Chapter 1

Results

1.1 Introduction

This chapter presents the findings of a machine learning experiment designed to predict vehicle maintenance costs, enabling a comparative analysis of efficiency between Electric Vehicles (EVs) and Internal Combustion Engine (ICE) vehicles. The experiment, detailed in the Jupyter Notebook experiment.ipynb, involved training and evaluating multiple regression models on a dataset containing 2000 records of vehicle attributes, with annual maintenance cost as the target variable. Performance was assessed using Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and R-squared (R^2) metrics, derived from 5-fold cross-validation and test set evaluations. Exploratory data analysis (EDA) visualizations, including histograms, box plots, and a correlation matrix, provide context for understanding differences between EVs and ICE vehicles. The results offer insights into model accuracy and the factors influencing maintenance costs, contributing to the broader objective of assessing vehicle efficiency.

1.2 Model Performance

Seven regression models were evaluated to predict annual maintenance costs: Linear Regression, Decision Tree, Random Forest, Gradient Boosting, CatBoost, LightGBM, and XGBoost. Performance was measured using MAE, RMSE, and R^2, calculated through 5-fold cross-validation on the training set and evaluation on a 20% test set. Table 1.1 summarizes the test set performance metrics for each model.

Linear Regression achieved the best performance, with the lowest test MAE of 3903.24 and the highest test R^2 of 0.6380, indicating that it explains approximately 63.8% of the variance in maintenance costs. The Decision Tree model performed the worst, with a test MAE of 5596.27 and an R^2 of 0.2327, suggesting limited predictive capability. Ensem-

Table 1.1: Model Performance Comparison on Test Set

Model	Test RMSE	Test MAE	Test R^2
Linear Regression	5626.91	3903.24	0.6380
Random Forest	5657.26	4050.87	0.6341
Gradient Boosting	5632.56	3977.80	0.6373
CatBoost	5949.83	4184.89	0.5953
LightGBM	6149.52	4302.86	0.5677
XGBoost (untuned)	6138.98	4368.90	0.5692
Decision Tree	8192.77	5596.27	0.2327

ble methods, including Random Forest, Gradient Boosting, CatBoost, LightGBM, and XGBoost, exhibited intermediate performance, with test MAEs ranging from 3977.80 to 4368.90 and R^2 scores from 0.5677 to 0.6373. These results suggest that Linear Regression effectively captures the underlying patterns in the data, likely due to predominantly linear relationships between features and the target variable.

1.3 Tuned XGBoost Model

Although Linear Regression outperformed other models, XGBoost was selected for hyperparameter tuning due to its robustness, ability to handle non-linear relationships, and capacity to provide feature importance insights. Hyperparameter tuning was conducted using GridSearchCV with a predefined grid of parameters, optimizing for negative MAE. The optimal hyperparameters identified were:

- colsample_bytree: 0.8
- learning_rate: 0.01
- max_depth: 3
- min_child_weight: 3
- n_estimators: 300
- subsample: 0.6

The tuned XGBoost model significantly improved over its untuned counterpart. Table 1.2 compares the performance metrics of the tuned and untuned XGBoost models.

Table 1.2: Performance Comparison of Untuned and Tuned XGBoost Models

Model	Test RMSE	Test MAE	Test R^2
XGBoost (untuned)	6138.98	4368.90	0.5692
XGBoost (tuned)	5642.49	3949.18	0.6360

The tuned XGBoost model achieved a test MAE of 3949.18, a substantial improvement over the untuned model's 4368.90, and an R^2 of 0.6360, closely approaching Linear Regression's 0.6380. These results indicate that hyperparameter tuning enhanced XGBoost's predictive accuracy, making it a competitive alternative to Linear Regression while offering additional interpretability through feature importance analysis.

1.4 Feature Importance

The tuned XGBoost model provides feature importance scores, which indicate the relative contribution of each feature to the model's predictions. Although exact numerical values for feature importances were not explicitly provided in the text outputs, the notebook includes a horizontal bar plot titled "XGBoost Feature Importances (Tuned Model)" generated using plt.barh. The plot displays the importance of each feature, sorted in ascending order, with feature names derived from the preprocessing pipeline, including numerical features (e.g., energy_consumption, co2_emissions_g_per_km) and one-hot encoded categorical features (e.g., vehicle_type_ICE).

Based on the dataset's context, features such as energy_consumption, co2_emissions_g_per_km, cost_per_km, and vehicle_type are likely among the most influential predictors of maintenance costs. For EVs, battery-related features (e.g., energy_storage_capacity) may play a significant role, while for ICE vehicles, engine-related features (e.g., mileage_km, torque_Nm) could be more prominent. The absence of specific importance scores limits precise interpretation, but the visualization suggests that these features drive the model's predictions, offering insights into the factors affecting maintenance costs.

1.5 Comparative Analysis of EV and ICE Vehicles

Exploratory data analysis (EDA) was conducted to compare key metrics between EVs and ICE vehicles, providing context for the model results. Two types of visualizations were generated: histograms and box plots.

1.5.1 Histograms

Figure 1.1 presents histograms comparing the distributions of nine metrics: energy consumption, CO_2 emissions, maintenance cost, cost per km, energy storage capacity, mileage, acceleration, torque, and lifespan.

Figure 1.1: Histograms Comparing EV and ICE Vehicle Metrics

Key observations from the histograms include:

- Energy Consumption: EVs peak around 15–20 kWh/100 km, while ICE vehicles peak around 10–15 L/100 km, indicating higher energy consumption for EVs.
- CO_2 Emissions: EVs show a sharp peak near 0 g/km, reflecting their zero-emission nature, while ICE vehicles peak around 100–150 g/km.
- Maintenance Cost: EVs have a lower median (2000–4000 units) compared to ICE vehicles (4000–6000 units).
- Cost per km: EVs peak at 0.10–0.15 units/km, lower than ICE vehicles at 0.20–0.25 units/km.
- Energy Storage Capacity: EVs have higher capacity (50–60 kWh) compared to ICE vehicles (20–40 L equivalent).
- Mileage: ICE vehicles show higher mileage (15,000–20,000 km) than EVs (10,000–15,000 km).
- Acceleration: EVs accelerate faster (6–8 seconds) than ICE vehicles (8–10 seconds).
- Torque: EVs exhibit higher torque (300–400 Nm) than ICE vehicles (200–300 Nm).
- Lifespan: ICE vehicles have a slightly longer lifespan (12–14 years) than EVs (10–12 years).

These findings suggest that EVs offer environmental and cost advantages, while ICE vehicles may have higher mileage and longevity.

1.5.2 Box Plots

Figure 1.2 presents box plots comparing the same metrics, reinforcing the histogram trends.

Figure 1.2: Box Plots Comparing EV and ICE Vehicle Metrics

The box plots confirm that EVs have lower median CO_2 emissions, lower maintenance costs, higher torque, and better acceleration, while ICE vehicles exhibit higher mileage and slightly longer lifespans. The narrower interquartile ranges for EVs in metrics like CO_2 emissions and torque indicate more consistent performance compared to the wider variability in ICE vehicles.

1.6 Correlation Analysis

A feature correlation matrix, shown in Figure 1.3, was used to examine relationships between the dataset's features.

Figure 1.3: Feature Correlation Matrix

Key correlations include:

- Strong Positive Correlations:
  - energy_consumption and co2_emissions_g_per_km (0.87): Higher energy consumption is associated with higher emissions, primarily for ICE vehicles.
  - co2_emissions_g_per_km and maintenance_cost_annual (0.73): Higher emissions correlate with higher maintenance costs.
  - maintenance_cost_annual and cost_per_km (0.65): Higher maintenance costs are associated with higher per-kilometer costs.
  - acceleration_0_100_kph_sec and torque_Nm (0.55): Higher torque is linked to faster acceleration.
- Strong Negative Correlations:
  - energy_storage_capacity and co2_emissions_g_per_km (-0.66): Larger energy storage capacities are associated with lower emissions.
  - co2_emissions_g_per_km and torque_Nm (-0.70): Higher emissions correlate with lower torque.
  - maintenance_cost_annual and energy_storage_capacity (-0.53): Larger energy storage capacities are linked to lower maintenance costs.
- Moderate Correlations:
  - energy_consumption and cost_per_km (0.78): Higher energy consumption is associated with higher per-kilometer costs.
  - lifespan_years and energy_consumption (0.40): Longer lifespans are weakly associated with higher energy consumption.

These correlations provide context for the model results, suggesting that features like co2_emissions_g_per_km and energy_storage_capacity are critical in understanding maintenance cost differences between EVs and ICE vehicles.

1.7 Discussion of Results

The results demonstrate that Linear Regression is the most accurate model for predicting vehicle maintenance costs, with a test MAE of 3903.24 and R^2 of 0.6380, indicating robust predictive capability. The tuned XGBoost model, with a test MAE of 3949.18 and R^2 of

0.6360, offers a competitive alternative and provides feature importance insights, which are valuable for interpreting the factors driving maintenance costs.

The EDA visualizations highlight significant differences between EVs and ICE vehicles. EVs exhibit lower CO_2 emissions, lower maintenance costs, and superior performance metrics (e.g., torque and acceleration), suggesting potential cost and environmental benefits. However, ICE vehicles show higher mileage and slightly longer lifespans, which may reflect their established technology and widespread use. The correlation analysis supports these findings, with strong relationships between emissions, maintenance costs, and energy storage capacity indicating that EVs' design (e.g., larger batteries) may contribute to lower operational costs.

Limitations of the study include the potential lack of representativeness in the dataset, which may not capture all vehicle types or operating conditions. The absence of specific feature importance values limits detailed interpretation of the tuned XGBoost model's predictors. Additionally, the dataset's historical nature may not account for future technological advancements in EV or ICE vehicle design. Future research could incorporate larger, more diverse datasets, additional features (e.g., driving conditions, maintenance frequency), or advanced modeling techniques to enhance predictive accuracy and generalizability.

In conclusion, this study provides a robust framework for predicting vehicle maintenance costs using machine learning, offering insights into the comparative efficiency of EVs and ICE vehicles. The findings suggest that EVs may offer economic and environmental advantages, particularly in terms of lower maintenance costs and emissions, although their operational characteristics differ from ICE vehicles. These results have implications for consumers, manufacturers, and policymakers seeking to promote sustainable transportation solutions.