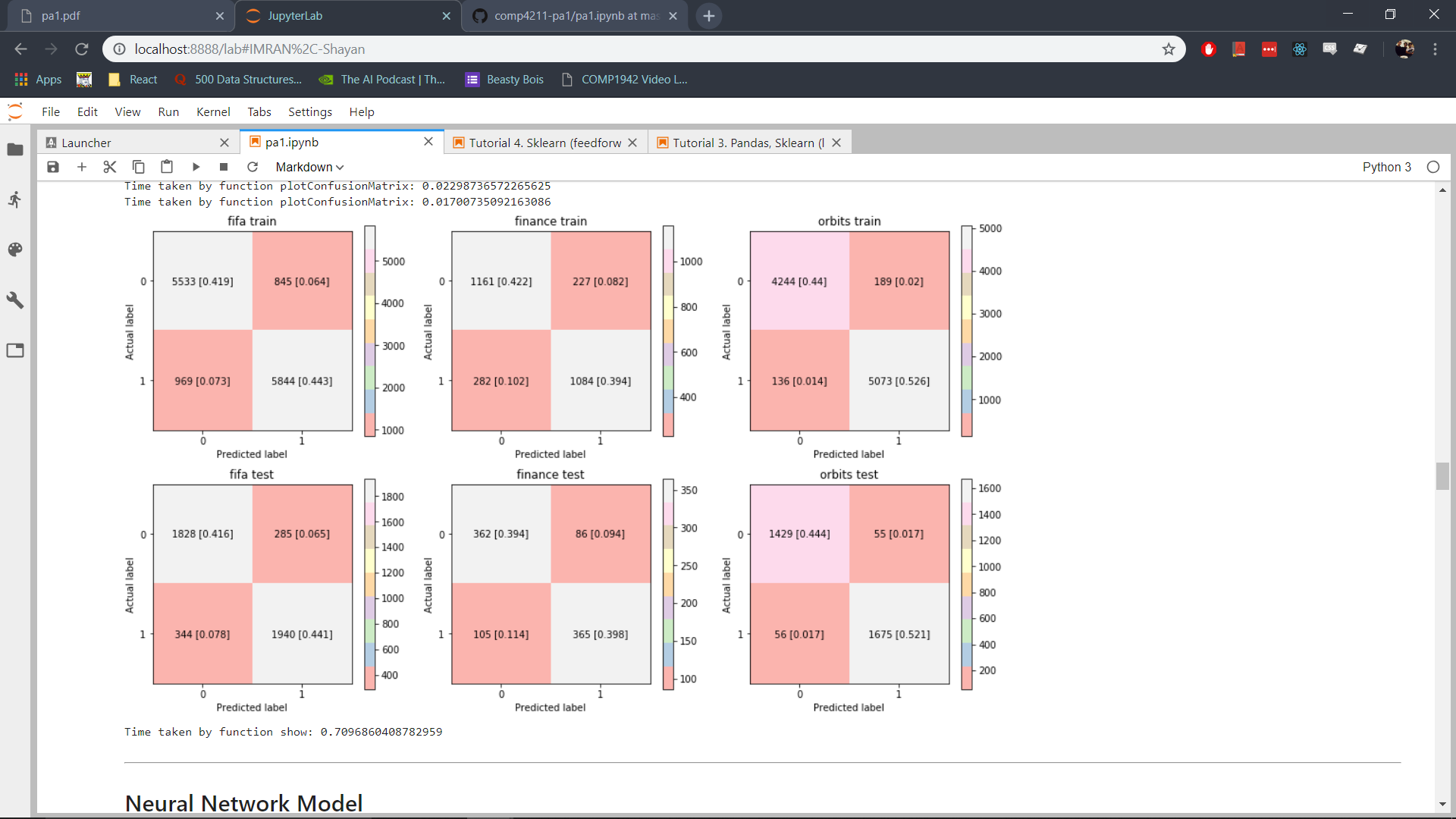
|  |  |
| --- | --- |
| Dataset | Time taken(seconds) |
| Fifa | 0.4976990222930908 |
| Finance | 0.21635055541992188 |
| Orbits | 1.5040926933288574 |

## Classification Metrics

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | Accuracy Training Set | Accuracy Test Set | AUC Training Set | AUC Test Set | Precision Test Set | Recall Test Set |
| Fifa | 0.862482 | 0.856948 | 0.862643 | 0.857254 | 0.871910 | 0.849387 |
| Finance | 0.815178 | 0.791939 | 0.815007 | 0.792316 | 0.809313 | 0.776596 |
| Orbits | 0.966293 | 0.965474 | 0.965628 | 0.965293 | 0.968208 | 0.967649 |

The logistic regression training results show that the Orbits dataset was modelled well with accuracy greater than 96% on both the training and test set. The model performs somewhat well for the Fifa dataset with accuracy on the training and test set higher than 85%. On Finance the accuracy was lower, 81% on the training set and 79% on the test set showing that there may not be enough training samples (2754 data points) to properly model the Finance dataset.

## Experiment Results Visualization



Time taken to plot the matrices:

|  |  |
| --- | --- |
| Function | Time taken(seconds) |
| Plot Fifa Train | 0.04197406768798828 |
| Plot Finance Train | 0.015993118286132812 |
| Plot Orbits Train | 0.02298736572265625 |
| Plot Fifa Test | 0.01798105239868164 |
| Plot Finance Test | 0.01699066162109375 |
| Plot Orbits Test | 0.01700735092163086 |
| plt.show() | 0.7096860408782959 |

# Empirical Study on Neural Network Model

## Model Settings

The neural network was modelled using sklearn.neural\_network.MLPClassifier. The following parameters are set for the model:

1. activation function = ‘relu’
2. random\_state = 0
3. solver = ‘sgd’ *(stochastic gradient descent)*
4. learning\_rate\_init = 0.01
5. learning rate= ‘adaptive’
6. tol = 1e-5

## Parameter Tuning using Cross Validation

Grid scores on development set:

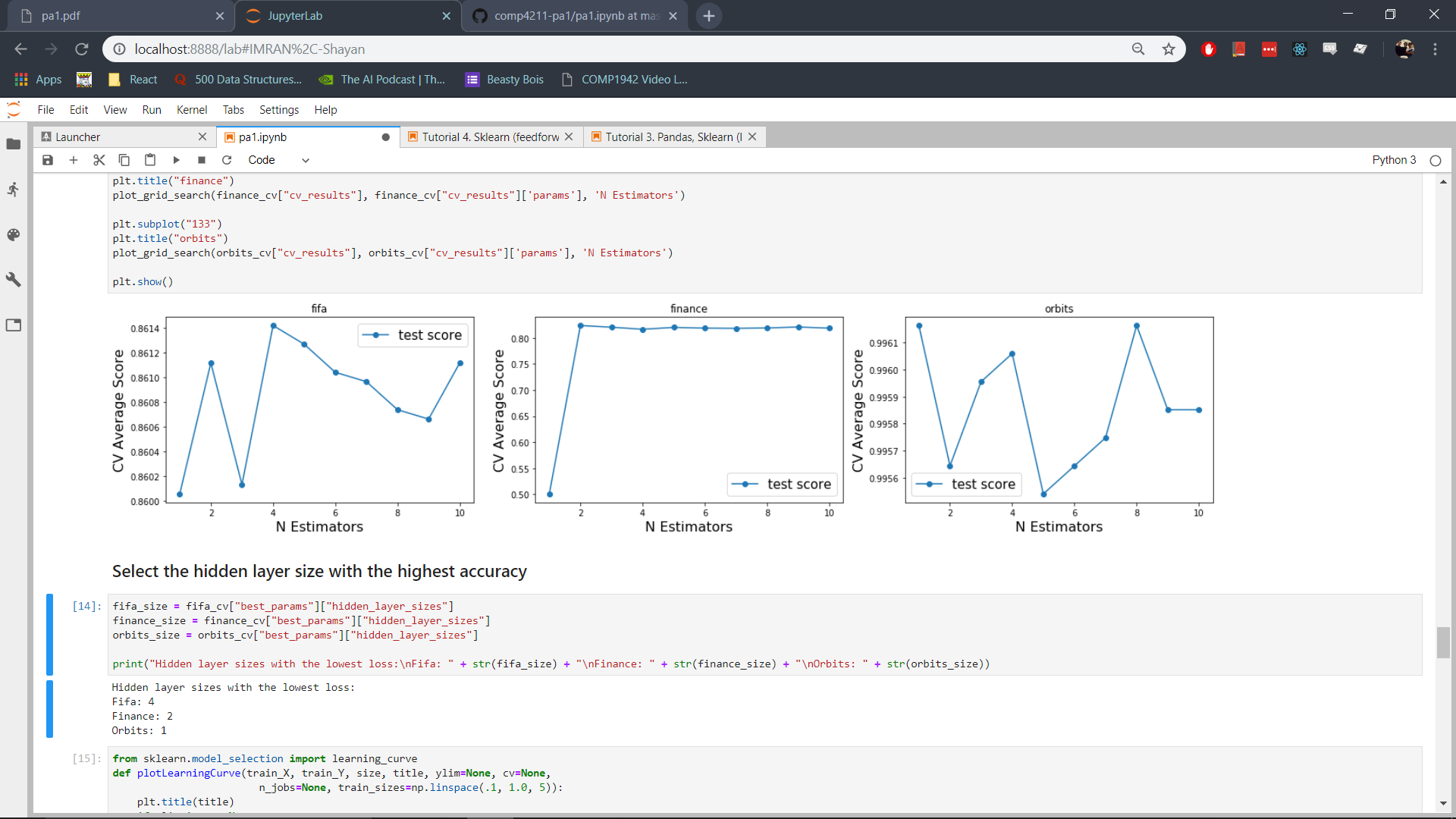
|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Fifa | | Finance | | Orbits | |
| H\* | Score | H\* | Score | H\* | Score |
| 1 | 0.860 (+/-0.012) | 1 | 0.501 (+/-0.007) | 1 | 0.996 (+/-0.004) |
| 2 | 0.861 (+/-0.013) | 2 | 0.824 (+/-0.014) | 2 | 0.996 (+/-0.005) |
| 3 | 0.860 (+/-0.011) | 3 | 0.821 (+/-0.010) | 3 | 0.996 (+/-0.004) |
| 4 | 0.861 (+/-0.014) | 4 | 0.817 (+/-0.008) | 4 | 0.996 (+/-0.004) |
| 5 | 0.861 (+/-0.013) | 5 | 0.820 (+/-0.006) | 5 | 0.996 (+/-0.004) |
| 6 | 0.861 (+/-0.013) | 6 | 0.819 (+/-0.011) | 6 | 0.996 (+/-0.004) |
| 7 | 0.861 (+/-0.013) | 7 | 0.818 (+/-0.012) | 7 | 0.996 (+/-0.004) |
| 8 | 0.861 (+/-0.013) | 8 | 0.819 (+/-0.014) | 8 | 0.996 (+/-0.004) |
| 9 | 0.861 (+/-0.012) | 9 | 0.821 (+/-0.010) | 9 | 0.996 (+/-0.004) |
| 10 | 0.861 (+/-0.012) | 10 | 0.819 (+/-0.016) | 10 | 0.996 (+/-0.004) |

Hidden layer sizes with the highest accuracy:

Fifa: 4

Finance: 2

Orbits: 1

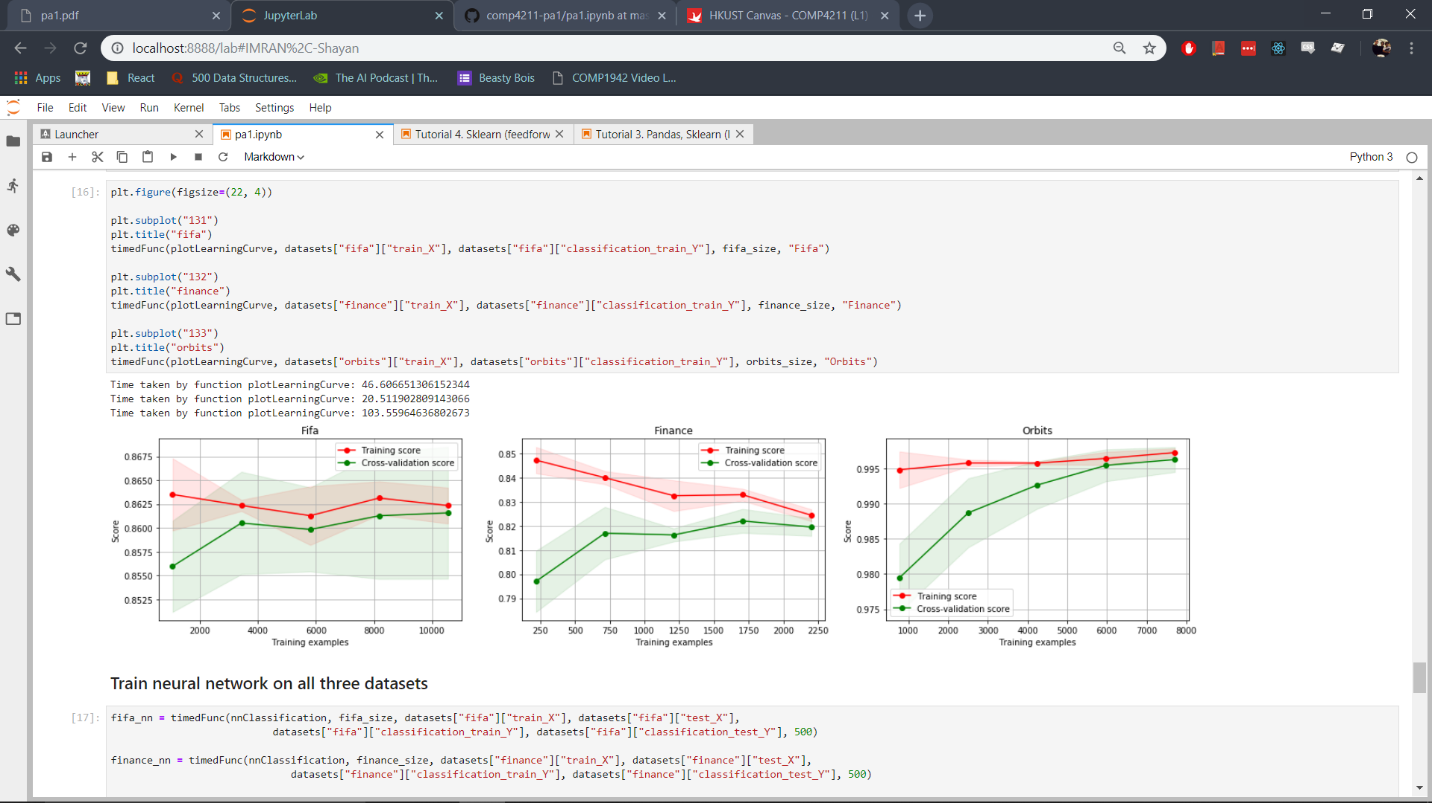


The time taken to perform cross validation is:

|  |  |
| --- | --- |
| Dataset | Time taken(seconds) |
| Fifa | 157.6978862285614 |
| Finance | 50.26172637939453 |
| Orbits | 370.3014750480652 |

## Neural Network Classification Metrics

The network is trained to ensure that overfitting does not occur and the results of the training are visualized below.



These learning curves show that overfitting does not occur because the cross validation score increases with increasing training examples when training is occurring. While training on the Fifa dateset the validation score increases while the training score decreases, this shows that the model is learning to generalize to perform better on newly observed data. For Orbits both the training score and the validation score increase over time showing that the model is performing well on new data. For Finance the validation curve also follow an upwards trend as more training data is used except for the last data point, which results in a very small decrease. The training score value also decreases for the last point in the Finance dataset so this decrease should not be a cause of concern for overfitting. All the models show a trend of the validation score and the test score reaching almost the same values which means that overfitting is not occurring in the model.

The time taken to train the neural networks to get validation and loss scores at different training example sizes is:

|  |  |
| --- | --- |
| Dataset | Time taken(seconds) |
| Fifa | 157.6978862285614 |
| Finance | 20.511902809143066 |
| Orbits | 103.55964636802673 |

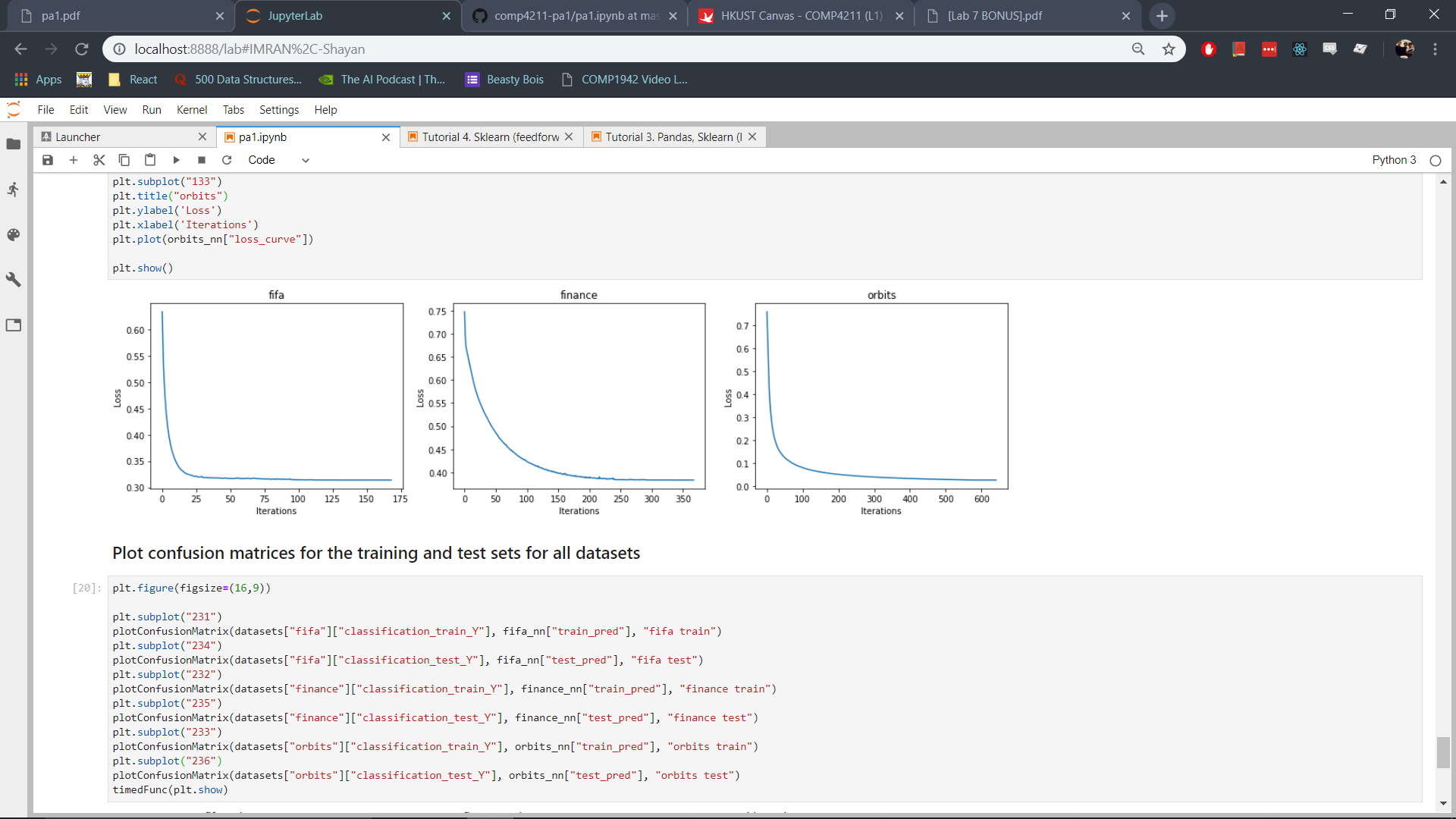
|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | Score Training Set | Score Test Set | AUC Training Set | AUC Test Set | Loss | Precision  Test  Set | Recall Test Set |
| Fifa | 0.862785 | 0.856266 | 0.862936 | 0.856385 | 0.314037 | 0.867765 | 0.853327 |
| Finance | 0.830792 | 0.811547 | 0.830717 | 0.811517 | 0.383500 | 0.817987 | 0.812766 |
| Orbits | 0.997200 | 0.996267 | 0.997139 | 0.965293 | 0.027944 | 0.997684 | 0.995378 |

The Orbits dataset was modelled very well with scores greater than 0.995 in both training and test sets resulting in very accurate predictions. The Fifa dataset also performs somewhat well with scores, precision and recall around 0.85 which means the model is able to classify a large number of values correctly. The Orbits dataset has a lower test score at around 0.81 signifying that more data may be needed for a better predictive model.

The time taken to train the neural networks is:

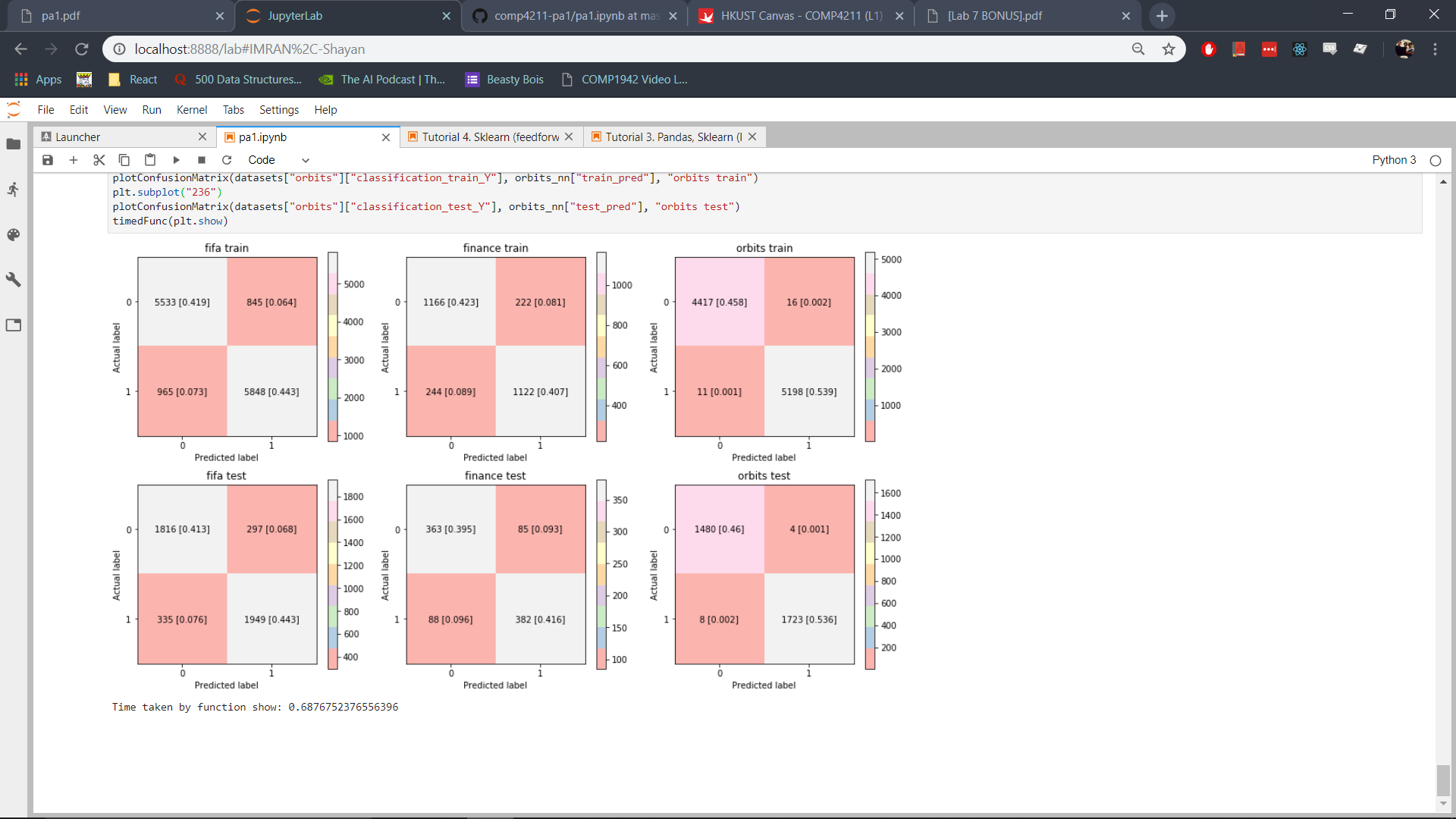
|  |  |
| --- | --- |
| Dataset | Time taken(seconds) |
| Fifa | 3.7313246726989746 |
| Finance | 1.3828811645507812 |
| Orbits | 5.812797546386719 |

## Experiment Results Visualization



The curves above represent the performance of the model over time, the loss of the model is on the y axis and the iterations on the x axis. All the models stabilized to an optimum value over time, with each reaching a different loss value in varied iterations.

For the Fifa dataset, the initial loss started at a value of greater than 0.6 but it experienced a steep decline, reaching around 0.32 in only 20 iterations and then stabilized to 0.314 after 150 iterations. The Finance dataset also experienced a decline in 150 iterations from 0.75 to 0.38 showing that there is not enough data to produce a model with high accuracy. The Orbits dataset on the other hand results in a very good model which results in a very low loss of 0.02794.



From the confusion matrices it is easy to see that the false positive and false negative values for Fifa are lower than the ones in the Finance dataset

# Computing Environment

OS: Windows 10 Home 64 Bit  
OS Version: 1803  
CPU: Intel Core i7-4720HQ  
GPU: Nvidia GeForce GTX 960M  
RAM: 16 GB  
Programming Language: Python 3.7.1  
Programming Environment: JupyterLab 0.35.3

# Comparative Analysis

It is difficult to compare the qualities of the linear regression model with the classification models as they are both performing different tasks. However, we can comment on the data given for both models and how appropriate it is for the tasks. It must be highlighted that the actual output data provided for both regression and classification is different, with regression giving a number in a range, while the binary classifiers produce either 0 or 1.

For the regression datasets, the Finance dataset works very well with linear regression, fitting the data almost perfectly, and the Fifa dataset is also performant. The Orbits dataset, however, produces comparatively poorer performance to the two as is seen by its R2 score of 0.69 while Finance has 1.0 and Fifa has 0.84. Analyzing the residuals vs fits curve also shows us that a non linear model may be more appropriate for the Orbits dataset.

For the classification datasets, however, the Orbits dataset performed very well in both logistic regression (an accuracy of 0.965) and neural network modelling (an accuracy of 0.996). This shows that the amount of data as well as the distribution of data for training of the model was adequate to produce impressive results during testing. The Fifa dataset once again produces good results with an accuracy of around 0.86 for both classifiers. The Finance dataset, however, does not produce results as good as those produced in regression. The accuracy for logistic regression is 0.79 while for the neural network model it is 0.81. This may be because the training data provided to train this model is not enough, so more data may be required for better accuracy. These results must be taken with a grain of salt however, as the performances on the test sets may not be indicative of values encountered in actual usage of the model. The performance of the model depends on the distribution of the training and test data, and whether it is a representative sample of the problem that is being modelled.

Comparison between regression and classification beyond this would be difficult to perform due to the large difference in the problem they are solving. We can, however, compare the performance of the logistic regression and the neural network models.

|  |  |  |
| --- | --- | --- |
|  | Logistic Regression Accuracy | Neural Network Accuracy |
| Fifa | 0.856948 | 0.856266 |
| Finance | 0.791939 | 0.811547 |
| Orbits | 0.965474 | 0.996267 |

|  |  |  |
| --- | --- | --- |
|  | Logistic Regression AUC | Neural Network AUC |
| Fifa | 0.857254 | 0.856385 |
| Finance | 0.792316 | 0.811517 |
| Orbits | 0.965293 | 0.965293 |

|  |  |  |
| --- | --- | --- |
|  | Logistic Regression Precision | Neural Network Precision |
| Fifa | 0.871910 | 0.867765 |
| Finance | 0.809313 | 0.817987 |
| Orbits | 0.968208 | 0.997684 |

|  |  |  |
| --- | --- | --- |
|  | Logistic Regression Recall | Neural Network Recall |
| Fifa | 0.849387 | 0.853327 |
| Finance | 0.776596 | 0.812766 |
| Orbits | 0.967649 | 0.995378 |