Hoang Nha Nguyen

04/25/2022

Office of Undergraduate Research

Honors Capstone Project

**RESEARCH ON TRANSPORTATION AND ENVIRONMENTAL FACTOR**

**Vehicles Greenhouse Gas Emissions Prediction**

1. ABSTRACT

The end-to-end data analytics project suggests integrating Machine Learning algorithms that deploy linear regression techniques to predict the amount of carbon dioxide (CO2) emitted by a vehicle. We want to live in a world that strives for zero carbon emissions and that in my project, Machine Learning models are deployed to predict the amount of carbon dioxide (CO2) a car produces based on its attributes with the hope of preventing the environmental disasters and global warming effects and assisting the consumers decision-making.

1. DATA EXPLORATION

We have 1067 instances (total number of car models) and 13 features that describes the car attributes and car behaviors, also knowns as rows and columns including:

1. **Year** – refers to model year when the vehicle was produced
2. **Make** – brand of the vehicle
3. **Model** – name of the car product or a range of products including:
   * 4WD/4X4 = Four-wheel drive
   * AWD = All-wheel drive
   * FFV = Flexible-fuel vehicle
   * SWB = Short wheelbase
   * LWB = Long wheelbase
   * EWB = Extended wheelbase
4. **Vehicle Class** – designation of automobile vehicle types that are not limited to compact, SUV, mid-size, station wagon, pickup truck, minivan, etc.
5. **Engine Size** – refer to as 'engine capacity' or 'engine displacement' and is the measurement of the total volume of the cylinders in the engine
6. **Cylinder** – the power unit of an engine; it's the chamber where the gasoline is burned and turned into power.
   * Transmission type - Gear transmission of the car including:
   * A = Automatic
   * AM = automated manual
   * AS = automatic with select shift
   * AV = continuously variable transmission
   * M = Manual
   * 3 – 10: number of gears
7. **Fuel Type** – energy sources by which the motor vehicles are powered including:
   * X = regular gasoline
   * Z = premium gasoline
   * D = Diesel
   * E = Ethanol
   * N = Natural Gas
8. **Fuel Consumption City** – represent fuel consumption while driving in inner-city conditions with heavy, stop-start, local traffic
9. **Fuel Consumption** – City and highway fuel consumption ratings are described in liters per 100 kilometers (L/100 km) unit and the combined rating (55% city, 45% highway) is in L/100 km unit and in miles per imperial gallon (mpg)
10. **CO2 Emissions** – the emissions of carbon dioxide (in grams per kilometer) for combined city and highway driving
11. **CO2 Ratings** – the emissions of carbon dioxide rated on a scale from 1 (worst) to 10 (best)
12. **Smog Ratings** – the emissions of smog-forming pollutants rated on a scale from 1 (worst) to 10 (best)
13. DATA VISUALIZATION

I’m interested in learning the data trends and patterns and this is an important step before conducting analysis. In the project, I mainly use Python to construct histograms, count plot, bar plot and correlation matrix heat map to detect any relationship and patterns of variables in the dataset and how well they interact with each other.

1. Correlation Matrix Heatmap

Graphical user interface, application

Description automatically generated

From the correlation matrix above, we can compute the correlation between all continuous variables. Since the variable “CO2Emissions” is “the target variable”, whereas other variables are independent variables, I’m expected to see any impact of other independent variables on the response variable “CO2Emissions”. There are positive relationships between ‘Engine Size’, ‘Cylinders’, ‘CT\_FC’, ‘HWG\_FC’, ‘CB\_FC’ and ‘CO2Emissions’, whose correlations show 0.82, 0.83, 0.97,0.93 accordingly. These car attributes are big contributors to the sources of CO2 emissions and should be taken into close consideration if CO2 gas emissions are expected to be addressed.

­­­­­­

1. Histograms

Graphical user interface, chart, application

Description automatically generated

1. Bar plot

Chart, bar chart

Description automatically generated

1. Pie chart

Chart, pie chart

Description automatically generated

1. Count plot

Chart, bar chart

Description automatically generated

1. DATA PREPROCESSING

It is detected that there are no missing values found within the continuous variables

Since there are variables in the dataset that contain categorical values, we need to convert the ‘object’ data type to ‘numerical’ data type to standardize the data types before we precede with data analysis.

To complete the analysis, we use python programming with importing the following libraries:

* NumPy
* Pandas
* Matplotlib
* Seaborn

Variables that contain numerical values consist of ‘Fuel Type’, ‘Vehicle Class’, ‘Transmission Type. Therefore, I need to get dummy variable for these variables. Once I obtained all numerical values. I created a new data frame to store all numerical values.

Now it comes to the step that we need to find correlations between all variables with the target variables. I use Pearson correlation to compute the relationships.

A picture containing chart

Description automatically generated

Graphical user interface, text, application

Description automatically generated

I’m creating a loop to assess all the correlations between all continuous variables with the target

variables. I will obtain variables whose correlations are greater than .5 or less than -.5 because they indicate positive correlation between the two.

Once I get the desired variables that correlate well with the ‘CO2Emissions’, I group all these variables into a new data frame that I’m ready to analyze.

1. DATA ANALYSIS

From the sklearn module, we will use the LinearRegression() method to create a linear regression object.

This object has a method called fit() that takes the independent and dependent values as parameters and fills the regression object with data that describes the relationship:



X = df3.drop('CO2Emissions', axis=1)

Y= df3['CO2Emissions']

X takes all the explanatory variables or independent variables

Y takes all the response variables or dependent variables.

I split the dataset into train and test dataset with train\_test\_split: x\_train, x\_test, y\_train, y\_test = train\_test\_split(X, Y , test\_size=0.2, random\_state=42)

Machine Learning Model Building

1. Linear Regression model

Graphical user interface, text, application, email

Description automatically generated

This is the output that the linear regression predicts:

A screenshot of a computer

Description automatically generated with medium confidence

I also compute the statistical measures of the model to test the accuracy of the model

Graphical user interface, text, application

Description automatically generated

1. Decision Tree Model

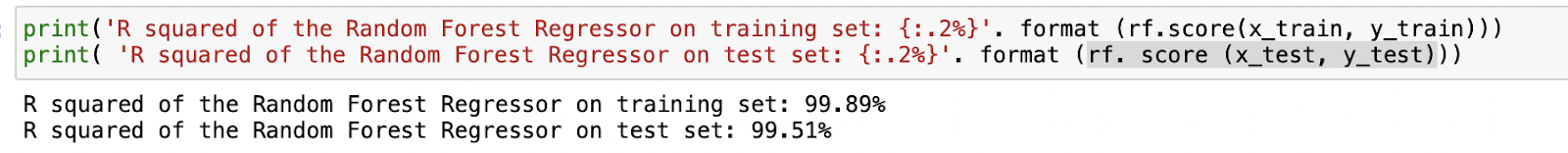
Text

Description automatically generated with low confidence

1. Random Forest Model

Text

Description automatically generated



Among the above three ML models, the decision tree model has the highest accuracy (99.66%) and lowest root mean squared error (1.26) in predicting the CO2 Emissions emitted by vehicles. So the decision tree model would give us the optimal output.

REFERENCES

<https://www.datacamp.com/tutorial/svm-classification-scikit-learn-python>

<https://machinelearningmastery.com/robust-regression-for-machine-learning-in-python/>

<https://stats.oarc.ucla.edu/stata/dae/multivariate-regression-analysis/>

<https://towardsdatascience.com/all-machine-learning-models-explained-in-6-minutes-9fe30ff6776a>