UC San Diego
RADY SCHOOL OF MANAGEMENT

Capstone Project: Used Car Price Prediction

Affaan Mustafa
MIT Applied Data Science Bootcamp
15 April 2022

Why is this an important problem to solve?

- The used car market in India is growing rapidly, and there is a huge demand for used cars.
- By developing an accurate pricing model for used cars, Cars4U, a tech start-up, can better serve its customers, help sellers set competitive prices, and facilitate transactions in the pre-owned car market.
- An effective pricing model will enable the company to devise profitable strategies using differential pricing and gain a competitive edge in the market.

What key questions need to be answered?

- What are the most significant factors that influence the price of a used car?
- How do these factors interact with each other, and how do they affect the price?
- Which machine learning models and techniques are most effective for predicting used car prices?
- How can we optimize the performance of the chosen model to increase prediction accuracy?
- How can the insights from the pricing model be translated into actionable strategies for Cars4U?

Data Dictionary

S.No. : Serial Number | Name : Name of the car which includes Brand name and Model name

Location : The location in which the car is being sold or is available for purchase (Cities)

Year : Manufacturing year of the car

Kilometers_driven : The total kilometers driven in the car by the previous owner(s) in KM

Fuel_Type : The type of fuel used by the car (Petrol, Diesel, Electric, CNG, LPG)

Transmission : The type of transmission used by the car (Automatic / Manual)

Owner : Type of ownership

Mileage : The standard mileage offered by the car company in kmpl or km/kg

Engine : The displacement volume of the engine in CC

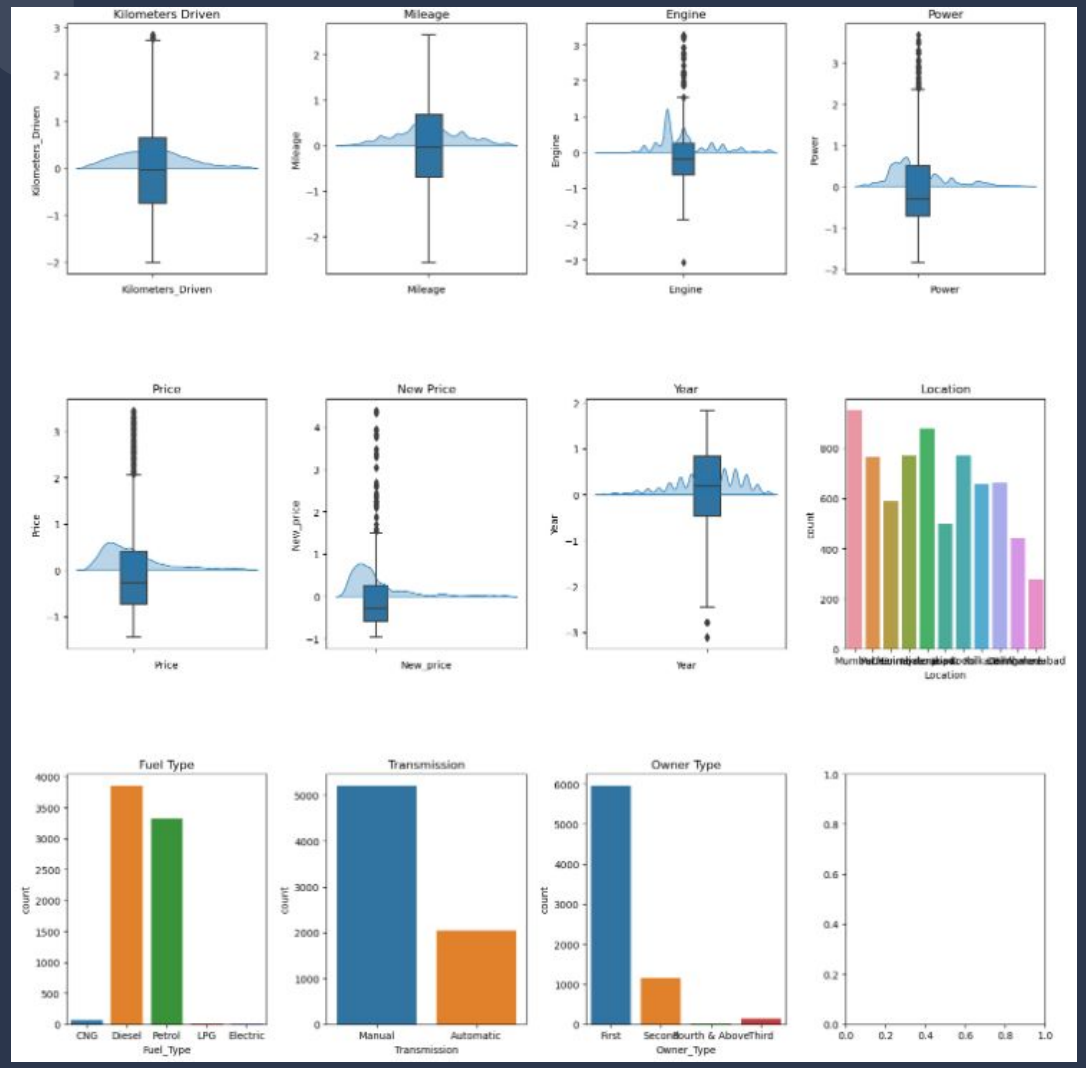
Power : The maximum power of the engine in bhp

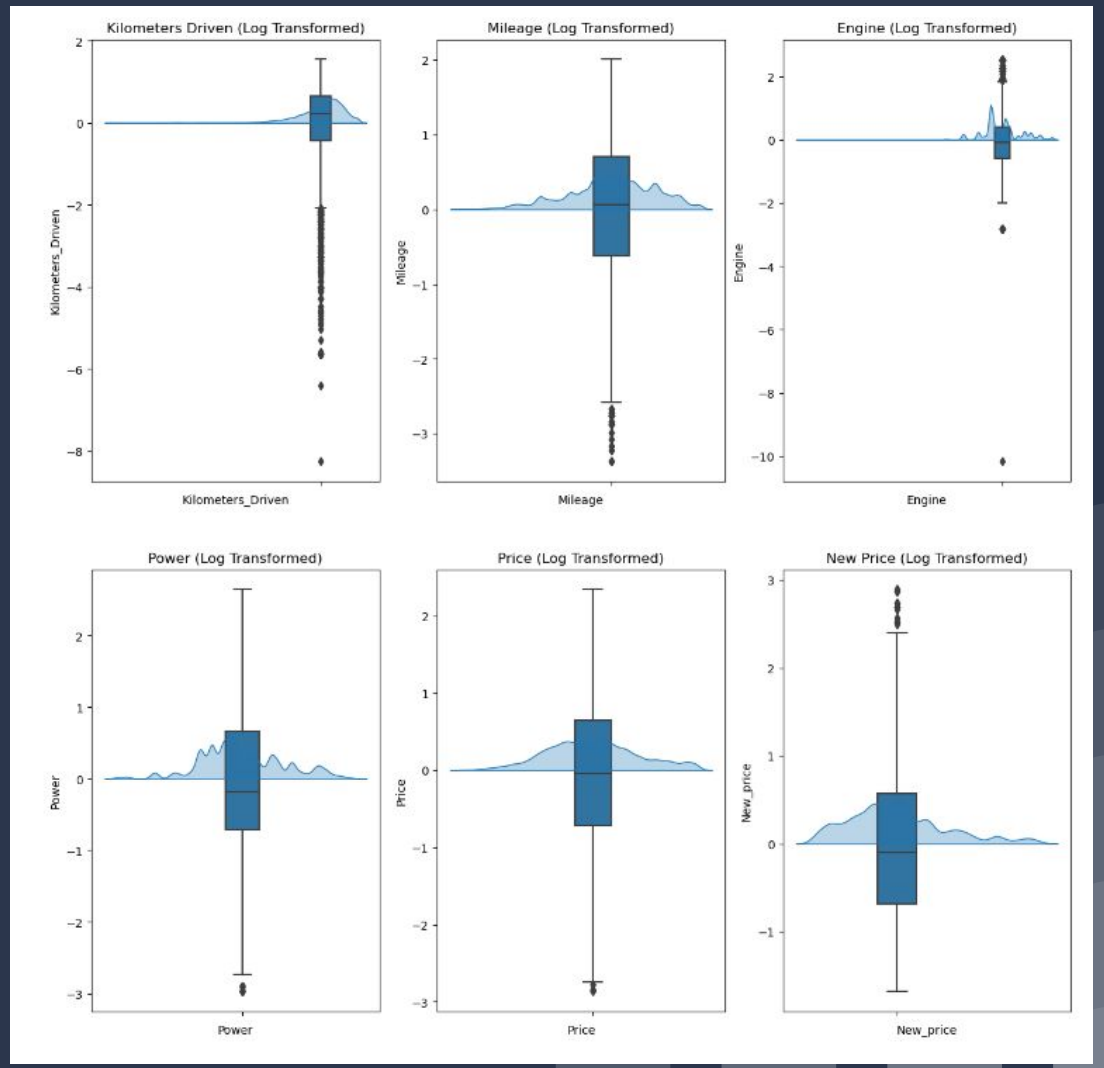
Seats : The number of seats in the car

New_Price : The price of a new car of the same model in INR 100,000

Price : The price of the used car in INR 100,000 (Target Variable)

Summary Statistics and Data Distribution





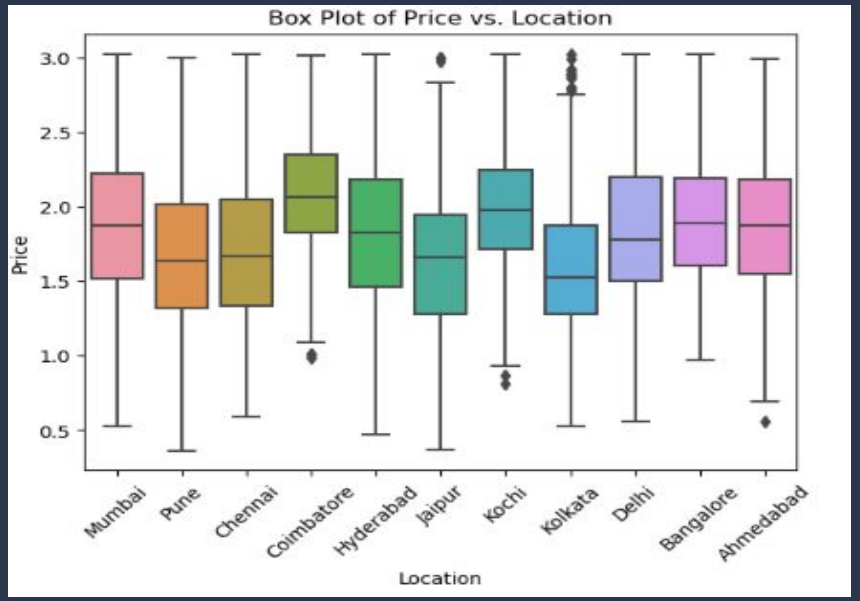
Bivariate Analysis and Preprocessing

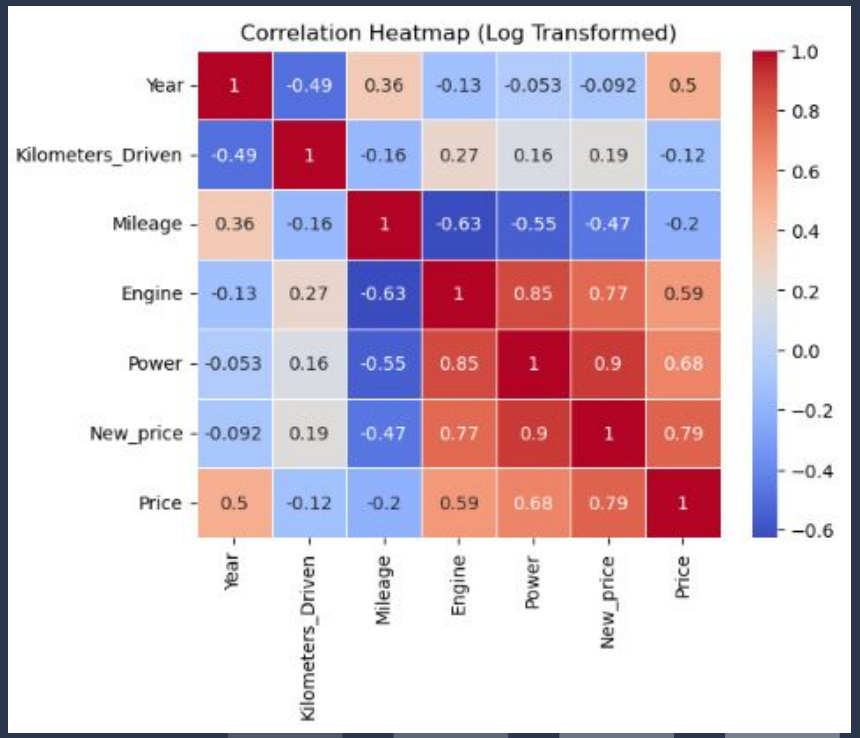
Further Feature Engineering and Data Preprocessing:

- Drop Name Column and Replace With Brand
- Apply Missing Value Treatment for Price, New Price, Power, Engine Mileage
  - Use KNN Imputation
  - Use Median Imputation

  Compare

Why IQR +- 1.5 for Outliers and not 3 Standard Deviations?





Model Comparison and Performance

Imputation Methods Compared:

- K - Nearest Neighbor
- Median Imputation

Algorithms Tested:

1) Linear Regression
2) Ridge / Lasso Regression
3) Decision Trees
4) Random Forest

Metric of Success:

- Lowest Mean Squared Error

- Training/Test Split of 75/25

```
Mean Squared Error for KNN imputed dataset using Linear Regression: 0.045270013058797486
Mean Squared Error for Median imputed dataset using Linear Regression: 0.044591836264260665
Mean Squared Error for KNN imputed dataset using Lasso Regression: 0.08929199548756574
Mean Squared Error for Median imputed dataset using Lasso Regression: 0.08869289505189333
Mean Squared Error for KNN imputed dataset using Ridge Regression: 0.06268257658035975
Mean Squared Error for Median imputed dataset using Ridge Regression: 0.06230539651489355
Mean Squared Error for KNN imputed dataset using Decision Tree: 0.0467363267746721
Mean Squared Error for Median imputed dataset using Decision Tree: 0.045105206015033594
Mean Squared Error for KNN imputed dataset using Random Forest Regression: 0.027068464286509192
Mean Squared Error for Median imputed dataset using Random Forest Regression: 0.025854380816171568
```

Random Forest Feature Importances:

	Feature	Importance
0	Power	0.534970
1	Year	0.355928
2	Engine	0.058799
3	Mileage	0.016932
4	Kilometers_Driven	0.007716
5	Transmission_Manual	0.006142

Key Findings and Insights

- Evaluation:

  Median imputation with Random Forest Regression technique has the lowest MSE of 0.0258 on the median imputed dataset.

  The performance of the imputation techniques (KNN vs. Median) seems to have a small effect on the MSE.

  Random Forest Regression generally outperforms other regression techniques (Linear, Lasso, Ridge, Decision Tree) on both KNN and Median imputed datasets.

- Tradeoff between Computational Exhaustion and Lowering the MSE

Further Improvements:

Trying other imputation techniques such as missForest, MICE, or Bayesian imputation.
Trying other regression techniques such as Gradient Boosting or Neural Networks.
Conducting further feature engineering to select relevant features and reduce noise in the dataset.
Increasing the size of the dataset if possible.
Combining multiple models using ensembling techniques such as Bagging or Boosting to improve prediction accuracy.

Business Recommendations and Next Steps

- The key actionable for stakeholders include:

  Using the proposed solution to develop pricing strategies that optimize profit and customer satisfaction.

  Supporting sellers in setting appropriate prices for their used cars. Helping buyers make well-informed purchasing decisions in the used car market.

  The expected benefits of implementing the proposed solution are accurate predictions Used Car Prices, which can help Cars4U in making informed decisions related to buying and selling used cars.

  The costs associated with implementing the solution include the time and resources required for feature engineering, ensembling, and further analysis. This includes compute complexity and if possible the cloud credits needed to run models like this. This balance between model accuracy and overhead costs will need to be optimized.

- The key risks and challenges include the potential for overfitting the model, which can lead to inaccurate predictions and poor decision-making. Additionally, the accuracy of the model may be affected by factors such as changes in the used car market or the availability of new data.

Conclusion

The used car market in India is growing rapidly, and an accurate pricing model is essential to help customers, sellers, and the tech start-up Cars4U make informed decisions in the pre-owned car market.

Through the development of a predictive pricing model using median imputation with Random Forest Regression, we were able to accurately estimate the price of used cars based on various factors such as mileage, brand, model, year, location, and more.

Our analysis revealed the most significant factors that influence the price of a used car, such as new price, power, year, engine, and kilometers driven.

Our proposed solution design using median imputation with Random Forest Regression, along with feature engineering and ensembling techniques, can provide accurate predictions for the missing values in the Price column and help in making informed decisions related to buying and selling used cars.

To further improve the performance of the model, some possible approaches could be trying other imputation and regression techniques, conducting further feature engineering, and increasing the size of the dataset if possible.

The proposed solution can give Cars4U a competitive edge in the market and help in developing profitable strategies using differential pricing.

Thank you for your time and attention 🙂