PROJECT REPORT DATA-SCIENCE FALL 2024

Predicting Impact of Vehicle Features on Emissions

Rabbi Awan, Ibrahim Moeed ,and Emad Rafique

Department of Computer Science, National University of Computer and Emerging

Sciences, Lahore November 27th, 2024

Abstract

The transportation sector significantly contributes to greenhouse gas emissions, primarily CO_2. This study explores the relationship between vehicle characteristics (e.g., engine size, fuel type, efficiency) and emissions using machine learning. Historical data spanning two decades was processed to train models, with the Random Forest Regressor outperforming others, achieving 98.7% accuracy and low error metrics (RMSE: 6.54 validation, 6.71 test). Its robustness and ability to handle non-linear relationships make it ideal for reliable predictions. The findings offer actionable insights for reducing emissions, supporting sustainable vehicle design and policymaking.

1 Introduction

With growing concerns about environmental sustainability and public health, analyzing vehicle emissions has become a priority for researchers, policymakers, and automakers. This project explores vehicle emissions data to identify key contributors such as engine size, fuel type, and efficiency metrics. By employing advanced data preprocessing, feature engineering, and machine learning models, the goal is to provide actionable insights into emissions patterns and predict emission levels effectively. This study highlights the importance of data-driven approaches in addressing environmental challenges related to vehicle emissions

2 Methodology

2.1 Data Collection :

- The dataset spans fuel consumption and vehicle attributes from 2000 to 2022, including features such as engine size, vehicle weight, fuel type, and CO_2 emissions.

2.2 Data Preprocessing :

- Removed missing values and duplicate entries to ensure data integrity.
- Standardized numerical features to maintain consistent scaling for machine learning algorithms.
- Encoded categorical variables, such as fuel type and transmission type, using label encoding for compatibility with predictive models.

2.3 Exploratory Data Analysis (EDA) :

- Conducted uni-variate analysis to examine the distribution of variables like engine size,

fuel consumption, and emissions using histograms and box plots.
- Performed bi-variate analysis, including scatter plots and heatmaps, to identify relationships between key features (e.g., engine size and emissions).
- Highlighted trends, such as larger engines and higher fuel consumption being associated with increased emissions, and efficiency being inversely correlated with emissions.

2.4 Dimensionality Reduction :

Feature Selection

- Recursive Feature Elimination (RFE) was used to identify key predictors of emissions, including engine size, efficiency, and combined fuel consumption.
- Feature Importance Analysis: Random Forest and XGBoost models were used to rank features by their contribution to emissions predictions, highlighting engine size, fuel consumption, and efficiency as the most influential variables.

2.5 Model Development :

- Linear Regression: Used as a baseline model for emissions prediction.
- Random Forest Regressor: Captured non-linear relationships and ranked feature importance.
- Extreme Gradient Boosting (XG Boost): Applied for robust predictions and handling feature interactions effectively.
- Decision Tree: Provided a simple yet interpretable model for understanding feature splits and their direct impact on emissions.

Evaluation Metrics

- Mean Absolute Error (MAE): Assessed average prediction error.
- R^2 (Coefficient of Determination): Evaluated the proportion of variance explained by the models.
- Root Mean Squared Error (RMSE): Measured prediction accuracy.

2.6 Tools and Libraries :

- Programming Language: Python.
- Libraries: Pandas and NumPy for data manipulation, Scikit-learn, Random Forest, Decision Tree, Extreme Gradient Boosting and Linear Regression for modeling, and Matplotlib and Seaborn for data visualization

3 Experiments

Key Visualizations:

Correlation Heatmap

- Visualizes the relationships between features like engine size, fuel consumption, and emissions to identify strong predictors and potential multicollinearity.

Trends of Variables with Emissions

- Scatter plots, box plot, hexbin plot, violin and strip plots, time graph, etc to illustrate the relationships between key variables (e.g., engine size, efficiency, fuel consumption) and emissions. These plots highlight how each variable impact emission levels.

Residual Histogram

- A histogram of residuals (differences between actual and predicted emissions) to assess the error distribution. A roughly normal distribution centered around zero indicates a well-fitted model.

Actual vs. Predicted Plot

- A scatter plot comparing the actual emission levels with the predicted levels, along with a reference diagonal line. This visualization assesses the alignment between predictions and ground truth, showcasing model performance.

Error Distribution

- A histogram or density plot of residuals to evaluate whether the errors are normally distributed and centered around zero, validating the assumptions of the regression model.

Insights

- Add observations derived from these visualizations, such as:
- High correlation between engine size and emissions.
- Residuals are randomly distributed, indicating a good model fit.
- Predicted vs. actual emissions align closely, reflecting strong model performance.

4 Results & Discussion

4.1 Results:

4.1.1 Model Performance

The Random Forest Regressor emerged as the most effective model for predicting vehicle emissions.

Performance metrics for the best model:

- Validation R^2: 0.987
- Test R^2: 0.986
- Validation RMSE: 6.54
- Test RMSE: 6.71
- The model effectively captured non-linear relationships between features and emissions, generalizing well to unseen data.

4.1.2 Feature Importance

Key predictors identified through feature importance analysis:

- Engine size
- Combined fuel consumption (L/100 km)
- Fuel type
- Efficiency
- Engine size showed the highest impact on emissions, highlighting its significance in the prediction task.

4.1.3 Insights from Visualizations

- **Correlation Heatmap:** Strong positive correlation between engine size and emissions, while efficiency showed an inverse relationship.

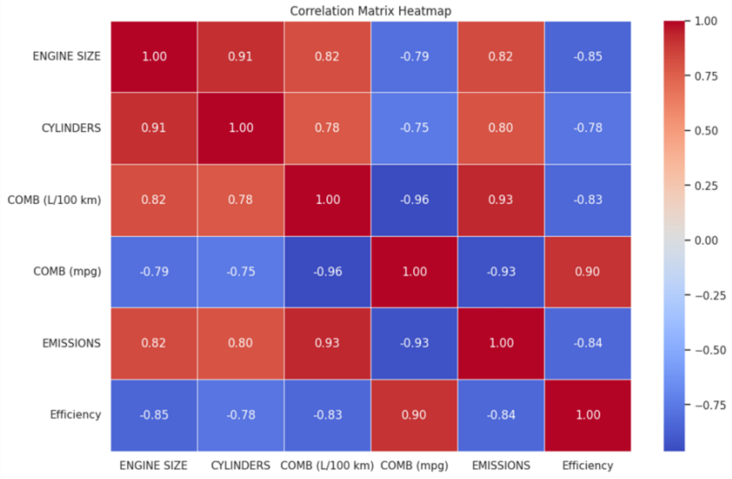

Fig.1 : The correlation matrix helps identify which variables are strongly correlated with emissions. Variables with high correlation values with emissions are likely to be useful for prediction

- **Trends Analysis:** Larger engines and higher fuel consumption are associated with increased emissions, whereas improved efficiency reduces emissions.

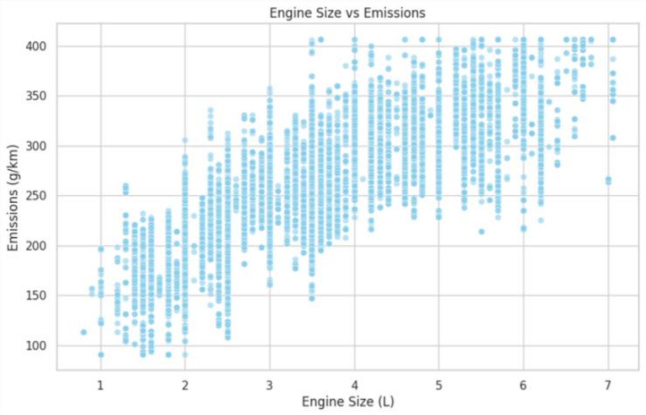

Fig.2: Scatter plot showing distribution among Engine Size and Emissions.

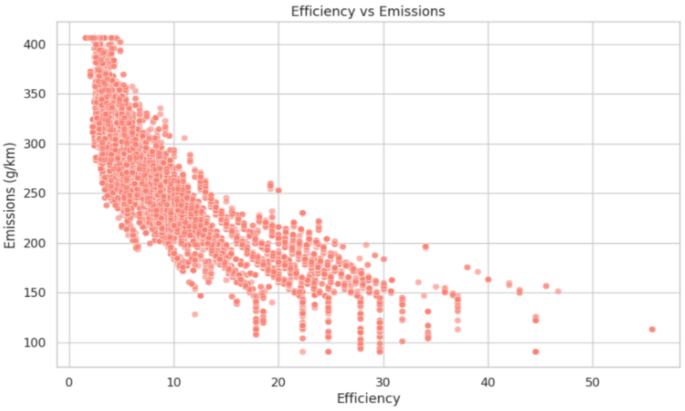

Fig.3 Scatter plot between Efficiency and Emission

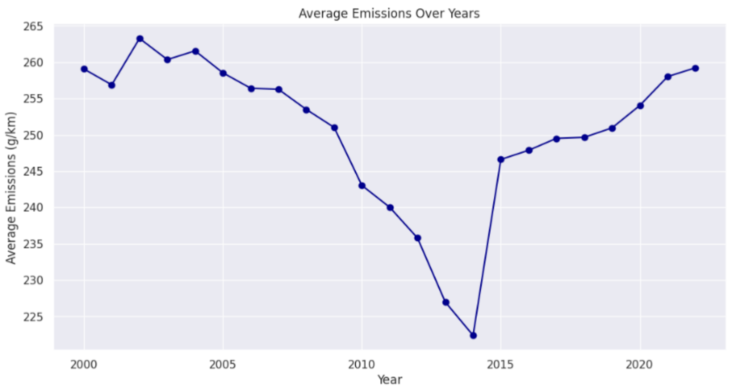

Fig.4 : Shows distribution of Emission over years

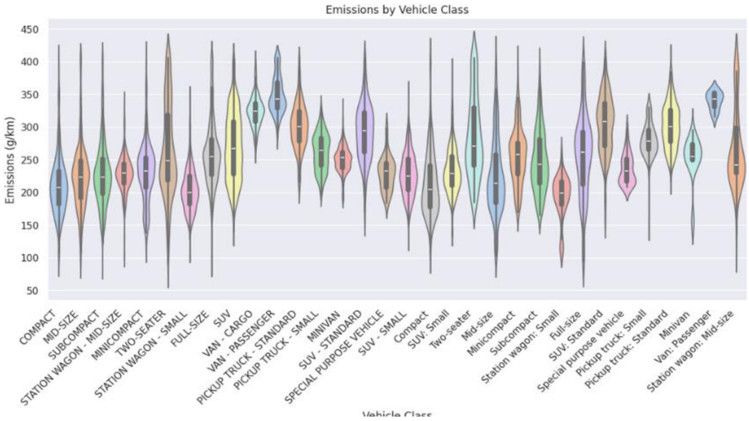

Fig 5: Violin plot of Emission by Vehicle class

- **Residual Histogram**: Residuals were symmetrically distributed, supporting model accuracy and no significant systematic errors.

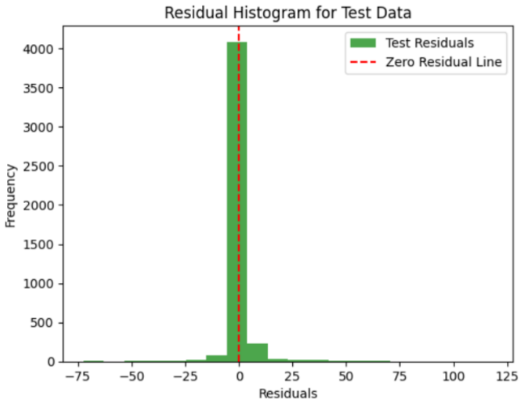

Fig .6 : The histogram shows that most residuals are close to zero, indicating good model performance.

- **Actual vs Predicted Plot**: Predicted values closely matched actual values, indicating strong model reliability.

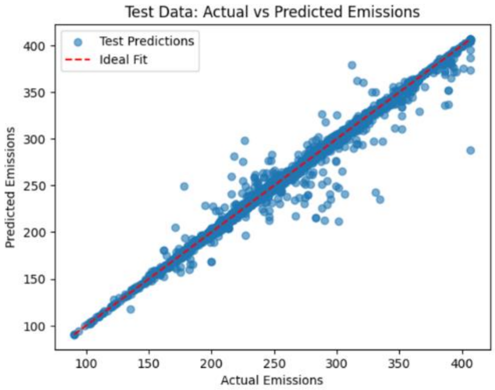

Fig.7 : The scatter plot shows a strong alignment of predicted emissions with actual emissions

4.2 Discussion

4.2.1 Interpretation of Results:

- The results confirm that vehicle attributes, particularly engine size, fuel consumption and efficiency, play a significant role in determining emissions.
- The model’s high accuracy and low error metrics indicate its robustness and suitability for real-world applications.

4.2.2 Implication:

- **For Automakers:** Insights into key predictors can guide the design of more fuel-efficient and environmentally friendly vehicles.
- **For Policymakers:** Data-driven evidence can support the development of regulations targeting high-emission vehicles.
- **For Consumers:** Provides transparency into how vehicle features impact emissions, enabling informed decisions when purchasing vehicles.

.

4.3.3 Challenges and Limitations:

- **Data Limitations:** The dataset, while comprehensive, may not fully represent global variations in vehicle features and emissions.
- **Model Assumptions:** The analysis assumes that the relationships between features and emissions remain consistent over time, which may not hold true for all scenarios.

5 Conclusion and Future Work

The study successfully demonstrated the critical impact of vehicle attributes, especially engine size and efficiency, on emissions. With the Random Forest Regressor achieving high accuracy and low error metrics, the model's robustness was validated for practical applications. These insights offer value to automakers for designing environmentally friendly vehicles, policymakers for implementing effective regulations, and consumers for making informed choices. This project contributes to understanding emission patterns and encourages a data-driven approach to reducing carbon footprints in the transportation sector.

Future Work:

1. **Incorporating Diverse Data Sources**: Future studies could explore datasets from global regions to account for variations in vehicle designs, regulations, and driving conditions, making the findings more generalizable.
2. **Dynamic Models**: Building time-series models to analyze changes in emission patterns over time and their correlation with evolving vehicle technologies and environmental policies.
3. **Expanding Features**: Including additional features such as weather conditions, road types, and maintenance records to enhance prediction accuracy.
4. **Real-time Applications**: Developing real-time predictive tools for monitoring and managing emissions for smart cities and intelligent transportation systems.
5. **Incorporating Diverse Data Sources**: Future studies could explore datasets from global regions to account for variations in vehicle designs, regulations, and driving conditions, making the findings more generalizable.

6. **Dynamic Models**: Building time-series models to analyze changes in emission patterns over time and their correlation with evolving vehicle technologies and environmental policies.
7. **Expanding Features**: Including additional features such as weather conditions, road types, and maintenance records to enhance prediction accuracy.
8. **Real-time Applications**: Developing real-time predictive tools for monitoring and managing emissions for smart cities and intelligent transportation systems.

References

[1] Dataset Available at: https://www.kaggle.com/datasets/ahmettyilmazz/fuel-consumption/data

[2] Scikit-learn: Machine Learning in Python Available at: https://scikit-learn.org/

[3] Matplotlib: Visualization with Python Available at: https://matplotlib.org/

[4] Breiman, L. (2001). "Random Forests." Machine Learning, 45(1), 5–32. Available at: https://link.springer.com/article/10.1023/A:1010933404324

[5] XGBoost Documentation Available at: https://xgboost.readthedocs.io/

[6] Scikit-learn: Linear Regression Implementation Available at: https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LinearRegression.html