"How Do Prices Vary from Economy to Luxury? Analysis of Used Car Prices"

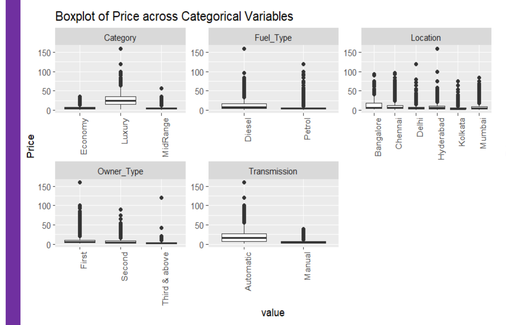
Sri Vaishno Sai Suren Prattipati, Sri Sai Yaswanth Reddy Rajoli
University at Albany, State University of New York

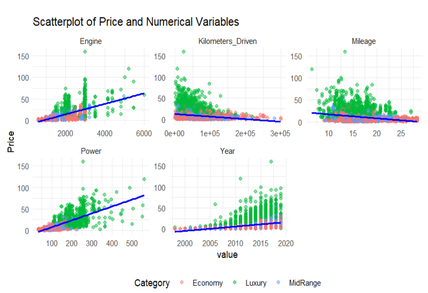
Motivation

The project delves into the used car market, aiming to elucidate how factors like brand, model age, and usage influence vehicle depreciation. This endeavor seeks to clarify pricing dynamics within the pre-owned market, making it more navigable for buyers and sellers alike. 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.

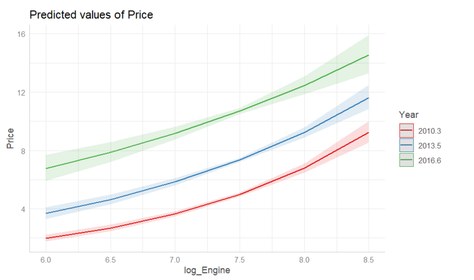
Employing a dataset encompassing a broad spectrum of vehicles, this analysis leverages statistical methods to uncover the determinants of used car prices in India. The dataset includes variables such as car make, year of manufacture, mileage, and selling price, providing a solid basis for exploring price trends. The findings aim to streamline the pre-owned car market, aiding buyers and sellers in understanding the complexities of car valuation and contributing to a more transparent and predictable buying & selling environment.

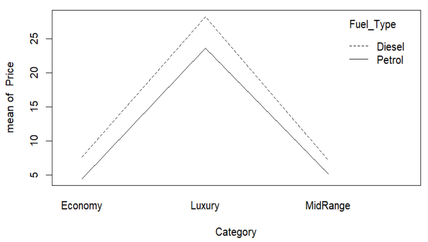
EDA



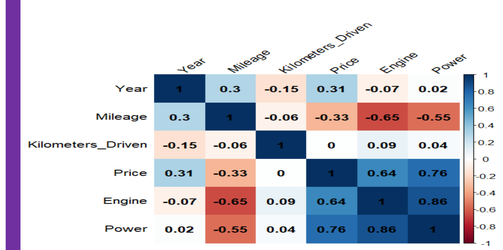


From the EDA, we observe that luxury cars, particularly diesel and automatic ones, fetch higher prices. Newer models with powerful engines and less usage, indicated by lower kilometers driven, also command premium pricing. While mileage shows a nuanced effect on price, first-owner cars are preferred, highlighting condition importance. Location shows less impact, with some exceptions of high outliers in certain cities. Overall, the data reflects a clear influence of car features and history on pricing.





The first plot demonstrates that cars with larger engines generally cost more, and this effect increases with newer models. The second plot indicates that diesel luxury cars are the most expensive, with fuel type influencing luxury car prices more than economy or mid-range cars.



Based on the correlations, we will consider variables and any interactions between them when building the model.

Model Building

After many experiments, the final model chosen is,

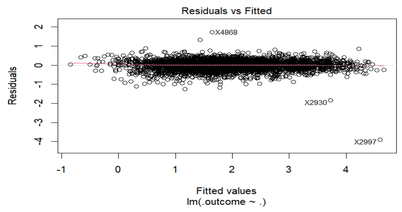
```
log(Price) ~ Category + Location + Transmission + log(Engine) *
    Year + log(Power) * Year + Engine * Power + Year * Mileage +
    log(Kilometers_Driven) * Year + log(Kilometers_Driven):Mileage +
    Category * Fuel_Type
```

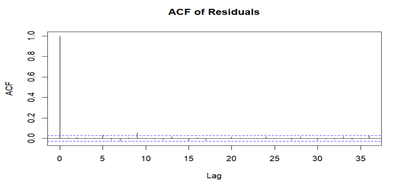
From the correlation plot, we consider interactions with high significance and log transformations for independent variables with skewed distribution.

Two models i.e., Linear Regression and Random Forest are trained with same data and model formula.

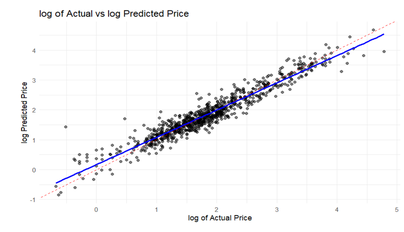
The transformed linear regression equation is undergone through backward selection, which is then validated with cross validation. The training results are:

```
Coefficients:
                                  Estimate Std. Error t value Pr(>|t|)
(Intercept)                     -2.110e+02  8.137e+01  -2.594 0.009527 **
CategoryLuxury                   5.172e-01  1.756e-02  29.456  < 2e-16 ***
CategoryMidRange                -6.413e-02  1.203e-02  -5.330 1.03e-07 ***
LocationChennai                 -1.161e-01  1.605e-02  -7.229 5.60e-13 ***
LocationDelhi                   -1.868e-01  1.704e-02 -10.967  < 2e-16 ***
LocationHyderabad                1.050e-02  1.781e-02   0.590 0.555335
LocationKolkata                 -3.756e-01  1.890e-02 -19.875  < 2e-16 ***
LocationMumbai                  -1.893e-01  1.607e-02 -11.777  < 2e-16 ***
TransmissionManual              -1.118e-01  1.150e-02  -9.723  < 2e-16 ***
`log(Engine)`                    9.925e+01  1.385e+01   7.165 8.92e-13 ***
Year                             1.035e-01  4.044e-02   2.558 0.010543 *
`log(Power)`                    -1.289e+02  1.047e+01 -12.317  < 2e-16 ***
Mileage                          3.021e+00  8.282e-01   3.647 0.000268 ***
`log(Kilometers_Driven)`        -2.013e+01  3.426e+00  -5.877 4.47e-09 ***
Fuel_TypePetrol                 -2.075e+01  1.423e-02 -14.588  < 2e-16 ***
`log(Engine):Year`              -4.922e-02  6.885e-03  -7.148 1.01e-12 ***
`Year:log(Power)`                6.417e-02  5.196e-03  12.351  < 2e-16 ***
`log(Engine):log(Power)`         6.510e-02  2.132e-02   3.053 0.002275 **
`Year:Mileage`                  -1.474e-03  4.088e-04  -3.607 0.000313 ***
`Year:log(Kilometers_Driven)`    1.001e-02  1.703e-03   5.878 4.42e-09 ***
`Mileage:log(Kilometers_Driven)` -5.478e-03  1.435e-03  -3.819 0.000136 ***
`CategoryLuxury:Fuel_TypePetrol` 8.836e-02  2.728e-02   3.239 0.001209 **
`CategoryMidRange:Fuel_TypePetrol` 8.580e-02 1.632e-02   5.259 1.51e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2461 on 4888 degrees of freedom
Multiple R-squared:  0.9188,    Adjusted R-squared:  0.9184
F-statistic:  2513 on 22 and 4888 DF,  p-value: < 2.2e-16
```





The testing results (in log scales):

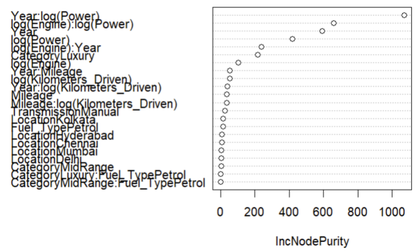


RMSE: 0.24783
MAE: 0.1854551

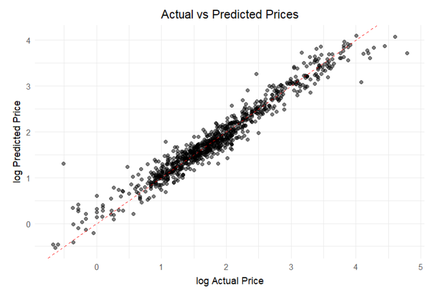
The second model is Random Forest with 500 trees. The Training results are:

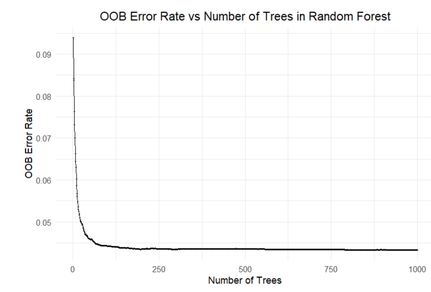
```
randomForest(x = model.matrix(bmod)[, -1], y = log(df$Price),
                     Type of random forest: regression
                           Number of trees: 500
No. of variables tried at each split: 10

          Mean of squared residuals: 0.0436518
                    % Var explained: 94.11
```

The Variable importance plot is as follows:



The testing results are (in log scales:





Discussion

The diagnostic plots for the linear regression model reveal a random scatter and no significant autocorrelation in 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	0.0614197	0.0419461
RMSE	0.2478300	0.2048075
MAE	0.1854551	0.1465463
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