ABHIVRUDDHI JOURNAL

ISSN No. 2584-1025

# $CO_2$ -Emission: A Predictive Modeling with Diverse Regression

Name: Ayush Mangesh Atre; Omkar Sanjay Chavan and Ashok Kumar*

Affiliation: MIT Art, Design & Technology University , Loni Kalbhor, Maharashtra

Email id: ayushatre2001@gmail.com ; omkarchavan27999@gamil.com; ashok.kumar@mituniversity.edu.in

***Abstract*—In present scenario $CO_2$ Emission is one of the most important factors for the pollution. In this research paper, we applied different modeling approach to find most important factors due to $CO_2$ Emission. In this study we applied statistical tools and techniques to understand the relationship between various factors and their impact on $CO_2$ emission. We observed the $CO_2$ emission by heavy and light-duty vehicles and studied the factors responsible for the growth of $CO_2$ emission like, Cylinders, Fuel consumption in cities and highways, Fuel Types, Vehicle class, etc. We applied multiple linear regression model and estimated the coefficients. Also, we compared classical and machine learning approach and recorded the results. To address multicollinearity, we performed statistical technique such as Principal Component Analysis and recorded the better results.**

**Keywords— Factor Analysis, Decision Tree, Random Forest Regression, KNN Regression, Principal Component Analysis and Multiple Linear Regression.**

## I. INTRODUCTION

The complexity of environmental issues arising from mobility has increased in tandem with the rapid urbanisation trend due to notable global climate shifts. Passenger cars account for twenty to thirty percent of worldwide greenhouse gas (GHG) emissions and seventy-five percent of carbon dioxide emissions. There are also more used automobiles on the road now, despite strict fuel and greenhouse gas regulations. Older cars now consume more natural resources and emit more air pollutants due to the rise in vehicles and vehicle miles travelled (VMT). A lot of daily operations depend on the transport sector, such as the movement of passengers and the sustained supply of goods. More than half of the oil produced worldwide is consumed by the transportation sector, hastening the depletion of fossil fuel supplies. The transport sector is to blame for the steady rise in fuel prices. The transportation industry has grown globally. Because of this, the transportation industry is accountable for about 25% of all $CO_2$ emissions caused by human activity globally. Light-duty automobiles are included in the transportation section. Any mobile device used primarily for people and goods transportation and having a gross vehicle weight classification of less than or equal to 10,000 pounds is considered a light-duty vehicle. Cars, vans, SUVs, and pickup trucks are a few examples [4]. These days, it's easier to see how energy consumption impacts a nation's ability to expand economically and how human health is impacted by carbon footprints. Consequently, during the past three decades, federal policymakers' primary concern has been national economic challenges. Known also as a greenhouse gas, carbon dioxide (CO2) is a major gas that traps heat. The combustion of fossil fuels such as coal, oil, and natural gas, as well as wildfires and natural occurrences like volcanic eruptions, produce it.The principal greenhouse gases in the Earth's atmosphere are ozone (O3), carbon dioxide (CO2), methane (CH4), nitrous oxide (N2O), and water vapour. Every year, a typical passenger car emits 4.6 metric tonnes of carbon dioxide. This amount may vary depending on the fuel type, fuel efficiency, and annual mileage of a vehicle. The study looks at recent trends in $CO_2$ emissions from various automakers and vehicle types. This study offers a comprehensive, comprehensive analysis of current light-duty vehicles' fuel efficiency and carbon dioxide emissions. This paper also covers the process of getting the dataset ready for analysis. In order to carry out and achieve the stated research objectives, an analytical and predictive study was conducted utilising 7384 light-duty trucks that were found from 2017 to 2021 from the government of Canada dataset. Two statistical techniques are used in this study to analyse the data: (1) an inferential statistical evaluation based on different vehicle attributes to ascertain the association among all data set features, and (2) a descriptive statistical examination to assess light-duty automobiles' fuel efficiency and CO2 emissions.

## II. METHODOLOGY

### *A. Factor Analysis*

A method for condensing a large number of variables into a smaller number of factors is factor analysis. In essence, factor loading is the factor and variable's correlation coefficient. Factor loading displays the variance on that specific factor that can be attributed to that variable. Greater factor loading indicates that the variable's variance is sufficiently extracted by the factor. Factor analysis is used to compress or aggregate data from an old variable that has undergone significant modification to a small number of new variables known as factors that retain the majority of the original variable's information[2].

ISSN No. 2584-1025

*B. Multiple Linear Regression*

Multiple linear regression is a method of statistics for predicting a variable's outcome based on the values of two or more variables. Sometimes referred to as simply multiple regression, it is a modification of linear regression. We forecast the value of the variable that is dependent, or the one we are attempting to predict, utilising the independent or explanatory components[1][3].

Let Y be the dependent variable and $X_1$, $X_2$, ..., $X_p$ be p independent variables. Then the multiple regression model can be written as:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots \ldots \ldots + \beta_p X_p + \epsilon$$

*C. Lasso Regression*

- A regularization term that penalizes the absolute size of the coefficients is introduced by Lasso regression. The lambda (α) parameter determines the penalty's strength.
- Fit the Lasso regression model to your data after you've determined the ideal α value. As a result of the algorithm's minimization of the sum of squared errors plus the penalty term, certain coefficients will be precisely zero.
- Determine which characteristics have non-zero coefficients after the model has been fitted. These are the features that have been chosen to add to the model.

*D. Principal Component Analysis*

Principal Component analysis is used to reduce the large dimension of the data.

Capturing the majority of the data variance, the Principal Components represent a linear pattern. They possess both magnitude and direction. Data are projected onto lower-dimensional space orthogonally, or perpendicularly, to form principal components[2].

Principal component analysis, or PCA, is a statistical procedure that converts a set of correlated variables into a set of variables that are uncorrelated by an orthogonal transformation.

*E. Machine Learning Techniques*

- Support Vector Machine

An input-output pair (X, y) dataset serves as the foundation for SVM regression, with X denoting the input features and y denoting the corresponding output values. Feature scaling involves normalizing or standardizing the input features to make sure they have a comparable scale. This step helps keep some features from dominating the training of the model because of their bigger magnitude. Building the Model: Selecting a Kernel: Pick an appropriate kernel function. Typical options include radial basis function (RBF), polynomial, linear, etc. Within the modified feature space, the kernel function calculates the similarity between pairs of data points. The regularization parameter (C), kernel parameters. The model's generalizability and complexity are influenced by these parameters. The SVM algorithm maximizes the margin by identifying the ideal hyperplane that best fits the data[5].

- Decision Tree

Pick a Feature: Decide which feature divides the data into subsets the best. Divide up the Data: Based on the chosen feature, segment the dataset into smaller groups. The Recursive Process: Proceed with each subset in turn, taking into account a distinct feature at each stage. End Conditions: Establish stopping parameters, such as a minimum number of samples in a leaf or a maximum depth. Make Predictions: Based on the average of the target values within that node, assign a forecast value to every terminal node (leaf). Construct Tree: Build the tree until the necessary halting conditions are satisfied. Expectation: Predict the average target value of each leaf as you work your way through the tree from the root to a new input[5].

- Random Forest Regression

Sampling utilizing bootstraps: Select training data subsets at random using replacement. Create Decision Trees: Using a random subset of features at each split, construct a decision tree for each subgroup. Total Predictions: For regression, average the predictions from each tree to obtain the final prediction. Collective Power: Robustness and generalization are enhanced by the ensemble of various trees. To build a more reliable and accurate regression model, the main concept is to aggregate the predictions of several different decision trees[5].

- KNN Regression

Dataset for Input: KNN regression starts with a dataset that has continuous target values (y) that correlate to input features (X). Normalization: To guarantee that every feature contributes equally to the distance computations, normalize or scale the features. By using normalization, characteristics with bigger scales are kept out of the distance calculations. Selecting a value for K indicates how many of the closest neighbors to take into account when making a prediction. The performance of the model is greatly influenced by this variable. Choose a distance metric to quantify the similarity or separation between data points in the feature space, such as the Manhattan distance, Euclidean distance, etc. How the closest neighbors are located is determined by the distance metric[5].

## III. Empirical Analysis

we utilized Factor Analysis to distill a smaller set of variables from a larger pool, discerning the primary factors that exert the most influence on vehicle $CO_2$ emissions. By employing Principal Component Analysis (PCA), we effectively mitigated multicollinearity between these variables, enhancing the robustness of our analysis. Furthermore, we evaluated various machine learning techniques, considering the loss function is the Root Mean Square Error (RMSE).

RMSE is particularly suitable for regression tasks, indicating the average deviation between predicted and actual

ISSN No. 2584-1025

values. This evaluation enabled us to identify the most suitable machine learning algorithms for our specific task, ensuring accurate predictions and insightful analysis of $CO_2$ emissions from vehicles.

The dataset used in the present research documents the precise variations in a vehicle's CO2 emissions based on a number of variables. The dataset was acquired through the Canadian government's official open data website. This includes information spanning a whole seven years.
There are 12 columns and 7385 rows in all.
This dataset contains the following features/variables.
Make, Model, Vehicle Class, Engine Size(L), Cylinders, Transmission, Fuel Type, Fuel Consumption City (L/100 km), Fuel Consumption Hwy (L/100 km), Fuel Consumption Comb (L/100 km), Fuel Consumption Comb (mpg), $CO_2$ Emissions(g/km).

### *A. Data Visualization*

- Boxplot:

The boxplot's main uses are to show whether a distribution is skewed and whether any potentially outlier observations /unusual observations may be found in the data set.

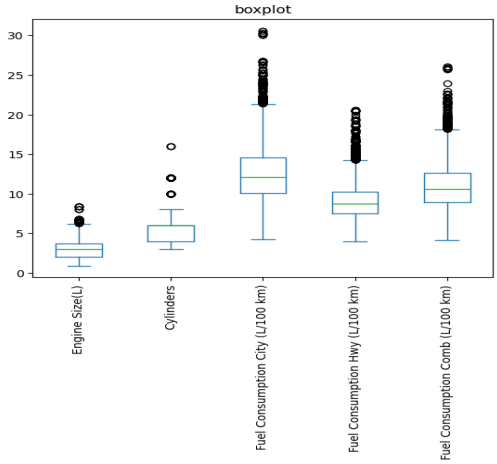


Fig 1. Boxplot

The graph shows there are potential outliers for each of the above features. The distribution of Cylinders is negatively skewed.

- Bar Graph

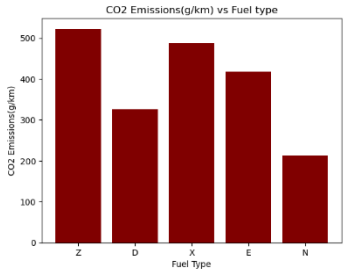


Fig 2. $CO_2$ Emission vs Fuel type

The Fuel type Z (Premium gasoline) and Fuel type X (Regular gasoline) have the high $CO_2$ Emission. The Fuel type N (Natural gas) have the lowest $CO_2$ Emission.

### *B. Factor Analysis*

**Table.1: Factor Analysis**

| Variables | Factor 1 | Factor 2 | Factor 3 | Factor 4 | Factor 5 |
|---|---|---|---|---|---|
| Engine Size(L) | 0.4653 | 0.8406 | 0.1078 | 0.0089 | -0.0670 |
| Cylinders | 0.4040 | 0.8697 | 0.1136 | -0.0236 | 0.0515 |
| Fuel Consumption City (L/100 km) | 0.8365 | 0.5118 | 0.1244 | -0.1385 | 0.0044 |
| Fuel Consumption Hwy (L/100 km) | 0.9046 | 0.3997 | 0.0396 | 0.1252 | 0.0118 |
| Fuel Consumption Comb (L/100 km) | 0.8700 | 0.4790 | 0.0969 | -0.0488 | 0.0065 |
| Fuel Consumption Comb (mpg) | -0.7948 | -0.4144 | -0.3431 | 0.0554 | 0.0475 |
| $CO_2$ Emissions(g/km) | 0.6959 | 0.5880 | 0.3367 | 0.0404 | 0.0543 |

Fuel Consumption City (L/100 km), Fuel Consumption Hwy (L/100 km), Fuel Consumption Comb (L/100 km) the value of Factor 1 for each variable is 0.8365, 0.9045, 0.8699 is nearly close to 1 which is quite high. So fuel consumption is the most influential factor for $CO_2$ Emission

### *C. Lasso Regression*

**Table 2. Lasso Regression**

| Sr.No. | Features | Coefficient |
|---|---|---|
| 0 | Engine Size(L) | 5.483273 |
| 1 | Cylinders | 6.213025 |
| 2 | Fuel Consumption City (L/100km) | -1.648136 |
| 3 | Fuel Consumption City (L/100km) | 5.026321 |
| 4 | Fuel Consumption Comb (L/100km) | 0 |
| 5 | Fuel Consumption Comb (mpg) | -4.887064 |

Fuel consumption Comb (L/100 km) is one feature that should not be included in the model, according to the result of the lasso regression feature selection process. This is because the feature's value is exactly equal to zero.

### *D. Principal Component Analysis*

**Table 3. Principal Component Analysis**

| 0 | 1 | 2 | 3 | 4 |
|---|---|---|---|---|
| 6.6269 | -0.3268 | -0.1522 | -0.6029 | -0.0985 |
| 2.3887 | -1.1928 | -0.3585 | -0.4909 | -0.3854 |
| 21.2695 | 3.7235 | 2.8261 | 0.5833 | 0.3275 |
| -2.4705 | -0.1786 | -0.8667 | 0.1387 | -0.0322 |
| -0.4247 | 0.3757 | -0.7017 | 0.1495 | -0.0441 |
| ... | ... | ... | ... | ... |
| 3.4979 | -1.1787 | -0.2113 | -0.2584 | -0.0632 |
| 2.2936 | -1.4039 | 0.0749 | -0.0968 | -0.0343 |
| 0.3133 | -1.9832 | -0.2123 | -0.1171 | -0.0291 |
| 2.2936 | -1.4039 | 0.0749 | -0.0968 | -0.0343 |
| -0.7652 | -2.213 | -0.2366 | -0.3257 | -0.0243 |

ISSN No. 2584-1025

Principal Components are the representation of eigenvectors the first principal component explains the maximum variation of the data set followed by the other principal components the values belonging to each principal component represent the weightage of a particular variable in the principal components.

**Table 4. Cumulative Variance**

| Principal Components | PC1 | PC2 | PC3 | PC4 | PC5 |
|---|---|---|---|---|---|
| Cumulative Variation | 86.3% | 95.67% | 97.9% | 99.3% | 100% |

Variance Explained by principal components = $\frac{\lambda_i}{\text{Total Variation}} * (100\%)$

On observing the cumulative variation, we can see that up to 4 principal components total variation explained is 99.3%. So, we can use up to 4 principal components for further analysis.

Machine Learning Techniques

**Table 5. Machine Learning Techniques**

| ML Techniques | Root Mean Square Error (RMSE) |
|---|---|
| SVM Regression | 12.6358 |
| Decision Tree Regression | 15.2043 |
| KNN Regression | 10.4539 |
| Random Forest | 9.2180 |

*E. Ordinary Least Square (OLS) Model*

Then the multiple regression model can be written as:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots \ldots \ldots + \beta_p X_p + \epsilon$$

OLS Model after PCA

$Y=247.6968-6.5903(PC1)+5.4261(PC2)\ 2.4041(PC3)+6.7825(PC4)+\varepsilon$

## CONCLUSION

In this study we observed that fuel consumption is the most influential factor for $CO_2$ Emission. From Principal Component Analysis we removed multicollinearity and built a regression model that would forecast $CO_2$ emissions by utilizing the technique of dimensional reduction. By using this method, we were able to improve the precision and effectiveness of our predictive model by accurately capturing the complex relationships between these multidimensional factors. Mainly we used four machine learning models that we carefully assessed for our research were K-Nearest Neighbors (KNN), Random Forest, Decision Tree Regression, and Support Vector Machines (SVM). The Random Forest model is the best fit model for our dataset, with the lowest Root Mean Squared Error (RMSE) among the approaches studied. Also for further study, we can apply Bayesian approach to forecast the $CO_2$ emission.

## ACKNOWLEDGMENT

We thank Prof. Dr. Virendra Shete (Director, MIT School of Engineering & Science) and Dr. Vinayak Dhumale (Head, Department of Applied Science & Humanities) MIT ADT University, Pune for providing their support and necessary facilities. Also, I would like to thank Dr. Pratibha Jadhav, Dr. Ashok Kumar and Prof. Rohit Raskar for their valuable suggestions and encouragement. We would like to thank all our colleagues for their support and valuable comment in this manuscript.

## REFERENCES

[1] Montgomery, Douglas C., Elizabeth A. Peck, and G. Geoffrey Vining, "*Introduction to linear regression analysis*", John Wiley & Sons, 2021.

[2] Bhuyan, Kartick Chandra. *Multivariate analysis and its applications*. New Central Book Agency, 2005.

[3] Gupta, S. C., and V. K. Kapoor. *Fundamentals of mathematical statistics*. Sultan Chand & Sons, 2020.

[4] Natarajan, Yuvaraj, Gitanjali Wadhwa, K. R. Sri Preethaa, and Anand Paul. "Forecasting carbon dioxide emissions of light-duty vehicles with different machine learning algorithms." *Electronics* 12, no. 10 (2023): 2288.

[5] Pradhan, Manaranjan, and U. Dinesh Kumar. *Machine Learning with Python*. Wiley, 2019.