Project Report on Multiple Regression model applied on Fuel Consumption Dataset

Submitted by –
Soham Sandeep Mangore

Index

1. Data Cleaning process and justifications

a) Dropping unwanted columns and applying Encoding for categorical data

The given dataset was already cleaned and had no missing values. However, there were few categorical columns like [‘MAKE’, ‘MODEL’, ‘VEHICLECLASS’, ‘TRANSMISSION’, ‘FUELTYPE’] which needed to be converted into numerical.

I first dropped the unwanted columns from the dataset . The columns MODELYEAR, MAKE, MODEL, and VEHICLECLASS are dropped as they are less useful for numerical analysis.

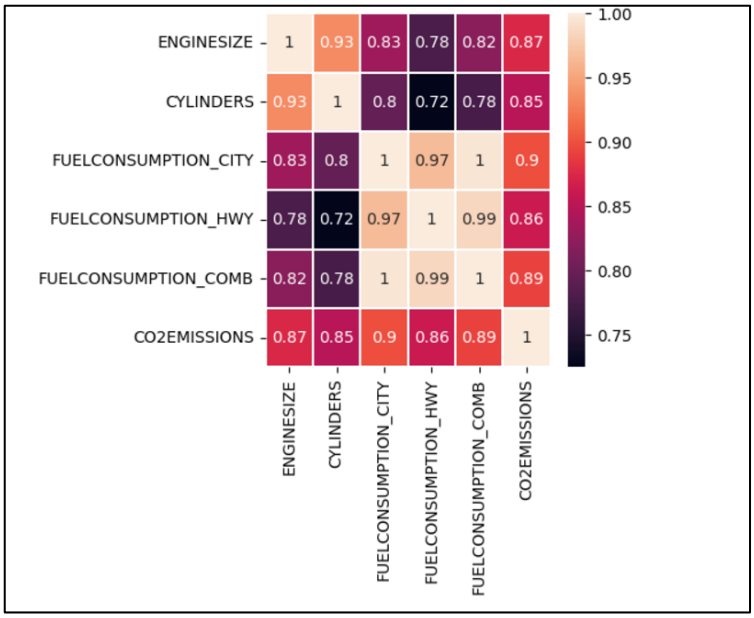
Furthermore, I Drop FUELCONSUMPTION_CITY and FUELCONSUMPTION_HWY because FUELCONSUMPTION_COMB already captures the overall fuel consumption. I prefer to work with FUELCONSUMPTION_COMB so dropped FUELCONSUMPTION_COMB_MPG.

I applied one-hot encoding on the columns 'TRANSMISSION', 'FUELTYPE' and converted them into numerical values.

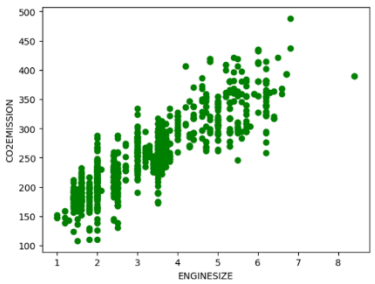
b) Standardization

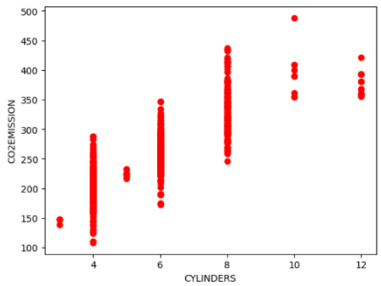
Standardization is more suitable for this dataset because features like ENGINE_SIZE and CO2_EMISSIONS follow a Gaussian-like distribution. Standardization improved the models by reducing the error metrics (MAE, MSE, and RMSE).

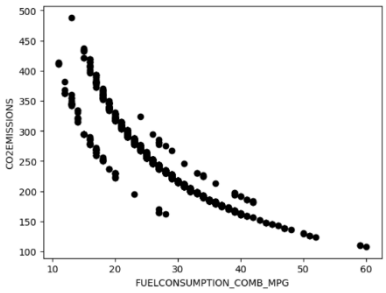
2. Correlation analysis with visualizations



- **FUELCONSUMPTION_CITY** and **FUELCONSUMPTION_COMB** have a correlation of 0.99, indicating they are very strongly positively related.
- **CO2EMISSIONS** has high correlations with **ENGINESIZE** (0.87), **CYLINDERS** (0.85), and other fuel consumption features.
- **CO2EMISSIONS** is strongly correlated with fuel consumption variables, especially **ENGINESIZE**
- **CYLINDERS** also show strong correlations with fuel consumption and emissions, as the number of cylinders typically affects engine performance.
- Thus, it can be inferred that **fuel consumption** and **CO2 emissions** are influenced by the **size** and **type** of the engine and its efficiency.







3. Model training and evaluation results

Multiple regression models were trained and evaluated on the scaled dataset. Among them, Random Forest and XGBoost delivered the best performance. The dataset was split into a 1:3 ratio for training and testing, with a random_state of 0. The models were trained using the following parameters:

1) Random Forest: n_estimators=100, random_state=42
2) XGBoost: n_estimators=100, random_state=42
3) SVR: kernel='rbf'
4) Linear Regression: Default parameters.

All the models were trained and tested on scaled data

4. Evaluation

Following evaluation metrics were used to evaluate the models :

1. **Mean Absolute Error (MAE):** Indicates the average error in predicting CO2 emissions. Lower MAE in models like XGBoost (0.08) shows it provides closer predictions compared to others.
2. **Mean Squared Error (MSE):** Highlights larger errors more due to squaring. XGBoost, with the lowest MSE (0.01), suggests fewer significant deviations in CO2 predictions.
3. **Root Mean Squared Error (RMSE):** Shows the standard error in predictions. XGBoost's RMSE (0.11) demonstrates better overall prediction accuracy compared to Random Forest (0.12).
4. **R^2 Score**: Represents how well the features explain CO2 emissions variability. XGBoost (0.99) captures the relationship more effectively than Random Forest (0.98).
5. **Residual Sum of Squares (RSS):** Measures the total error. XGBoost's lowest RSS (4.62) confirms it minimizes prediction errors better than other models.

Overall, The XGBoost model demonstrated superior performance in predicting CO2 emissions, with the lowest error metrics across all evaluations: MAE (0.08), MSE (0.01), RMSE (0.11), and RSS (4.62).

5. Comparative analysis of the three regression models

No.	Model	MAE	MSE	RSME	R^2	RSS
1	Random Forest	0.09	0.02	0.12	**0.98**	5.37
2	XGBoost	0.08	0.01	0.11	**0.99**	4.62
3	SVR	0.10	0.02	0.14	0.98	6.63
4	Linear Regression	0.11	0.02	0.14	0.98	6.77

In the comparative analysis, XGBoost performed the best with the lowest MAE (0.08), MSE (0.01), and RMSE (0.11), along with the highest R^2 (0.99), indicating strong predictive accuracy and minimal errors. Random Forest followed closely with competitive results (MAE: 0.09, MSE: 0.02, RMSE: 0.12, R2: 0.98), showcasing its robustness. SVR and Linear Regression had slightly higher error values and lower R^2, suggesting they were less effective at prediction compared to the ensemble models. The RSS values also reflect this, with XGBoost achieving the lowest (4.62), making it the most accurate model for this dataset.