Report: Modeling Canadian CO2 Emissions in Cars

Capstone Project

Prasan Ratnayake (Zone 24x7)
Data Science Engineer

Introduction

The transportation sector stands as a significant contributor to Canada's carbon footprint, with vehicles accounting for a substantial portion of CO_2 emissions. Understanding and mitigating these emissions have become critical in the pursuit of sustainable environmental practices.

The dataset utilized in this study encapsulates a comprehensive official record of CO_2 emissions data sourced from a myriad of cars, encompassing various makes, models, and characteristics. Spanning 7385 records and 12 columns, this dataset offers a rich repository of information crucial for assessing and predicting CO_2 emissions within the Canadian automotive landscape

The relevance of this analysis transcends the confines of a mere predictive modeling exercise. At its core, this study seeks to address the pressing need for a deeper understanding of the factors influencing CO_2 emissions in cars. By leveraging machine learning techniques and extensive data analysis, we aim to unravel patterns, correlations, and predictive insights that can significantly impact policymaking, environmental initiatives, and consumer awareness.

Data

The dataset comprises 7385 records and 12 columns, providing essential insights into CO2 emissions factors for different cars. Abbreviations used in the dataset to describe features are detailed in the Data Description sheet, enhancing the understanding of the data.

Url :[Dataset]()

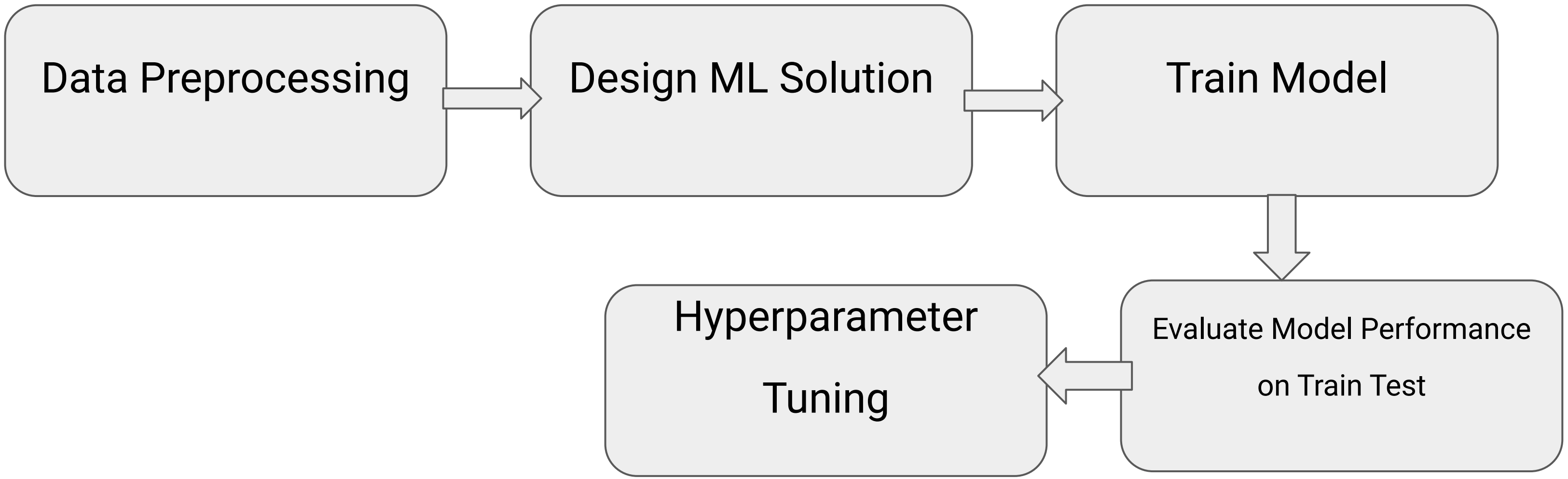
Data Columns

Key columns such as 'Make', 'Model', 'Vehicle Class', 'Engine Size (L)', 'Cylinders', 'Transmission', 'Fuel Type', and fuel consumption metrics ('Fuel Consumption City', 'Highway', and 'Combined') form the backbone of this dataset, contributing vital information about CO2 emissions.

Data

	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	ILX	COMPACT	2.0	4	AS5	Z	9.9	6.7	8.5	33	196
1	ACURA	ILX	COMPACT	2.4	4	M6	Z	11.2	7.7	9.6	29	221
2	ACURA	ILX HYBRID	COMPACT	1.5	4	AV7	Z	6.0	5.8	5.9	48	136
3	ACURA	MDX 4WD	SUV - SMALL	3.5	6	AS6	Z	12.7	9.1	11.1	25	255
4	ACURA	RDX AWD	SUV - SMALL	3.5	6	AS6	Z	12.1	8.7	10.6	27	244

Methodology



Methodology

Data Preprocessing

Before model development, a rigorous data preprocessing phase was executed. This phase involved meticulous cleaning, handling missing values, categorical variable encoding, and feature scaling. The objective was to ensure the dataset's readiness for subsequent modeling tasks.

Model Development and Hyperparameter Tuning

A variety of machine learning techniques were employed to predict CO2 emissions effectively. Through systematic experimentation, hyperparameter tuning, and cross-validation, the best-performing model was identified.

Best Model and Accuracy

The XGBoost algorithm emerged as the most accurate model, achieving an impressive R-squared score of 99.07%. Notably, the Random Forest model closely followed with a score of 98.89%. Other models, such as Support Vector Machines and Linear Regression, also contributed valuable insights.

Results

	Model	R-squared Score
2	XGBoost	99.068980
1	Random Forest	98.893750
3	Support Vector Machines	93.962746
0	Linear Regression	89.002191

Discussion

Understanding CO2 Emissions Patterns

The analysis of CO2 emissions in Canadian cars uncovered intricate patterns and relationships among various vehicle features and emission levels. By exploring factors such as engine specifications, fuel types, and consumption metrics, we gained valuable insights into the dynamics influencing emissions. Notably, features like engine size, cylinder count, and fuel consumption emerged as significant contributors to CO2 emissions, corroborating existing knowledge in the automotive domain.

Impact of Machine Learning Techniques

The application of diverse machine learning algorithms unveiled promising avenues for predicting CO2 emissions. Among these, XGBoost and Random Forest models demonstrated exceptional accuracy, surpassing traditional linear models and providing deeper insights into the nonlinear relationships within the dataset. The high accuracy achieved by these models underscores their potential as robust predictive tools for assessing and managing CO2 emissions in cars.

Discussion

Significance for Policy and Sustainability Initiatives

The findings from this analysis hold immense implications for policymakers, environmentalists, and industry stakeholders. Accurate predictive models enable informed decision-making regarding emission regulations, fuel efficiency standards, and incentivizing eco-friendly vehicle designs. Additionally, understanding emission patterns can aid consumers in making environmentally conscious choices when purchasing vehicles, fostering a shift towards greener transportation practices.

Limitations and Future Directions

While our analysis provided significant insights, several limitations warrant consideration. The dataset's completeness and representativeness may influence model performance, and the inclusion of additional data sources could enhance predictive accuracy. Furthermore, exploring temporal trends and integrating real-time emissions data could bolster the models' predictive capabilities in a dynamic automotive landscape.

Discussion

Conclusion and Recommendations

In conclusion, the modeling of CO_2 emissions in Canadian cars represents a crucial step towards understanding, predicting, and mitigating environmental impacts. The insights gained from this analysis not only contribute to the academic discourse but also offer actionable insights for policymakers and industry stakeholders. Moving forward, continuous refinement of models and data collection methodologies stands imperative for developing robust, adaptive models capable of addressing evolving emission challenges.