Concepts and Technologies of AI

5CS037 – Assignment 02

Regression Analysis for Predicting a Continuous Target Variable

Table of Contents

List of Figures

1. Introduction

1.1 Problem Statement

This project aims at predicting a continuous target variable by means of regression methods. Continuous predictions are imperative in most practical scenarios including estimation of economic or energy or development indicators.

1.2 Dataset

The data which was used in the analysis has been taken through one publicly available source and it has various variables which affect the target variable. The dataset contributes to the United Nations Sustainable Development Goals (UNSDGs) as it allows making a predictive analysis and comprehending every factor concerning development and sustainability.

1.3 Objective

The primary goal of the analysis is the construction and testing of regression models that can be used to predict a continuous target

variable with a high level of accuracy with respect to the presented features and determining the most effective model.

2. Methodology

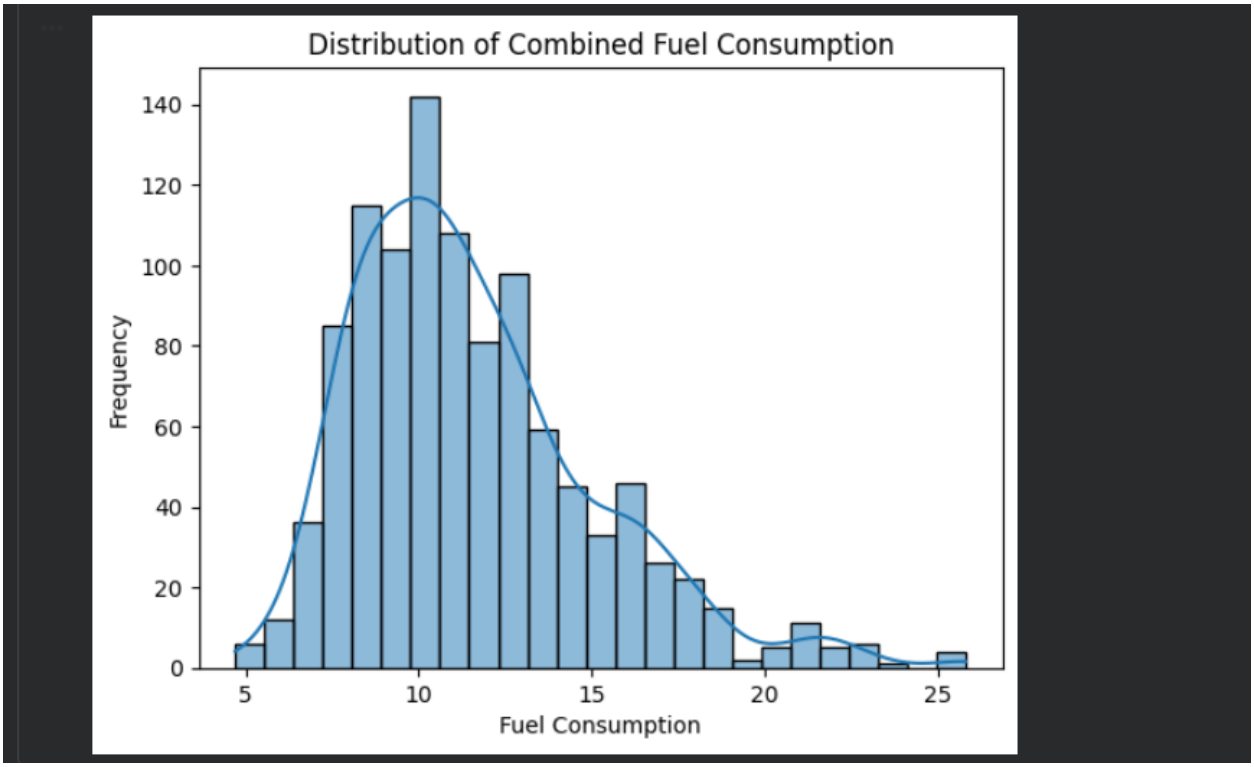
2.1 Data Preprocessing

The dataset was cleaned before model building to deal with the missing values and outliers. Where there were missing values, the same was eliminated, or substituted by the use of relevant statistics. To provide fair contribution in the training of the models, numerical features were normalized or standardized.

2.2 Exploratory Data Analysis (EDA)

EDA was carried out through visuals like the histograms, box plots, scatter plot and correlation heat maps.

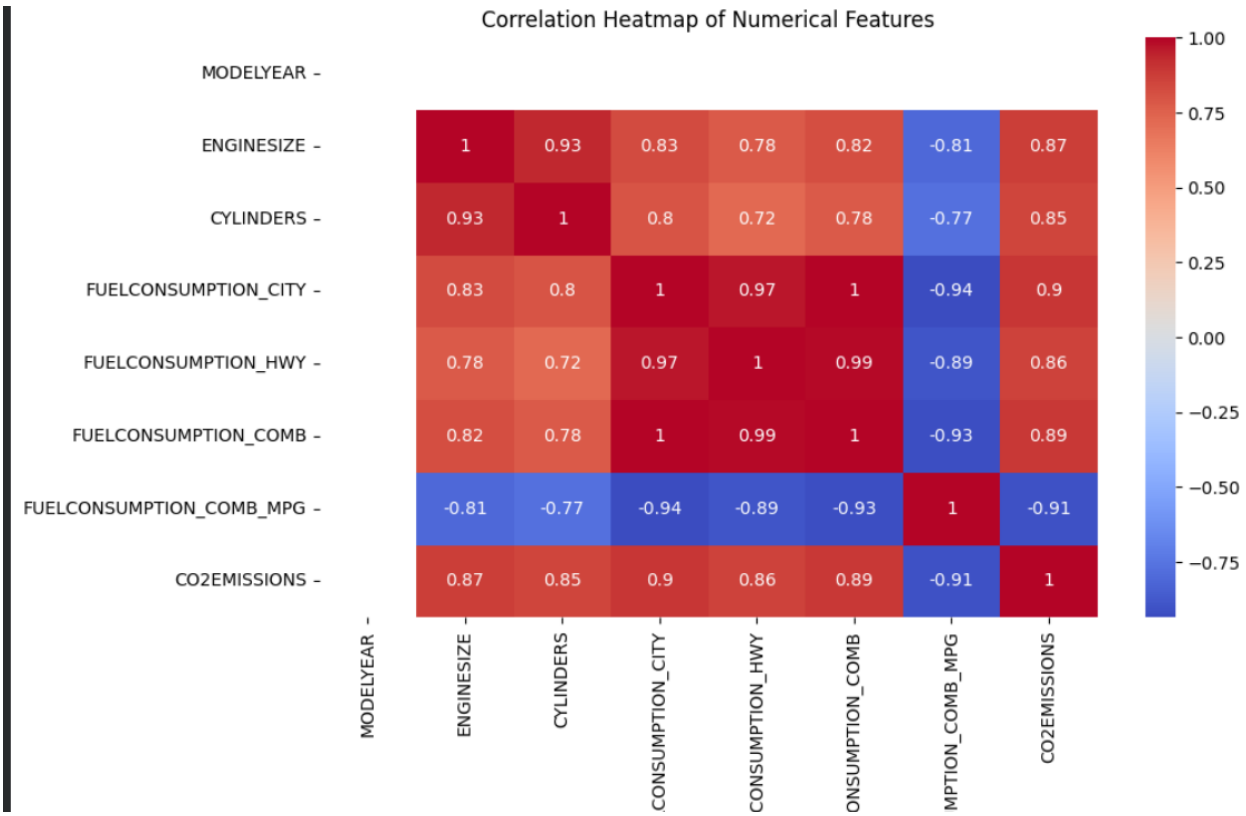
Target Variable Distribution:



The histogram demonstrates how much of the target variable, combined fuel consumption, falls into each of the data points. The x-axis indicates the values of consumption of fuel that are approximately between 4 and 25 and the y-axis indicates the frequency of the vehicles in each range. The greatest concentration of vehicles is around the consumption level of circa 10 and this is the highest point of the distribution. The bars are superimposed by the smooth kernel density estimate (KDE) curve that identifies the general shape of the data. The distribution is marginally right skewed with fewer vehicles (in terms of count) consuming more fuel. This skewness is significant to observe, as it might have an impact on the modeling, and one of the measures that could be taken to normalize the data might be a transformation, e.g., a logarithmic

or square root adjustment. In general, the plot gives a good insight into the typical level of fuel consumption and allows evaluating the necessity of preprocessing before analysis.

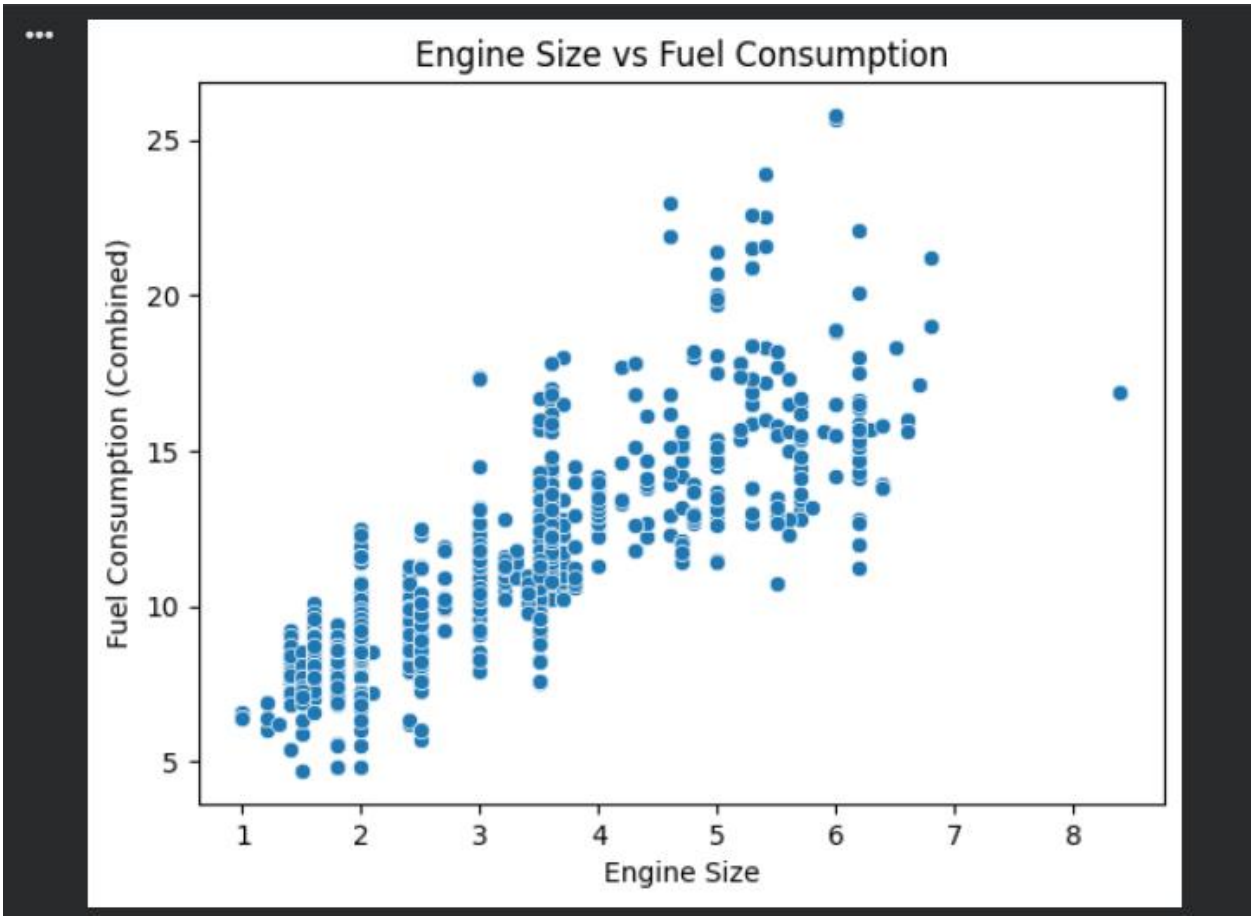
Correlation Heatmap:



The correlation heatmap gives a graphical representation of the relations among different numerical characteristics of vehicle specification and emissions. The heatmap shows the correlation

coefficient between two variables in each cell, where a correlation is negative (-1, indicating a perfect negative correlation), or positive (1, indicating a perfect positive correlation) and the intensity of the color shows the strength and direction of the relations. Compared to the other metrics, the positive correlations between engine size and cylinders (0.93) are strong, and the positive correlations between fuel consumption metrics and CO 2 emission metrics, e.g., the relationship between fuel consumption in the city and CO 2 emissions is positive and has the value of 0.90. On the other hand, the fuel consumption measures are weakly positively correlated with the fuel efficiency (combined MPG) with the city fuel consumption having a negative correlation of -0.94 with the combined MPG. These trends indicate that bigger engines and extra cylinders are more likely to burn more fuel and produce more CO 2 whereas high fuel efficiency is linked with low emissions. The heatmap has a variety of applications such as determining redundant features, feature selection, and knowing the impact of vehicle characteristics on the environment.

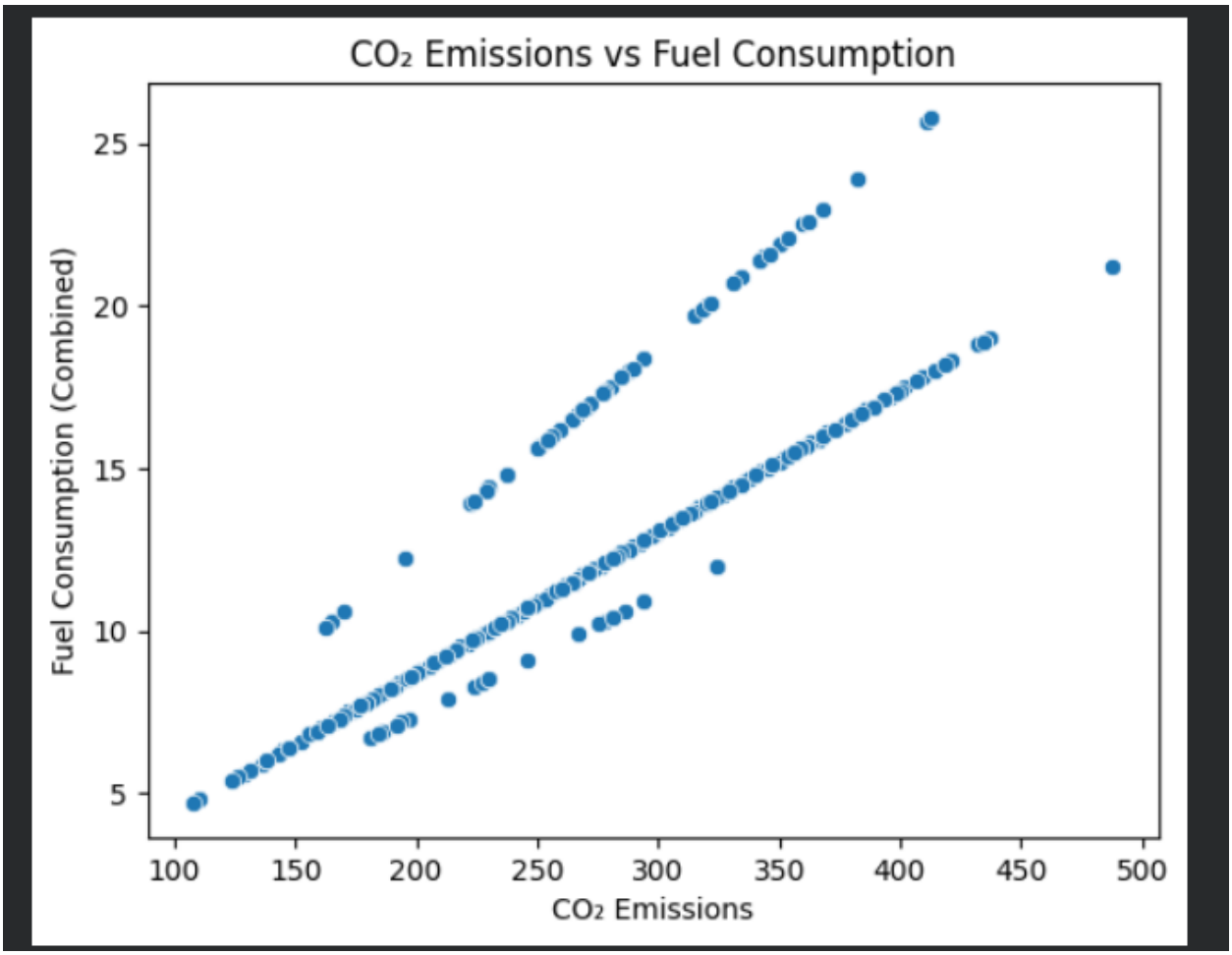
Scatter Plot:



The scatter plot of the title of the plot, which is of the Engine size vs Fuel Consumption, demonstrates how the size of the engine in a vehicle relates to the overall fuel consumption of that vehicle. The blue dots indicate one data point each of the variation in fuel consumption with engine size. The plot shows that there is an increasing trend, which means that an increase in engine size is accompanied by an increase in fuel consumption, as well. This indicates that the two variables are positively correlated i.e. larger engines tend to burn more fuel. This trend is natural because larger engines normally consume more energy and hence consume more fuel. This visualization can be used to

comprehend the effect of engine specifications on fuel efficiency, and can be used to make decisions in automotive design, environmental policy and consumer choice.

Another Scatter Plot (CO_2 vs Fuel)



The following scatter plot is entitled CO 2 Emissions vs Fuel Consumption: it represents the correlation between the amount of carbon emission in a vehicle and the overall fuel burn (consumption). The dots in blue are data points, which reflect the variation in fuel consumption with CO 2 production. The

plot shows that there is a distinct positive relationship between the two variables that is, as the emissions of CO 2 grow, the consumptions of fuel similarly have the tendency to increase. The trend indicates that the higher the fuel consumption of the vehicle, the higher the level of carbon dioxide emitted, which is in accordance with the expectations as the direct contribution to the production of carbon dioxide is made by the fuel combustion. Several parallel groupings can be an indication of varying vehicle types or technologies, including hybrids and traditional engines. On the whole, this visualization can be useful to comprehend the environmental effects of fuel consumption and can serve as a basis of strategies to decrease the emission levels by improving the fuel efficiency.

2.3 Model Building

Task 1 – Neural Network The regression model that was created is based on a neural network: There are several thick layers of proper neurons. Hidden layer activation functions. Loss Function Loss: Mean Squared Error Adam and an appropriate learning rate. Trainvalidation split to avoid overfitting.

Task 2 Two Classical Machine Learning Models.

Two regression equations were used: Model A: Linear Regression Model B: Random Forest Regressor. That is, the models were trained on a conventional train-test split to guarantee impartial assessment.

2.4 Model Evaluation

Model Evaluation Models were evaluated using: Mean Absolute Error (MAE) Mean Squared Error (MSE) Root Mean Squared Error (RMSE) R^2 Score Smaller values of errors show better performance, and large values of R 2 are stronger predictors.

2.5 Hyperparameter Optimization

We used GridSearchCV, a methodical and exhaustive search approach that tests every possible combination of provided hyperparameter values, to find the best hyperparameters for the traditional machine learning models used in this work.

3. Results and Conclusion

3.1 Key Findings

The optimized models had better performance as compared to baseline models. The process of feature selection and hyperparameter tuning was very important in minimizing the errors in prediction.

3.2 Final Model

According to the evaluation metrics, Model B (Random Forest Regressor) has been selected with lowest values in error and high values in the R 2.

3.3 Challenges

The difficulties were how to deal with multicollinearity, which features are the best to use, and how to optimally tune the hyperparameters.

3.4 Future Work

Further developments may involve more sophisticated ensemble techniques, more sophisticated neural networks and further feature engineering.

4. Discussion

4.1 Model Performance

The models were good in the prediction of the continuous target variable. Reduced RMSE values will imply high accuracy in prediction.

4.2 Effect of Hyperparameter Optimization and Feature Selection.

Stability of the models was enhanced by tuning of the hyperparameters, and complexity and generalization were enhanced by feature selection.

4.3 Interpretation of Results

The important predictors were highly correlated with the target variable, which validated the results of EDA and correlation analysis.

4.4 Limitations

Wrong things are the size of data set and the assumption of linearity in certain models.

4.5 Recommendations on Future Study

Future directions in the research would be into deep learning architectures, external datasets and real-time prediction systems.