Estimating transfer times of large datasets for scientific computing

IRIS -HEP Fellow - Oleksii Brovarnyk

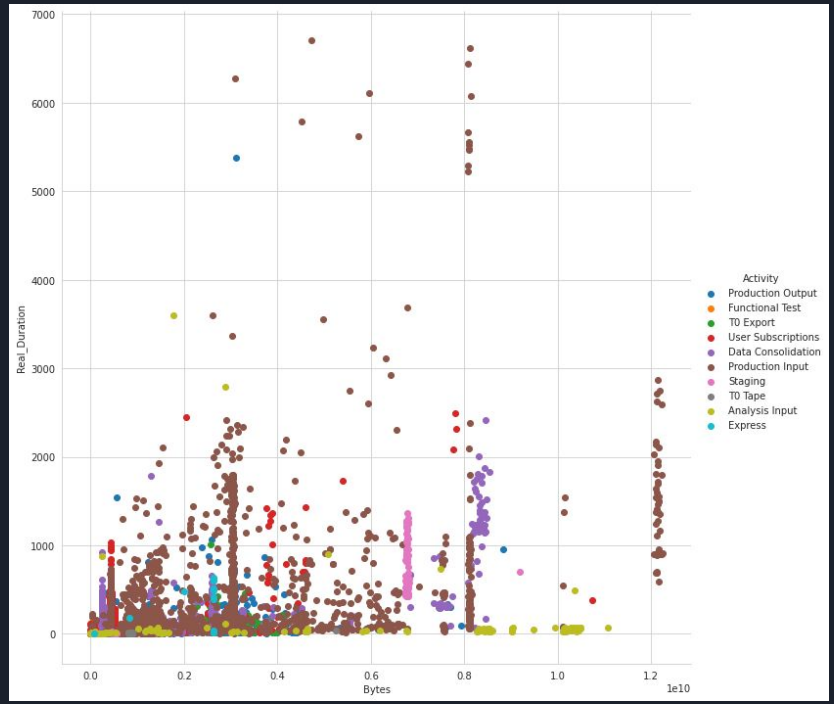
Mentor - Mario Lassnig (CERN)

Project description

This project will continue the existing research of the Rucio team on **the estimation of the duration of file transfers** for large scale sciences. The distributed data management environment for scientific experiments forms a complex ecosystem with dynamic interactions between users and data centers. Rucio's central role as the data management system, and the rich amount of data gathered about the transfers and data rules life cycles will help in creating machine learning for transfer time estimation.

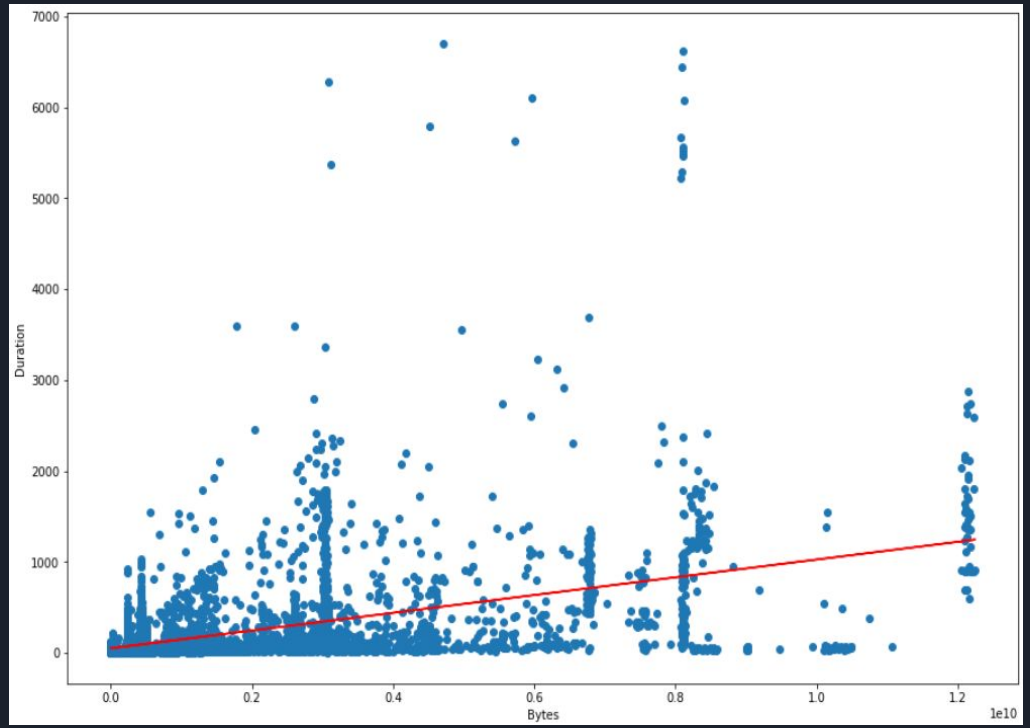
Project description

To start, the research needs data on successful file transfers. This is based on the Rucio events **(event_type: transfer-done)** which are available in the Rucio Elasticsearch instance. The idea is then to generate **time series** from this data, including event metadata like **"started_at"** (when did the transfer start), **"transferred_at"** (when did the transfer finish), **"bytes"** (how big was the file), and many more. There are more than 30 different variables that can be used for this model.



Results achieved

- To start the estimation, the inputs for a first model was the **number and size of files**, and the outputs was the **duration in seconds**. From this, a first **linear regression** model was learned, which should answer the question: For a given dataset (number of files, and gigabytes), **how long will it take for this dataset to finish transferring**.
- After this first model, we enter into a cycle, which we use **to add (or remove) variables** from the available input data to see **which ones improve the prediction the best**, and validate against the history for correctness. We will then repeat this cycle as often as possible to get the best possible prediction.

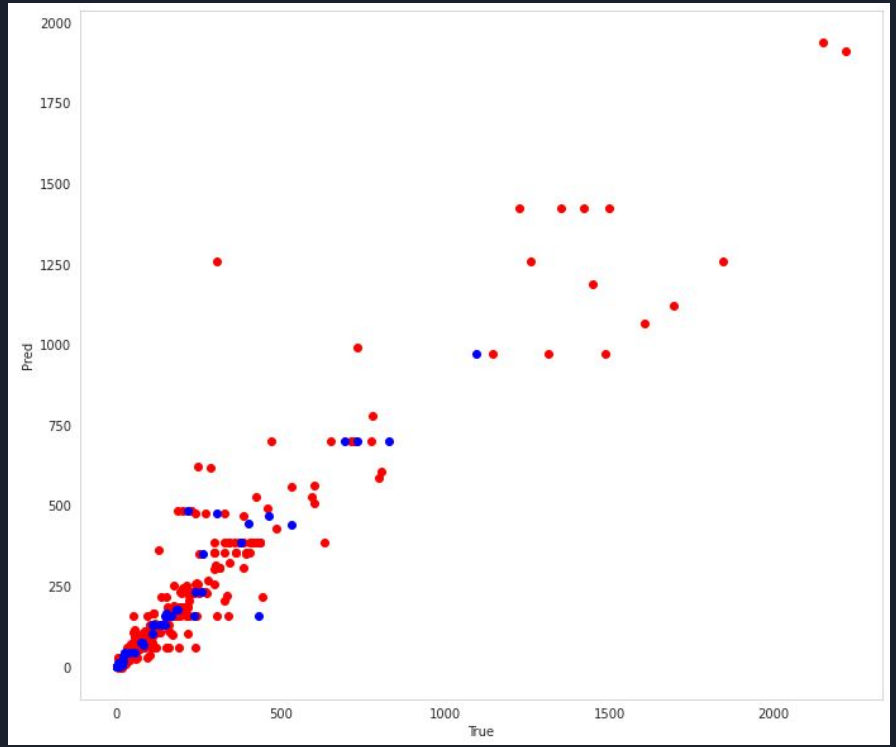


Results achieved

We made a model with tensorflow and keras and achieved a good result in predicting the data.

The **R2_Score metric shows 93 percent** model accuracy, and the **root-mean-square error shows 60 seconds**, which is a good result.

At the moment we are still working on improving the results.



Thanks for listening!