**Kaggle competition: Mercedez-Benz greener manufacturing**

**Goal:**

* How do different settings of car features (and their resulting combinations) affect the time taken for reliability and safety testing?
* Make a predictive model that takes in car features, and their respective settings, to predict testing time.

**Data attributes:**

* Anonymised
* Only binary and categorical
* Small data set relative to features
* Data set seems high qual.

**Summary:**

* Abstract like summary of points below

**Challenges faced:**

* Small data set 400:4000, features to samples. Danger of overfitting.

**Feature selections/engineering:**

* Most important features?
  + X0,
* How are you selecting features?
* Are you make any important transformations?
* Are you finding any interesting interactions between features?

**Modelling and Training:**

* What models are you considering?
* What training methods are you using?
* Are you ensembling models, and if so, how did you weight?

**Lessons learnt / interesting findings:**

* **Why is feature selection/extraction important?**
* Feature selection is important for three mean reasons:

1. Curbs complexity (variance reduction by shrinking the hypothesis space),
2. Increases accuracy (by getting rid of irrelevant features, which we could consider as noise, and that may confuse the algorithm),
3. Reduces training time (lass data, less training time)

* **Feature selection vs dimensionality reduction?**
* Feature selection is selecting a subset of features based on usefulness with respect to predicting the target.
* Dimensionality reduction works by combining multiple features into a smaller set of features (e.g. principal components) that explain most of the variability of the original, larger feature set.
* **Why is it important to consider feature correlations?**
* Because we may have redundant features, which make a good case for dimensionality reduction.
* **Why are feature interactions important?**
* Because there may be an interdependence between features and the target. Specifically, if the effect of one feature on the target changes depending on the level of another feature.
* Note, dimensionality reduction (e.g. PCA) and finding feature interactions (e.g. factor analysis) are related, but not the same thing.
* **Clustering…**
* Useful for finding patterns in our data that share similar characteristics.
* **Describe the attributes of your data!**

Other useful bits:

* Univariate scatter plots, against a running index are useful for finding outliers.
* loc access / filter rows and columns by labels (e.g. index labels) or booleans (conditions).
* iloc access /filter rows and columns by integers / array concepts (see <https://www.analyticsvidhya.com/blog/2020/02/loc-iloc-pandas/>)

**References:**

1. First place competition solution: <https://www.kaggle.com/c/mercedes-benz-greener-manufacturing/discussion/37700>
2. EDA notebook: <https://www.kaggle.com/sudalairajkumar/simple-exploration-notebook-mercedes>