CONCLUSIONS

In our project as well as exploring the data through “Tableu” we used unsupervised Machine Learning for clustering and supervised Machine Learning and Deep Learning for making predictions.

Exploring the data through “Tableu” we identified some relationships between car properties and “CO2 Emissions” from data, such as identifying average CO2 Emission according to car model or finding most polluting fuel types.

Using unsupervised Machine Learning we have trained the model by using similarities in data by K-Means method and created sample clusters according to CO Emissions and Fuel Consumptions. Considering the purpose of an analysis more clusters can be done by changing the variables in the data.

Both of our models (multiple linear regression and logistic regression) have high accuracy, short training time, and make good predictions.

1. With multiple linear regression by changing multiple properties of a car, CO2 emissions can be predicted with 88% accuracy.
2. With logistics regression by changing the threshold value of CO2 emissions it can be predicted/tested with 98% accuracy and classified as lower or higher from that specific value.

Our models have high accuracy and less training time because:

There is no irrelevant information in the data set.

We don’t have any redundant features in our model.

We don’t have any random data value.

We have enough continuous numeric data.

We have well defined categorical data with valuable information.

The features are highly correlated to target variable.

We haven’t dropped the outliners deliberately because every data has its importance with its own value in regression analysis also the car values are not random and belong to real existing cars. This also increases our model’s efficiency by recognizing cars in similar data sets. For instance, we have analyzed Canada’s car properties and CO2 emissions, but our models can also give high accuracy in other countries’ data sets in same topic.

We did not include Fuel Consumption Comb (L/100 km) and Fuel Consumption Comb (mpg) columns to our models since they are combination of Fuel Consumption City and Hwy features (55% city, 45% Hwy).

|  |  |
| --- | --- |

RECOMMENDATIONS

In our project as well as exploring the data through “Tableu” we wanted to use supervised, unsupervised Machine Learning and Deep Learning as tools. They all created valuable outcomes to solve real life problems and make predictions related to “CO2 Emissions and cars’ properties” so they can be used for making predictions in different data sets.

However, during Logistics Regression modelling in Deep Learning we have realized that the task is not too complex for deep learning since it reaches the high accuracy with min epoch and even without in need of a second layer. Since deep learning is about non - linearity and complexity, we think that this is due to the highly linear nature of our data and high correlations between dependent and independent variables. Thus, instead of using Deep Learning, supervised Machine Learning application can be well enough for this type of linear model – data relationship.