Aditya Jindal
Email : adityaji@iitk.ac.in | adijindal30@gmail.com

Please answer the following:

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

In this problem we are given data which one probably collected through car servicing. And so in the features we are given with various parameters for cars like odometer reading, car , inspection time, battery health, etc.

Here are the steps which I took for this problem:

- Import necessary libraries which you can get from **requirements.txt**
- Import the data, and visualize it, and see the necessary statistics.
- Remove the data which has >20% missing values
- Convert categorical features to one hot encode form, calculate used time using registration year & Inspection date year, etc.
- Select the dependent and independent data as X,y
- Split the data into test & train, and fit models like linear regression, Random forest, K Means etc.
- Visualize confusion metric for each label
- Perform outlier analysis using KNN.

2. Basics:

a. How well does your model work?

	Precision	Recall	F1
weighted avg	**0.44**	**0.45**	**0.45**

As we see from the weighted avg score, our model did quite good, given that we use only ½ of the features given.

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

I use simple linear regression as a baseline, which gives a score of weighted f1 of 0.32, and so we can surely say our model works well.

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

Confusion Matrix, and weighted F1 score. Because F1 score include both precision and recall.

d. What issues did you encounter?

First issue which is most common in classification problem is the issue of very low number of classes for class=1,3.

Also the data is missing for >½ the features.

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

The features like odometer reading, “used time” are some of the important character which we can also relate from the real life that these features are responsible for good/bad health of a patient.

Also features like ‘fuel_type_Electric', 'fuel_type_Hybrid', 'fuel_type_Petrol', 'fuel_type_Petrol + CNG', 'fuel_type_Petrol + LPG' are also important.

3. Next steps:

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

Well if we have more data on engine transmission comment, we can easily do some sort of sentiment analysis which gives score on each comment, which we can further use as a feature of rating prediction.

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

As I describe in part a, we have 4 features related to comments, so we can produce a sentiment score for each text, and for missing comments we can just give 0 score as neutral sentiment for this task.