1.            Briefly describe your approach to this problem and the steps you took

Aim: Classification Task to figure out the Engine Ratings on based on different engine features and status of car during inspection.

(1) Lot of categorical variables sparse in nature. 70+ Features and highly unbalance multi-class classification problem.

(2) Start the analysis with finding out the null count in all features. Majorly all variables have more than 50% null count. Dropping them will not be a good idea and fortunately, it is possible to fill in most of the blanks because 'NA' in the data below means 'None' or 0. Seeing the csv files, we can recognise the facts.

(3) Feature Engineering: Using Registration Date and Inspection Date calculate the Duration (in months )it took for the vehicle to report the issues and come for inspection .

(4) Detecting outliers on the Continuous Variable using Box plot and perform log transformation and outlier removal through the 95% confidence interval below.

(5) To handle highly unbalanced classes we use RandomOverSampler.

(6) Benchmark using XGBoost and then see the feature importance.

(7) Run Grid Search Over Parameter space to find optimal parameters.

(8) Mention cross validation as a means to evaluate the model.

2.            Basics:

a.       How well does your model work?

To validate a classification model we need to use different performance metrics. There's no relative measure of correctness. Instead we are either correct or not on a predicted label.

To measure the performance of a classification model we use a confusion matrix to get class wise performance.

b.       How do you know for sure that’s how well it works?

Use model to predict on cross-validation set and monitor deviation. It performs consistent on train and validation set.

c.        What stats did you use to prove its predictive performance and why?

Since it is a multi-class classification we generate classification report and use metric like precision and recall which give a better picture.

d.       What issues did you encounter?

Highly unbalanced classes makes it difficult to build unbiased model.

Sparse dataset as most of features have more than 50% null values.

e.       What insights did you obtain from this data? For example: What features are important? Why? What visualisations help you understand the data?

Observed from graphs the impact of odometer readings, fuel type and engineTransmission\_exhaustSmoke from model feature importance plot.

3.            Next steps:

a.            What other data (if any) would have been useful?

I assume information about the vehicle like brand/model , fuel capacity , geographical location car past driving history and some description about them would have been helpful.

b.            What are some other things you would have done if you had more time?

(1) Model Ensembling (2) Feature Engineering (3) Feature Correlations (4) Improved way to fill missing values (5) bayesian optimisation for parameter tuning