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

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

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.