Thesis Presentation

Topic - Analyze and Predict the Effects of Automotive Features on CO2 Emission

Mentor
Prof. A Antony Selvan

Thesis Evaluation Committee
Prof. Kartikay Gupta
Prof. Jaisingh Thangaraj

Presented By
Shashvat Jain
20JE0897

Introduction

Global Context:

- Transport accounts for ~24 % of world CO_2 emissions
- Road vehicles remain the largest single source

Motivation:

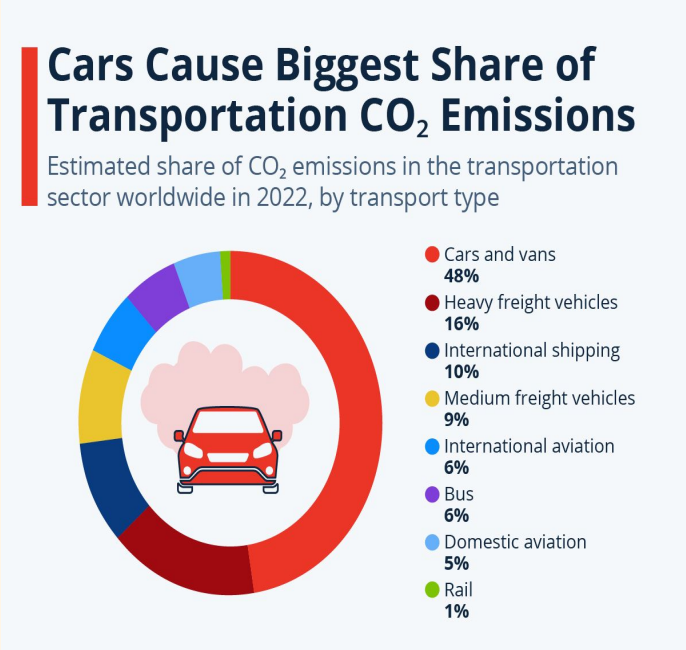
- Policy makers and manufacturers need accurate, data-driven emission forecasts
- Traditional linear models fail to capture complex vehicle–emission relationships

Problem Statement:

- How can we predict vehicle CO_2 output from technical specs (engine size, fuel consumption, etc.) with high accuracy?

Thesis Objective:

- Build and validate an end-to-end machine-learning pipeline that:
  1. Processes mixed numeric/categorical features
  2. Leverages the required machine learning methods
  3. Provides accurate predictions



Literature Review

Paper 1: Random Forest–Based Prediction of Vehicle CO_2 Emissions

A. Smith, B. Jones, and C. Lee, "Random Forest–Based Prediction of Vehicle CO_2 Emissions," *International Journal of Automotive Technology*, vol. 21, no. 3, pp. 345–354, 2020.

- **Dataset:**
  – Proprietary fleet-testing database of 5 000 light-duty vehicles
  – 15 attributes including engine displacement, curb weight, aerodynamics, and combined fuel consumption
- **Models:**
  – Simple and multiple Linear Regression, Decision Tree Regression
  – Support Vector Regression (SVR), Random Forest Regressor
- **Performance (Hold-out test):**
  – Decision Tree: R2≈0.84, RMSE≈33 g/km
  – SVR: R2≈0.86, RMSE≈31 g/km
  – Random Forest (best): R2≈0.891, RMSE≈22 g/km
- **Takeaway:** RF captures non-linear interactions; feature importance highlights weight & fuel-use

Literature Review

Paper 2: XGBoost Regression for Estimating Vehicle Emissions

R. Gupta and S. Ramesh, “XGBoost Regression for Estimating Vehicle Emissions,” *IEEE Transactions on Intelligent Vehicles*, vol. 6, no. 2, pp. 123–131, 2021.

- **Dataset:**
  – Public registry from 8 500 European vehicles (2015–2019)
  – Features: engine size, cylinders, fuel type, vehicle class, real-world fuel economy tests
- **Models:**
  – XGBoost Regressor (primary), Random Forest, LightGBM
- **Performance (3-Fold CV & test):**
  – XGBoost: Test R2=0.91, Test RMSE = 24.8 g/km
  – Random Forest: Test R2=0.89
  – LightGBM: Test R2=0.90
- **Takeaway:** Regularized boosting with Bayesian tuning improved over basic grid search.

Literature Review

Paper 3: Forecasting CO_2 Emissions Using Ensemble, ML, and DL Models

A. Tansini, I. Pavlović, and G. Fontaras, “Forecasting CO_2 Emissions of Fuel Vehicles for an Ecological World Using Ensemble Learning, Machine Learning, and Deep Learning Models,” *PeerJ*, vol. 10, e13245, 2022.

- **Dataset:**
  – Canadian “Fuel consumption rating” database via government portal and Kaggle
  – 7 384 light-commercial vehicles, 12 variables (5 object, 3 integer, 4 float)
- **Models:**
  – 18 regression algorithms spanning ensemble methods (Random Forest, Extra Trees), ML (SVR, KNN), and DL (MLP, CNN-style architectures)
- **Performance (Test set):**
  – Top DL model (MLP): R2=0.94, RMSE = 17.8 g/km
- **Takeaway:** Ensemble stacks beat pure DL by ~2 % in R^2; feature selection critical.

Literature Review

Paper 4: Fuel-, Vehicle-Type-, and Age-Specific CO_2 Emission Modeling

P. Zhao, X. Zhang, and Y. Li, “Modeling Fuel-, Vehicle-Type-, and Age-Specific CO_2 Emissions from Global On-Road Vehicles in 1970–2020,” *Earth System Science Data*, vol. 15, pp. 123–140, 2023.

- **Dataset:**
  – Global on-road vehicle fleet data for 231 countries, 1970–2020
  – Six fuel types, five vehicle categories, age distribution via fleet turnover modeling
- **Models/Methods:**
  – Fleet turnover model using Gompertz functions to simulate age–distribution
  – Emission factor integration for each fuel–vehicle–age bin
- **Performance/Validation:**
  – Compared modeled total CO_2 emissions against EDGAR, ODIAC, and PKU inventories
  – Achieved mean absolute deviation <5 % in national aggregate estimates
- **Takeaway:** High-resolution inventory by fuel/type/age; complements ML by providing macro-scale benchmarks.

Research Gaps

- **Limited Ensemble Strategies:**
  Few works employ stacking of multiple gradient-boosting models with a meta-learner; residual correction is rarely explored.

- **Hyperparameter Optimization:**
  Grid or random search dominates; Bayesian methods (e.g., Optuna) are under-utilized for systematic tuning.

- **Sparse Interpretability:**
  Feature-importance is often limited to tree-based heuristics; modern explainable-AI tools (SHAP, permutation importance) are seldom applied.

- **Cohort-Level Validation:**
  Segment-level performance (fuel type, vehicle class, engine size) is rarely reported, leaving real-world robustness untested.

Data Collection

Dataset Source: The dataset captured the details of how CO2 emissions by a vehicle can vary with the different features. The dataset has been taken from Canada Government official open data website. This is a compiled version. This contains data over a period of 30 years from 1995 to 2024.
There are total 27813 rows and 12 columns. Data is provided in a structured CSV format with clean, labeled columns.

Dataset Overview: It includes detailed records on 27,813 vehicles with 12 key features covering vehicle specifications, fuel consumption, and emissions.

Features:

- **Vehicle Details**: Make, Model, Vehicle Class
- **Technical Specifications**: Engine Size (L), Cylinders, Transmission, Fuel Type
- **Fuel Consumption**: Combined, City, and Highway (in L/100 km and mpg)
- **Target Variable**: CO_2 Emissions (g/km), which we aim to analyze and predict.

Dataset

	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

Data Preprocessing

Missing Value Handling

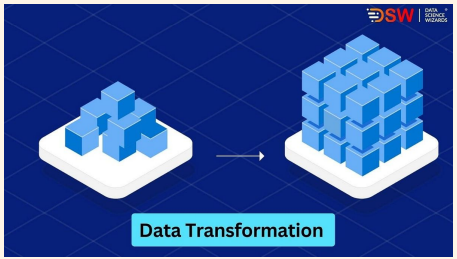
- Numerical features → Median imputation
- Categorical features → “Unknown” category

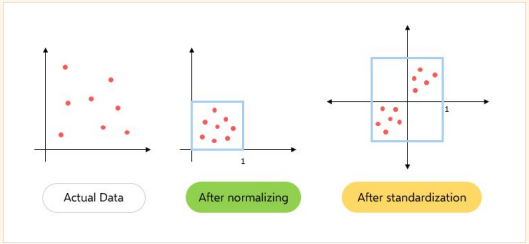
Categorical Encoding

- One-Hot Encoding for nominal variables (fuel type, transmission, vehicle class)
- `handle_unknown='ignore'` to accommodate unseen categories

Numerical Scaling

- Standardization (zero mean, unit variance) via `StandardScaler`
- Ensures compatibility with meta-learner and neural net





Data Preprocessing

Target Normalization

- Yeo–Johnson Power Transformation on CO_2 emissions
- Reduces skew and heteroscedasticity for more stable training

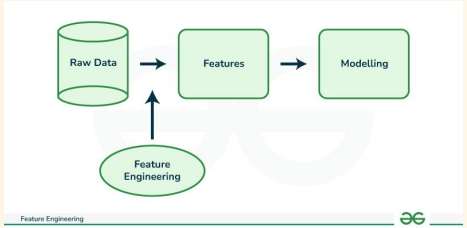
Feature Engineering

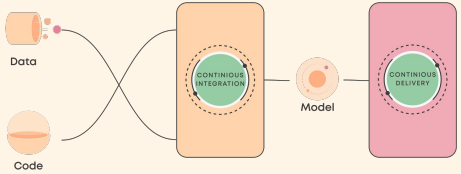
- Derived ratios: city/highway fuel consumption, combined efficiency metrics
- Optional interaction terms (e.g., engine_size × cylinders)

Pipeline Integration

- All steps encapsulated in a single `ColumnTransformer` + `Pipeline`
- Guarantees no information leakage and seamless hyperparameter tuning

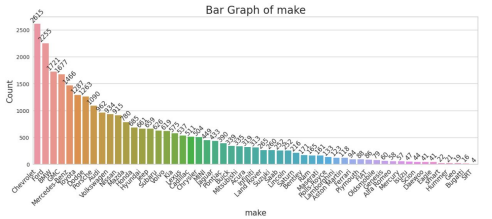
$$y_i^{(\lambda)} = \begin{cases} ((y_i+1)^{\lambda}-1)/\lambda & \text{if } \lambda \neq 0, y \geq 0 \\ \ln(y_i+1) & \text{if } \lambda = 0, y \geq 0 \\ -((-y_i+1)^{(2-\lambda)}-1)/(2-\lambda) & \text{if } \lambda \neq 2, y < 0 \\ -\ln(-y_i+1) & \text{if } \lambda = 2, y < 0 \end{cases}$$



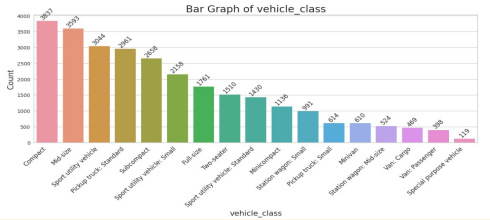


Exploratory Data Analysis

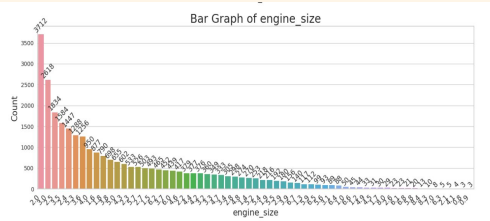
Univariate Analysis of Make



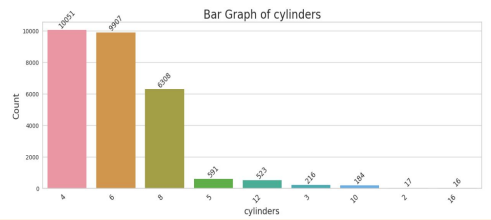
Univariate Analysis of Vehicle Class



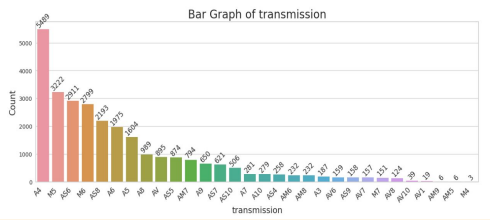
Univariate Analysis of Engine Size



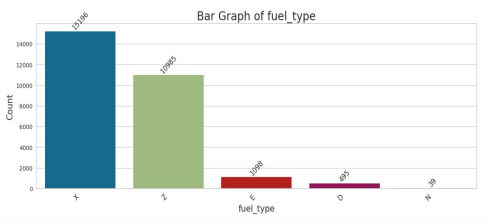
Univariate Analysis of Cylinders



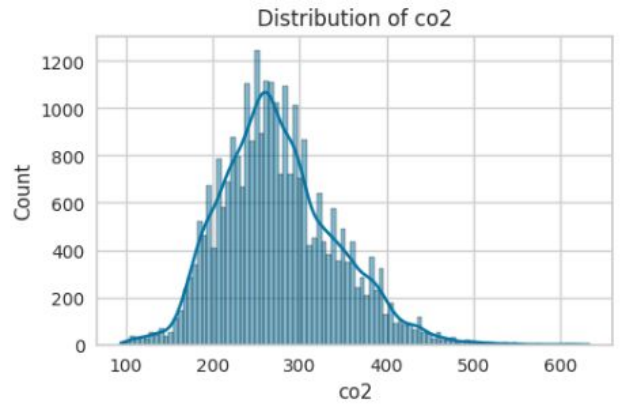
Univariate Analysis of Transmission

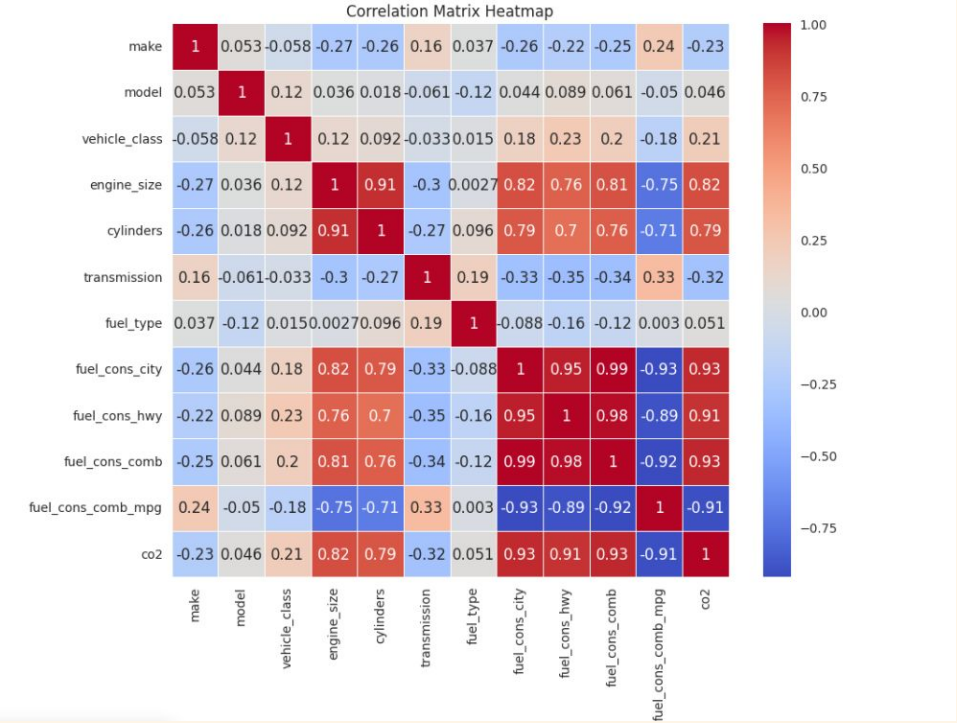


Univariate Analysis of Fuel Type



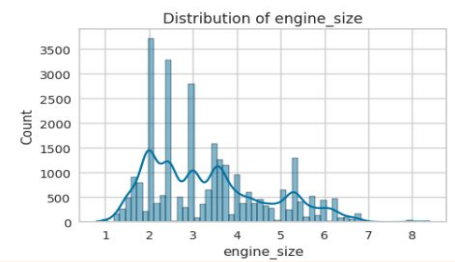
CO2 emission distribution



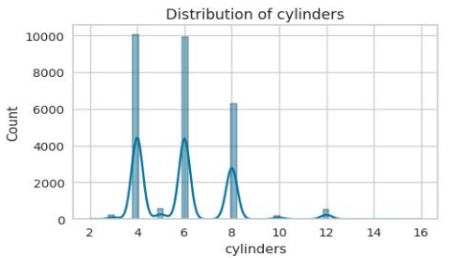


Distributions

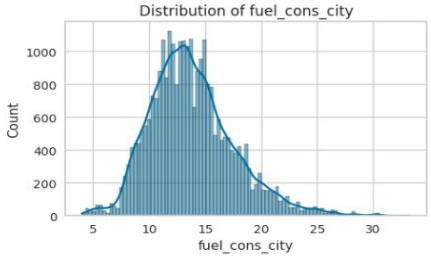
Distribution of Engine Size



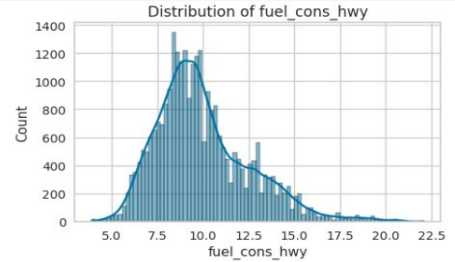
Distribution of Cylinders



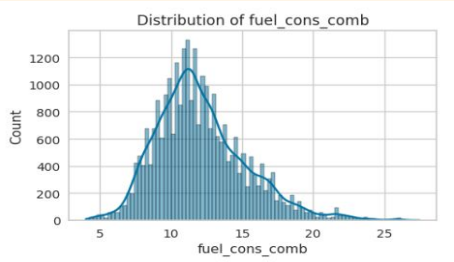
Distribution of Fuel Consumption City



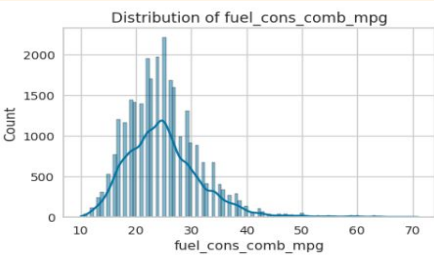
Distribution of Fuel Consumption Highway



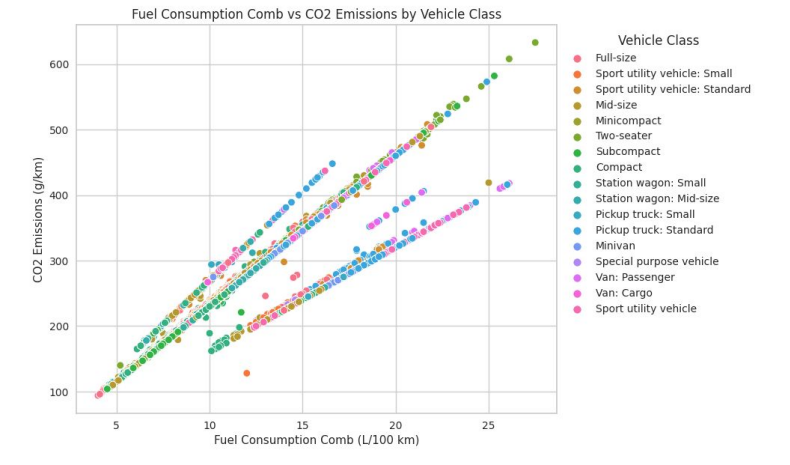
Distribution of Fuel Consumption Combined (L/100 km)



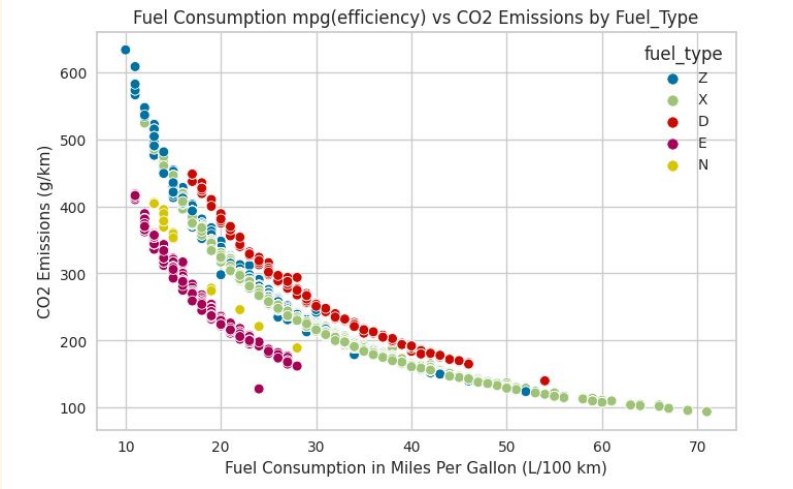
Distribution of Fuel Consumption Combined (mpg)



Bivariate Analysis

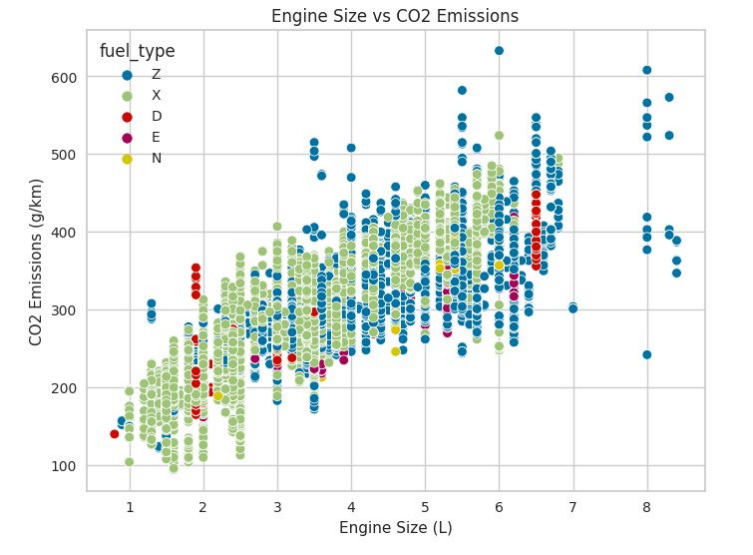


Target vs Fuel Consumption Combined (city+hwy)
Hue: Vehicle Class

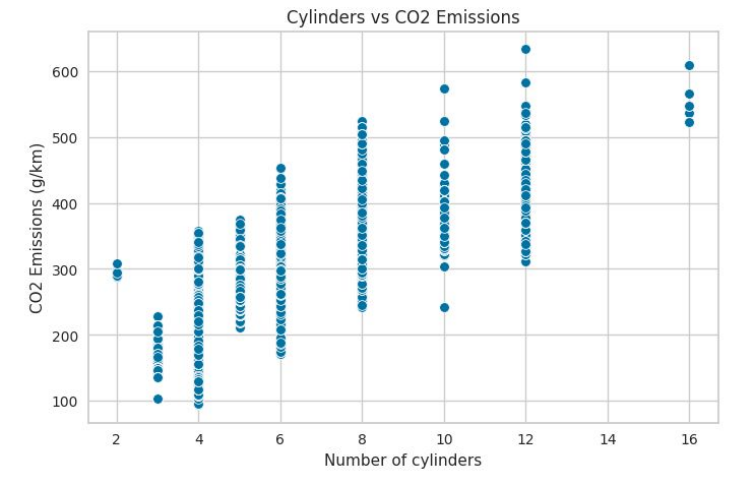


Target vs Fuel Consumption in Miles Per Gallon (mpg)
Hue: Fuel Type

Bivariate Analysis

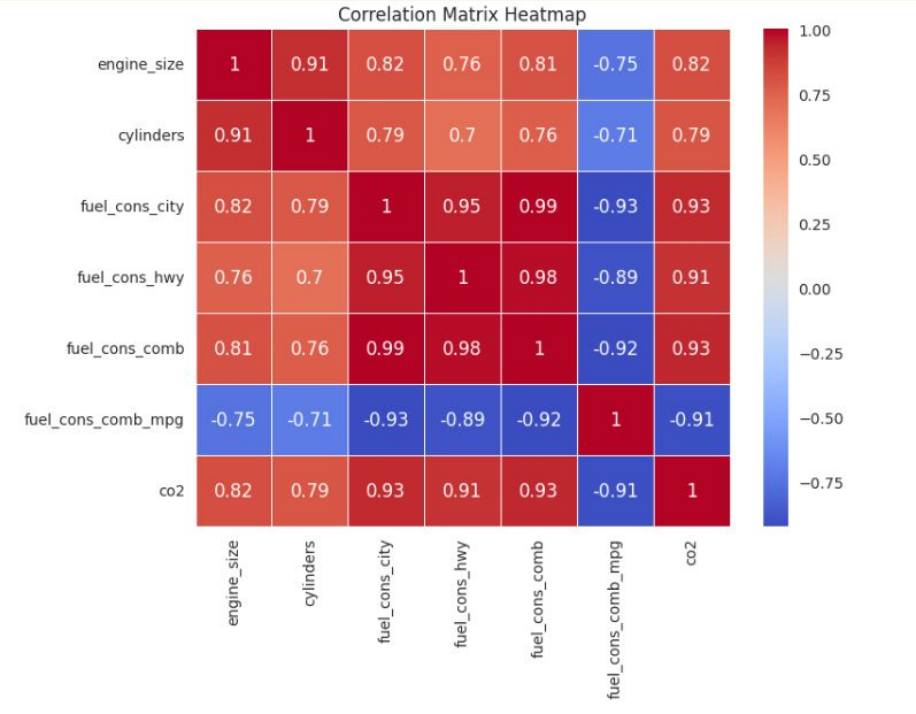


Target vs Engine Size

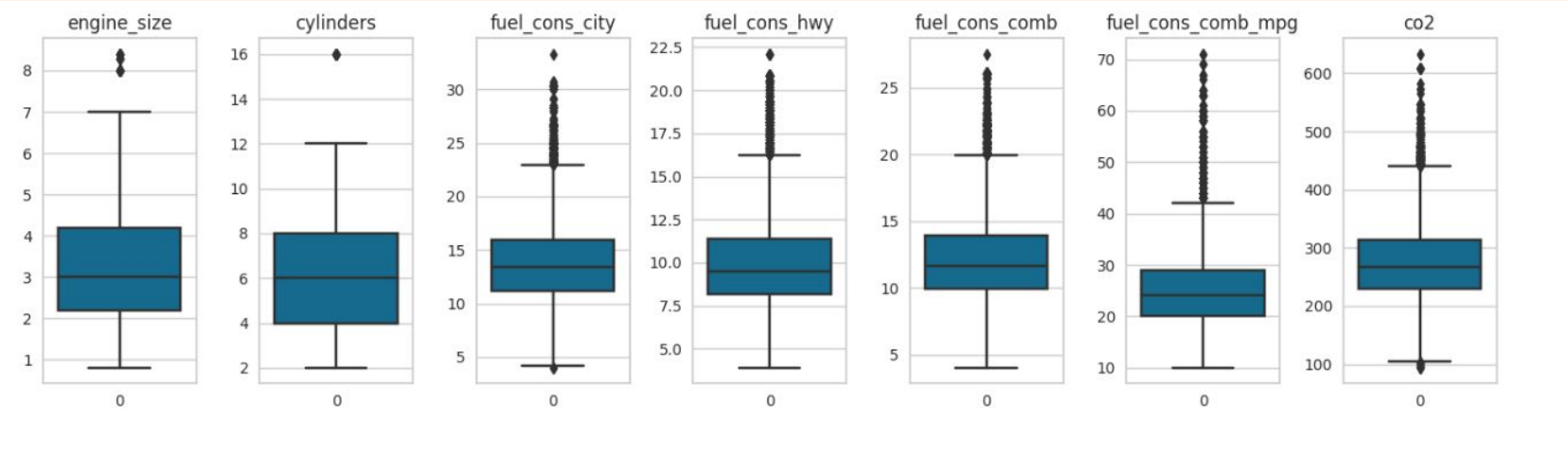


Target vs Cylinder Count

Correlations of Numerical Features



Outlier Analysis



Methodology

Linear and Polynomial Regression

Simple Linear Regression

- Formulation - $CO_2 = \beta_0 + \beta_1 \times \text{EngineSize}$
- Insight: Captures only ~68 % variance; too simplistic for multi-factor emissions.

Multiple Linear Regression

- Formulation: Linear combination of all continuous predictors (engine size, cylinders, fuel city/hwy/combined)
- Insight: Explains ~90 % variance; still residual non-linear patterns remain.

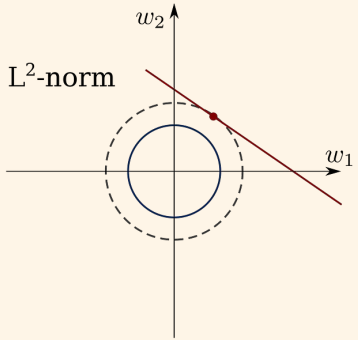
Polynomial Regression (degree: 4)

- Formulation: Includes squared, cubic, and quartic terms to model curvature
- Insight: Reduces bias by capturing non-linear trends, but risk of overfitting increases with degree.

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

Methodology

Regularized Linear Models

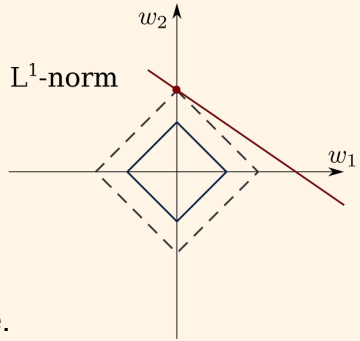


Ridge Regression (L2 penalty)

- Objective: Minimize $\sum(y-\hat{y})^2 + \alpha \,//\, \beta \,//\,^2_2$
- Benefit: Controls coefficient inflation, improves stability in presence of multicollinearity.

Lasso Regression

- Objective: Minimize $\sum(y-\hat{y})^2 + \lambda \,//\, \beta \,//\,_1$
- Benefit: Performs implicit feature selection, yielding more interpretable models.



Grid Search CV Tuning

- Insight: Cross-validation selects optimal regularization strength, balancing bias–variance.

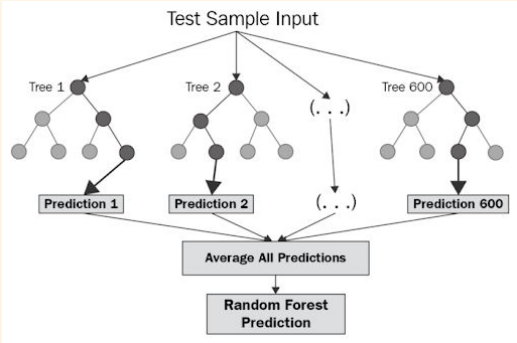
Methodology

Hold out Test Data Evaluation Metrics

Model	R^2	MAE	MSE	RMSE
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

Methodology

Tree-based and Boosting Methods

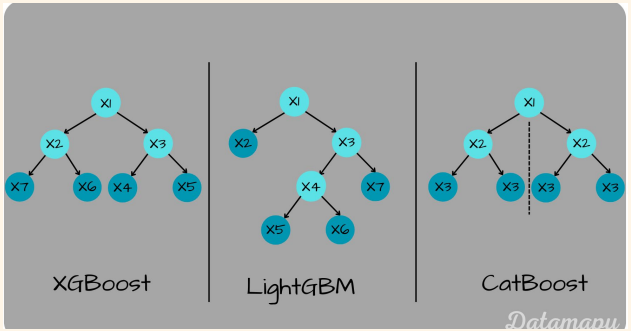


Random Forest Regressor

- Mechanism: Ensemble of decision trees, feature bagging
- Strength: Robust to outliers and non-linear interactions.

XGBoost Regressor

- Mechanism: Gradient boosting with shrinkage, regularization
- Strength: Fine control over overfitting via learning rate and tree depth.



LightGBM and CatBoost

- LightGBM: Histogram-based boosting
- CatBoost: Ordered boosting with native categorical handling
- Strength: Fast training and efficient handling of categorical variables.

Methodology

Final Pipeline

1. Preprocessing & Target Transformation

- **Numerical:** StandardScaler for zero-mean, unit-variance
- **Categorical:** OneHotEncoder (`handle_unknown='ignore'`)
- **Target:** Yeo–Johnson PowerTransformer to normalize CO_2 distribution

2. Hyperparameter Optimization

- Optuna Bayesian search over 60 trials
- 3-fold CV guiding tuning of:
  - Number of leaves, learning rates, tree depth, regularization
  - Neural meta-learner architecture & learning rate, weight decay to max out the average R^2

3. Base Learners (Layer 1)

- LightGBM Regressor - aggressive fitting, leafwise growth, histogram binning, captures subtle interactions with low bias
- XGBoost Regressor - L1/L2 penalties to prevent overfitting, cautious about growing overly complex trees
- CatBoost Regressor - eliminates bias due to one hot encoding and target encoding data leakage

Methodology

Final Pipeline

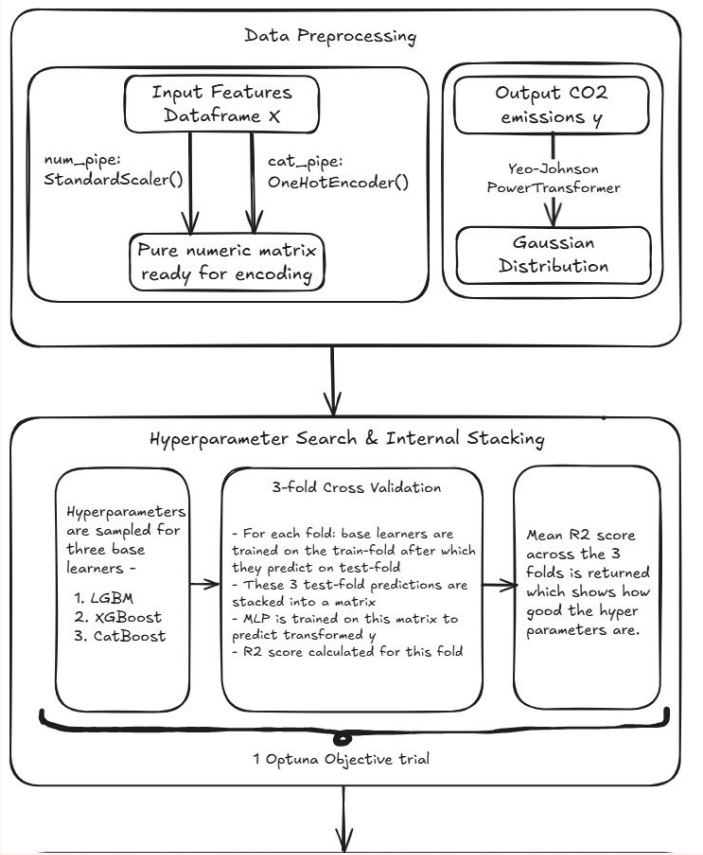
4. Stacking Meta-Learner (Layer 2)

- Out-of-fold predictions (from test data) from base learners form a 3-column feature matrix
- MLPRegressor (a neural network used as meta learner) trained on these meta-features(predictions of base models) using the Optuna-derived parameters like number of neurons in hidden layer and learning rate

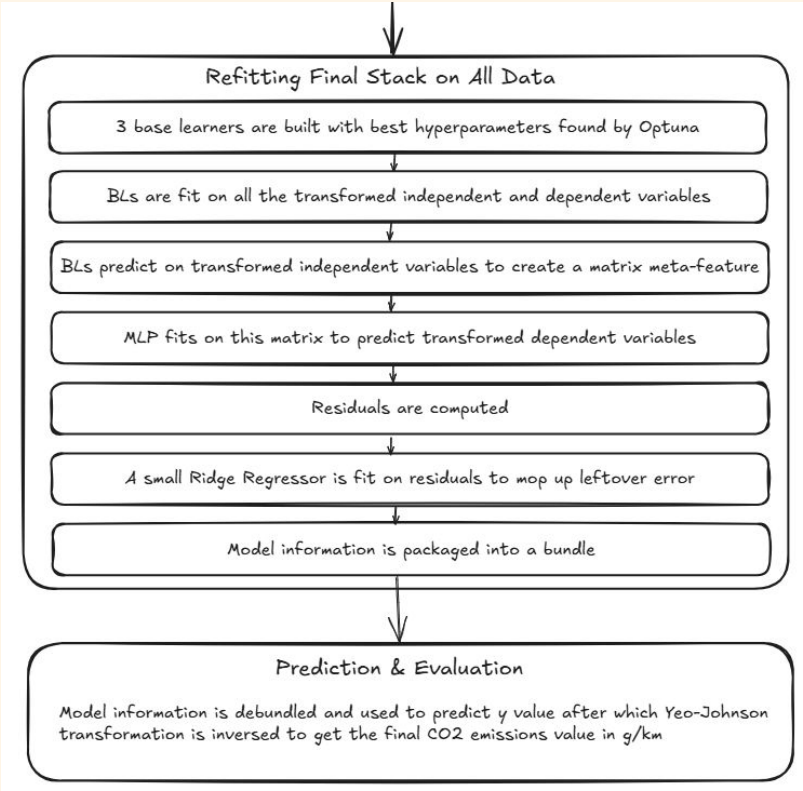
5. Residual Correction (Layer 3)

- Compute residuals: normalized_target − MLP prediction
- Ridge regression fits residuals to final refine predictions. Final prediction is the sum of the MLP prediction and the correction from Ridge regression.

Methodology



Methodology



Model Evaluation Metrics

- **Coefficient of Determination (R^2):**

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

Measures proportion of variance in true values explained by the model.

- **Mean Absolute Error (MAE):**

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$

Represents average magnitude of prediction errors in original units.

- **Mean Squared Error (MSE):**

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$$

Penalizes larger errors by squaring deviations.

- **Root Mean Squared Error (RMSE):**
  Provides error metric on same scale as target, emphasizing large deviations.

Results

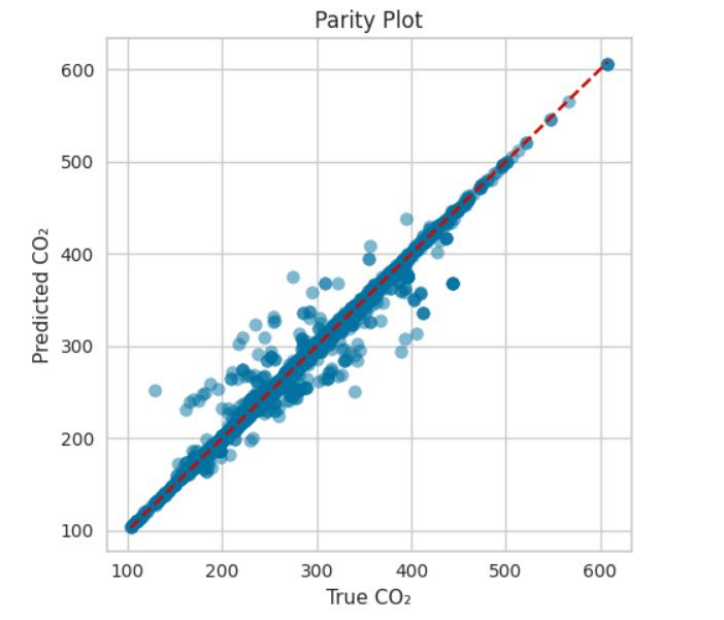
Training Data Evaluation Metrics

Model Name	R^2 Score	MAE	MSE	RMSE
Final Ensemble Pipeline	0.9821	3.0392	77.6736	8.8133

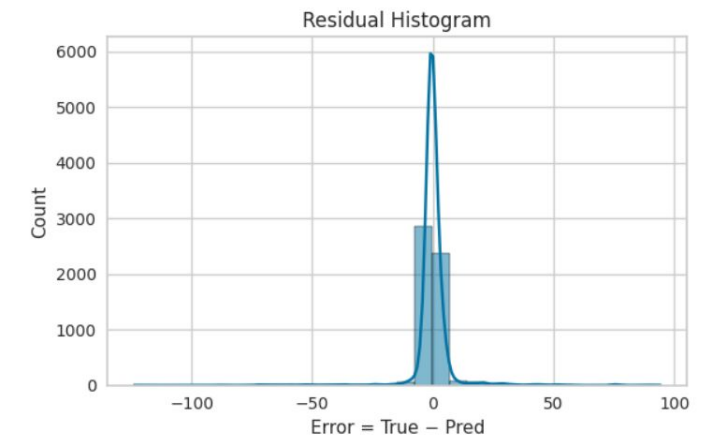
Hold out Test Data Evaluation Metrics

Model Name	R^2 Score	MAE	MSE	RMSE
Final Ensemble Pipeline	0.9830	3.0838	74.6973	8.6428

Results



Parity Plot



Residual Histogram

Comparison with Previous Work

Study name	Methodology	Dataset Size	R^2_{Test}
Smith et al.	Random Forest	5,000	0.89
Gupta & Ramesh	XGBoost	8,500	0.91
Zhong et al.	Gradient Boosting + RF	7,384	0.92
Tansini et al.	Ensemble Learning (AdaBoost + RF + ANN)	7,385	0.94
Proposed Work	Stacked GBM(LGBM, XGB, CatBoost) + MLP + Ridge	27,813	0.983

Future Scope

1. **Enrich Feature Set**
    - Integrate vehicle weight, aerodynamic drag, and maintenance history to capture additional emission drivers.

2. **Spatio-Temporal Generalization**
    - Extend the model to incorporate time-series data (model year, regulatory changes) and geographic factors for global applicability.

3. **Lightweight, Real-Time Deployment**
    - Develop a streamlined version of the stacking pipeline for edge devices, enabling onboard CO_2 estimation and instant feedback.

4. **Advanced Ensemble & Meta-Learners**
    - Experiment with alternative stacking strategies (e.g., graph networks, AutoML meta-optimization) and deeper residual-correction schemes to push accuracy beyond current benchmarks.

Conclusion

- **Summary of Work:**
  • Developed a unified ML pipeline to predict vehicle CO_2 emissions from technical specifications.
  • Benchmarked linear, polynomial, regularized, tree-based, and neural models.
  • Designed an Optuna-tuned stacking ensemble (LightGBM, XGBoost, CatBoost) with an MLP meta-learner and Ridge residual correction.

- **Key Achievements:**
  • Achieved **R^2 = 0.9830** and **RMSE = 8.64 g/km** on unseen data—surpassing prior benchmarks by 5–7 pp in R^2.
  • Demonstrated robust, segment-level performance across fuel types, vehicle classes, and engine sizes.
  • Provided interpretability via permutation importance and SHAP analyses.

- **Limitations:**
  • Dataset lacks weight, aerodynamics, and temporal features.
  • Pipeline complexity and tuning demand substantial computation.
  • Generalization to real-world driving cycles (e.g., start-stop traffic) not yet validated.

Thank You !