Introduction

In the dynamic realm of the automotive industry, the ability to accurately predict vehicle sales prices holds paramount importance for dealerships, manufacturers, and customers alike. This paper delves into the pivotal role of understanding and forecasting vehicle sales prices, elucidating its multifaceted significance across various stakeholders. Through a comprehensive analysis, this report explores the compelling reasons driving the need for precise price predictions and outlines the goals of constructing a robust regression model and identifying key variables in predicting vehicle sales prices.

Price predictions are crucial for automotive manufacturers and dealerships, guiding production volumes to meet market demand and optimizing inventory costs. These insights enable streamlined production schedules and responsive adjustments to market dynamics. Additionally, accurate pricing models empower dealerships to set competitive prices, attracting customers while ensuring profitability. For consumers, reliable pricing information facilitates informed decision-making, empowering them to navigate the market confidently. With comprehensive pricing insights, customers can evaluate vehicle value, compare prices, and negotiate effectively for optimal deals, fostering satisfaction and trust in the purchasing process.

The vehicle sales dataset used in this study is sourced from Kaggle and contains 558,837 observations with 16 variables. Among these variables, the manufacturing year stands as a pivotal predictor, ranging from years 1982 to 2015. Additionally, the odometer reading serves as a critical indicator of usage and wear, with lower mileage vehicles typically commanding higher prices due to their perceived reliability and longevity. The condition rating, ranging from 1 to 50, provides insights into a vehicle's overall state of preservation and mechanical integrity, impacting its market value accordingly. Furthermore, the Manheim Market Report (MMR) value offers an estimate of a vehicle's market worth based on comprehensive market data, serving as a valuable benchmark for refining price predictions and aligning them with prevailing market conditions.

In addition to these key variables, other factors such as vehicle make, model, trim, body type, transmission, state, color, and interior features contribute to the complex interplay of determinants shaping vehicle sales prices. Through meticulous analysis of these variables, this report endeavors to unravel the intricate relationships between predictors and sales prices in the automotive market, thereby enhancing our understanding of pricing determinants and facilitating informed decision-making for stakeholders across the industry.

Methods

Data Pre-Processing

To manage the computational burden, a random sample of 10,000 vehicles was drawn from the original dataset. The sample was then pre-processed to ensure data integrity and facilitate model training. Variables such as Vehicle Information Number, Seller, and Saledate were deemed irrelevant and subsequently dropped.

Categorical predictors, including make, model, trim, body, transmission, color, and interior, were converted into dummy variables to enable their inclusion in the modeling process. Meanwhile, numeric predictors such as year, odometer, condition, and MMR were standardized to a mean of 0 and standard deviation of 1 to ensure uniformity in scale across variables.

Missing values were identified in critical columns such as make, model, trim, body, transmission, and condition. To maintain data quality, all vehicles with missing values were dropped from the dataset. This pre-processing step resulted in a final dataset comprising 8,442 vehicles and 13 variables for subsequent analysis.

Analysis

An elastic net regression model, specifically Lasso regression with ridge regularization, was employed to address multicollinearity and prevent overfitting. Elastic net regression introduces penalty terms to the regression coefficients, effectively shrinking coefficients of less important predictors to zero. The model involved tuning for the penalty and mixture to optimize model performance.

A random forest model was constructed to capture nonlinear relationships and interactions between predictors. This ensemble learning technique aggregates predictions from multiple decision trees to improve overall prediction accuracy. The model was trained using a grid search approach to identify optimal hyperparameters such as the number of variables randomly sampled as candidates at each split (mtry), the minimum number of data points required to split a node (min_n), and the number of trees in the forest (trees).

A stochastic gradient boosting model was utilized to iteratively minimize prediction errors by building a sequence of weak learners. This ensemble learning technique combines the strengths of gradient boosting with the efficiency of stochastic gradient descent. The model's hyperparameters, including mtry, trees, min_n, tree_depth, learn_rate, loss_reduction, sample_size, and stop_iter, were tuned using a grid search approach to optimize model performance.

An 80/20 training-test split was employed for each model, with the models undergoing rigorous evaluation using 5-fold cross-validation to ensure reliable estimation of performance metrics such as RMSE, MAE, and R^2 on the test dataset. Finally, the best-performing model was selected based on its performance metrics for further analysis and interpretation.

Results

Exploratory Analysis

Exploratory data analysis revealed insightful patterns in the relationship between vehicle attributes and their selling prices. The analysis began by examining average selling prices across different vehicle attributes, providing valuable insights into consumer preferences and market trends.

The bar plots (Figures 1 and 2) illustrating the top 10 makes and body types by average selling price revealed significant variations in pricing across different categories. For example, luxury makes such as Ferrari and Bentley commanded higher selling prices compared to economy brands like Ford and Chevrolet. On the other hand, vehicles such as Mega Cab, G Convertible, Transit Van, Double Cab, and Crewmax Cab emerged as the top five in terms of average selling price, indicating a strong demand for commercial vehicles. The bar plots visually highlighted these disparities, offering a clear understanding of the market dynamics.

Figure 1: Bar Plot of top 10 makes by average selling price

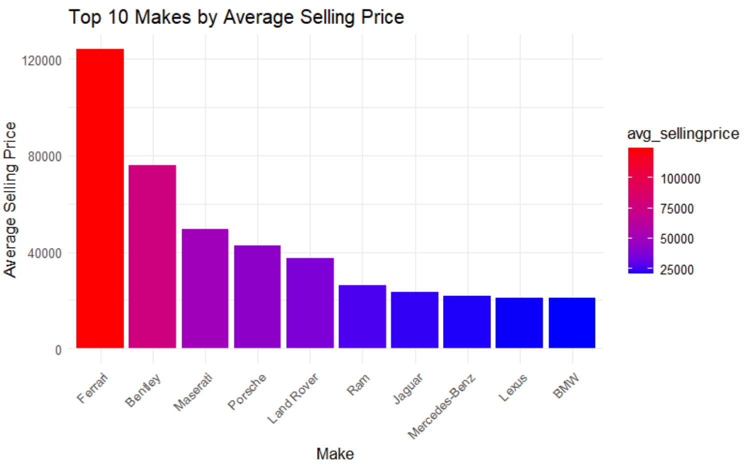
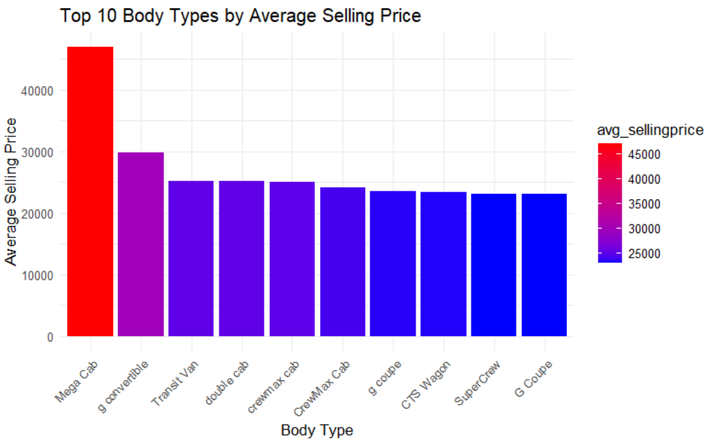
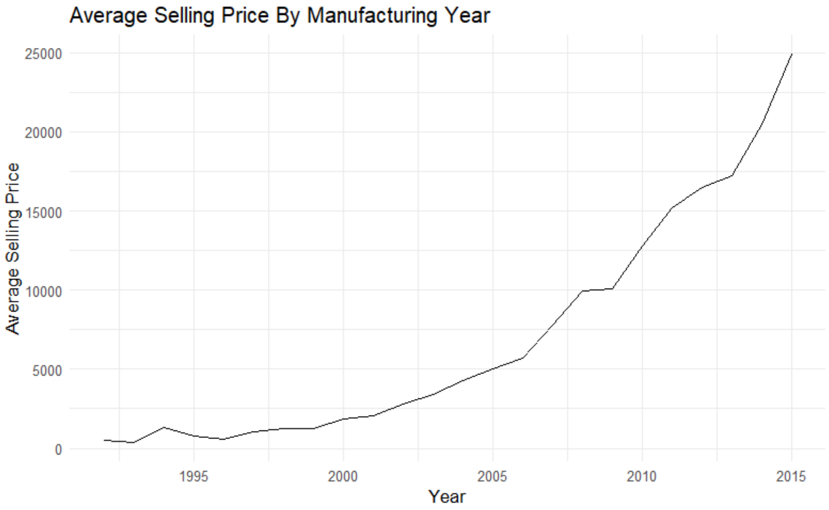


Figure 2: Bar Plot of top 10 body types by average selling price



Additionally, the time series plot depicting the average selling price by manufacturing year showcased trends in price fluctuations over time. The plot revealed that newer model years generally exhibited higher average selling prices, reflecting the depreciation of vehicles as they age. The time series plot provided valuable insights into the temporal evolution of vehicle prices, allowing for a deeper understanding of market trends and highlighting the importance of considering the age of vehicles in predicting their selling prices.

Figure 3: Time Series Plot of average selling price by manufacturing year



Model Evaluation and Comparison

Hyperparameter tuning was conducted for each model to identify the optimal configuration that maximizes predictive performance. Table 1 presents the optimal hyperparameters for the Penalized Regression Model, including the penalty and mixture values. Similarly, Table 2 and Table 3 outline the optimal hyperparameters for the Random Forest Model and the Stochastic Gradient Boosting Model, respectively.

Table 1: Optimal hyperparameters of the Penalized Regression Model

Parameter	Optimal Value
Penalty	4.66e-08
Mixture	0.4905

Table 2: Optimal hyperparameters of the Random Forest Model

Parameter	Optimal Value
Trees	682
Min_n	29
Mtry	1064

Table 3: Optimal hyperparameters of the Stochastic Gradient Boosting Model

Parameter	Optimal Value
Trees	1348
Min_n	17
Mtry	627
Tree Depth	4
Learn Rate	0.0037
Loss Reduction	1.799e-07
Sample Size	0.8556
Stop Iter	9

Upon analyzing the performance metrics of each model (Table 4), it is evident that the Random Forest Model outperformed the other two models in terms of root mean square error (RMSE) and mean absolute error (MAE), with values of 2087.2015 and 1028.7197, respectively. The Random Forest Model also achieved the highest R-squared value of 0.9520, indicating a better fit to the data compared to the other models.

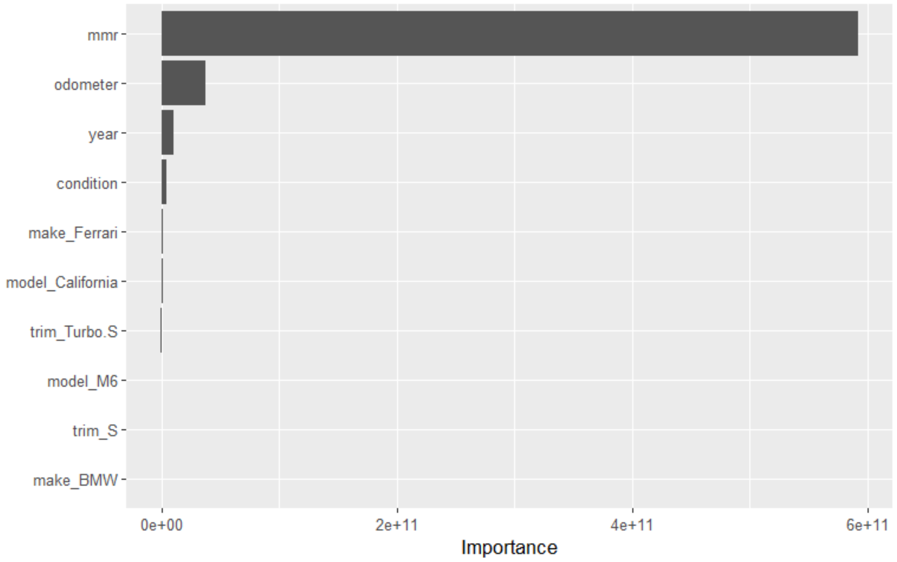
Table 4: Performance metrics of the models

Model	RMSE	R-Square	MAE
Penalized Regression	2208.6648	0.9464	1166.6057
Random Forest	2087.2015	0.9520	1028.7197
Stochastic Gradient Boosting	2153.4236	0.9496	1102.3722

While the Penalized Regression Model demonstrated competitive performance with an RMSE of 2208.6648 and an R-squared value of 0.9464, it exhibited slightly higher error metrics compared to the Random Forest Model. The Stochastic Gradient Boosting Model also performed well but showed slightly inferior results compared to the Random Forest Model, with an RMSE of 2153.4236 and an R-squared value of 0.9496.

Further analysis using the Random Forest model unveiled critical insights into the importance of predictor variables in predicting vehicle sales prices. The variable importance plot (Figure 4) revealed that the Manheim Market Report (MMR) and odometer reading emerged as the two most important variables, followed by the manufacturing year and condition rating. This finding underscores the significance of the age, condition, and usage of a vehicle along with its current market value in predicting its selling price.

Figure 4: Variable Importance Plot of the Random Forest Model



In summary, based on the performance metrics evaluated, the Random Forest Model emerges as the most effective model for predicting vehicle sales prices in this analysis. Its superior performance in terms of predictive accuracy makes it the preferred choice for forecasting sales prices in the automotive market.

Conclusion

In conclusion, my analysis highlights the Random Forest model as the most effective predictor of vehicle sales prices, demonstrating superior accuracy compared to alternative models. Notably, the Random Forest model outperformed others, showcasing its potential to offer precise predictions crucial for stakeholders in the automotive industry. Through my analysis, I identified the Manheim Market Report and vehicle mileage as the most influential variables in predicting selling prices. These insights underscore the importance of considering market data and vehicle usage in pricing strategies.

However, it's essential to recognize certain limitations in my approach. While training the model on a random sample helped manage the computational burden, it may not fully capture the complexity of the automotive market. Moreover, there's a risk of overfitting, and the possibility of imprecise predictions, as indicated by the RMSE of $2087. To address these limitations, future research could leverage the entire dataset and employ additional feature engineering techniques. Incorporating data from external sources, such as consumer sentiment and vehicle history reports, could further enhance predictive accuracy.

The implications of my model extend beyond predictive analytics, offering valuable insights for stakeholders. Dealerships and online platforms could leverage dynamic pricing strategies informed by my model's predictions, optimizing pricing decisions in response to market fluctuations. Moreover, my findings advocate for fair pricing practices, empowering customers to make informed decisions in the automotive market. By embracing data-driven approaches and incorporating advanced analytics, stakeholders can navigate the complexities of vehicle pricing, driving innovation and competitiveness in the industry.