Predictive Modeling of Car Prices: Unveiling Patterns and Enhancing Accuracy in the Automotive Marketplace

Group 3

Nithya PS
Abiraj R
K.Lakshmi
Pankaj Kumar
Aanchal
Naveen

December 15, 2023

Introduction

The automotive industry, characterized by constant innovation and consumer diversity, undergoes a continuous evolution in response to technological advancements and shifting market demands. Understanding the factors that influence the pricing of automobiles is crucial for manufacturers, dealers, and consumers alike. In this context, the dataset under consideration encompasses a comprehensive array of attributes associated with various cars, serving as a rich source for predictive modeling.

The dataset integrates a multitude of features, encapsulating both quantitative and qualitative aspects of the vehicles. Parameters such as location, year of manufacture, kilometers driven, fuel type, transmission mechanism, owner type, mileage, engine specifications, power output, color, seating capacity, number of doors, and new price collectively contribute to the intricate tapestry of information. The pivotal objective is to harness machine learning techniques to decipher patterns within this data and construct models capable of predicting the prices of these diverse automobiles.

Before delving into the complexities of machine learning models, a preliminary exploration of the dataset's descriptive statistics is imperative. This statistical analysis serves as a compass, guiding us through the distribution of the target variable — 'Price.' The mean, median, minimum, and maximum values offer a bird's-eye view of the central tendencies and outliers within the pricing structure. Simultaneously, a correlation analysis unveils potential relationships between numerical features, providing insights into interdependencies that may influence subsequent regression models.

This meticulous examination not only reveals the inherent characteristics of the data but also lays the groundwork for informed decision-making in the modeling phase. The exploration of these statistics is not a mere formality but a crucial step in deciphering the language of the dataset.

Desciptive Statistics

The dataset under consideration comprises 5961 rows and 15 columns, offering a rich tapestry of information about various cars. The dataset's structure, denoted as (5961, 15), signifies the presence of 5961 data points (cars) and 15 different features. Each row represents a unique car entry, while the columns encompass a diverse set of attributes, including 'Year,' 'Kilometers Driven,' 'Seats,' 'Number of Doors,' and 'Price.'

The statistical measures provide a glimpse into the central tendencies and variability within key features. The 'Year' column indicates that the cars in the dataset span from 1998 to 2019, with an average manufacturing year of approximately 2013. The 'Kilometers Driven' feature exhibits a wide range, from a minimum of 171 km to a maximum of 6,500,000 km, highlighting the diversity in usage. The 'Seats' column suggests that most cars have five seats, while the 'Number of Doors' tends to be four in the majority of cases.

The 'Price' feature, our target variable, showcases a mean value of 9.53, a median (50th percentile) of 5.66, and a standard deviation of 11.21. The substantial difference between the mean and median suggests potential skewness in the price distribution, which can be further explored graphically. The range of prices extends from a minimum of 0.44 to a maximum of 160, indicating a diverse pricing spectrum within the dataset.

The unique car models in the 'Name' column present a varied and extensive list, ranging from popular brands like Maruti and Hyundai to luxury options like Audi and BMW. The presence of such a diverse set of models underscores the dataset's comprehensiveness and the representation of a wide array of car types and classes.

While the count of entries for each feature suggests some variability, it is essential to note potential missing values. For instance, 'Seats' has 5956 non-null entries, indicating a small number of missing values. This observation highlights the importance of data cleaning and imputation processes before undertaking further analyses or model building.

In summary, the descriptive statistics provide a foundational understanding of the dataset's structure, key features, and potential areas for exploration. Further visualizations and modeling efforts will unveil deeper insights into the relationships between these features and the pricing dynamics of the cars.

Insights

In analyzing the relationship between owner type and kilometers driven, a clear pattern emerges, revealing distinct driving behaviors among different owner categories. Notably, vehicles under the "Third" owner type consis-

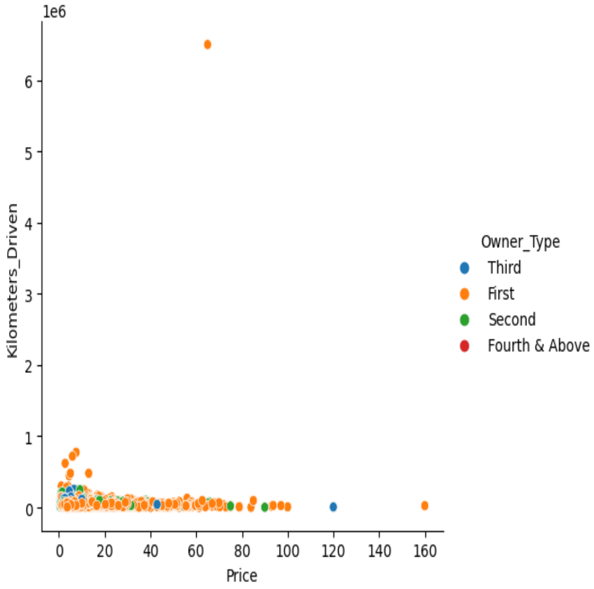

Figure 1: Kilometeres Driven VS Owner Type

tently exhibit the highest kilometers driven, as indicated by consistently elevated values on the y-axis across all price ranges. Conversely, cars under the "First" owner type generally have the lowest values on the y-axis, reflecting comparatively lesser distances covered. Furthermore, an intriguing trend unfolds as the graph progresses along the x-axis, showcasing an upward slope indicating an increase in kilometers driven with higher car prices. This suggests that, irrespective of owner type, more expensive cars tend to be driven greater distances. However, the non-linear nature of the lines on the graph implies that factors beyond owner type contribute to this phenomenon, hinting at a nuanced interplay of variables influencing the mileage of vehicles in the diverse automotive landscape.

Determining whether the listed price of new cars is a challenging task due to the many factors that drive a used vehicles predict on the Indian market today. Deciding whether a new cars are worth than the second-hand cars is very difficult task for sellers as well. Several factors including Transmission, fuel types, year, manual, owner type, etc.,can influence the actual worth of car. In order to get a better understanding of the data, we plotted bargraphs.To analyse the project to which our features are linearly related to price, we plotted the price against fuel type and year for a particular model.

The heatmap visually encapsulates the correlations between numerical

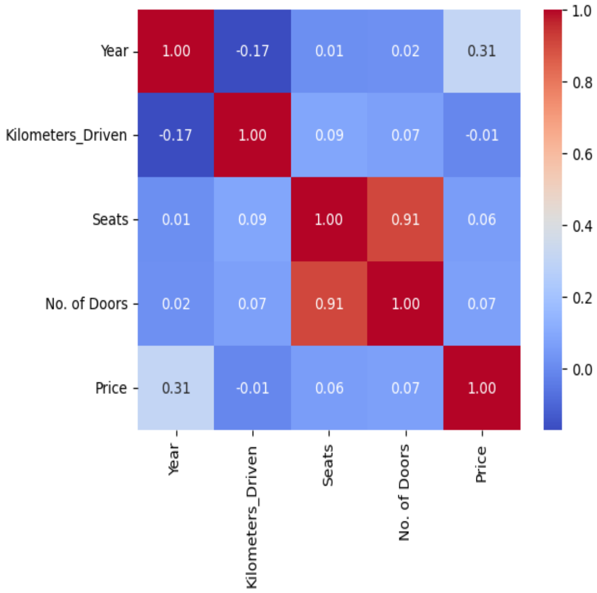

Figure 2: Heatrmap

variables in the car dataset, revealing insightful patterns. Kilometers Driven exhibits a positive correlation with Price and the number of doors, suggesting that higher-mileage cars tend to be pricier and possess more doors, potentially associated with commercial usage. Conversely, a negative correlation between Kilometers Driven and Year implies that older cars accumulate higher mileage, aligning with the expected wear over time. A weak positive correlation emerges between Seats and Price, indicating that cars with more seats may be associated with higher prices. Additionally, the positive correlation between the number of doors and Price suggests that cars with increased door count tend to be more expensive, possibly due to perceived luxury or spaciousness. Notably, negative correlations between Year and both Seats and the number of doors hint at evolving car design trends over time. Overall, the heatmap provides a comprehensive overview of the nuanced relationships within the dataset, guiding our understanding of how these variables interact in shaping the automotive landscape.

Predictive Modeling

In the pursuit of predicting car prices based on a comprehensive dataset, two distinct regression models were employed: Linear Regression and Random

Forest Regressor. Each model brings its unique strengths and characteristics to the table, offering a nuanced perspective on the relationship between various features and the target variable, 'Price.'

Linear Regression Model:

The Linear Regression model, a foundational approach in predictive modeling, assumes a linear relationship between the predictor variables and the target variable. In the context of this analysis, the model seeks to establish a linear equation that best fits the relationship between the selected features and the observed car prices. The R-squared value, a key metric in regression analysis, serves as a measure of how well the model explains the variance in the target variable. The initial application of Linear Regression yielded an R-squared value of approximately 0.48.

While Linear Regression provides a straightforward interpretation of the relationship between individual predictors and the target variable, it assumes a linear correlation, which may not capture complex non-linear patterns present in the dataset. This limitation becomes evident when dealing with intricate relationships between features that deviate from linearity.

Random Forest Regressor Model:

To address the limitations of Linear Regression, a Random Forest Regressor, a more sophisticated and flexible model, was introduced. The Random Forest model comprises an ensemble of decision trees, each contributing to the overall prediction. This ensemble approach allows the model to capture non-linear relationships, interactions, and intricate patterns within the data. The Random Forest Regressor demonstrated a remarkable improvement in predictive performance, yielding an R-squared value of approximately 0.83.

The Random Forest model's ability to handle complex relationships, outliers, and non-linearities makes it a powerful tool in predicting car prices. It excels in scenarios where the relationship between predictors and the target variable is intricate and might not conform to a simple linear pattern. Moreover, the Random Forest model is inherently robust to overfitting, providing a balance between accuracy and generalization.

Model Comparison and Recommendation:

Comparing the two models, the Random Forest Regressor clearly outperforms the Linear Regression model in terms of predictive accuracy. The substantially higher R-squared value for the Random Forest model suggests that it better captures the underlying patterns and complexities within the dataset, leading to more accurate predictions of car prices. The Random Forest's capacity to handle non-linear relationships and interactions makes it well-suited for the diverse and intricate nature of the automotive dataset.

Conclusion

As we conclude this expedition into the intricate world of predictive modeling for car prices, the journey has not only revealed patterns and insights but also unveiled the dynamic interplay of methodologies in the automotive landscape. The contrasting performances of Linear Regression and the Random Forest Regressor have underscored the importance of selecting a model that aligns with the intricacies of the dataset.

The Linear Regression model, a stalwart in the realm of regression analysis, provided us with an initial glimpse into the linear relationships within the data. However, its limitations in capturing non-linear patterns became evident as we delved deeper into the nuances of the automotive dataset. The Random Forest Regressor, with its ensemble of decision trees, emerged as a formidable contender, flexing its muscles in deciphering the intricate dance between various features and car prices. The substantial improvement in the R-squared value demonstrated its prowess in navigating the complex and non-linear landscape of car pricing.

In this pursuit of prediction, we've not only weighed the merits of models but have also delved into the heartbeat of the automotive industry. The diverse range of car models, the evolution of manufacturing years, and the kaleidoscope of features have painted a vivid picture of a market in constant flux. It's a marketplace where consumer preferences, technological advancements, and market trends converge to shape the price tags on our four-wheeled companions.