Feature engineering

Checking the description of the features, we learn that a jet is a type of pseudo particle that is created as a result of the collision of other particles. We decide to categorise the data depending on the number of jets so that for each subset the corresponding undefined features dependeing on the jet number can be removed. Moreover, for every subset there are also Nan values in the feature *DER\_mass\_MMC* and we replace them by the corresponding mean. In conclusion, we end up with four data-sets without any Nan values.

In addition, we also extend our feature vector by adding a polynomial basis of degree d for increasing the performance of the linear regression

|  |  |  |
| --- | --- | --- |
| Features | Jet 0 | Jet 1 |
| *DER\_deltaeta\_jet\_jet* | *Undef.* | *Undef.* |
| *DER\_mass\_jet\_jet* | *Undef.* | *Undef.* |
| *DER\_lep\_eta\_centrality* | *Undef.* | *Undef.* |
| *PRI\_jet\_leading\_pt* | *Undef.* |  |
| *PRI\_jet\_leading\_eta* | *Undef.* |  |
| *PRI\_jet\_leading\_phi* | *Undef.* |  |
| *PRI\_jet\_subleading\_pt* | *Undef.* | *Undef.* |
| *PRI\_jet\_subleading\_eta* | *Undef.* | *Undef.* |
| *PRI\_jet\_subleading\_phi* | *Undef.* | *Undef.* |

Undefined features in the data set per jet and per features