Car Price Prediction Project

Submitted by:

Neha

ACKNOWLEDGEMENT

The sources that helped me out in the completion of this project are, I usually take reference from my personal notes, internet sources and some websites and all of these sources helped me out for the completion of this project.

Introduction

Business problem Framing

The demand for used cars has certainly increased from pre-pandemic times. The used vehicle market is facing supply constraints due to three factors: customers are holding on to their used vehicles and not selling them as they were doing before the pandemic, exchanges have been impacted due to continued challenges in the new car market, and repossessions have virtually stopped since April due to the ongoing loan moratorium. Due to supply constraints and unpredictable nature of the lockdown, sales were impacted during the pandemic period.

“The used car market is a sunrise industry and whenever there is an economic downturn people tend to gravitate towards used cars.”

This mismatch between demand and supply has led to a 6-7 per cent increase in the buying price of used cars, while the selling price has increased by a bit more.

Conceptual Background of the Problem

The concept behind the problem is, With the covid 19 impact in the market, we have seen lot of changes in the car market. Now some cars are in demand hence making them costly and some are not in demand hence cheaper. One of our clients works with small traders, who sell used cars. With the change in market due to covid 19 impact, our client is facing problems with their previous car price valuation machine learning models. So, we are preparing new car price valuation models from new data of used cars.

For this we should know, which variables are important to predict the price of cars, how do these variables describe the price of the cars.

Review of Literature

India is one of the world's fastest growing automobile markets and is poised to become the third largest passenger's car market in 2020. The recorded sales

growth of 4 wheelers like passenger car & utility vehicle has also risen up to 7.87 % and 6.25% respectively.

45 variables identified here after extensive literature review and this factor may be taken care by the dealer salesperson to showcase the car to customers and highlighting these benefits for acceptance and deal closure. The purchases of 4-wheelers accommodate all stages of buying process.

Considering the above factors Sales of pre-owned cars may increase in the aftermath of the covid-19 pandemic as commuters will likely prefer private conveyance but financial constraints may hamper purchases of new vehicles, according to a survey conducted by Cars 24.

The automotive and mobility industries have certainly been among the hardest hit during the COVID-19 pandemic. The picture is improving, however. Car dealerships are getting busier, and many are eagerly seeking more inventory to sell. Overall mobility is picking up steadily, although not to pre-COVID-19 levels. Usage of shared mobility services and public transit is picking up significantly, while regions where many commuters and their employers accept the practicalities of working from home are recovering more slowly.

Now, as economies return to some semblance of normal, automotive OEMs, car dealers, and government officials need to know how long full recovery may take and what the “next normal” could look like.

To help answer such questions, we are continuing to regularly survey consumers in the United States, the United Kingdom, Germany, France, Italy, Japan, and China on their mobility behaviors and plans around car buying and servicing. Our survey looks at both current consumer sentiment and anticipated future behavior as economies find the next normal.

Car buying and servicing

Globally, consumers' intent to purchase cars is close to pre-COVID-19 levels, fueled by positive outlooks.

- Intent to purchase new and used cars over the next 12 months is almost back to pre-COVID-19 levels (new cars at 94% versus pre-COVID-19 levels and up by 7% over September 2020; used cars at 97% versus pre-COVID-19 levels, up by 1% compared to September 2020).
- There are significant increases in purchase intent for EVs, particularly in Europe and China, motivated by government incentives and by increased consciousness about sustainability.
- Prospective buyers are less inclined to want to interact with sellers at car dealerships. That decline in preference is falling across all regions and age groups especially for consumers between 55 and 70 years of age, who now consider online buying as a relevant alternative to visiting dealers.
- Interest in buying cars entirely online remains flat at 59% globally with regional variation.

- The outlook for aftermarket services continues to improve. In the last few months, more customers have been getting maintenance and repairs done

rather than waiting. The next month shows significant uptake in net intent.

Mobility

- About 51% of global respondents still state that they intend to travel less than before the COVID-19 pandemic. However, mobility is picking up gradually and at different rates, with the fastest recovery in the United States.
- Regular use of public transport has picked up significantly compared to late 2020. Shared modes of transport (especially micro mobility services) are now above pre-COVID-19 levels.
- Public transport and shared mobility modes are considered more or less safe again with regard to COVID-19 infection.
- The frequency of commuting trips is recovering at different rates. Around the world, expectations differ about commuting patterns and workplace scenarios in the next normal.

- Respondents are largely in favor of greener mobility infrastructure. Almost half (49%) indicate that

current green initiatives should be amplified and accelerated.

Motivation for the Problem Undertaken

My only objective behind this project is to use data Analytics and predict the car prices and help the car traders to increase their overall revenue, profits, by focusing on changing trends in car sales and purchases after covid –19.

Thus, we can accordingly manipulate the strategy of the firm that will yield high returns.

Analytical Problem Framing

Mathematical/Analytical Modeling of the Problem

Mathematical /Analytical functions and methods are used such as df.shape () to check out the number of rows and columns in

the dataset. df.dtypes() is used to check out the data type of the columns we have in the dataset. df.isnull() is used to check whether null values are present or not, we have also checked it

by using Heatmap from seaborn. And the other one is checking out the statistical summary of the dataset by using df.describe() method. We got lots of fresh information from statistical summary is like:

1. We can see that in the count function many columns have different values, which clearly means null values are present in the columns.

2. The difference in mean and median is almost similar.

3.From the count we can see that we have no missing values.

4.We can see that we have small difference in 75% percentile and max in columns like Name, Model, Km_driven,Discount, Discount_price, Sale_Price.

Data Sources and their Formats

We have scraped the data from multiple CAR selling websites the saved the data in a csv file. Then we have imported the dataset, in the dataset we have 5340 rows and 10 columns. We can see the snapshot of the dataset below:

	Unnamed: 0	Name	Model	Transmission	Km_driven	Owner	Fuel	Discount	Discount_price	Sale_price
0	0	2020 Renault Kwid	1.0 RXT Opt AT Automatic	Automatic	3,252	1st Owner	Petrol	₹5,000	4,31,699	4,36,699
1	1	2013 Maruti Alto 800	LXI Manual	Manual	13,807	1st Owner	Petrol	₹11,000	2,16,899	2,27,899
2	2	2020 Maruti New Wagon-R	VXI 1.0 Manual	Manual	3,865	1st Owner	Petrol	₹5,000	5,16,399	5,21,399
3	3	2019 Maruti Swift	VXI Manual	Manual	3,362	1st Owner	Petrol	₹5,000	6,20,699	6,25,699
4	4	2020 Maruti Baleno	SIGMA 1.2 K12 Manual	Manual	5,645	1st Owner	Petrol	₹6,000	5,75,299	5,81,299
...	...	...	...	...	...	...	...	...	...	...
5335	5335	2016 Maruti Wagon R 1.0	VXI	Manual	33,463	1st Owner	Petrol	₹6,000	3,88,699	3,94,699
5336	5336	2015 Hyundai Eon	SPORTZ	Manual	15,346	3rd Owner	Petrol	₹20,000	3,01,299	3,21,299
5337	5337	2012 Maruti Wagon R 1.0	VXI	Manual	38,553	1st Owner	Petrol	₹12,000	3,15,199	3,27,199
5338	5338	2018 Tata Tiago	XT 1.2 REVOTRON	Manual	11,253	1st Owner	Petrol	₹7,000	4,86,799	4,93,799
5339	5339	2017 Maruti Alto 800	VXI	Manual	21,239	1st Owner	Petrol	₹37,000	3,20,999	3,57,999

5340 rows × 10 columns

Data Preprocessing Done

Following are the steps followed for data cleaning:

Missing Values if we have missing values then treat them by using suitable method, but we have no missing data in our dataset.

Label Encoding, we have used label encoding on the columns which were in string format.

Removing the outliers, we have few outliers present in the dataset, we can remove these outliers by using zscore or IQR, but as we have few outliers, we can ignore them.

Removing the highly correlated columns, we have from the correlation matrix that few columns are highly negatively correlated, thus we have dropped those columns.

Removing Skewness, we have observed that skewness is present in many columns, Soo we have removed the skewness by using power transform method.

Data Inputs-Logic–Output Relationships

The input columns of the data frame express the features provided in the car, so that by the help of these features we can predict the Sale price of the car. The price of the house completely depends on the features.

Hardware and Software Requirements and Tools used

The hardware device used is only the pc. In software devices so many applications and libraries are used to for the completion of the project. I have used Jupyter software for using jupyter notebook. The libraries used are Selenium, pandas, NumPy, matplotlib, seaborn these libraries are used for analysis purpose. And from sklearn power transform is used for removing skewness.

Model/s Development and Evaluation

- **Identification of possible problem-solving approaches (methods)**

Methods are used such as df.shape () to check out the number of rows and columns in

the dataset. df.dtypes() is used to check out the data type of the columns we have in the dataset. df.isnull() is used to check whether null values are present or not, we have also checked it by using Heatmap from seaborn. And the other one is checking out the statistical summary of the dataset by using df.describe() method. We got lots of fresh information from statistical summary is like:

1. We can see that in the count function many columns have different values, which clearly means null values are present in the columns.

2. The difference in mean and median is almost similar.

3.From the count we can see that we have no missing values.

4.We can see that we have small difference in 75% percentile and max in columns like Name, Model, Km_driven,Discount, Discount_price, Sale_Price.

- **Testing of identified Approaches**

We have used different regression models listed below:

LinearRegression()

DecisionTreeRegressor()

SVR()

KNeighborsRegressor()

Lasso(alpha=0.0001)

Ridge(alpha=0.0001)

Run and Evaluate Selected models

```
model=[lm,dtr,svr,knr,ls,rd]

for m in model:
    m.fit(x_train,y_train)
    pred=m.predict(x_test)
    print('error:',m)
    print('Mean absolute error:',mean_absolute_error(y_test,pred))
    print('Mean squared error:',mean_squared_error(y_test,pred))
    print('Root Mean Squared error:',np.sqrt(mean_squared_error(y_t
```

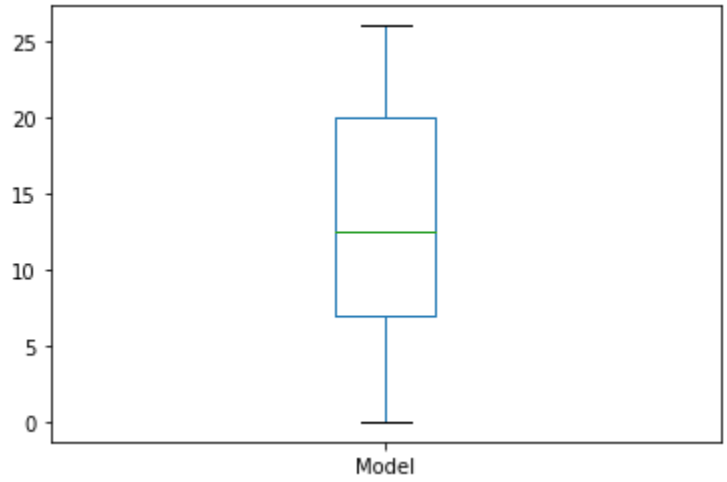
These are the models used to evaluate.

- **Key Metrics for success in solving problem under consideration**

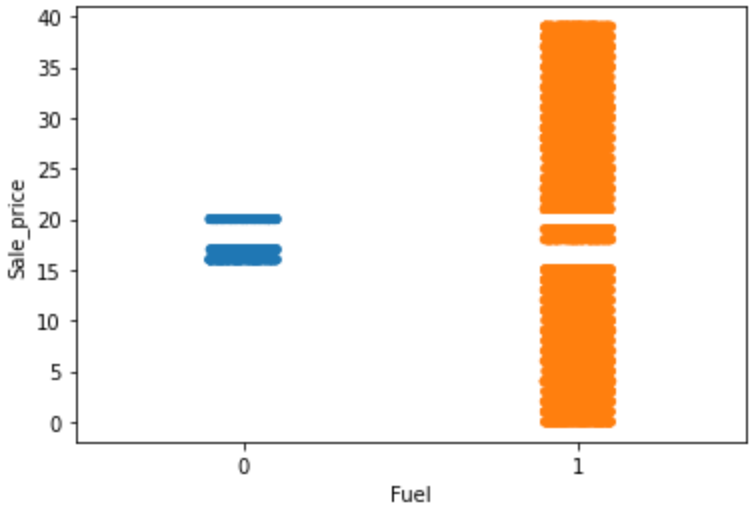
It's a regression problem thus we have used r2_score, mean absolute error mean squared error, and root means squared error as the evaluation metrics.

- **Visualization**

Using Boxplot, we have used a boxplot for each of the columns. Here we have a boxplot of model column in this plot, we can see that we have a maximum value of 25 and a minimum of 0. Where Q1, Q3, and median is lying at 6, 20, 13

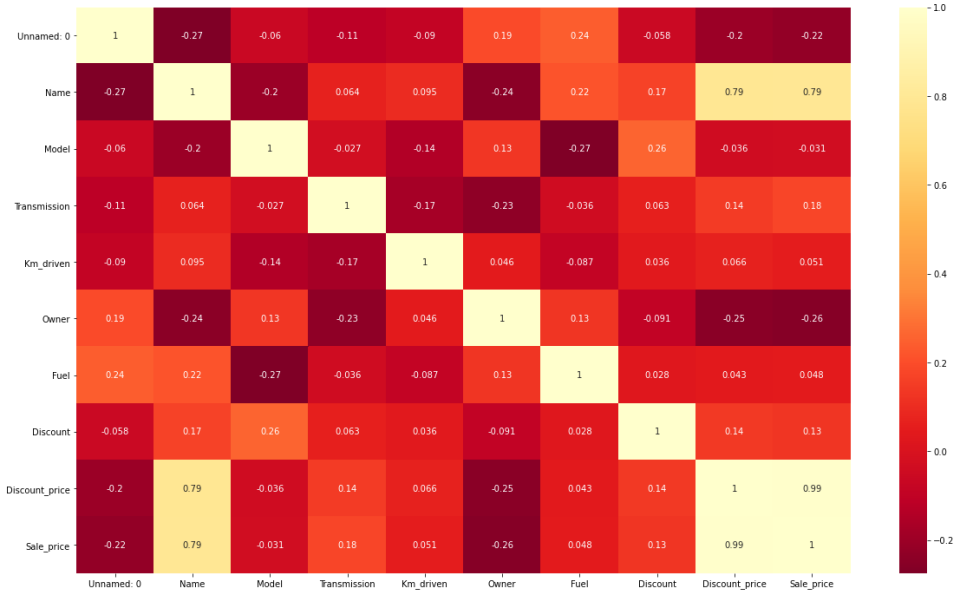


Using Strip plot We have used a strip plot for bivariate analysis to see the relation of each column with the target column. Here is a strip plot of the column fuel. We can clearly see the sale price of each type of fuel from the plot. WE can see that the highest sale price is for Diesel cars.



Multivariate Analysis

In the correlation matrix, we can see the relation of each column with all the other columns. We can see the correlation matrix in the plot. We have also used a pair plot to see the graphical relation of each column with the other columns.



Key Observations:

Now we can clearly identify the correlation of independent variable with the target variable"SaLe_Price".

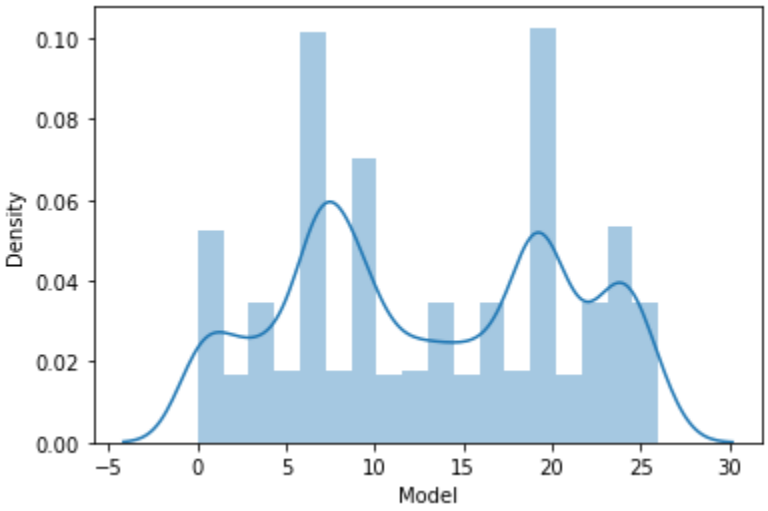
Light shades are highly positively correlated.

Sale_price is negatively correlated with Unnamed: o and Owner column.

Sale_price is positively correlated with discount price an name column.

Skewness

Here we have checked the skewness of each column of the dataset. Here we have a skewness plot of the column model, here we can see that all the curves carry skewness, not normally distributed we will remove the skewness later on using the appropriate method.



Till here visualization of the data is done.

Interpretation of the Results

From the complete analysis of used car price prediction, we are having a trained machine learning model by using which we can make the prediction for car prices.

Conclusion

The conclusion comes out to be we have complete Machine learning model for car price valuation for the car price traders, who was facing problems after the changes in market due to the Covid-19 pandemic.

This will help out the industry facing problem in this aspect. Thus the business problem in this aspect is resolved.