How Do Prices Vary from Economy to Luxury?

Analysis of Used Car Prices

Abstract:

The automotive sector has experienced significant growth in the past decade, with increased automobile manufacturing and sales expanding the used car market substantially. Addressing this growth, this project develops a robust price prediction model using the R programming language, incorporating advanced machine learning techniques such as linear regression and random forest. Unlike prior studies that relied on simpler models, this approach thoroughly analyzes a broad array of factors affecting car prices, including make, model, year, mileage, fuel type, and transmission, across a substantial dataset. The models are designed to enhance market transparency by aiding investors, customers, and businesses in making informed decisions. Moreover, the project addresses challenges like data quality, categorical variable handling, missing data, and model scalability. The goal is to provide a comprehensive system that simplifies car valuation, promoting a more transparent and predictable marketplace for used cars.

Introduction:

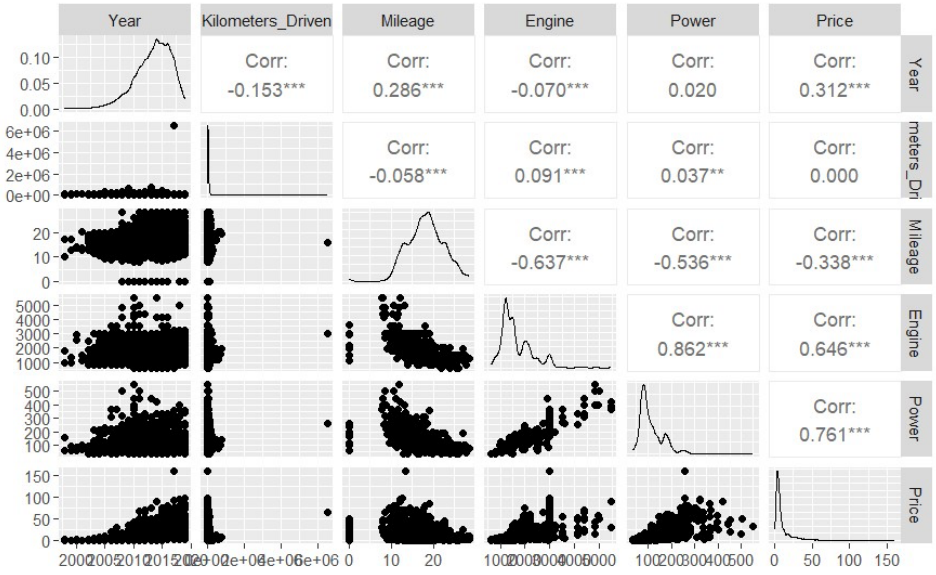
The project delves into the used car market in India, aiming to elucidate how factors like brand, model age, and usage influence vehicle depreciation. This analysis seeks to clarify pricing dynamics within the used car market, making it more navigable for buyers and sellers alike. Employing a dataset encompassing a broad spectrum of 6019 vehicles with variables such as location, year of the car model, kilometres driven, and price, this study leverages statistical methods, including multiple regression techniques and random forest models, to uncover the determinants of used car prices. By dissecting price distribution across various car segments, the study offers insights into the depreciation trends that impact car valuations, facilitating more informed decision-making for all market participants. This report will also touch upon the challenges encountered in the analysis, such as data completeness and model accuracy, thus ensuring a robust approach to understanding market trends.

Name	Location	Year	Kilometers	Fuel_Type	Transmissi	Owner_Ty	Mileage	Engine	Power	Seats	Price
Maruti Wagon R LXI CNG	Mumbai	2010	72000	CNG	Manual	First	26.6 km/kg	998 CC	58.16 bhp	5	1.75
Hyundai Creta 1.6 CRDi SX Option	Pune	2015	41000	Diesel	Manual	First	19.67 kmpl	1582 CC	126.2 bhp	5	12.5
Honda Jazz V	Chennai	2011	46000	Petrol	Manual	First	18.2 kmpl	1199 CC	88.7 bhp	5	4.5
Maruti Ertiga VDI	Chennai	2012	87000	Diesel	Manual	First	20.77 kmpl	1248 CC	88.76 bhp	7	6
Audi A4 New 2.0 TDI Multitronic	Coimbatore	2013	40670	Diesel	Automatic	Second	15.2 kmpl	1968 CC	140.8 bhp	5	17.74
Hyundai EON LPG Era Plus Option	Hyderabad	2012	75000	LPG	Manual	First	21.1 km/kg	814 CC	55.2 bhp	5	2.35
Nissan Micra Diesel XV	Jaipur	2013	86999	Diesel	Manual	First	23.08 kmpl	1461 CC	63.1 bhp	5	3.5
Toyota Innova Crysta 2.8 GX AT 8S	Mumbai	2016	36000	Diesel	Automatic	First	11.36 kmpl	2755 CC	171.5 bhp	8	17.5

Exploratory Data Analysis:

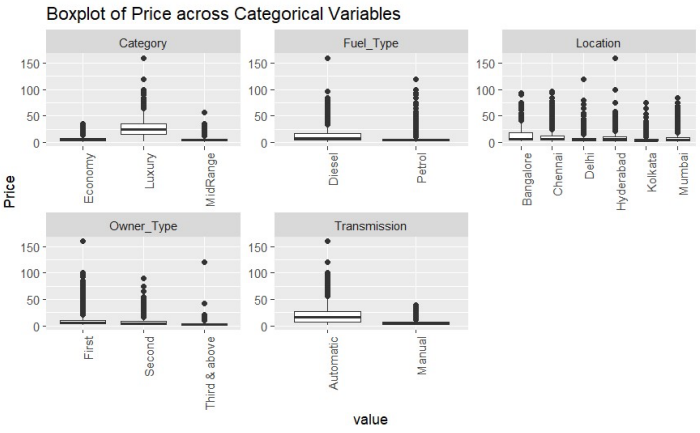
Before conducting exploratory data analysis on the used car dataset, several preprocessing steps were undertaken to refine the data for accurate modelling. Initially, the units for the engine size, power output, and mileage were standardized by removing textual units to ensure uniformity

across these measurements. Additionally, car brand names were extracted from the full vehicle names, and these brands were manually categorized into segments of economy, mid-range, and luxury based on perceived market positioning. To simplify the analysis, vehicles powered by CNG and LPG were excluded from the dataset, and multiple locations were consolidated into broader categories to reduce the complexity of geographical influences. Finally, the dataset was split into distinct training and testing sets to validate the effectiveness of the predictive models developed in subsequent stages. This meticulous data cleaning and preparation set the stage for a robust analysis aimed at understanding price variations in the used car market.

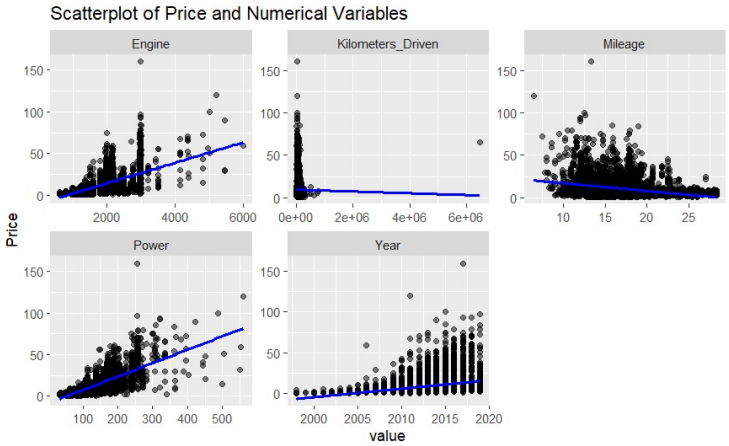


The above scatterplot matrix with histograms and correlation coefficients indicates varying degrees of linear relationships among the year, mileage, engine size, power, and price of cars. For instance, engine size and power have a strong positive correlation, while mileage negatively correlates with engine size, suggesting larger engines are less fuel-efficient. The histograms show that variables like Kilometers Driven and Price are right-skewed, indicating the presence of high-value outliers. Log transformations would likely be beneficial for Price due to its skewness, Kilometers Driven to compress the range and normalize the distribution, and potentially for Engine and Power if their relationship with price is nonlinear or if their distributions are skewed. These transformations can help stabilize variance, achieve normality, and improve model performance when these variables are used as predictors in regression models.

Categorical Variables :

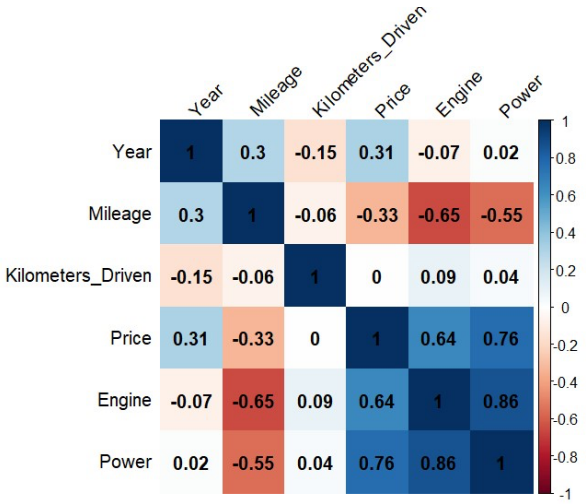


The boxplot shows the price distribution across different categories in the data. Luxury cars have a wider price range and higher median prices compared to economy and midrange categories, highlighting the premium placed on luxury cars. Also, diesel cars generally have higher prices than petrol cars. Manual cars tend to be cheaper than automatic cars, and first-owner cars fetch higher prices, possibly due to better-assumed conditions and lower risk.



Numerical Variables - The scatterplot shows that engine size and power display a positive correlation with price, indicating that cars with larger engines and higher power outputs are priced higher. The year of manufacture also shows a positive trend, with newer cars tending to be more expensive. Kilometers driven does not show a strong linear relationship with price, suggesting that other factors may influence price more than the car's usage.

Correlation - The strong positive correlation between engine and power (0.86) and a moderately positive correlation between power and price (0.76), both of which are logical given that a more powerful engine typically commands a higher price. There is a notable negative correlation between engine and mileage (-0.65), and between power and mileage (0.55), suggesting that more powerful and larger engines tend to be less fuelefficient.



Interactions - Some possible interactions that can be observed from the EDA are; the interaction between log(Engine) and Year captures how the value of engine size depreciates over time, affecting the price. The negative correlation between Mileage and both Engine and Power indicates that fuel efficiency impacts price differently across engine sizes and power outputs, justifying the Mileage by log(Kilometers_Driven) interaction. Lastly, the Category by Fuel_Type interaction can explain how the impact of fuel type on price varies across different car segments, capturing nuances in how luxury and economy categories are valued differently based on fuel type. These interactions provide a thorough understanding of the dynamic interplay between car features and their combined impact on the price in the used car market.

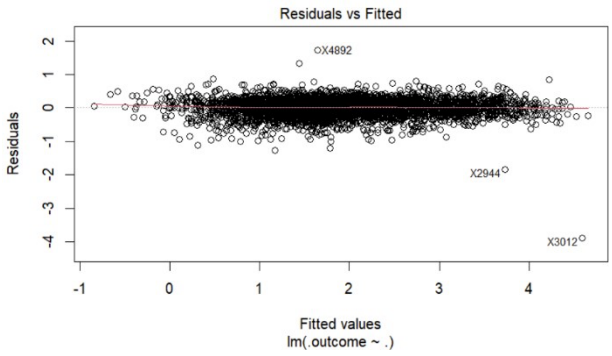
Linear Regression:

By taking all the trends and interactions from EDA, the final linear regression model can be formulated as follows:

$$log(Price) = \beta_0 + \beta_1(Category) + \beta_2(Location) + \beta_3(Transmission) + \beta_4(log(Engine)) + \beta_5(Year) + \beta_6(log(Power)) + \beta_7(Mileage) + \beta_8(log(Kilometers\_Driven)) + \beta_9(Category * Fuel\_Type) + \beta_{10}(log(Engine) * Year) + \beta_{11}(log(Power) * Year) + \beta_{12}(log(Engine) * log(Power)) + \beta_{13}(Year * Mileage) + \beta_{14}(log(Kilometers\_Driven) * Year) + \beta_{15}(log(Kilometers\_Driven) * Mileage) + \varepsilon$$

Even after applying backward selection, weighted linear regression and crossvalidation the model's parameters and outcomes remain unchanged. Hence, we will consider the cross-validation model which is validated on 1000 folds of training data. The final training results are:

```
Linear Regression

4935 samples
   9 predictor

No pre-processing
Resampling: Cross-Validated (1000 fold)
Summary of sample sizes: 4931, 4929, 4929, 4929,
Resampling results:

  RMSE       Rsquared   MAE
  0.2256722  0.9319338  0.1845008
```




```
Coefficients:
                                    Estimate Std. Error t value Pr(>|t|)
(Intercept)                       -1.928e+02  7.819e+01  -2.466 0.013684 *
CategoryLuxury                     5.254e-01  1.745e-02  30.112  < 2e-16 ***
CategoryMidRange                  -6.408e-02  1.204e-02  -5.325 1.06e-07 ***
LocationChennai                   -1.143e-01  1.597e-02  -7.157 9.45e-13 ***
LocationDelhi                     -1.853e-01  1.696e-02 -10.921  < 2e-16 ***
LocationHyderabad                  1.088e-02  1.773e-02   0.614 0.539437
LocationKolkata                   -3.741e-01  1.883e-02 -19.868  < 2e-16 ***
LocationMumbai                    -1.876e-01  1.600e-02 -11.726  < 2e-16 ***
TransmissionManual                -1.124e-01  1.152e-02  -9.756  < 2e-16 ***
`log(Engine)`                      9.543e+01  1.369e+01   6.969 3.61e-12 ***
Year                               9.432e-02  3.886e-02   2.427 0.015261 *
`log(Power)`                      -1.262e+02  1.046e+01 -12.062  < 2e-16 ***
Mileage                            2.958e+00  7.698e-01   3.842 0.000124 ***
`log(Kilometers_Driven)`          -2.028e+01  3.410e+00  -5.948 2.91e-09 ***
Fuel_TypePetrol                   -2.034e-01  1.397e-02 -14.560  < 2e-16 ***
`log(Engine):Year`                -4.730e-02  6.807e-03  -6.949 4.16e-12 ***
`Year:log(Power)`                  6.281e-02  5.191e-03  12.099  < 2e-16 ***
`log(Engine):log(Power)`           6.078e-02  2.127e-02   2.857 0.004295 **
`Year:Mileage`                    -1.443e-03  3.798e-04  -3.800 0.000147 ***
`Year:log(Kilometers_Driven)`      1.008e-02  1.695e-03   5.949 2.88e-09 ***
`Mileage:log(Kilometers_Driven)`  -5.388e-03  1.367e-03  -3.940 8.25e-05 ***
`CategoryLuxury:Fuel_TypePetrol`   8.230e-02  2.719e-02   3.027 0.002480 **
`CategoryMidRange:Fuel_TypePetrol` 8.443e-02  1.632e-02   5.173 2.39e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2466 on 4912 degrees of freedom
Multiple R-squared:  0.9191,    Adjusted R-squared:  0.9188
F-statistic:  2538 on 22 and 4912 DF,  p-value: < 2.2e-16
```

The final Regression equation is, log(Price) = -192.8 + 0.5254*CategoryLuxury - 0.06408*CategoryMidRange – 0.1143*LocationChennai – 0.1853*LocationDelhi + 0.01088*LocationHyderabad – 0.3741*LocationKolkata - 0.1876*LocationMumbai – 0.1124*TransmissionManual + 95.43*log(engine) + 0.0943*Year – 126.2*log(power) + 2.958*Mileage – 20.28*log(Kilometers_Driven) – 0.2034 * Fuel_TypePetrol + (significant interaction terms) + ε

Interpretation - Some example interactions are for every 1 per cent change in engine capacity, the price of the car changes by 95.4%; for every 1 unit change in a year, the price of the car changes by 9.89%; moving from economy category to luxury category, the price of car increases by 69.1%; The prices of used cars from location banglore to location delhi decreased by 17.57%, which is due to evidence of higher tax in Karnataka (i.e., the state in which Bangalore city is located) than in Delhi and to location Chennai increased by 12.10%, which was due to supply and demand in the Chennai.

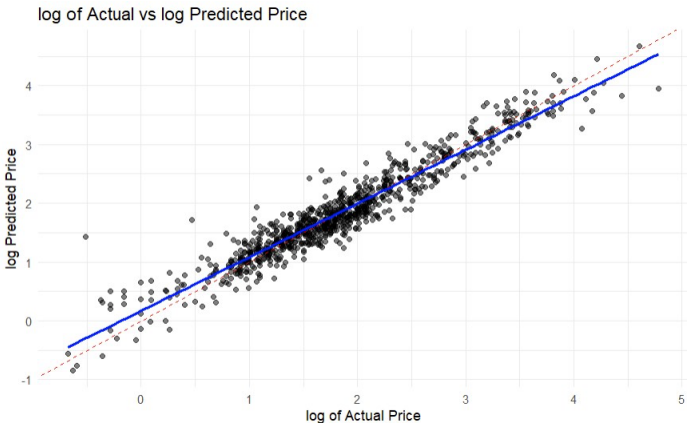
Outliers - We can identify some outliers in the above residuals vs fitted values plot. Below is the justification for some of these outliers.

1) An outlier from Chennai is an economy category car from 2015 with over 100,000 kilometres driven, which is unusually high indicating extensive use in a short period, such as heavy commercial use or long-distance travel, setting it apart from typical personal use vehicles.
2) Another one is a MidRange vehicle from Bangalore, 2015, that has a recorded mileage of 0. This could be a data processing error or the vehicle's mileage was not available, which would make it an outlier in terms of data completeness.

3) A luxury car from Kolkata, in 2012, recorded a very low odometer reading of just 7,000 kilometres, paired with a high-power output of 226.6 bhp. The low usage and high performance might suggest a premium or sports model that is rarely driven. It also has a high price of 35 Lakhs (3.5 million), which could be justified by its low mileage and high performance, indicating that it's a high-value, wellmaintained vehicle (an Audi TT 40).

Predictions – These are the testing results obtained from the model:

The points clustered along the blue line suggest that for many observations, the model's predictions are close to the actual values. However, as the log of actual prices increases, the spread of points suggests that the model's predictions become less precise. This might be due to higher variability in the higher price range or model limitations in capturing the price dynamics at the upper end of the market.



The testing RMSE and MAE values are:

```
RMSE: 0.2492986
MAE: 0.186164
```

Random Forest:

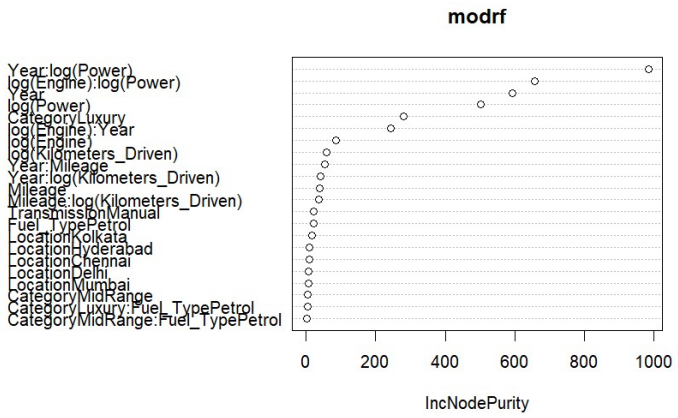
A Random Forest is also implemented to the same model equation and results are obtained as follows:

The model explained about 94.17% of the variability in data the training MSE is 0.0436 and the training RMSE is 0.2088833.

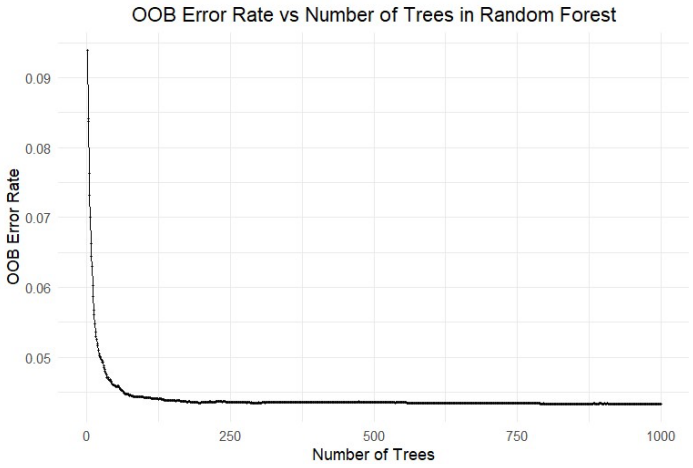
```
               Type of random forest: regressior
                     Number of trees: 250
No. of variables tried at each split: 10

          Mean of squared residuals: 0.04363222
                    % Var explained: 94.17
```

The variable importance plot is as follows:

The most important variables are Year, log(power), log(engine), and luxury category excluding interaction terms.



To hyper-tune the random forest algorithm we will plot the Out of bag error rate vs the of trees. The desired value obtained from the plot is 250 trees.



The testing results are as follows:

The points are quite tightly clustered around the line, which suggests that the model has a good level of predictive accuracy. However, there seems to be a slight trend where the model underpredicts the lower log prices and overpredicts the higher log prices, indicated by the points deviating from the line at both ends. This pattern could suggest that the model may be less accurate at the extremes of the price range.

The testing MSE is 0.0425 and the RMSE is 0.206.

Discussion:

The diagnostic plots for the linear regression model reveal a random scatter and no significant autocorrelation in the ACF plot concluding a strong linear relationship between the log of actual and predicted prices with some outliers. Overall, these diagnostics indicate a reasonable model fit.

The Random Forest model demonstrates strong predictive performance with key predictors like Year: log(Power) and log(Engine): log(Power) showing significant importance. The log actual vs. log predicted prices are closely clustered around the diagonal, indicating accurate predictions. The model's OOB error rate stabilizes, suggesting an optimal number of trees and efficient performance without the need for further complexity.

	Metric	Linear_Regression	Random_Forest
MSE	MSE	0.0614197	0.0419461
RMSE	RMSE	0.2478300	0.2048075
MAE	MAE	0.1854551	0.1465463
R2	R2	0.9168807	0.9436703

The Random Forest model surpasses the Linear Regression in predictive accuracy, with lower MSE, RMSE, and MAE values, and a higher R-squared, indicating accurate predictions around log actual values and a better capture of variance, thus handling complex patterns.

While the variable importance plot indicates that year, engine size, and power are key predictors of used car prices, the category, especially luxury versus economy, still influences value. Generally, luxury vehicles command higher prices due to superior features and brand prestige, which can be intertwined with the model's highlighted features, like power. The category's impact might be indirect, as luxury attributes are reflected in correlated variables, such as engine specifications. Thus, the category does affect pricing, but its role may be embedded within other significant predictors identified by the model.

References:

The project title, "How Do Prices Vary from Economy to Luxury?" remains pertinent, as it encapsulates the essence of how vehicle classification, including distinctions such as luxury, potentially impacts pricing, even if this effect is mediated through other variables in your model. Although your model highlights other variables as more significant, with an R2 score of 0.94 indicating a very strong predictive capability, it surpasses Sharma and Sharma's[1] R2 score of 0.86. This implies that while the category variable isn't a direct dominant factor in your model, it indirectly enhances the model by influencing key features like engine size and power, which are indicative of a car's luxury status and thus affect its pricing.

Reference-1 Reference-2