**Approach for the competition**

**EDA notes:**

* Skewness of dependent column (Price)
* Possible outliers in price, Levy, Mileage attributes (Data points that lie outside the 99th percentiles).
* Finally, by visualization of the outliers, managed to delete some from Price and Levy.
* Idea regarding duplicates in dataset.
* Matching ID’s in Train and Test set
* Low value counts on some production year attributes
* Having Engine volume and type, on single column
* Reason for special character in Levy column.

**Pre-processing steps:**

* Replacing the Missing values in Levy column with different strategy (MIN, MAX, MEAN, -1, 0)
* Replacing with 0 value worked well for all the models based on trial and error method
* Separating Engine volume and Engine Type
* Removing the outliers (based on trial and error)
* Tried scaling the data (didn’t work)

**Modelling:**

* Checking the performance across all the regression models
* Checking the performance across all the regression models by taking log transformation of dependent variable (Price)
* Selecting best performing models on baseline.
* Tuning the best performance models (Did not work well after tuning, so manually tuned for few parameters only)
* Selecting cross-validation strategy.
* Feature Engineering (aggregation features with group by)
* Making the subset of features for different models based on cross validation score
* Keep track of CV score (since for small variation in local CV, there was large variation on Public Leader-board)
* Selected 3 models’ Random forest, Extra Trees, Light GBM
* Made 3 subset of best features for 3 models.
* Check the feature importance plots each time.
* Check performance of individual models on Public LB (70% test data).
* Blend the models with weights based of Individual performance on Public LB

Due to limited submissions(3/day) couldn’t try some things

**Things for further try:**

* Frequency encoding the data
* Other possible transformations for dependent variable
* Categorical-to- Categorical feature creation.
* Tuning models for different subset of features.
* Generating predictions at each fold