Predicting Engine Rating

Insights Summary

The problem consisted of finding the prediction ratings for the given dataset on the basis of present features and the condition of the engine. It also consisted of finding the outliers and analysing them.

Data Cleaning

The dataset required cleaning in multiple columns. The dataset initially consisted of 73 columns out of which, fifty four columns were observatory columns for columns related. The related columns consisted of problems to the major column consisting of the main variable in a classified form, namely “Yes” or “No”. Data Cleaning is a very important part of solving the problem as it leads us into exponentially greater or smaller directions in our process.

Data wrangling

This section consisted of finding the missing or null values and fixing the columns present with the most underlying null or NaN values. NaN values are always treated before building the model as they can deviate us from finding the appropriate result or an accurate prediction model. Hence, data wrangling is also a major part of developing a model just like Data Cleaning/Preparation.

Exploratory Data Analysis

In order to start understanding the (linear) relationship between an individual variable and the price. We can do this by using "regplot", which plots the scatterplot plus the fitted regression line for the data.Most of the classified variables which were classified in “Yes” or “No” were changed to dummies (0 and 1). They revealed that they almost had overlapping relations with the dependent variable which helped us determine that they were not at all good predictors of the dependent variable.The pearson correlation heatmap also show that except for year and odometer_readings no other variable had a good relation with dependent variable. Boxplots also revealed that Electric engines necessarily had a rating lower than all the

Model Development

Model development consisted of using the cleaned dataset to define and identify the appropriate Machine Learning prediction model to which it should be fed. Model development is a long process consisting of defining all the anomalies and understanding why our data is being fed into that model which is decided. Here we had our target variable with discrete float values from 0.0 to 5.0. The idea of building the model here consisted of getting predictions for discrete values. These kinds of problems as well as most of the problems can also be solved with regression. So, hereby I used normal regression models just to get an understanding of how the variables are affecting the output of the model.

Outlier Analysis and Detection

Outliers can be overwhelming or they can be useful for a data scientist while dealing with problems like these. This dataset consisted of many outliers in the continuous variable columns in the independent variable like “odometer_reading” as well as categorical variables like the “fuel_types” column. This dataset also consisted of outliers in the dependent variable column “rating_egineTransmission”. Usage of correlation statistics, boxplots, scatterplots, distplots revealed various insights into the versatility of these variable affecting our solution. “odometer_readings “ consisted of numerous outliers with a mean in the lower quartile range of about 76,400. “rating_engineTransmission” consisted of outliers below the rating 3 and above the rating 4.5.