Machine Learning

Fall semester 2020-2021

Final project – Task2 – Report

Anastasia Foudouli

Fuel Consumption Prediction

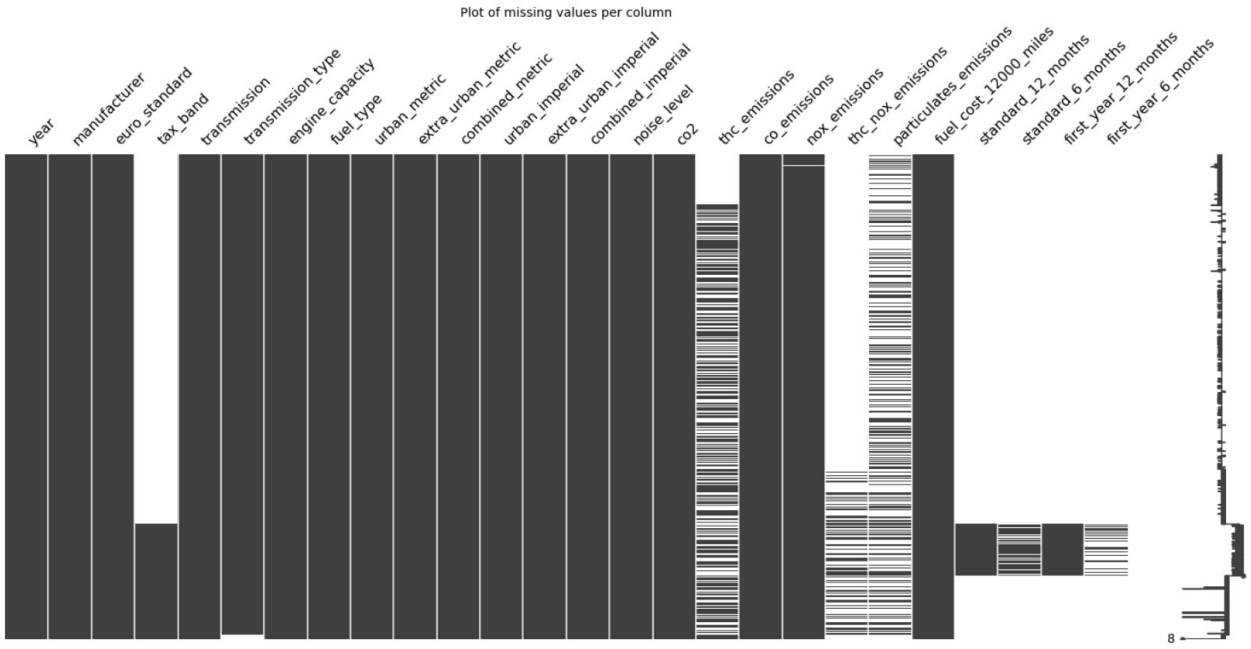
Dataset: Car fuel consumptions and Emissions 2000-2013

Goal: Predict fuel cost per 12000 miles given the attributes of a specific car

Preprocessing

The first task was to find any issues with the dataset and process the data in a way that can benefit the model into finding a solution that would be able to produce minimal error in the predicted versus the actual values.

First thing that was noticed, is that there were many missing values.

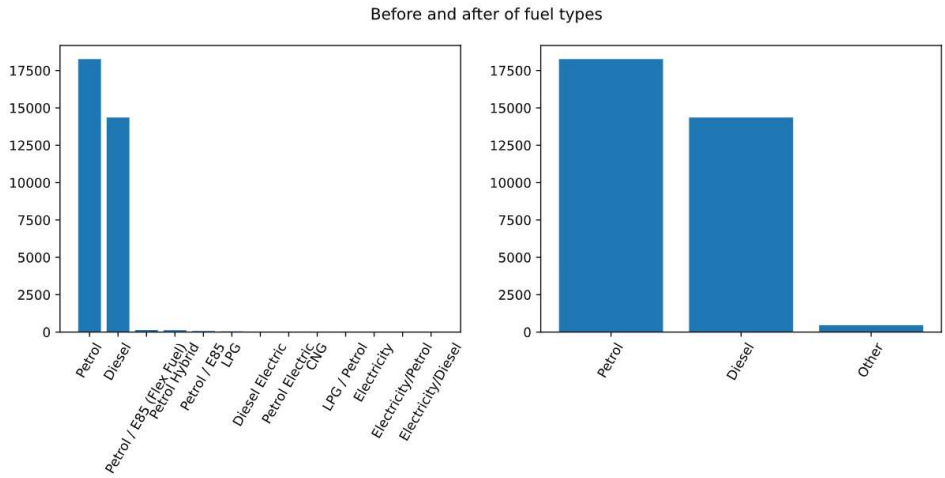


There are variables in the dataset, like “first year 6 months” or “thc nox emissions” where over 80% of the entries is missing. All variables with over 50% of missing entries were *dropped* completely, as trying to interpolate in some way would probably lead to a distribution that is not close to the actual one.

Also, there were 10 missing values in the target variable. These entries were also removed from the dataset.

For the rest of the missing entries, these were filled as follows:

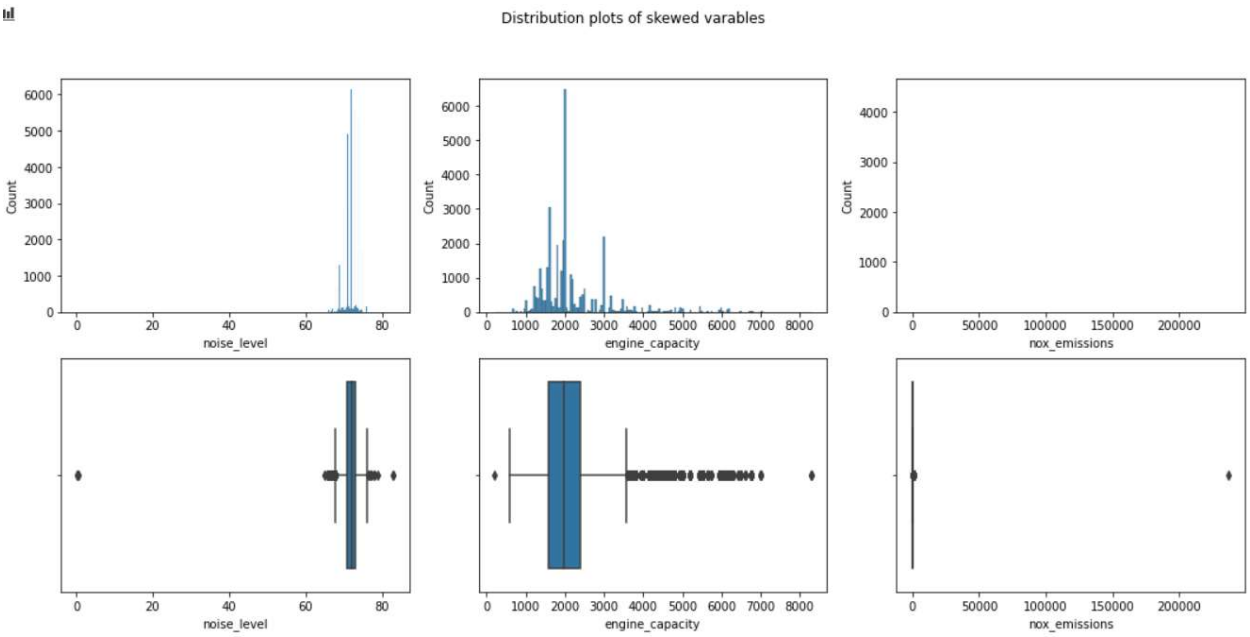
- 'manufacturer', 'transmission', 'fuel type' missing values replaced by mode
- 'transmission type' missing values were filled with mode of transmission type per transmission category
- Numeric missing values were filled using a KNN imputer



The categorical values had very high cardinality and each entry would appear a limited number of times. Therefore, it was decided to use only transmission and fuel types from the categorical variables as predictors for the model. As for fuel type, there were many different values of fuel type with very low number of appearances. They were mapped together to the best suiting category.

Next, a check for outlying values was conducted by observing how skewed a distribution is. Noise level, engine capacity and NOx emissions had the most skewed distributions. For noise level and NOx emissions outliers were revealed.

Noise level values are usually around 70 in this dataset and also the values are integers. The values on the far left of the distribution were like 0.2 or 0.3. It was considered that they were wrong entries, and they were replaced with 72, 73 etc. For NOx emissions, there was one value far above the rest, and it looked like the value was multiplied by 1000. So, the value was replaced by itself divided by 1000.



Modelling

The data were split using an 80-20 train test split.

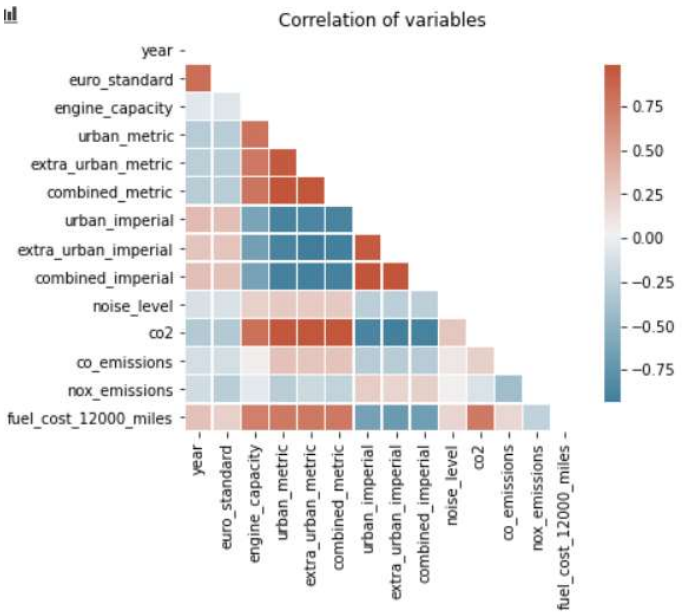
After that the data were scaled in [0,1] and multiple models were trained using 5-fold cross validation to check for overfitting. All the models were trained on their default parameters.

The following results were achieved:

Model	CV mean RMSE	Test RMSE
Linear Regression		154.28
Ridge Regression	159.81	154.49
Support Vector Regression	570.44	244.29
KNN Regressor	73.79	65.70
Random Forest Regression	40.49	29.66
XGBoost Regression	35.55	27.14

XGBoost performed better that the rest of the model, so hyperparameter tunning was conducted. The best model achieved RMSE of 24.27 on test set.

Last, it needs to be mentioned that not all variables are highly correlated with the target variable and that there is also collinearity in the dataset. Keeping these variables, lead to models that are more complex and are adjusting to noise rather than generalizing.



Boruta feature selection technique was used to drop a few of the features. The method outputs Noise level, CO emissions and Manual transmission type to be dropped. Dropping these features did lead to a model that was able to achieve a lower RMSE score on both the mean of cross validated sets and the test set (35.22 and 21.51 respectively).