**CASE STUDY CROSS COMPARISON**

***EXECUTION TIME***

All the mentioned time are in milli seconds

All the models were ran on the same machine in jupyter notebook.

The execution time were changing every time we execute it but it was not varying by a huge margin hence margin. We have taken the most recent ones for this comparison.

CASE STUDY 1 – Dataframes\_Defaults

CASE STUDY 2 – Dataframes\_Handling\_imbalanced\_data

CASE STUDY 3 – RDD

|  |  |  |  |
| --- | --- | --- | --- |
|  | CASE STUDY 1 | CASE STUDY 2 | CASE STUDY 3 |
| Logistic Regression | 1579ms | 4634ms | L1 – 1795ms L2 – 247ms |
| Decision Tree | 6882ms | 4730ms | 1481ms |
| Random Forest (500 trees) | 22949ms | 22098ms | 70520ms |
| Gradient-Boosted Tree | 9647ms | 7213ms | 29419ms |
| Naïve Bayes | -- | -- | 438ms |
| Linear SVM | -- | -- | L1 – 2735ms L2 – 1831ms |

*Analysis:*

It was observed that RDD version of Logistic regression with L2 regularization was taking very less time for execution compared to its Data frame version. Interestingly when we ran added weighted column to the data from in case study 2 logistic regression took more time for execution.

Clearly for decision tree it’s RDD version was taking the least time compared to it’s data frame siblings.

Thinks were a bit different for Random Forest and Gradient Boosted trees, for them data frames were having lesser execution time than their RDD versions.

Another interesting observation from the table above is that among RDD models if we have L1 and L2 versions, always L2 has lesser execution time than L1.

**NOTE:** Comparison between various models, results and observations are included in Jupyter Notebook itself. A final conclusion for each case study has also been included in the respective notebook files.

CASE STUDY 2 has a comparison with CASE STUDY 1 and CASE STUDY 3 has a comparison with CASE STUDY 2 and CASE STUDY 1 in it’s conclusions. Hence CASE STUDY 3’s conclusion is the final cross comparison conclusion.