Please answer the following:

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

1. By analysing the data, we get to know Dataset has number of missing values.
2. Did Data pre-processing by handling missing values and then worked on categorical data using one –hot encoding technique
3. Used XGBoost as Machine learning model for training and testing data
4. Perform hyperparameter tuning on the best model using negative\_mean\_absolute\_error as performance metric

2.            Basics:

a.       How well does your model work?

* My model worked well as we have passed best estimators to XGBoost Regressor model by hyper-tuning it using random search and Used XGBoost model with the best parameters for prediction of rating.

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

* This model works well for sure because when we calculated mean\_absolute\_error for our model, it came out to be 0.0001469 which is much better i.e, our error is quite low.

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

* As our model, Predicting Engine Rating , is a regression problem, we use negative mean absolute error. The Mean Absolute Error (or MAE) is the average of the absolute differences between predictions and actual values. It gives an idea of how wrong the predictions were.

d.       What issues did you encounter?

* Lots of missing values are there which needed to be handled for best fit model
* Handled categorical data using One- hot encoding for multivariable
* Highly imbalanced data
* Using XGBoost for approx. 27k data required some time computation

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

* As we have to handle lot of missing values here, we have to choose features that have enough data to train our model better
* Using heatmap visualization, we get to know about features/columns that have lots of missing values . So we removed those features with more than 50% missing values
* By visualizing ratings-plots, we came to know that our data is highly imbalanced. So we can’t use Accuracy as our performance metric

3.            Next steps:

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

* If reviews along with rating would have been given, Predicted rating would be much interpretable as we could have justify our rating by categorizing positive and negative reviews

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

* I would have done more pre-processing and feature Engineering of Data to get best fit model
* I would have tried Neural Network to train our model which could have lead to better performance