Analysing Used Car Pricing using Regression Analysis

Abstract

In the modern world, where pollution is rampant because of cars and inflation is on the rise, used car market has emerged as a highly competitive market. Pricing always plays crucial role in attracting customers and ensuring profitability of used car dealership, it is essential for us to set the right prices for our cars to gain a competitive edge. To achieve this, we need to understand the factors that influence customer pricing decisions and identify reliable predictors of sale prices. The problem at hand is to determine if the mileage of cars is a good predictor of their sale prices and if the relationship between mileage and price differs across various car brands. The goal is to develop a data-driven approach that helps the dealership set optimal prices to maximize profitability and customer satisfaction.

Approach and Motivation

Our approach involves collecting and preprocessing data on various features and sale prices of used cars. We then conduct exploratory data analysis (EDA) to identify patterns and relationships between these features and prices for each car brand. Next, we employ multiple linear regression, considering mileage as a predictor variable, to build a pricing model for the cars. We evaluate the performance of the regression models using metrics such as R-squared and root mean squared error (RMSE). This approach aligns with related work in the field of pricing optimization using regression analysis and leverages the power of data-driven insights for decision-making.

Dataset

For this project, we are using the dataset on used car sales with various features such as Sale Price, Mileage, Year of Manufacturing, Engine Power, Engine Size, New Price of Cars, No. of Seats, Fuel type, Type of Transmission, Location and Type of Fuel used. The data provided comprises of *7253 rows and 13 columns*.

Pre-Processing

In order to get better understanding of this data, data preprocessing, feature engineering and feature scaling. For e.g., missing values were found in columns **Mileage**, **Power**, **Seats**, **Price** and **New_Price**. The rows with missing values were dropped but because the New_Price column had more than 6000 values missing, the whole column was dropped.

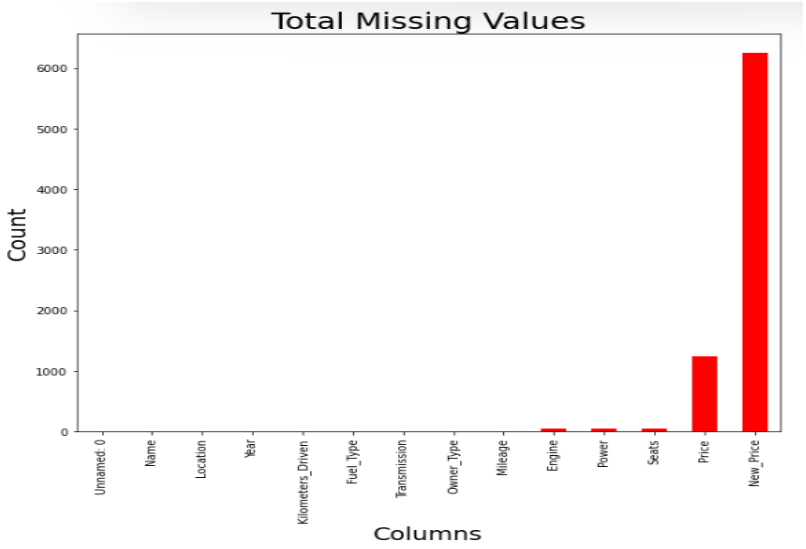

Fig. Graph depicting no. of missing values.

Also, the Column Unnamed: 0 was dropped because of irrelevant data. So, after the first round of preprocessing, we have been left with *5975 rows and 12 columns.*

In the next round, the numerical data was extracted from **Mileage, Engine** and **Power** columns and the null values in Power and Mileage were replaced by the mean values of the respective columns. Also, the outliers were identified using boxplots and handled accordingly.

Using Feature Engineering, in case of Engine, Power and Mileage, the values of these data columns had both numeric and categorical data. In order to properly analyse the data, we had to extract the numeric data and then convert the remaining values to float64 data type.

The name of the cars from column **Name** were selected such that only the name of the brand remains, thus creating a new column **Brand** and dropping the previous column Name. This helped in removing unnecessary and irrelevant data from the dataset and make it concise.

We also used feature scaling for accurate analysis of the data. During univariate analysis of data, it was found that the columns **Price**, **Engine** and **Power** were heavily skewed. So, to normalize the data in order to get more accurate results of regression analysis, log normalization was applied to the respective columns. Log normalization, also known as logarithmic transformation, is a data preprocessing technique used to transform skewed or highly variable data into a more normalized or symmetric distribution. It involves applying the natural logarithm (base e) or other logarithmic functions to the data.

The main purpose of log normalization is to reduce the impact of extreme values and make the data more suitable for certain statistical analyses or machine learning algorithms that assume a normal distribution or require linear relationships between variables.

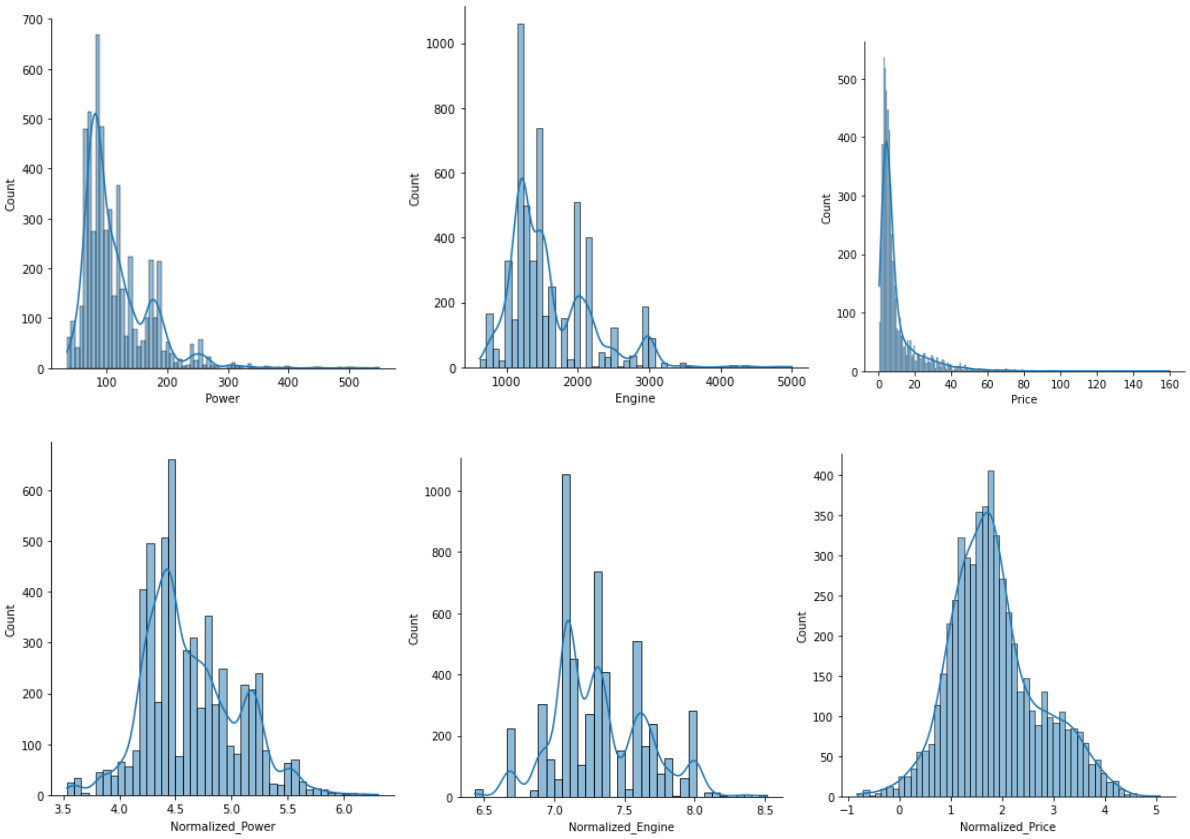

Fig. (Top): Power, Engine & Price without Feature Scaling, (Bottom) After Feature Scaling

Data Analysis was conducted to gain insights into the relationship between mileage and sale prices for each brand of the car. Although outliers were present, maximum of the brands displayed a linear growth in the price of car with more mileage. For e.g. When Maruti's mileage and price were analyzed, a clear growing trend was displayed, indicating that increasing mileage lead to increase in price of the car. The same trend was displayed in case of Hyundai, Honda, etc.

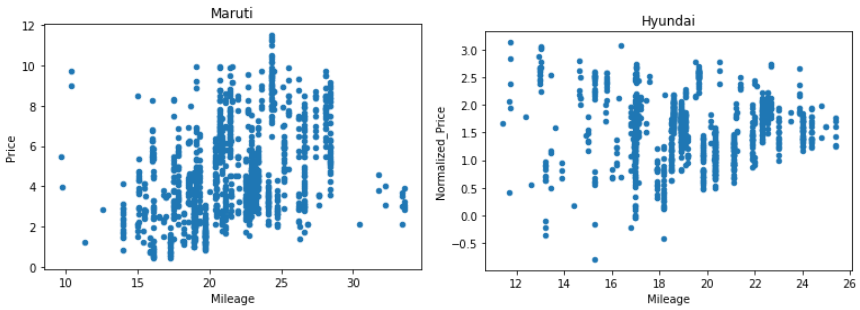

Fig. Maruti and Hyundai Price and Mileage trends

Although most of the brands displayed increase in price with increase in mileage, there are some companies like Land Rover, Renault or Chevrolet that displayed interesting and unique

patterns. The reason for this might be that people don't consider Mileage as a big variable when buying cars of the mentioned brands.

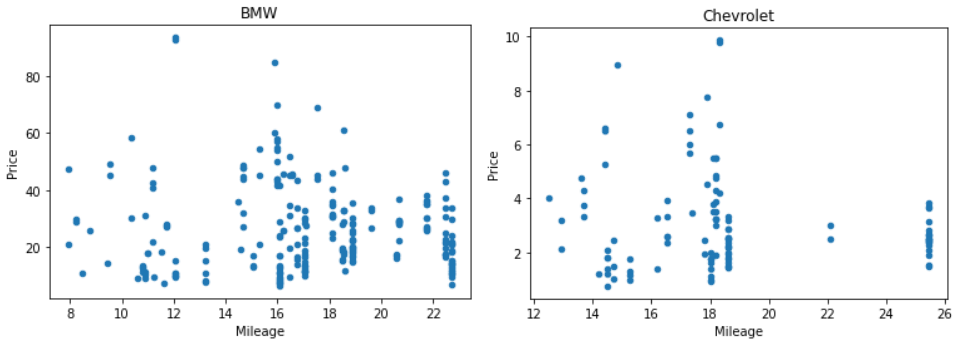

Fig. BMW and Chevrolet Price and Mileage trends

Analysing Linearity in the Dataset

To completely analyse the complete data, various visualisations were created for bivariate analysis between the important features of the used cars and the Normalized Price. On some features like Engine, Power and Year, a fair degree of linearity was visible.

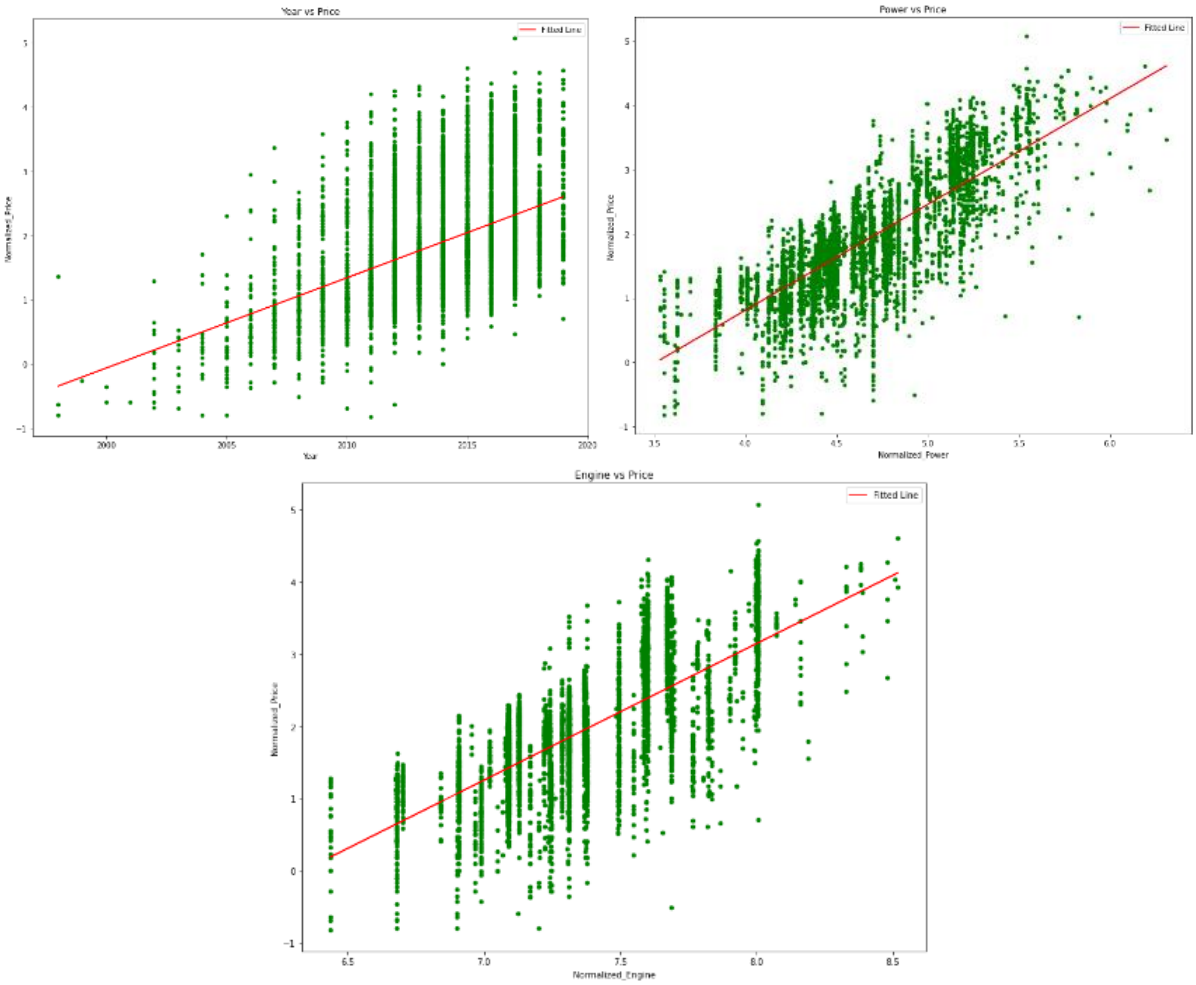

Fig. Scatter Plots to determine trends between important features.

Methodology

We primarily used Multiple Linear Regression to analyse the price of the used cars using features such as Year, Power and Engine of the cars (as a linear trend is clearly visible during Exploratory Data Analysis). Multiple linear regression is a statistical modeling technique used to analyze the relationship between multiple independent variables and a single dependent variable. It extends the concept of simple linear regression, which only considers one independent variable, to incorporate multiple predictors simultaneously.

The model was created by importing the necessary libraries, creating 2 data frames x and y with independent and dependent variables, respectively. After that, training and test splits were created. The data was divided in 80:20 ratio – 80% for training set and 20% for test set.

The coefficients for the dependent variables were calculated and it was found that Power, Engine and Year generated positive coefficients thus showing positive relationship between the variables.

On the other hand, using the coefficients for Mileage led to a negative value, thus establishing negative relationship between the price and the mileage of the used cars. Taking the latter into account, it was decided that the dependent features should be the ones with positive coefficients so as to get the best accuracy for the model.

So, the model was trained using the training set and then the model was tested on the unseen data simulating real world situations. On visualizing the trend between the actual and predicted values, it was clear that a linear trend could be observed thus displaying the accuracy of the model.

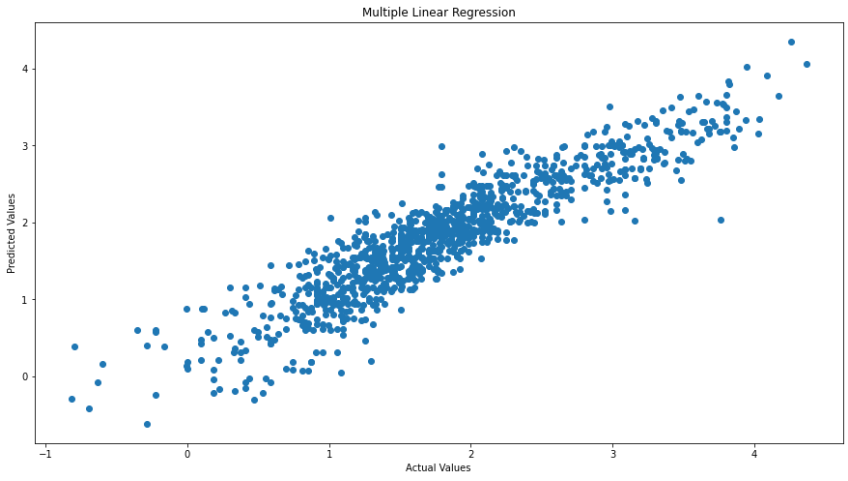

Fig. Plot depicting trend between actual and predicted values.

Result

The results of the model have been quantified in terms of the R^2 score, although we have also used Mean Absolute Error and Mean Squared Error to further show the relationship between the variables.

Type of Metric used	Score on testing data
R^2 Score	84%
Mean Absolute Error	0.26
Mean Squared Error	0.34

Findings

Based on the analysis of the Used Car dataset, the mileage of the car is not a good predictor of the price. There are many other factors that affect the price of a car, such as the make and model, the year of manufacture, the condition of the car, and the demand for that particular car. During the EDA, it was clear by observing the Scatter plot of Mileage vs Price that Mileage alone cannot be considered a good predictor for valuing the Price of used cars.

It is clear that there is no clear relationship between the two variables. There are cars with high mileage that are expensive, and there are cars with low mileage that are cheap. One reason why mileage is not a good predictor of price is that it is not the only factor that affects fuel efficiency. The make and model of the car, the weight of the car, and the type of fuel that the car uses all affect fuel efficiency.

Another reason why mileage is not a good predictor of price is that the price of a car can depreciate over time, even if the car has low mileage. This is because the car will become less valuable as it gets older, even if it is in good condition.

Form our early analysis of the dataset, we find high collinearity between the features such as Power of Engine, Size of Engine and the Year of Manufacturing. So, these features were used as independent variables during The Regression analysis resulting in 84% accuracy of the model prediction thus stemming their importance as important variables for analysis of used car prices. So, it is clear that when buying used cars, customers tend to focus more on Engine Power, Size and they Year of Manufacturing.

Limitations

Some of the limitations that were faced during the analysis of this dataset were:

- The data provided was missing values on many of the columns and was also filled with null values. To overcome this, 2 columns were dropped completely and many were required to be filled with values comparing to the statistical methods such as mean and median.
- Many of the columns' data was highly skewed, so the particular data values had to undergo feature scaling using log normalization. Also outliers had to be handled in order for accurate results.
- The values for some brands like Isuzu, Force, Fiat, etc. have so little data that it is almost insignificant to analyse the particular data. So in future, it would be better if a lot more data is collected, especially use data from various other dealers too to better understand customer – dealer relationship and selling trends of cars.

Conclusion

- The analysis has shown that although mileage seems as a significant predictor of sale prices for used cars for some brands, other factors such as car age, brand, condition, and additional features also play a more significant role in determining sale prices. Considering these factors can further refine the pricing strategy and better meet customer preferences.
- The multiple linear regression model provides a reliable framework for understanding the relationships between predictor variables and sale prices. It allows for data-driven pricing decisions and improved profitability for the used car dealership.
- Gathering a more extensive dataset encompassing various car brands, models, and features can enhance the accuracy and robustness of the regression model. By including a larger number of data points, a better understanding of customer preferences can be achieved, leading to improved precision in pricing decisions.
- Incorporating market trends and economic factors that influence car prices, such as inflation, fuel prices, and industry-specific events, can enable the pricing strategy to adapt to changing market dynamics. This consideration will ensure that the pricing strategy remains relevant and aligned with the current market conditions.
- Actively monitoring customer feedback and reviews can provide valuable insights into their perception of pricing. By analyzing this feedback, the dealership can identify areas for improvement and make adjustments to the pricing strategy accordingly. This customer-centric approach aims to enhance customer satisfaction and generate positive reviews, thus strengthening the dealership's reputation in the market.

By considering these future steps, the used car dealership can continue to refine its pricing strategy, improve customer satisfaction, and establish itself as a prominent player in the market.

References

1. N. Monburinon, P. Chertchom, T. Kaewkiriya, S. Rungpheung, S. Buya and P. Boonpou, "Prediction of prices for used car by using regression models," 2018 5th International Conference on Business and Industrial Research (ICBIR), Bangkok, 2018, pp. 115-119.
2. API Reference — scikit-learn 1.2.2 documentation: Scikit-learn: Machine Learning in Python, Pedregosa et al., JMLR 12, pp. 2825-2830, 2011.
3. McKinney, W. (2018). Python for data analysis: Data wrangling with Pandas, NumPy, and IPython. O'Reilly Media.
4. Hastie, T., Tibshirani, R., & Friedman, J. (2009). The elements of statistical learning: Data mining, inference, and prediction. Springer.