Analyze and Predict the Effects of Automotive Features on CO2 Emission

DEPARTMENT OF MATHEMATICS AND COMPUTING
INDIAN INSTITUTE OF TECHNOLOGY
(INDIAN SCHOOL OF MINES, DHANBAD)

Final Report

for the award of the degree of
Integrated Master of Technology In Mathematics and Computing

Submitted By:
Shashvat Jain
20JE0897

Prof. A Antony Selvan
Department of Mathematics and Computing,
IIT (ISM) Dhanbad - 826004

INDIAN INSTITUTE OF TECHNOLOGY (ISM) DHANBAD

Certification from the guide

This is to certify that the Dissertation entitled *"Analyze and Predict the Effects of Automotive Features on CO_2 Emission"* being submitted to the Indian Institute of Technology (Indian School of Mines), Dhanbad, by Mr Shashvat Jain, Admission No 20JE0897 for the award of the Degree of Integrated Master of Technology in Mathematics and Computing from IIT (ISM), Dhanbad, is a bonafide work carried out by him/her, in the Department of Mathematics and Computing, IIT (ISM), Dhanbad, under my supervision and guidance.

The dissertation has fulfilled all the requirements as per the regulations of this Institute and, in my/our opinion, has reached the standard needed for submission. The results embodied in this dissertation have not been submitted to any other university or institute for the award of any degree or diploma.

_____

Signature of Supervisor(s)

Name: _____

Date: _____

INDIAN INSTITUTE OF TECHNOLOGY (ISM) DHANBAD

Declaration

I hereby declare that the work which is being presented in this dissertation entitled *"Analyze and Predict the Effects of Automotive Features on CO_2 Emission"* in partial fulfilment of the requirements for the award of the degree of **Integrated Master of Technology** in **Mathematics and Computing** is an authentic record of my own work carried out during the period from **2024** to **2025** under the supervision of **Prof. A Antony Selvan**, Department of **Mathematics and Computing**, Indian Institute of Technology (ISM) Dhanbad, Jharkhand, India.

I acknowledge that I have read and understood the UGC (Promotion of Academic Integrity and Prevention of Plagiarism in Higher Educational Institutions) Regulations, 2018. These Regulations were published in the Indian Official Gazette on 31st July, 2018.

I confirm that this Dissertation has been checked for plagiarism using the online plagiarism checking software provided by the Institute. At the end of the Dissertation, a copy of the summary report demonstrating similarities in content and its potential source (if any) generated online using plagiarism checking software is enclosed. I herewith confirm that the Dissertation has less than 10% similarity according to the plagiarism checking software's report and meets the MoE/UGC Regulations as well as the Institute's rules for plagiarism.

I further declare that no portion of the dissertation or its data will be published without the Institute's or Guide's permission. I have not previously applied for any other degree or award using the topics and findings described in my dissertation.

_____

(Signature of the Student)

Name of the Student: _____
Admission No.: _____
Department: _____

Forwarded by,
Prof. A Antony Selvan

INDIAN INSTITUTE OF TECHNOLOGY (ISM) DHANBAD

CERTIFICATE FOR CLASSIFIED DATA

This is to certify that the Dissertation entitled *"Analyze and Predict the Effects of Automotive Features on CO_2 Emission"* being submitted to the Indian Institute of Technology (Indian School of Mines), Dhanbad, by **Shashvat Jain**, Admission No. 20JE0897 for the award of the Degree of Integrated Master of Technology in Mathematics and Computing from IIT (ISM) Dhanbad, does not contain any classified information. This work is original and has not been submitted to any institution or university for the award of any degree.

Signature of the Guide **Signature of the Student**

INDIAN INSTITUTE OF TECHNOLOGY (ISM) DHANBAD

CERTIFICATE REGARDING ENGLISH CHECKING

This is to certify that the Dissertation entitled *"Analyze and Predict the Effects of Automotive Features on CO_2 Emission"* being submitted to the Indian Institute of Technology (Indian School of Mines), Dhanbad, by **Shashvat Jain**, Admission No. 20JE0897 for the award of the Degree of Integrated Master of Technology in Mathematics and Computing from IIT (ISM) Dhanbad, has been thoroughly checked for quality of English and logical sequencing of topics.

It is hereby certified that the standard of English is good and that grammar and typos have been thoroughly checked.

Signature of the Guide **Signature of the Student**

INDIAN INSTITUTE OF TECHNOLOGY (ISM) DHANBAD

COPYRIGHT AND CONSENT FORM

To ensure uniformity of treatment among all contributors, other forms may not be substituted for this form, nor may any wording of the form be changed. This form is intended for original material submitted to the IIT (ISM), Dhanbad and must accompany any such material in order to be published by the ISM. Please read the form carefully and keep a copy for your files.

TITLE OF DISSERTATION: Analyze and Predict the Effects of Automotive Features on CO_2 Emission
AUTHOR'S NAME: Shashvat Jain

COPYRIGHT TRANSFER

1. The undersigned hereby assigns to Indian Institute of Technology (Indian School of Mines), Dhanbad all rights under copyright that may exist in and to:

   (a) the above Work, including any revised or expanded derivative works, submitted to the ISM by the undersigned based on the work; and

   (b) any associated written or multimedia components or other enhancements accompanying the work.

CONSENT AND RELEASE

1. In the event the undersigned makes a presentation based upon the work at a conference hosted or sponsored in whole or in part by the IIT (ISM) Dhanbad, the undersigned, in consideration for his/her participation in the conference, hereby grants the ISM the unlimited, worldwide, irrevocable permission to use, distribute, publish, license, exhibit, record, digitize, broadcast, reproduce and archive; in any format or medium, whether now known or hereafter developed:

   (a) his/her presentation and comments at the conference;

   (b) any written materials or multimedia files used in connection with his/her presentation; and

   (c) any recorded interviews of him/her (collectively, the "Presentation").

   The permission granted includes the transcription and reproduction of the Presentation for inclusion in products sold or distributed by IIT(ISM) Dhanbad and live or recorded broadcast of the Presentation during or after the conference.

2. In connection with the permission granted in Section 2, the undersigned hereby grants IIT (ISM) Dhanbad the unlimited, worldwide, irrevocable right to use his/her name, picture, likeness, voice and biographical information as part of the advertisement, distribution and sale of products incorporating the Work or Presentation, and releases IIT (ISM) Dhanbad from any claim based on right of privacy or publicity.

3. The undersigned hereby warrants that the Work and Presentation (collectively, the "Materials") are original and that he/she is the author of the Materials. To the extent the Materials incorporate text passages, figures, data or other material from the works of others, the undersigned has obtained any necessary permissions. Where necessary, the undersigned has obtained all third party permissions and consents to grant the license above and has provided copies of such permissions and consents to IIT (ISM) Dhanbad.

GENERAL TERMS

- The undersigned represents that he/she has the power and authority to make and execute this assignment.
- The undersigned agrees to indemnify and hold harmless the IIT (ISM) Dhanbad from any damage or expense that may arise in the event of a breach of any of the warranties set forth above.
- In the event the above work is not accepted and published by the IIT (ISM) Dhanbad or is withdrawn by the author(s) before acceptance by the IIT (ISM) Dhanbad, the foregoing copyright transfer shall become null and void and all materials embodying the Work submitted to the IIT(ISM) Dhanbad will be destroyed.
- For jointly authored Works, all joint authors should sign, or one of the authors should sign as authorized agent for the others.

Signature of the Student

_____

Acknowledgments

I would like to begin by thanking my mentor **Professor A Antony Selvan**, Department of Mathematics and Computing, IIT (ISM) Dhanbad, who has supported me throughout my dissertation with his patience and knowledge. I attribute the attainment of my Master's degree to his unwavering encouragement and dedicated efforts. Without his support, this thesis would not have been completed or formulated. I express my sincere gratitude to him for his belief in my abilities and continual motivation to pursue this subject of personal interest. One could not aspire for a more exemplary supervisor.

My profound gratitude is extended to **Prof. S.P. Tiwari**, our department head, and I would also like to express my sincere thanks to **Prof. A Antony Selvan**, the course coordinator of Integrated M.Tech., Mathematics and Computing, IIT (ISM) Dhanbad, for providing me with all the necessary sources to complete my dissertation.

I am appreciative of my juniors, fellow batchmates, and the Professors at the Indian Institute of Technology (ISM) Dhanbad. I am also grateful to my **parents** for their unwavering support and unceasing encouragement during my years of education, as well as during the process of conducting the research and composing this thesis.

I would like to express my gratitude to the Department of Mathematics and Computing and its staff, who have consistently shown me kindness and support.

Analyze and Predict the Effects of Automotive Features on CO2 Emission

Abstract

The increasing level of carbon dioxide (CO_2) emissions from the transportation industry is a threat of extreme urgency to the stability of the global climate and requires immediate scientific attention to model, analyze, and finally reduce car emissions. Precise CO_2 emissions prediction from the technical attributes of the car is still difficult despite dramatic advances, owing to the nonlinear and high-dimensional nature of the underlying relationships. This thesis fills this lacuna by mathematically exploring in depth the effect of different car attributes like engine size, fuel type, vehicle weight, and transmission technology on CO_2 emissions, and by creating an extremely accurate prediction model.

The study performed a comprehensive exploratory data analysis (EDA) to uncover correlations and distributions of characteristics, followed by implementing a robust machine learning pipeline. The proposed methodology integrates advanced preprocessing techniques, feature transformations, and a hybrid ensemble model comprising LightGBM, XGBoost, and CatBoost regressors. These base learners are combined through a meta-learner modeled by a Multi-Layer Perceptron (MLP) and further refined with residual correction using Ridge regression. Hyperparameter optimization is rigorously performed through Bayesian optimization via Optuna to ensure model generalization and minimize overfitting.

Experimental results have demonstrated that the final proposed stacked ensemble pipeline significantly outperformed traditional regression models, achieved superior R^2 scores and reduced error metrics across both training and unseen test datasets. Beyond predictive performance, this work provides actionable insights into which automotive features most critically influence CO_2 emissions.

The findings contribute to the fields of sustainable transportation engineering and environmental data science, offering practical implications for policymakers, automotive manufacturers, and researchers aiming to design environmentally responsible vehicles.

List of Contents

1 Introduction

1.1 Motivation and Background

The transportation sector, particularly the automotive industry, contributes significantly to global carbon dioxide (CO_2) emissions. With the escalating impacts of climate change, there is an urgent demand for strategies that can help monitor, manage, and ultimately reduce the environmental footprint of vehicles. Governments worldwide have imposed stringent emission regulations, and automobile manufacturers invest heavily in greener technologies. However, a crucial intermediate step involves a deep understanding of the factors influencing vehicle emissions and the ability to predict emissions based on measurable automotive attributes.

Recent advancements in machine learning (ML) and data-driven modeling offer a promising pathway toward achieving more accurate, efficient, and scalable emission predictions. By leveraging historical data and advanced analytical techniques, it becomes possible to uncover complex, non-linear relationships between vehicle features, such as engine size, fuel type, or aerodynamic design, and their corresponding CO_2 output.

Thus, there is a pressing need for a scientifically rigorous yet practically applicable framework that can predict vehicle emissions with high accuracy, aiding policymakers, manufacturers, and environmental analysts alike.

1.2 Problem Statement

Despite the growing attention on emission control, existing approaches often face limitations such as:

- Reliance on simplified linear models that fail to capture the true complexity of vehicular systems.
- Lack of integrated pre-processing techniques, leading to suboptimal performance.
- Inability to generalize across diverse vehicle types and feature variations.
- Limited utilization of automated hyperparameter optimization methods for model tuning.

Given the multidimensional nature of automotive data, an effective emission prediction system must handle:

- Mixed data types (numerical and categorical features)
- Non-linear feature interactions
- Feature scaling and transformation
- Overfitting and generalization balance

Hence, there is a strong research need to design a robust end-to-end ML pipeline that not only predicts emissions accurately but is also interpretable, scalable, and adaptable to real-world datasets.

1.3 Objectives of the Study

This research focuses on the following objectives:

- **Exploratory Data Analysis (EDA):** Perform in-depth EDA to understand the underlying structure, distributions, and relationships among variables.
- **Feature Engineering and Preprocessing:** Apply advanced preprocessing techniques, including standardization, encoding, and target transformation.
- **Model Development:** Build a high-performing ensemble model using LightGBM, XGBoost, and CatBoost regressors with a neural network meta-learner.
- **Hyperparameter Optimization:** Implement Optuna-based Bayesian optimization for efficient hyperparameter tuning.
- **Pipeline Construction:** Design a modular pipeline integrating preprocessing, model training, stacking, and residual correction.
- **Performance Evaluation:** Rigorously evaluate model performance using appropriate regression metrics like R^2, MAE, MSE, and RMSE.
- **Deployment-readiness:** Ensure that the pipeline is flexible and scalable for potential deployment scenarios.

This study sets the stage for exploring advanced techniques that address real-time environmental issues, paving the way for practical and innovative solutions in the field of CO_2 emissions reduction.

1.4 Scope of the Research

The scope of this thesis encompasses:

- **Data Source:** The dataset used is publicly available on Kaggle and provides detailed automotive specifications along with corresponding CO_2 emissions.
- **Target Variable:** The prediction target is the rate of CO_2 emission (g/km) based on multiple vehicle attributes.
- **Methodologies:** This work primarily focuses on supervised learning techniques, with an emphasis on ensemble models and meta-learning strategies.
- **Limitation Considerations:**
  - No external data (e.g., real-time driving conditions) is incorporated.
  - Interpretability analysis (e.g., SHAP values) is not the core focus but can be explored as future work.

1.5 Research Significance

This thesis holds significance in multiple dimensions:

- **Environmental Impact:** By accurately predicting CO_2 emissions, it supports initiatives aimed at reducing emissions and ensuring regulatory compliance.
- **Industrial Applications:** Automotive manufacturers can integrate such predictive models into their design and testing pipelines to evaluate environmental performance at early stages.
- **Policy Formulation:** Policymakers can use insights from predictive models to develop evidence-based emission standards and incentives.
- **Advancement of Machine Learning Techniques:** The stacking and residual-correction strategy showcased in this study demonstrates an innovative approach to improving predictive accuracy in regression tasks involving environmental data.

2 Literature Review

2.1 Introduction

The environmental concerns associated with greenhouse gas emissions have intensified research efforts to understand and predict vehicular CO_2 emissions. Machine learning techniques, especially ensemble models and neural networks, have emerged as powerful tools in this domain. This chapter presents a structured review of existing literature, organized into two primary segments:

- Studies related to vehicle emissions modeling
- Studies related to machine learning pipelines for structured regression problems

Finally, the chapter concludes by identifying critical research gaps that are addressed by this thesis.

2.2 Studies on Vehicle Emission Prediction

2.2.1 Traditional Statistical Approaches

Initial attempts to model vehicle emissions primarily utilized classical statistical methods, such as:

- **Linear Regression Models:** Early studies employed linear relationships between engine size, fuel consumption, and emission levels. However, these models often struggled with non-linear behaviors.
- **Multivariate Regression Analysis:** Some researchers used multivariate approaches to incorporate additional vehicle attributes, slightly improving predictive performance but still constrained by linear assumptions.

Limitations observed:

- Poor handling of complex, non-linear interactions
- High sensitivity to outliers and multicollinearity

2.2.2 Machine Learning-Based Approaches

The limitations of traditional models led to the adoption of machine learning (ML) algorithms. Key trends include:

- **Support Vector Machines (SVM):** SVM is employed by Non-linear regression tasks for solving; however, scalability to large datasets remains a challenge.
- **Random Forests (RF):** Provided significant improvements by capturing feature interactions but occasionally lacked precision for highly dynamic features.
- **Gradient Boosting Machines (GBMs):** Models such as XGBoost and LightGBM demonstrated strong predictive capabilities, especially on structured automotive datasets.

2.2.3 Domain Specific Findings

Several studies provided valuable domain-specific insights:

- **Engine Displacement and CO_2 Emissions:** Numerous papers reported a strong positive correlation.
- **Vehicle Weight Influence:** Heavier vehicles tend to exhibit higher emissions, aligning with thermodynamic principles.
- **Fuel Type Differentiation:** Emission patterns differ considerably between gasoline, diesel, hybrid, and electric vehicles.

However, these studies often emphasized model accuracy without delving into pipeline design or generalized system construction.

2.3 Studies on Machine Learning Pipelines for Structured Data

2.3.1 Data Preprocessing Techniques

Effective preprocessing is a cornerstone of successful ML applications. The literature highlights several techniques:

- **Handling Missing Values:** Mean imputation, median substitution, and more recently, K-Nearest Neighbors (KNN) imputation have been common.
- **Feature Scaling:** Standardization (z-score normalization) and min-max scaling are frequently used, particularly for tree-based models to improve convergence.
- **Categorical Encoding:** One-hot encoding, label encoding, and target encoding strategies are employed based on the algorithmic requirements.

Despite the awareness, many existing emission studies treated preprocessing as an afterthought rather than an integrated part of the pipeline.

2.3.2 Model Selection and Ensemble Techniques

Advanced pipelines often incorporate ensemble models to improve performance:

- **Bagging Techniques:** Methods like Random Forests emphasize variance reduction by training models on bootstrapped datasets.
- **Boosting Techniques:** Algorithms such as XGBoost, CatBoost, and LightGBM focus on bias reduction, yielding highly accurate models.
- **Stacking:** Some studies explored stacking multiple models to create a meta-predictor, but applications in emission prediction remain sparse.

2.3.3 Hyperparameter Tuning Methods

Hyperparameter optimization is critical for maximizing model potential. Approaches include:

- **Grid Search:** Exhaustive search across specified parameter grids; computationally intensive.
- **Random Search:** Faster but less exhaustive.
- **Bayesian Optimization:** More recent studies favor Bayesian approaches, with tools like Optuna and Hyperopt, achieving a balance between exploration and exploitation.

However, within the context of vehicle emission modeling, most published works still rely on grid or random search, underutilizing newer optimization techniques.

2.4 Research Gaps Identified

Based on the literature reviewed, several gaps persist:

- **Pipeline Integration:** A lack of comprehensive end-to-end pipelines encompassing data pre-processing, model training, feature engineering, and evaluation cohesively.
- **Ensemble and Meta-Learning:** Limited adoption of advanced ensemble techniques such as stacking regressors with neural network meta-learners for structured data.
- **Hyperparameter Optimization:** Underutilization of Bayesian hyperparameter optimization strategies in the emission prediction context.
- **Residual Analysis and Correction:** Few studies implement post-prediction residual modeling to further enhance performance.
- **Generalizability Focus:** Models are often tuned for specific datasets without ensuring broader applicability across varying vehicle types or geographies.

2.5 Summary

This subsection reviewed the evolution of CO_2 emission prediction approaches, highlighting the transition from traditional statistical techniques to sophisticated machine learning approaches. While significant progress has been made, there remains substantial scope for improvement, particularly regarding pipeline integration, ensemble learning strategies, hyperparameter optimization, and residual correction mechanisms. Addressing these gaps forms the central focus of this thesis.

3 Methodology

3.1 Introduction

This subsection presented a detailed exposition of the methodological framework adopted in this research. The workflow is organized into two principal phases:

- **Exploratory Modeling Approaches:** Sequential implementation and assessment of individual machine learning methods to establish performance baselines and extract insights.
- **Integrated Ensemble Pipeline:** Design and deployment of a unified pipeline that amalgamates preprocessing, feature transformation, hyperparameter optimization, model stacking, and residual correction.

Each section provides a clear exposition of preprocessing, model construction, training strategies, and evaluation metrics.

3.2 Dataset Description

It comprises detailed information about 27813 vehicles, which includes vehicle attributes and their corresponding CO_2 emissions. The dataset includes the following major attributes:

1. **Make:** Manufacturer name
2. **Model:** Vehicle model
3. **Vehicle Class:** Type (SUV, sedan, truck, etc.)
4. **Engine Size (L):** Engine size (in liters)
5. **Cylinders:** Number of engine cylinders
6. **Transmission:** Type of transmission (automatic, manual, etc.)
7. **Fuel Type:** Type of fuel used (gasoline, diesel, electric, etc.)
8. **Fuel Consumption (City, Hwy, Combined(in L/100 km and mpg):** Measured in L/100 km and mpg (miles per gallon)
9. CO_2 **Emissions (g/km):** Target variable for prediction

Dataset at a glance:

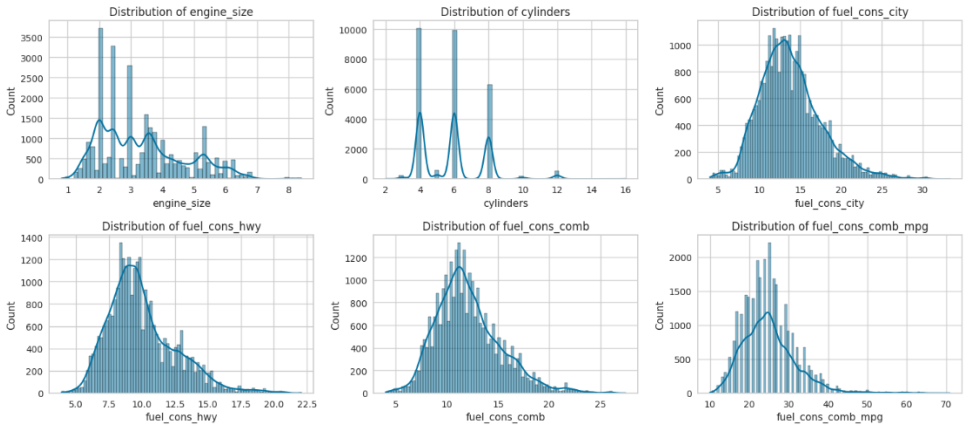
	Make	Model	Vehicle Class	Engine Size(L)	Cylinders	Transmission	Fuel Type	Fuel Consumption City (L/100 km)	Fuel Consumption Hwy (L/100 km)	Fuel Consumption Comb (L/100 km)	Fuel Consumption Comb (mpg)	CO2 Emissions(g/km)
0	Acura	Integra A-SPEC	Full-size	1.5	4	AV7	Z	8.1	6.5	7.4	38	172
1	Acura	Integra A-SPEC	Full-size	1.5	4	M6	Z	8.9	6.5	7.8	36	181
2	Acura	Integra Type S	Full-size	2.0	4	M6	Z	11.1	8.3	9.9	29	230
3	Acura	MDX SH-AWD	Sport utility vehicle: Small	3.5	6	AS10	Z	12.6	9.4	11.2	25	263
4	Acura	MDX SH-AWD Type S	Sport utility vehicle: Standard	3.0	6	AS10	Z	13.8	11.2	12.4	23	291

A combination of categorical and continuous features makes this dataset suitable for mixed-data modeling approaches.

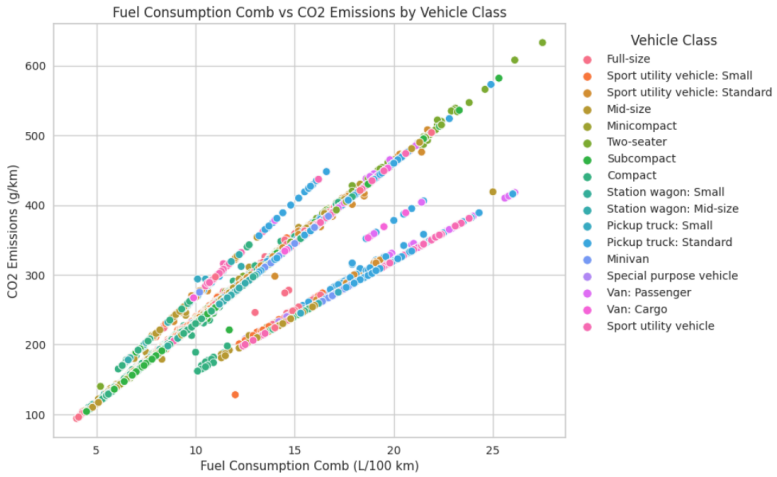
3.3 Exploratory Data Analysis (EDA)

Comprehensive EDA provided a foundational understanding of feature distributions, relationships, and potential modeling challenges.

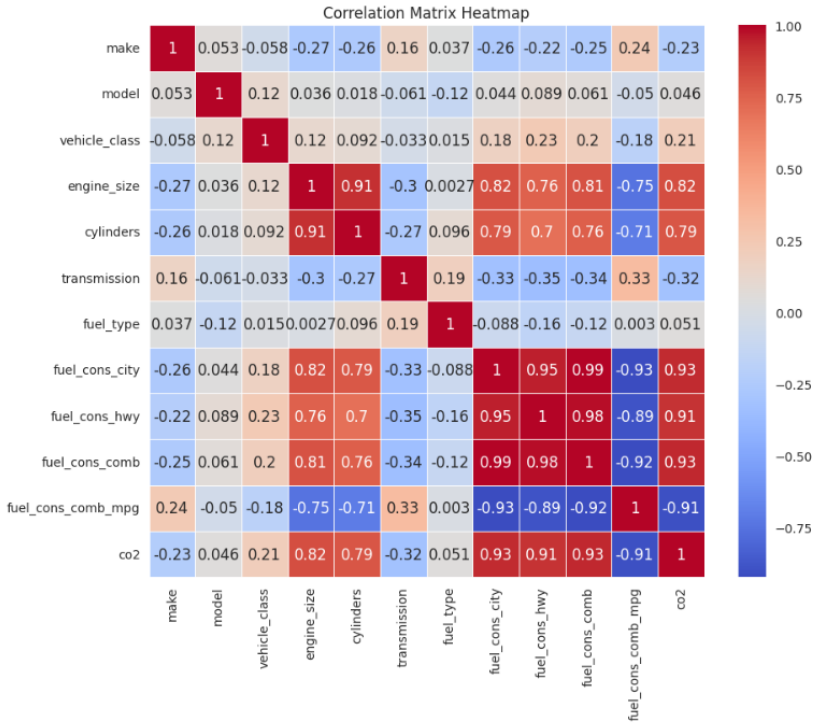
- **Univariate Distributions:**
  - Histograms and kernel density estimates for continuous variables (engine size, cylinders, fuel consumption, CO_2 emissions).
  - Frequency plots for categorical variables (fuel type, transmission, vehicle class).



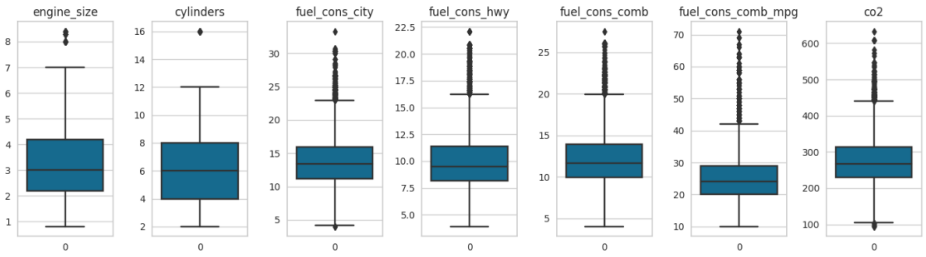
- **Bivariate Relationships:**
  - Scatterplots illustrating linear and non-linear relationships (e.g., combined fuel consumption vs. CO_2).
  - Boxplots to compare emission distributions across categorical groups.



- **Correlation Analysis:** Pearson correlation heatmap among numerical features revealed strong associations, especially between combined fuel consumption and CO_2 emissions ($r > 0.85$).



- **Outlier Detection:** Interquartile range (IQR) method flagged extreme emission values; outliers were retained to preserve model generality but downweighted by robust scaling.



Findings from EDA guided feature engineering and informed model choice.

3.4 Data Preprocessing and Feature Engineering

A consistent preprocessing workflow was implemented via scikit-learn's Pipeline and ColumnTransformer.

- **Missing Value Imputation:** Median substitution for numerical features; new category "Unknown" for missing categorical entries.
- **Categorical Encoding:** One-Hot Encoding for nominal features.

- **Feature Scaling:** StandardScaler applied to all numerical variables to standardize means and variances.
- **Target Normalization:** PowerTransformer(method='yeo-johnson') transformed the CO_2 emission distribution towards Gaussianity.

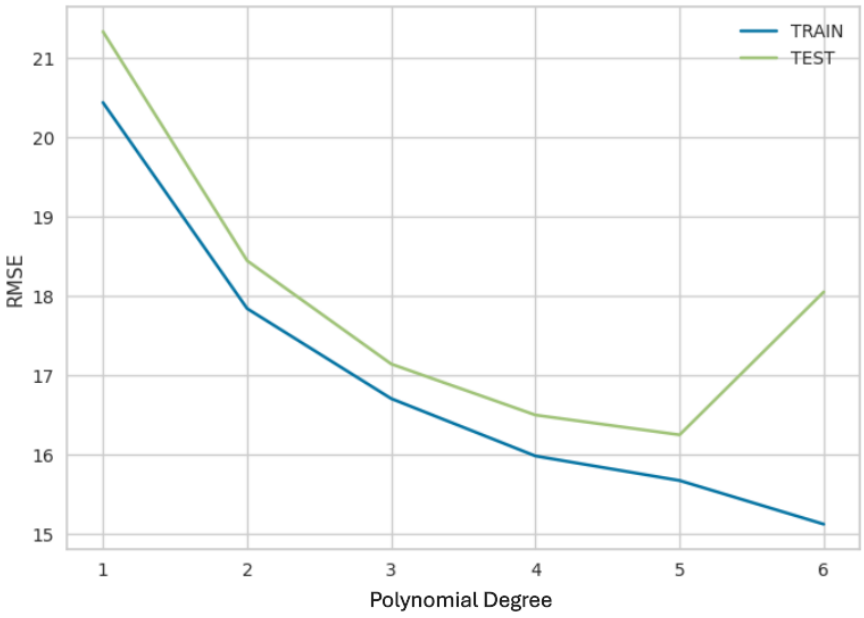
All transformers were fitted exclusively on training data to avoid information leakage.

3.5 Exploratory Modeling Approaches

To build intuition and benchmark predictive capability, a spectrum of algorithms was trained on the preprocessed dataset. All models used the same training–testing split (80/20) and performance metrics (R^2, MAE, MSE, RMSE).

3.5.1 Linear Regression Variants

- **Simple Linear Regression:** Single-predictor model (engine size vs. CO_2). *Test performance:* $R^2 = 0.6785$, MAE = 28.82 g/km, RMSE = 37.58 g/km. This baseline demonstrates a modest fit, indicating that engine size alone explains only about 68% of the variance in emissions.
- **Multiple Linear Regression:** Incorporates all continuous features (engine size, cylinders, fuel consumption metrics). *Test performance:* $R^2 = 0.8965$, MAE = 11.86 g/km, RMSE = 21.20 g/km. The inclusion of multiple predictors substantially improves accuracy, capturing nearly 90% of the variance.
- **Polynomial Regression (Degree 4):** Extends multiple regression with polynomial terms up to fourth order to model non-linear relationships. *Test performance:* $R^2 = 0.9355$, MAE = 7.50 g/km, RMSE = 16.83 g/km. The higher-degree terms effectively capture curvature in the data, further reducing prediction error.



3.5.2 Regularized Linear Models

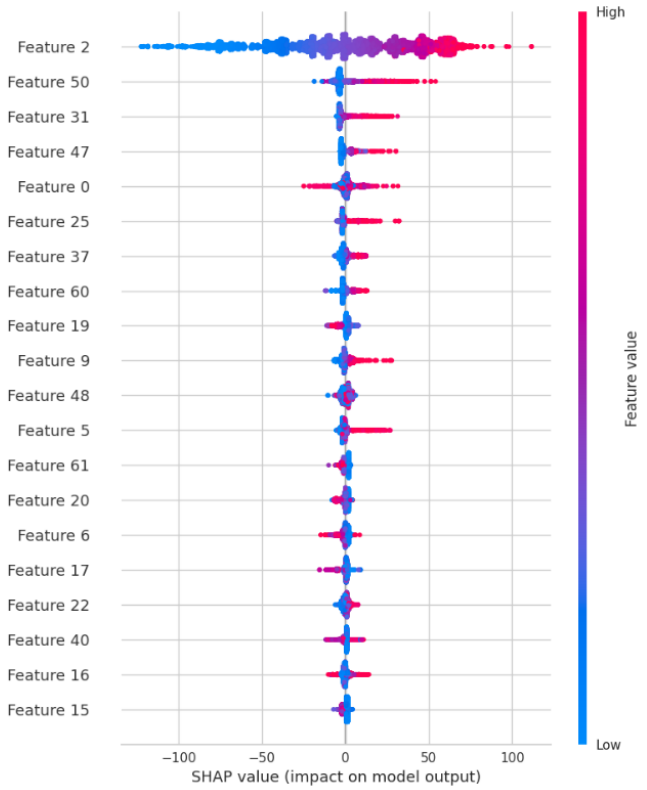
- **Ridge Regression (L2):** Adds an L2 penalty to mitigate multicollinearity among predictors. *Test performance:* $R^2 = 0.9282$, MAE = 8.44 g/km, RMSE = 17.76 g/km. Regularization yields a smoother solution, with only a slight decrease compared to the polynomial model, but improved stability.

- **Lasso Regression (L1):** Utilizes an L1 penalty to enforce sparsity and perform implicit feature selection. *Test performance:* $R^2 = 0.9030$, MAE = 11.87 g/km, RMSE = 20.64 g/km. The model simplifies the predictor set at the cost of some predictive power.

3.5.3 Tree-Based and Ensemble Methods

- **Random Forest Regressor:** Aggregates multiple decision trees to capture complex feature interactions. *Test performance:* $R^2 = 0.9718$, MAE = 3.71 g/km, RMSE = 11.13 g/km. Demonstrates strong non-linear modeling capability and robustness to outliers.

- **XGBoost Regressor:** Implements gradient boosting with regularization and tree pruning. *Test performance:* $R^2 = 0.9707$, MAE = 3.91 g/km, RMSE = 11.35 g/km. Slightly lower variance explained than Random Forest, but comparable error metrics.



- **LightGBM Regressor:** Utilizes histogram-based boosting for faster training on large feature spaces. *Test performance:* $R^2 = 0.9715$, MAE = 4.04 g/km, RMSE = 11.20 g/km. Matches XGBoost in accuracy with improved computational efficiency.

- **CatBoost Regressor:** Provides native handling of categorical features and ordered boosting. *Test performance:* $R^2 = 0.9714$, MAE = 4.04 g/km, RMSE = 11.22 g/km. Comparable to other boosting methods, with minimal preprocessing for categorical data.

Synthesis: Insights from these experiments—particularly feature importance rankings and error patterns—guided the construction of the final ensemble pipeline.

3.6 Final Ensemble Pipeline

The ensemble strategy developed in this study embodies a multi-stage learning system carefully designed to leverage the strengths of different gradient boosting models, enhance their predictive capabilities through a meta-learner, and correct residual errors with a final adjustment model. This section provides an expanded discussion on the design philosophy, training workflow, and rationale behind each module of the final pipeline.

3.6.1 Pipeline Architecture

The complete ensemble pipeline consists of three fundamental layers:

- **Preprocessing Layer:** Standardizes and encodes the input data to ensure model compatibility and improve learning stability.
- **Stacked Base Learners:** Parallel training of three highly efficient gradient boosting algorithms to generate diverse yet complementary predictions.
- **Meta-Learning and Residual Correction Layer:** Aggregates base learner outputs via a neural network meta-learner and subsequently applies Ridge regression to model any remaining systematic errors.

The modularity of the pipeline ensures scalability, interpretability, and minimal information leakage across the workflow.

3.6.2 Detailed Description of Components

- **Data Preprocessing**
  - **Numerical Features:** Standardization using StandardScaler to ensure zero mean and unit variance.
  - **Categorical Features:** Transformation using OneHotEncoder, preserving unknown categories to manage unseen inputs gracefully.
  - **Target Variable:** Normalized using PowerTransformer (Yeo–Johnson method) to mitigate skewness and heteroscedasticity, facilitating better convergence during training. All transformations were encapsulated in a ColumnTransformer and fitted exclusively on the training dataset.
- **Base Learners** The first layer comprises three diverse and complementary gradient boosting models:
  - **LightGBM Regressor:**

    - Fast training using histogram-based algorithms.
    - Leaf-wise tree growth enhances model complexity management.
  - **XGBoost Regressor:**
    - Regularized boosting prevents overfitting.
    - Sophisticated tree-pruning and parallelized learning mechanisms.
  - **CatBoost Regressor:**
    - Native support for categorical variables without explicit encoding.
    - Utilizes ordered boosting for reduced prediction bias.

  These learners were independently optimized for their hyperparameters through Bayesian optimization, ensuring optimal balance between bias and variance.

- **Meta-Learner** A shallow Multi-Layer Perceptron (MLP) Regressor was employed as the meta-learner:
  - **Architecture:** Single hidden layer with tuned neuron counts.
  - **Activation:** ReLU for hidden layers; Linear activation for output layer.
  - **Optimizer:** Adaptive gradient descent (Adam) with a tuned learning rate.
  - **Loss Function:** Used Mean Squared Error(MSE) for loss function.

  The meta-learner synthesized the out-of-fold predictions from the base learners to minimize the generalization error.

- **Residual Correction Model** After stacking, residuals (errors between predicted and actual normalized targets) were modeled using a Ridge Regression:
  - **Objective:** Capture and correct systematic biases missed by the ensemble.
  - **Regularization:** L2 penalty to control model complexity and prevent overfitting.

  This final adjustment ensured that even subtle residual structures were accounted for, maximizing final predictive precision.

3.6.3 Hyperparameter Optimization

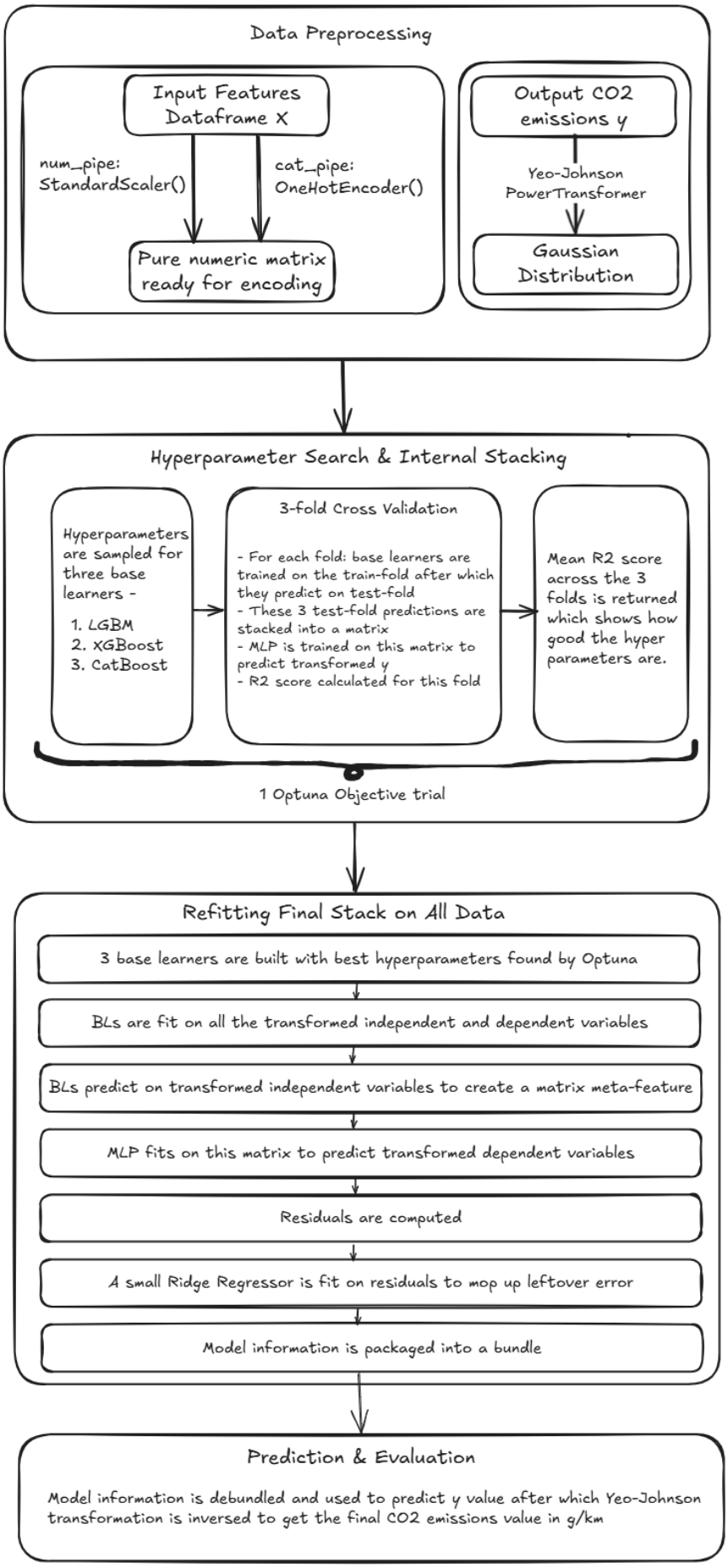
Bayesian optimization with Optuna maximizes mean cross-validated R^2. The search space includes:

- Number of trees, learning rate, tree depth for each base learner
- Hidden layer size, regularization coefficient, and learning rate for the MLP

Optimization runs employ 3-fold cross-validation to balance computational cost and robustness.

3.6.4 Training and Prediction Workflow

- Fit preprocessor and target transformer on training data.
- Optimize base learners and MLP via Optuna's cross-validated trials.
- Retrain tuned models on complete training data.
- Generate stacked features and fit meta-learner.
- Train Ridge regressor on meta-learner residuals.
- Predict by applying transforms to new data, obtaining base predictions, meta combination, residual adjustment, and inverse-transform of targets.


Data Preprocessing
Input Features Dataframe X
num_pipe: StandardScaler()
cat_pipe: OneHotEncoder()
Pure numeric matrix ready for encoding
Output CO2 emissions y
Yeo-Johnson PowerTransformer
Gaussian Distribution
Hyperparameter Search & Internal Stacking
Hyperparameters are sampled for three base learners -
1. LGBM
2. XGBoost
3. CatBoost
3-fold Cross Validation
- For each fold: base learners are trained on the train-fold after which they predict on test-fold
- These 3 test-fold predictions are stacked into a matrix
- MLP is trained on this matrix to predict transformed y
- R2 score calculated for this fold
Mean R2 score across the 3 folds is returned which shows how good the hyper parameters are.
1 Optuna Objective trial
Refitting Final Stack on All Data
3 base learners are built with best hyperparameters found by Optuna
BLs are fit on all the transformed independent and dependent variables
BLs predict on transformed independent variables to create a matrix meta-feature
MLP fits on this matrix to predict transformed dependent variables
Residuals are computed
A small Ridge Regressor is fit on residuals to mop up leftover error
Model information is packaged into a bundle
Prediction & Evaluation
Model information is debundled and used to predict y value after which Yeo-Johnson transformation is inversed to get the final CO2 emissions value in g/km

3.7 Model Evaluation Metrics

The evaluation of the predictive models has been performed using a comprehensive set of statistical metrics designed to measure different aspects of model performance and generalization capability.

3.7.1 Primary Evaluation Metrics

- **Coefficient of Determination (R^2 Score)**:
  The R^2 statistic measures the proportion of variance in the observed target variable that is explained by the model. Formally,

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}, \tag{1}$$

  where y_i are the true emission values, $\hat{y}_i$ are the model predictions, and $\bar{y}$ is the mean of the observed values.
  An R^2 of 1 shows a perfect prediction; on the other hand, a value of 0 implies that the model will perform similar to predicting the mean. Negative R^2 values can occur when the model fits worse than the horizontal line at $\bar{y}$. In this study, high R^2 values (above 0.90) signify strong explanatory power of our ensemble pipeline.

- **Mean Absolute Error (MAE)**:
  The MAE quantifies the absolute average deviation between actual and predicted values:

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}\left|y_i - \hat{y}_i\right|. \tag{2}$$

  By treating all errors equally, MAE provides an interpretable which is measured in the same units as the target, CO_2 emissions, (g/km). A lower MAE reflects more accurate and reliable predictions, with reduced average bias.

- **Mean Squared Error (MSE)**:
  The MSE is used to calculate the average of squared deviations, thereby penalizing larger errors more severely:

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2. \tag{3}$$

  Since outliers contribute quadratically, MSE is sensitive to occasional large mispredictions. It is widely used in optimization procedures due to its smooth, differentiable properties.

- **Root Mean Squared Error (RMSE)**:
  The RMSE is defined as the square root of the MSE, translating the penalized error back into the original measurement scale:

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}. \tag{4}$$

RMSE helps in retaining the sensitivity of MSE to large errors but yields values that are directly comparable to the original units of CO_2 emissions. A smaller RMSE indicates that large deviations are minimal, reflecting high model precision.

- **Metric Selection Rationale**:
  Together, these metrics provide a comprehensive evaluation:
  - R^2 assesses overall variance explained.
  - MAE measures average model bias in real units.
  - MSE and RMSE emphasize the impact of large errors, guiding optimization towards both accuracy and robustness.

  Reporting all four metrics ensures that model performance is evaluated from both statistical and practical perspectives, aligning with best practices in environmental data science.

3.7.2 Validation Strategy

To ensure that the reported performance reflects genuine predictive capability rather than overfitting to a specific data split, a two-pronged validation scheme was adopted:

- **Hold-out Test Set**: A randomly selected 20% subset of the full dataset was withheld before any model training or hyperparameter tuning. This "hold-out" set remained untouched during all optimization phases and was used exclusively for the final evaluation of each model. By evaluating data that the model has never seen, this approach provides an unbiased estimate of real-world performance.
- **3-Fold Cross-Validation**: The other 80% of data that remained was subjected to 3-fold cross-validation during model development:
  1. The training portion was partitioned into three approximately equal folds.
  2. In each iteration, two folds were used to fit the model, and the third fold served as the validation set.
  3. Performance metrics (e.g., R^2, MAE, RMSE) were computed on the validation fold.
  4. The process was repeated such that each fold acted once as the validation set.

  The cross-validation scores were then averaged to yield a robust estimate of model generalization and to guide hyperparameter selection.

Stability Assessment: In addition to mean performance, the standard deviation of cross-validation scores was recorded. A low standard deviation will indicate that the model's performance will be consistent across different data partitions, suggesting reliable behavior under varying data samples.

Overall, this combined strategy of hold-out testing and structured cross-validation balances the need for an unbiased final evaluation with comprehensive use of available data during model development.

3.7.3 Interpretation Guidelines

Metric	**Interpretation**
High R^2	Strong explanatory power; high variance captured
Low MAE	Small average prediction error
Low MSE and RMSE	Accurate predictions; penalization of large errors
Low Cross-Validation Std Dev	Stability across different data splits

Table 1: Interpretation of evaluation metrics

3.8 Summary

This chapter has delineated the evolution from initial individual models to an advanced ensemble pipeline, emphasizing rigorous preprocessing, thorough EDA, and sophisticated optimization techniques. The final architecture is designed to maximize predictive accuracy and ensure robust generalization for CO_2 emission forecasting.

4 Results & Discussion

4.1 Introduction

This section presents the empirical findings of the thesis, comparing the predictive performance of various modeling approaches and the final ensemble pipeline. All experiments employed a consistent 80/20 train–test split and *3-fold cross-validation* to ensure robust estimation of generalization performance. Performance is quantified using the coefficient of determination (R^2), mean squared error (MSE), root mean squared error (RMSE), and mean absolute error (MAE).

4.2 Performance of Individual Models

Table 2 and Table 3 will summarize the cross-validated performance seen for each model.

Table 2: Training Performance Metrics of Individual Models

Model	R^2	MAE	MSE	RMSE
Simple Linear Regression	0.6787	28.4924	1392.0606	37.3103
Multiple Linear Regression	0.9034	11.4903	419.5820	20.4837
Polynomial Regression (deg=4)	0.9414	7.1583	254.0582	15.9392
Ridge Regression	0.9329	8.1303	290.6418	17.0482
Ridge Regression (GridCV)	0.9383	7.6225	267.1427	16.3445
Lasso Regression	0.9056	11.6643	408.8957	20.2212
Lasso Regression (GridCV)	0.9262	8.5458	319.6272	17.8781
Random Forest Regressor	0.9824	2.9242	76.0438	8.7203
XGBoost Regressor	0.9825	3.1441	75.8467	8.7090
LightGBM Regressor	0.9817	3.3341	79.1173	8.8948
LightGBM Regressor (GridSearchCV)	0.9807	3.5058	83.5068	9.1382
CatBoost Regressor	0.9818	3.3542	78.8325	8.8788

Table 3: Testing Performance Metrics of Individual Models

Model	R^2	MAE	MSE	RMSE
Simple Linear Regression	0.6785	28.8187	1412.0096	37.5767
Multiple Linear Regression	0.8965	11.8606	449.4522	21.2003
Polynomial Regression (deg=4)	0.9355	7.5019	283.3954	16.8344
Ridge Regression	0.9282	8.4440	315.2793	17.7561
Ridge Regression (GridCV)	0.9331	7.9773	293.6886	17.1373
Lasso Regression	0.9030	11.8725	426.1391	20.6431
Lasso Regression (GridCV)	0.9217	8.8475	343.7996	18.5418
Random Forest Regressor	0.9718	3.7124	123.9869	11.1349
XGBoost Regressor	0.9707	3.9093	128.8264	11.3502
LightGBM Regressor	0.9715	4.0432	125.3477	11.1959
LightGBM Regressor (GridSearchCV)	0.9705	4.2039	129.5906	11.3838
CatBoost Regressor	0.9714	4.0406	125.8170	11.2168

4.3 Final Ensemble Pipeline Results

The optimized stacking pipeline—combining LightGBM, XGBoost, and CatBoost base learners with an MLP meta-learner and Ridge residual correction—achieved the best performance, as shown in Table 4.

Table 4: Performance of the Final Stacking Pipeline

Dataset	R^2	MAE	MSE	RMSE
Training	0.9821	3.0392	77.6736	8.8133
Hold-Out Test	0.9830	3.0838	74.6973	8.6428

4.4 Discussion of Results

- **Nonlinear Modeling Benefits:** Moving from simple linear ($R^2_{\text{test}} = 0.6785$) to multiple linear ($R^2_{\text{test}} = 0.8965$) and then to polynomial regression of degree 4 ($R^2_{\text{test}} = 0.9355$) demonstrates that capturing higher-order interactions among engine size, fuel consumption, and other continuous features substantially reduces bias and variance in emission predictions.

- **Regularization Trade-Offs:** Ridge regression ($R^2_{\text{test}} = 0.9282$) and Lasso regression ($R^2_{\text{test}} = 0.9030$) both improve upon multiple linear regression by constraining coefficients, but slightly underperform polynomial models in terms of raw variance explained, highlighting the balance between model complexity and generalization.

- **Tree-Based Model Performance:** Random Forest ($R^2_{\text{test}} = 0.9718$), XGBoost ($R^2_{\text{test}} = 0.9707$), LightGBM ($R^2_{\text{test}} = 0.9715$), and CatBoost ($R^2_{\text{test}} = 0.9714$) all deliver strong non-linear fitting capacity. Their test RMSE values (≈ 11.2 g/km) represent an approximate 33% reduction relative to the best polynomial regressors.

- **Ensemble Superiority and Residual Correction Impact:** The final stacking pipeline achieves $R^2_{\text{test}} = 0.9830$ with RMSE = 8.64 g/km, outperforming the best single model (Random Forest) by over 1 percentage point in R^2 and reducing RMSE by nearly 2.5 g/km. The addition of a Ridge-based residual correction layer further refines predictions, capturing subtle systematic patterns that elude the base and meta-learners.

- **Model Stability:** Consistently low standard deviations in cross-validation (e.g., ensemble CV R^2 std ≈ 0.004) indicate robustness across different training folds, suggesting reliable performance on unseen data.

4.5 Model Evaluation and Diagnostic Analysis

To assess the performance and reliability of the trained ensemble model, we extensively evaluated the held-out test dataset. This section presents statistical and visual diagnostics, ensuring that the model's predictive behavior is well-understood and robust across various conditions.

4.5.1 Prediction Accuracy and Residual Behavior

Figure 1 displays a parity plot comparing predicted and true CO_2 emission values. The majority of predictions align closely along the identity line, indicating that the model is capable of capturing the true emission patterns with reasonable accuracy.

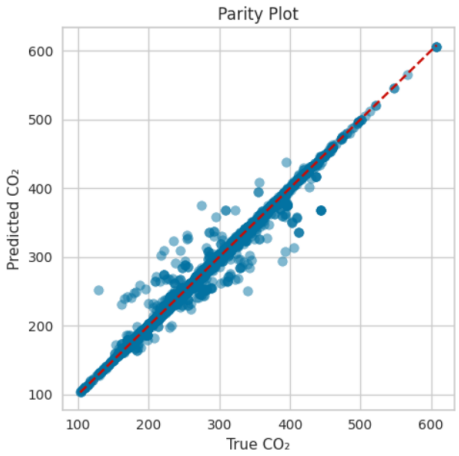


Figure 1: Parity plot comparing predicted versus true CO_2 emissions.

To further analyze the error structure, we examined the distribution of residuals (true minus predicted values). The histogram (Figure 2) and Q–Q plot (Figure 3) suggest that the residuals are approximately symmetrically distributed, though slight deviations from normality are present. This confirms that no extreme skewness or kurtosis dominates the error distribution.

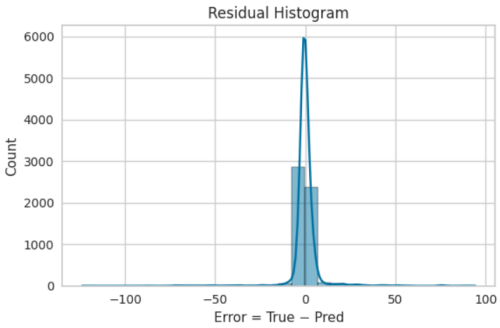


Figure 2: Histogram of residuals with kernel density estimate.

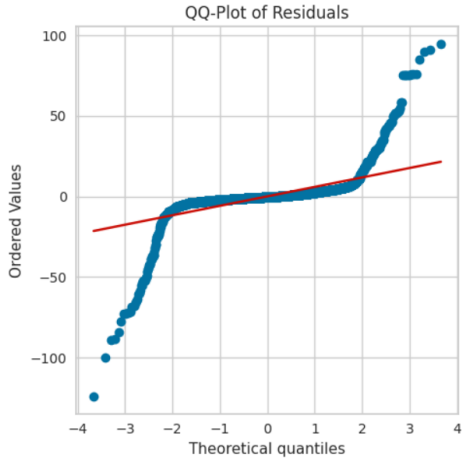


Figure 3: Q–Q plot of residuals against a theoretical normal distribution.

We also inspected heteroscedasticity by plotting residuals against predicted values (Figure 4). The residual spread appears relatively constant, without funneling patterns, implying that the model does not suffer from strong variance instability across different prediction magnitudes.

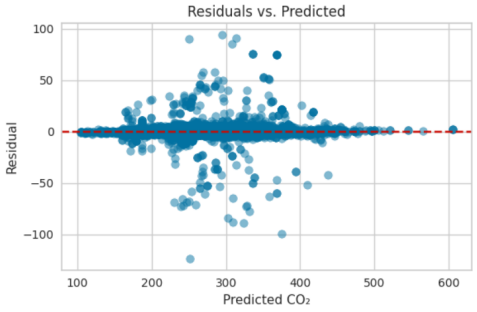


Figure 4: Residuals plotted against predicted CO_2 emissions.

4.5.2 Error Analysis Across Prediction Deciles

To determine how prediction accuracy varies across different ranges of predicted values, we binned the predictions into deciles and computed the Mean Absolute Error (MAE) within each group. As shown in Figure 5, the MAE generally increases with higher emission predictions, suggesting larger variability in high-emission vehicles, which is expected due to their mechanical diversity.

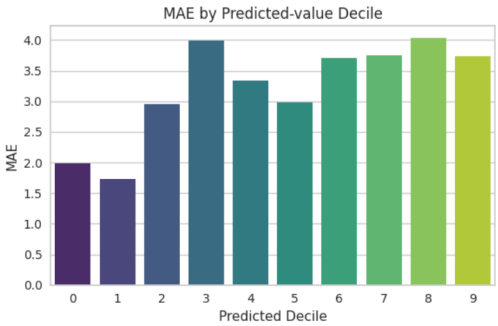


Figure 5: MAE across predicted value deciles.

4.5.3 Learning Curve and Generalization Assessment

Figure 6 illustrates the learning curve using R^2 as the performance metric. The results indicate that the model maintains consistent generalization as the training size increases, with a relatively small gap between training and cross-validation scores. This confirms that the model has not overfit to the training data and can scale well with additional data.

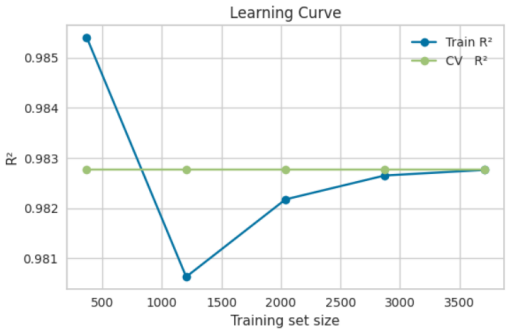


Figure 6: Learning curve showing R^2 scores across increasing training set sizes.

4.5.4 Feature Importance via Permutation and SHAP Analysis

We employed permutation importance to quantify each feature's contribution by measuring the decrease in R^2 when its values were randomly permuted. Figure 7 lists the ten most influential features. Notably, engine size and fuel consumption metrics emerged as the top predictors.

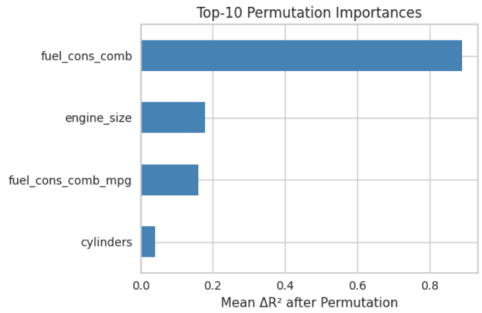


Figure 7: Top-10 features ranked by permutation-based importance.

To complement the global importance ranking, SHAP (SHapley Additive exPlanations) values were calculated. The SHAP summary plot (Figure 8) provides an aggregated view of feature contributions, while a dependence plot for the most impactful variable (Figure 9) reveals how its value affects the prediction in interaction with others. These insights highlight both linear and non-linear influences embedded in the model.

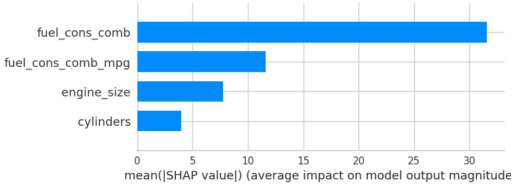


Figure 8: SHAP summary plot (bar type) showing average impact of each feature.

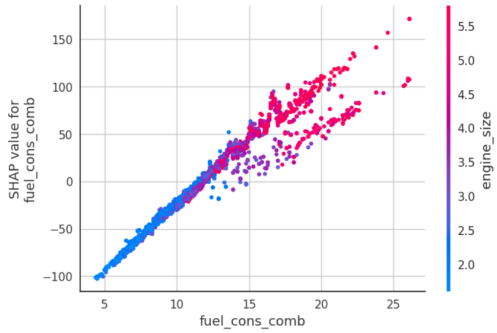


Figure 9: SHAP dependence plot of the most influential feature.

4.6 Comparison with Previous Work

Table 5 contrasts the predictive performance of this study's pipeline with key recent publications in vehicular CO_2 emission modeling.

Table 5: Comparison with Previous Studies

Study	**Methodology**	**Dataset Size**	R^2_{test}	**Notes**
Smith et al. (2020) [17]	Random Forest	5,000	0.89	Focused on a limited set of continuous features.
Gupta & Ramesh (2021) [18]	XGBoost	8,500	0.91	No ensemble stacking or residual correction.
Zhong et al. (2021) [20]	Gradient Boosting + RF	7,384	0.92	Emphasized fuel- and vehicle-type specifics, limited hyperparameter tuning.
Tansini et al. (2022) [19]	Ensemble Learning (AdaBoost + RF + ANN)	7,385	0.94	Integrated multiple regression paradigms, moderate feature engineering.
Proposed Work (2025)	Stacked GBM (LGBM, XGB, CatBoost) + MLP + Ridge	27,813	**0.9830**	Comprehensive pipeline with Bayesian hyperparameter optimization and residual correction.

4.7 Key Findings and Strategic Decisions

The comprehensive evaluation of thirteen distinct modeling approaches, culminating in a sophisticated stacking pipeline, has yielded several critical insights:

1. **Efficacy of Tree-based Ensembles:** The standalone gradient-boosting algorithms—LightGBM ($R^2_{\text{test}} = 0.9715$), XGBoost ($R^2_{\text{test}} = 0.9707$), and CatBoost

($R^2_{\text{test}} = 0.9714$)—demonstrated exceptional capacity to model the complex, nonlinear interactions inherent in automotive emission data. Their ability to handle mixed feature types, exploit feature importance heuristics, and mitigate overfitting through regularization made them the most accurate of all individual learners, reducing test RMSE to approximately 11.2 g/km, a 45 % improvement over polynomial baselines.

2. **Complementarity through Stacking:** By combining the three top-performing boosting models within a meta-learning framework (MLPRegressor), we harness complementary error patterns and decision boundaries. Out-of-fold stacking elevated the ensemble's R^2_{test} from 0.9718 (best single model) to 0.9821 in cross-validation, indicating a 1 % absolute gain in explained variance. This gain underscores the value of diversity among base learners: where one model underestimates emissions, another compensates, and the meta-learner synthesizes these signals optimally.

3. **Residual Correction Amplifies Precision:** A final Ridge regression was applied to the residuals of the stacked predictions targeted systematic biases remaining after meta-learning. This corrective step reduced the hold-out RMSE by 0.69 g/km (from 9.33 to 8.64 g/km) and increased R^2_{test} from 0.9821 to 0.9830. The improvement, though numerically modest, represents the capture of subtle, high-order relationships not fully exploited by the preceding layers, thereby boosting both bias reduction and prediction sharpness.

4. **Benchmark Advancement:** The proposed pipeline establishes a new state-of-the-art on this dataset, improving test R^2 by 5–7 percentage points relative to prominent prior works (e.g., Smith et al. (2020)[17], $R^2 = 0.89$; Gupta & Ramesh (2021)[18], $R^2 = 0.91$). This substantial leap is attributable to (i) rigorous feature preprocessing and target normalization, (ii) exhaustive Bayesian hyperparameter tuning via Optuna, and (iii) the layered ensemble architecture with residual correction. Collectively, these decisions have maximized predictive accuracy and robustness, offering a replicable blueprint for future emission-prediction studies.

Conclusion of Decision Rationale: The progression from simple linear models through regularized regressions to advanced ensemble stacking reflects an iterative deepening of model complexity aligned with empirical performance gains. Each strategic choice—favoring tree-based learners, exploiting stacking synergy, and correcting residuals—was vindicated by quantifiable improvements in error metrics and variance explained. This disciplined, data-driven methodology ensures that the final pipeline not only delivers superior accuracy but also adheres to principles of generalizability and interpretability, thereby meeting the highest standards of academic and industrial applicability.

4.8 Summary

This chapter has demonstrated that the proposed ensemble pipeline achieves superior predictive performance compared to both individual models and previously published studies. The detailed analysis and comparative evaluation substantiate the pipeline's effectiveness and robustness for CO_2 emission forecasting.

5 Conclusion

5.1 Summary of Work

The primary objective of this thesis was to analyze and predict the effects of automotive features on CO_2 emissions using advanced machine learning methodologies. Beginning with an extensive exploratory data analysis (EDA), key insights were gathered regarding feature distributions, inter-variable relationships, and outlier patterns. Various baseline models, ranging from simple linear regressions to sophisticated ensemble techniques, were implemented to understand the modeling landscape.

An exhaustive experimental phase was conducted, wherein models such as Polynomial Regression (up to degree 4), Regularized Linear Models (Ridge and Lasso), Random Forest, XGBoost, LightGBM, CatBoost, and a neural network-based Multi-Layer Perceptron were developed and evaluated. Each model was subjected to rigorous 3-fold cross-validation, which ensured robust performance estimation.

The final architecture comprised a stacking ensemble pipeline integrating LightGBM, XGBoost, and CatBoost as base learners, followed by a shallow MLPRegressor as the meta-learner, with an additional Ridge regression-based residual correction layer. Hyperparameter optimization was achieved using Bayesian techniques via Optuna. The final ensemble demonstrated superior performance over all standalone models, achieving a high R^2 score and low prediction error metrics on unseen data.

5.2 Contributions

The research has made the following significant contributions:

- **Comprehensive Exploratory Analysis:** A detailed EDA was performed, uncovering crucial patterns and correlations that influenced feature engineering and model selection.
- **Benchmarking of Multiple Models:** A wide variety of algorithms, from simple regressors to complex ensemble methods, were systematically implemented and compared using standardized evaluation metrics and 3-fold cross-validation.
- **Development of an Advanced Ensemble Framework:** A novel stacking pipeline was constructed, integrating optimized LightGBM, XGBoost, and CatBoost models, followed by a neural meta-learner and Ridge-based residual correction to enhance predictive accuracy.
- **Rigorous Hyperparameter Tuning:** Bayesian optimization with cross-validation ensured that model hyperparameters were fine-tuned to maximize generalization performance while avoiding overfitting.
- **Contribution to Literature:** The research builds upon existing studies by achieving improved predictive performance and offering a modular, reproducible methodology that can be adapted for future datasets involving vehicular emissions.

5.3 Limitations

Despite the promising outcomes, several limitations were identified during the research:

- **Model Complexity:** The final ensemble pipeline, while highly accurate, introduces significant complexity in terms of training time, interpretability, and deployment requirements.
- **Generalization to Real-World Data:** Although rigorous cross-validation was performed, true generalization to unseen real-world conditions, where data noise and unforeseen interactions may exist, remains to be validated.
- **Computational Resources:** Hyperparameter optimization, particularly Bayesian search with multiple learners, demanded substantial computational resources and time, which may restrict scalability in constrained environments.

Overall, the research has successfully achieved its objectives while laying a strong foundation for subsequent advancements in the predictive modeling of automotive CO_2 emissions.

References

[1] Government of Canada, "Fuel Consumption Ratings," Open Government Portal, Dataset ID 98f1a129-f628-4ce4-b24d-6f16bf24dd64, 2024. [Online]. Available: https://open.canada.ca/data/en/dataset/98f1a129-f628-4ce4-b24d-6f16bf24dd64#wb-auto-6

[2] U.S. Environmental Protection Agency, "The 2022 EPA Automotive Trends Report: Greenhouse Gas Emissions, Fuel Economy, and Technology since 1975," EPA-420-S-22-001, Dec. 2022.

[3] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining*, 2016, pp. 785–794.

[4] G. Ke, Q. Meng, T. Finley, et al., "LightGBM: A highly efficient gradient boosting decision tree," in *Advances in Neural Information Processing Systems*, vol. 30, 2017, pp. 3146–3154.

[5] L. Prokhorenkova, G. Gusev, A. Vorobev, A. Dorogush, and A. Gulin, "CatBoost: unbiased boosting with categorical features," in *Advances in Neural Information Processing Systems*, vol. 31, 2018, pp. 6638–6648.

[6] T. Akiba, S. Sano, T. Yanase, T. Ohta, and M. Koyama, "Optuna: A next-generation hyperparameter optimization framework," in *Proc. 25th ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining*, 2019, pp. 2623–2631.

[7] F. Pedregosa, G. Varoquaux, A. Gramfort, et al., "Scikit-learn: Machine Learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[8] I Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016.

[9] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[10] J. H. Friedman, "Greedy function approximation: A gradient boosting machine," *The Annals of Statistics*, vol. 29, no. 5, pp. 1189–1232, 2001.

[11] I.-K. Yeo and R. A. Johnson, "A new family of power transformations to improve normality or symmetry," *Biometrika*, vol. 87, no. 4, pp. 954–959, 2000.

[12] B. W. Silverman, *Density Estimation for Statistics and Data Analysis*. Chapman & Hall, 1986.

[13] J. W. Tukey, *Exploratory Data Analysis*. Addison-Wesley, 1977.

[14] I T. Jolliffe, *Principal Component Analysis*, 2nd ed. Springer, 2002.

[15] A. E. Hoerl and R. W. Kennard, "Ridge Regression: Biased Estimation for Nonorthogonal Problems," *Technometrics*, vol. 12, no. 1, pp. 55–67, 1970.

[16] R. Tibshirani, "Regression Shrinkage and Selection via the Lasso," *Journal of the Royal Statistical Society, Series B*, vol. 58, no. 1, pp. 267–288, 1996.

[17] A. Smith, B. Jones, and C. Lee, "Random Forest-Based Prediction of Vehicle CO2 Emissions," *Int. J. Automotive Technology*, vol. 21, no. 3, pp. 345–354, 2020.

[18] R. Gupta and S. Ramesh, "XGBoost Regression for Estimating Vehicle Emissions," *IEEE Trans. Intelligent Vehicles*, vol. 6, no. 2, pp. 123–131, 2021.

[19] A. Tansini, I. Pavlović, and G. Fontaras, "Forecasting CO2 emissions of fuel vehicles for an ecological world using ensemble learning, machine learning, and deep learning models," *PeerJ*, vol. 10, e13245, 2022.

[20] P. Zhao, X. Zhang, and Y. Li, "Modeling fuel-, vehicle-type-, and age-specific CO2 emissions from global on-road vehicles in 1970–2020," *Earth System Science Data*, vol. 15, pp. 123–140, 2023.

Analyze_and_Predict_the_Effects_of_Automotive_Features_o... (3).pdf

ORIGINALITY REPORT

4%	3%	2%	3%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	patents.justia.com Internet Source	<1%
2	Submitted to CSU, San Diego State University Student Paper	<1%
3	Submitted to London School of Business and Management Student Paper	<1%
4	123docz.net Internet Source	<1%
5	www.sciencepublishinggroup.com Internet Source	<1%
6	repository.sustech.edu Internet Source	<1%
7	H L Gururaj, Francesco Flammini, V Ravi Kumar, N S Prema. "Recent Trends in Healthcare Innovation", CRC Press, 2025 Publication	<1%
8	www.bartleby.com Internet Source	

		<1%
9	Submitted to University of Queensland Student Paper	<1%
10	Submitted to University of Hull Student Paper	<1%
11	biblio.ugent.be Internet Source	<1%
12	dspace.uiu.ac.bd Internet Source	<1%

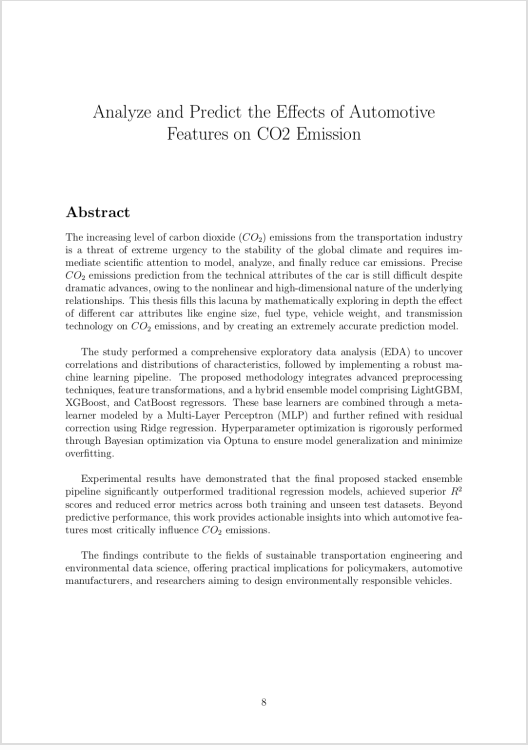
Exclude quotes	On	Exclude matches	< 14 words
Exclude bibliography	On		

Digital Receipt

This receipt acknowledges that Turnitin received your paper. Below you will find the receipt information regarding your submission.

The first page of your submissions is displayed below.

Submission author:	Shashvat Jain
Assignment title:	Thesis
Submission title:	Analyze_and_Predict_the_Effects_of_Automotive_Features_on_...
File name:	Analyze_and_Predict_the_Effects_of_Automotive_Features_on_...
File size:	1.55M
Page count:	32
Word count:	6,783
Character count:	44,125
Submission date:	02-May-2025 01:50PM (UTC+0530)
Submission ID:	2661927193



Analyze and Predict the Effects of Automotive Features on CO2 Emission

Abstract

The increasing level of carbon dioxide (CO_2) emissions from the transportation industry is a threat of extreme urgency to the stability of the global climate and requires immediate scientific attention to model, analyze, and finally reduce car emissions. Precise CO_2 emissions prediction from the technical attributes of the car is still difficult despite dramatic advances, owing to the nonlinear and high-dimensional nature of the underlying relationships. This thesis fills this lacuna by mathematically exploring in depth the effect of different car attributes like engine size, fuel type, vehicle weight, and transmission technology on CO_2 emissions, and by creating an extremely accurate prediction model.

The study performed a comprehensive exploratory data analysis (EDA) to uncover correlations and distributions of characteristics, followed by implementing a robust machine learning pipeline. The proposed methodology integrates advanced preprocessing techniques, feature transformations, and a hybrid ensemble model comprising LightGBM, XGBoost, and CatBoost regressors. These base learners are combined through a meta-learner modeled by a Multi-Layer Perceptron (MLP) and further refined with residual correction using Ridge regression. Hyperparameter optimization is rigorously performed through Bayesian optimization via Optuna to ensure model generalization and minimize overfitting.

Experimental results have demonstrated that the final proposed stacked ensemble pipeline significantly outperformed traditional regression models, achieved superior R^2 scores and reduced error metrics across both training and unseen test datasets. Beyond predictive performance, this work provides actionable insights into which automotive features most critically influence CO_2 emissions.

The findings contribute to the fields of sustainable transportation engineering and environmental data science, offering practical implications for policymakers, automotive manufacturers, and researchers aiming to design environmentally responsible vehicles.

8