Topic: Comparative Analysis of Vehicle Price Prediction using Gradient Boost, Random Forest and SVM.

Name: Shubh Gupta || BUID: U42785653

Abstract:

This report presents a comprehensive analysis of vehicle price prediction using advanced machine learning techniques. The project's primary objective is to develop a predictive model that can accurately estimate a vehicle's price based on various features such as make, model, engine size, body style, and more. Utilizing a well-curated dataset comprising numerous vehicle attributes, the study embarks on a detailed exploration of data characteristics, followed by rigorous preprocessing to render the data suitable for machine learning applications.

Key methodologies employed in this study include the use of popular machine learning models: Gradient Boosting, Random Forest, and Support Vector Machine (SVM). Each model was meticulously trained and tested on the processed dataset, with performance evaluated based on Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE). The analysis revealed insightful findings on the models' predictive capabilities, with Gradient Boosting demonstrating superior performance.

Introduction:

In an era where data-driven decision-making is paramount, the automotive industry stands at the forefront of leveraging advanced analytics for strategic advantage. This project is situated at the intersection of machine learning and automotive economics, with the aim to develop a sophisticated model capable of accurately predicting vehicle prices. Such

predictions are not merely academic; they hold substantial practical value in guiding manufacturers, dealerships, and consumers through the complexities of the automotive market. The project leverages a rich dataset encompassing a wide spectrum of vehicle attributes, including but not limited to make, model, engine size, and body style. The underlying hypothesis is that a meticulously crafted machine learning model can unravel the intricate patterns within this data, offering precise estimations of vehicle prices.

To navigate this challenge, the study adopts a methodical approach, implementing and scrutinizing a suite of advanced machine learning algorithms such as Gradient Boosting, Random Forest, and Support Vector Machine (SVM). These models were selected for their robustness and versatility in handling high-dimensional and non-linear data structures. The efficacy of each model is rigorously evaluated using Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE), metrics that provide a quantitative measure of prediction accuracy. This comparative analysis is not merely a technical exercise but a quest to glean deeper insights into the factors that drive vehicle pricing. It is anticipated that the outcomes of this study will not only contribute to the academic discourse in predictive analytics but also offer tangible, actionable insights to stakeholders in the automotive sector, enhancing their capacity for informed decision-making in a competitive market landscape.

Dataset Description:

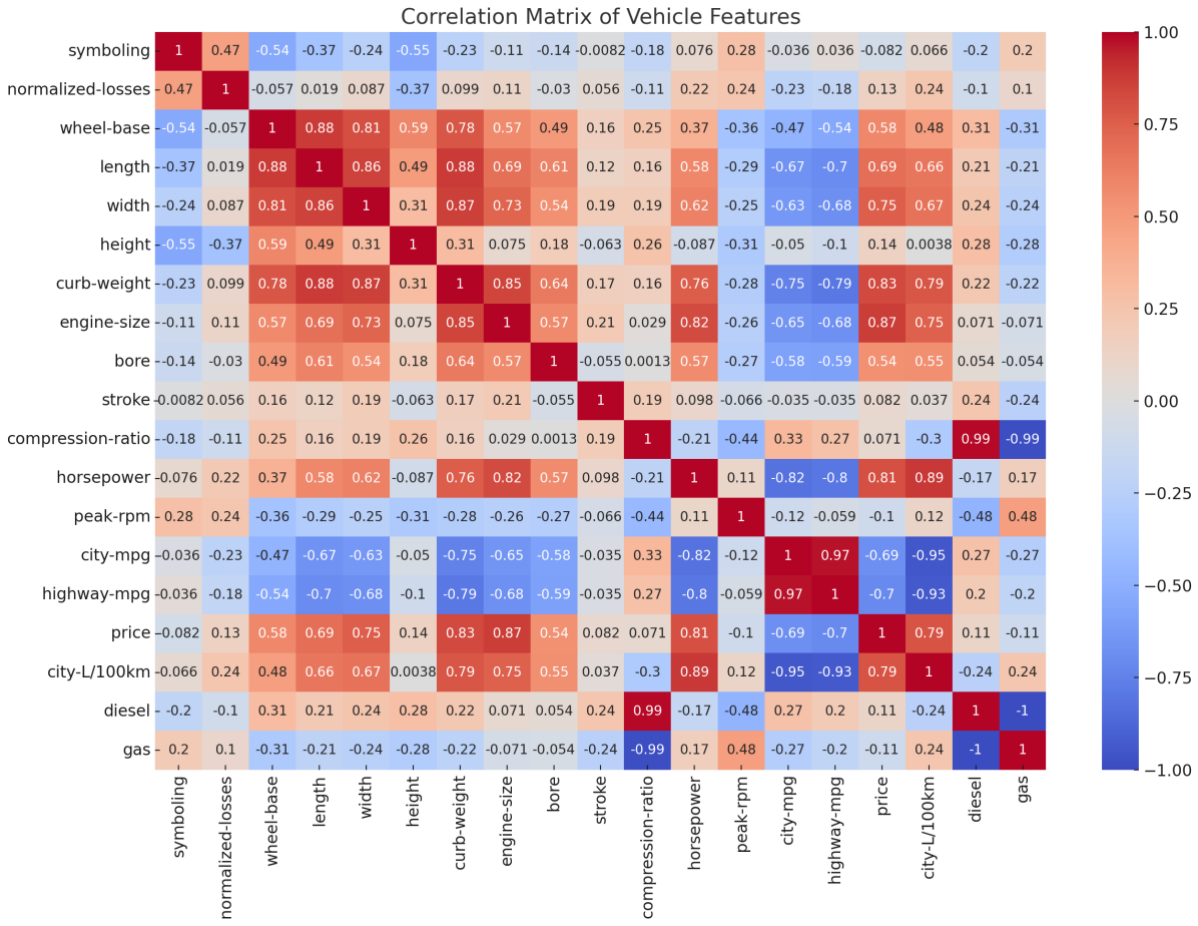
The foundation of this study is a comprehensive dataset meticulously curated to capture the multifaceted nature of vehicle pricing. Comprising an extensive range of variables, the dataset encapsulates critical vehicle characteristics such as make, model, engine size, body style, wheelbase, horsepower, and more, totaling over 20 distinct attributes. Each record in the dataset represents an individual vehicle, characterized by these features, with the target variable being the vehicle's market price. The dataset's diversity in vehicle types and features ensures a robust analytical framework, allowing for a nuanced exploration of the various factors influencing vehicle pricing.

Prior to analysis, the dataset underwent a rigorous preprocessing phase. This included handling missing values, encoding categorical variables for computational tractability, and normalizing numerical features to a standard scale. These steps were crucial in preparing the dataset for the application of machine learning algorithms, ensuring that the subsequent analysis was grounded on a clean, consistent, and representative data foundation.

Methodology:

The methodology adopted in this project is structured and comprehensive, encompassing the following key stages:

1. **Model Selection**: Central to the study were three machine learning models—Gradient Boosting, Random Forest, and Support Vector Machine (SVM). These models were chosen for their proven effectiveness in regression tasks and their ability to handle complex, multi-dimensional data.



2. **Model Training and Validation**: The models were trained on a partition of the dataset, allowing them to learn and adapt to the underlying patterns within the data. A separate subset of the data was reserved for testing, providing an unbiased evaluation of each model's predictive performance.

3. **Performance Evaluation**: To assess the models' effectiveness, two primary metrics were employed:

    - **Root Mean Squared Error (RMSE)**: This metric provides a measure of the model's prediction error magnitude, with lower values indicating higher accuracy.

    - **Mean Absolute Error (MAE)**: MAE offers a straightforward average of absolute errors, giving a clear indication of the typical prediction error size.

4. **Comparative Analysis**: The models' performances were juxtaposed to identify which model most accurately predicts vehicle prices. This comparison was crucial not only in determining the best-performing model but also in understanding the strengths and limitations of each approach within the context of vehicle price prediction.

This methodical approach ensured a thorough and objective analysis, leading to results that are both reliable and insightful for applications in the automotive industry.

Result Analysis

Model	RMSE	MAE
Gradient Boosting	2,562.63	1,648.28
Random Forest	2,973.10	1,934.26
SVM	11,104.09	7041.45

The comparative analysis of the Gradient Boosting, Random Forest, and Support Vector Machine (SVM) models yielded insightful findings. Gradient Boosting emerged as the top performer, exhibiting the lowest values in both Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE). This model's effectiveness can be attributed to its ability to iteratively correct errors from previous predictions, thereby refining its accuracy with each step. The Random Forest model, while not as precise as Gradient Boosting, still demonstrated commendable performance. Its ensemble approach, which aggregates predictions from multiple decision trees, proved effective in capturing the complex relationships within the dataset. In contrast, the SVM model, although a powerful tool for classification tasks, showed comparatively higher error metrics in this specific regression scenario. This could be due to its sensitivity to the high dimensionality and varied nature of the dataset.

These results underscore the importance of model selection in machine learning tasks. While all three models are capable in their own right, their performance can vary significantly based on the characteristics of the dataset and the nature of the prediction

task. The superiority of Gradient Boosting in this context highlights its robustness and adaptability, particularly in dealing with complex datasets like those involving vehicle price prediction.

Conclusion:

This project underscores the effectiveness of machine learning in predicting vehicle prices, with Gradient Boosting emerging as the most accurate model. The study highlights the critical role of selecting appropriate models based on dataset characteristics, with different models showing varying degrees of success. While Gradient Boosting demonstrated superior performance, it's important to recognize the inherent limitations tied to data quality and model selection.

The insights gained hold practical value for stakeholders in the automotive industry, aiding in pricing strategies and market analysis. Future developments could involve incorporating more diverse data and applying these models in real-world scenarios for further refinement. Overall, this project not only contributes to the field of predictive analytics but also exemplifies the potential of machine learning in enhancing data-driven decision-making in the automotive sector.