Project Report: Predicting Car Prices

Kota Sai Varun <121910301003@gitam.in>
Jyothi Prakash <121910310017@gitam.in>

Abstract

Determining the price of a used car is a challenging task, due to the many factors that drive a used vehicle's price on the market. The focus of this project is to develop a machine learning model that can accurately predict the price of a car based on its features, in order to make informed purchases. We implemented a Linear Regression model on a dataset consisting of information of used cars. Conventional linear regression yielded satisfactory results..

Motivation

Deciding whether a used car is worth the posted price when you see listings online can be difficult. Several factors, including mileage, selling price, fuel type, year, etc. can influence the actual worth of a car. From the perspective of a seller, it is also a dilemma to price a used car appropriately. Based on existing data, the aim is to use machine learning algorithms to develop models for predicting used car prices.

Dataset

For this project, we are using the dataset on used cars listed on www.cardekho.com available on Kaggle. The features available in this dataset are Name, Year, Selling Price, Kilometers driven, fuel, transmission and owner.

Pre-processing

In order to get a better understanding of the data, we did feature engineering for the variables and plotted the boxplot of the data. We noticed that the dataset had many outliers, primarily due to large price sensitivity of used cars.

```
In [3]: df.head()
```

Out[3]:

	Car_Name	Year	Selling_Price	Present_Price	Kms_Driven	Fuel_Type	Seller_Type	Transmission	Owner
0	ritz	2014	3.35	5.59	27000	Petrol	Dealer	Manual	0
1	sx4	2013	4.75	9.54	43000	Diesel	Dealer	Manual	0
2	ciaz	2017	7.25	9.85	6900	Petrol	Dealer	Manual	0
3	wagon r	2011	2.85	4.15	5200	Petrol	Dealer	Manual	0
4	swift	2014	4.60	6.87	42450	Diesel	Dealer	Manual	0

```
In [8]: df.drop(['Year','Current_Year'] ,axis=True,inplace=True)
        df.head()
```

Out[8]:

	Selling_Price	Present_Price	Kms_Driven	Fuel_Type	Seller_Type	Transmission	Owner	Total_Years
0	3.35	5.59	27000	Petrol	Dealer	Manual	0	6
1	4.75	9.54	43000	Diesel	Dealer	Manual	0	7
2	7.25	9.85	6900	Petrol	Dealer	Manual	0	3
3	2.85	4.15	5200	Petrol	Dealer	Manual	0	9
4	4.60	6.87	42450	Diesel	Dealer	Manual	0	6

Fig : (Top) Dataframe of raw data. (Bottom) Dataframe after feature engineering

Certain features such as car names were dropped during training as these were unique to each vehicle, thus adding no value to the training process.

Data Visualization:

1. Box Plot:

A box plot (or box-and-whisker plot) shows the distribution of quantitative data in a way that facilitates comparisons between variables or across levels of a categorical variable. The box shows the quartiles of the dataset while the whiskers extend to show the rest of the distribution, except for points that are determined to be “outliers” using a method that is a function of the inter-quartile range.

Example:

```
seaborn.boxplot(x = "Fuel_Type",y = "Selling_Price")
```

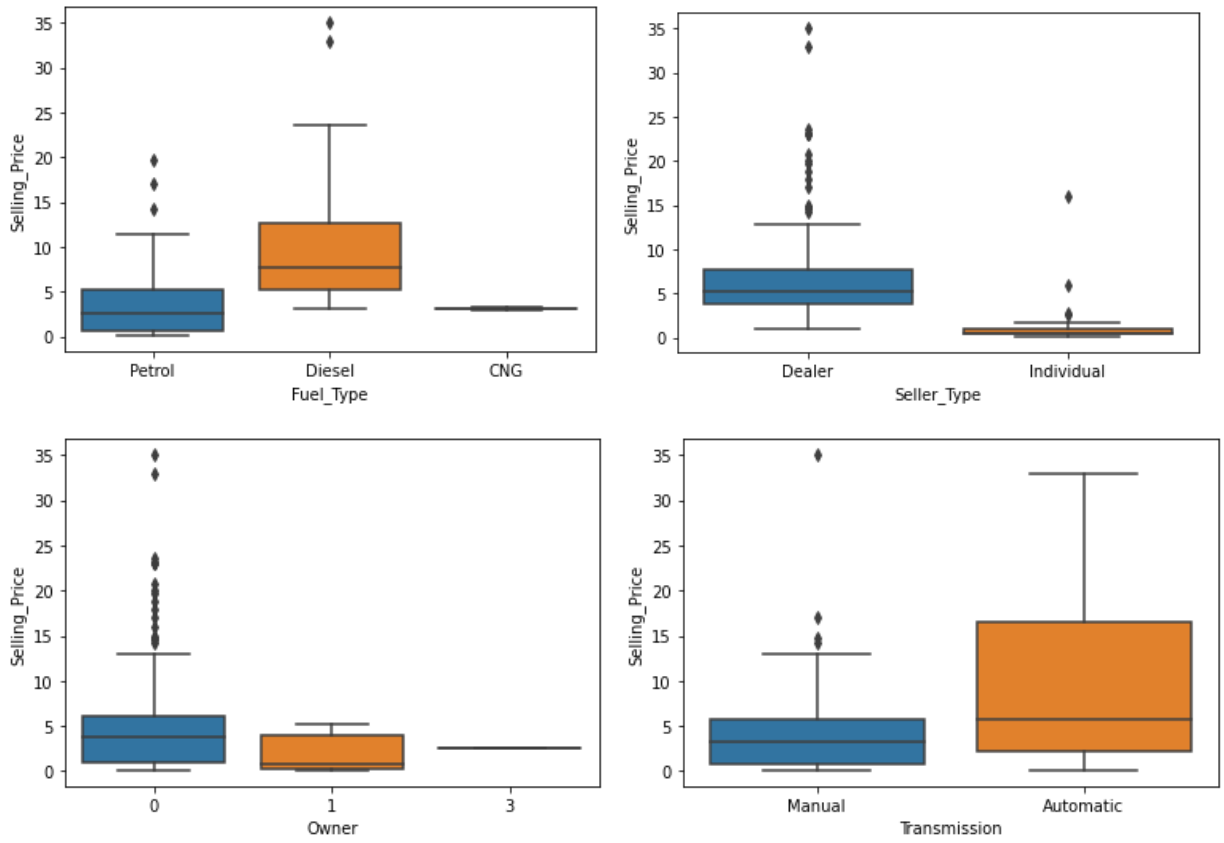


Fig: (Top-Left) Selling Price vs Fuel_Type . (Top-Right) Selling_Price vs Seller_Type
(Bottom-Left) Selling_Price vs Owner (Bottom-Right) Selling_Price vs Transmission

Analysing Linearity in Dataset

To analyze the degree to which our features are linearly related to price, we plotted joint plots and pair plots. There seemed to be a fair degree of linearity for these features.

Joint Plot:

Draw a plot of two variables with bivariate and univariate graphs.

This function provides a convenient interface to the JointGrid class, with several canned plot kinds. This is intended to be a fairly lightweight wrapper; if you need more flexibility, you should use JointGrid directly.

Example:

$$seaborn.jointplot(x = "Total\_Year", y = "Selling\_Price")$$

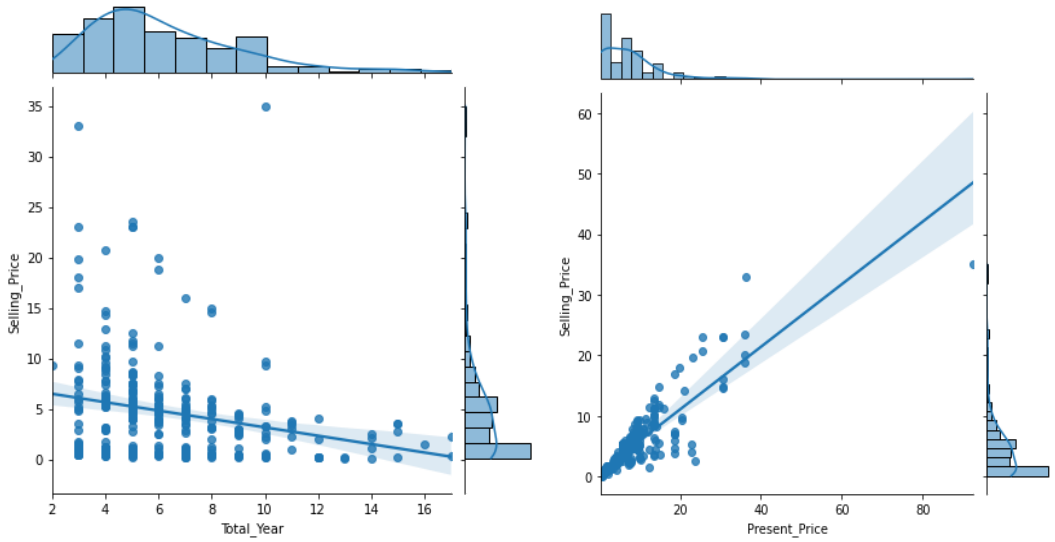


Fig: (Left) Selling Price vs Year joint plot. (Right) Selling Price vs Current Price joint plot.

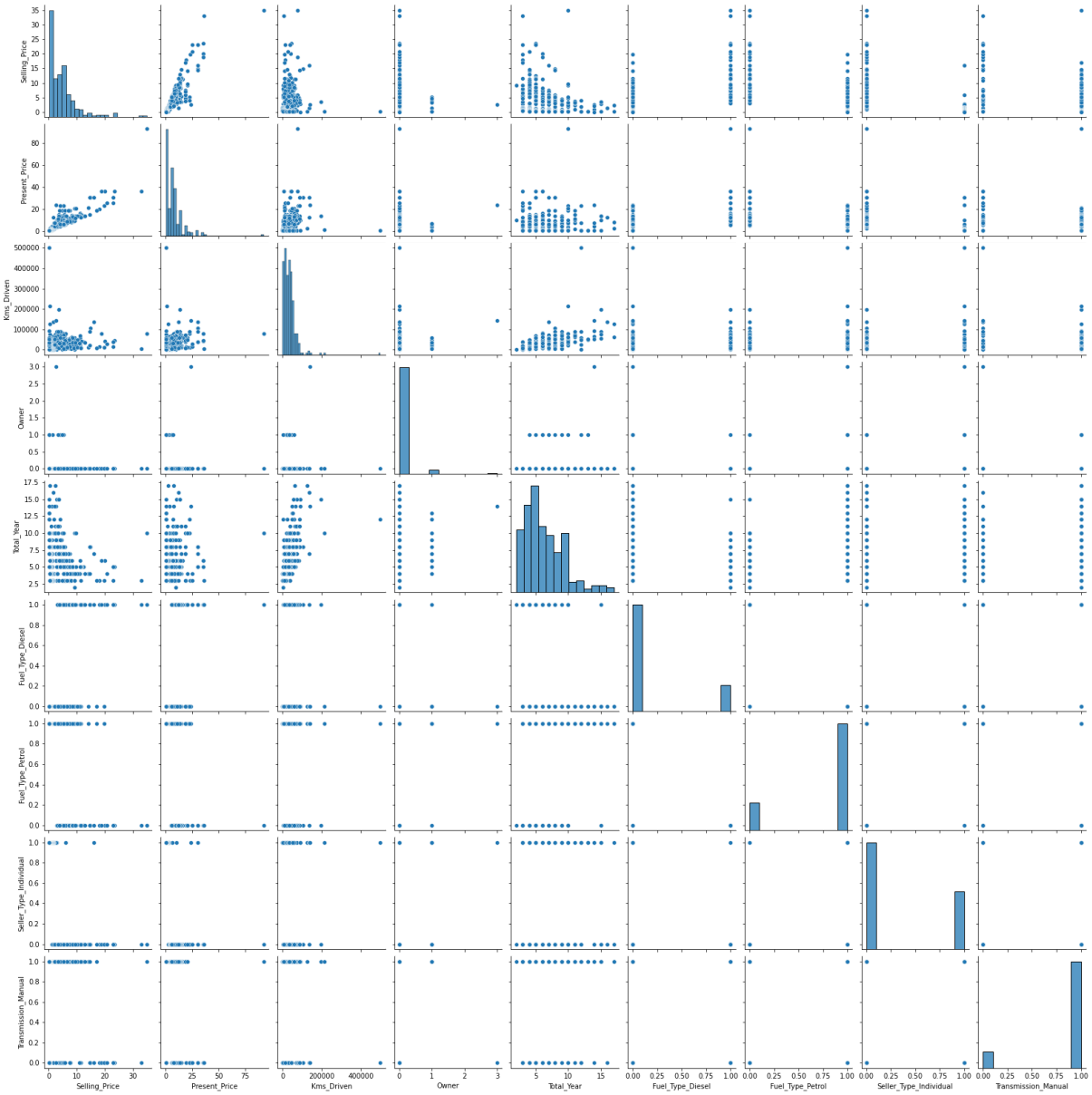
Outcome:

1. From the first plot we can come to a conclusion that the total number of years a car is inversely proportional to the selling price.
   i.e., If the car is old then the selling price will also decrease.

2. From the second plot we can come to a conclusion that present price is directly proportional to the selling price.
   i.e., If the car present price in showroom is more than the sell price

Pairwise Plot:

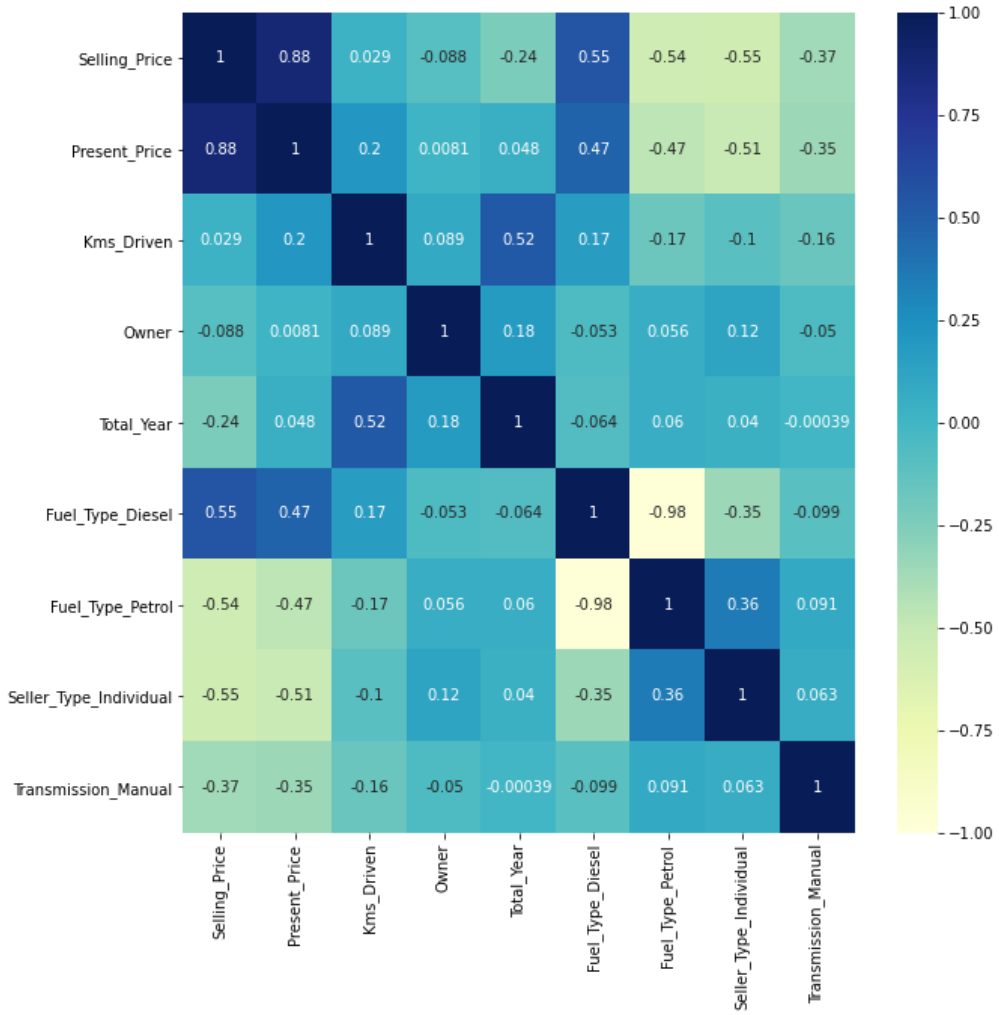
Plot pairwise relationships in a dataset.
By default, this function will create a grid of Axes such that each numeric variable in data will be shared across the y-axes across a single row and the x-axes across a single column. The diagonal plots are treated differently: a univariate distribution plot is drawn to show the marginal distribution of the data in each column.



Heatmap Plot:

Plot rectangular data as a color-encoded matrix.

This is an Axes-level function and will draw the heatmap into the currently-active Axes if none is provided to the ax argument. Part of this Axes space will be taken and used to plot a colormap, unless cbar is False or a separate Axes is provided to cbar_ax.



Outcome:

1. We can find that Present_price is highly correlated with the Selling_price.

i.e., A positive change in the Present_price will show a significant positive impact on the Selling_price

2. We can find that Fuel_Type_Petrol is negative correlated with the Selling_price

i.e., Fuel_Type_Petrol is showing a negative impact on the Selling_price

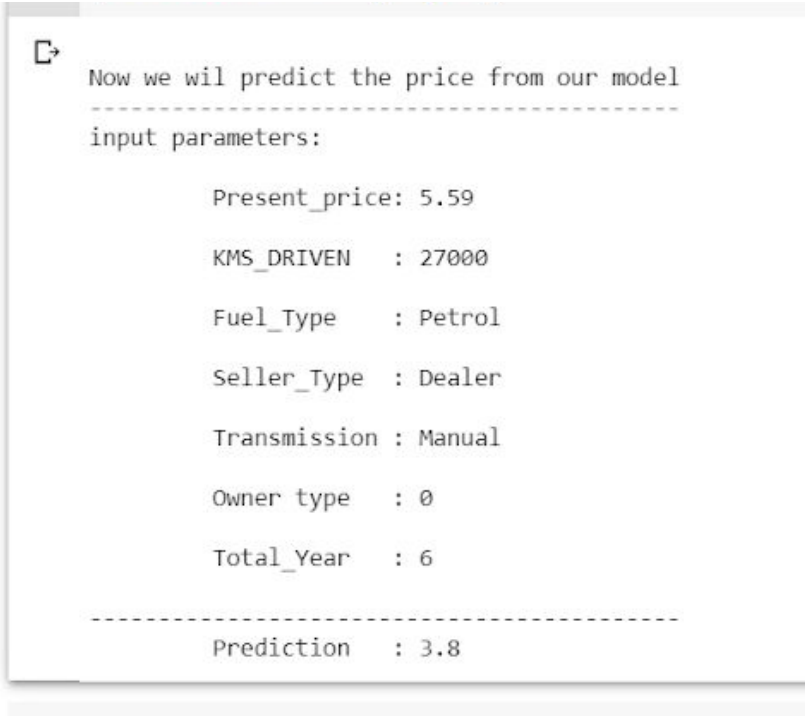
Methodology

Multiple Linear Regression
Multiple Linear Regression was chosen as the first model due to its simplicity and comparatively small training time.

Results

The Final Result of the model is shown below

For the Given input The output value is: $model.predict([[5.59, 27000, 0, 6, 0, 1, 0, 1]])$



Future Work:

For better performance, we plan to remove the outliers, use adaptive learning rates and better models.

Contributions

All team members contributed equally towards this project.
Link to project repository : https://github.com/SIG-ML-adv/car-price-predictionm-cardekho

References

1. https://www.kaggle.com/nehalbirla/vehicle-dataset-from-cardekho
2. www.cardekho.com