Appala Suneel Varri - 1002111813

**Professor:** Ramakrishna Koganti

**Used Cars Price Prediction**

**Introduction**

The usage of cars has been exponentially increasing in our day-to-day lives. Many individuals are now interested in purchasing used vehicles at some point in life due to their affordability and financial status. Introducing Machine Learning algorithms will help businesses and customers to predict the price of a used car in an estimated range.

**Problem Definition**

When customers want to sell their vehicle and are looking for an estimate of the resale value of the vehicle, they would generally reach out to the company workshop or check in third-party applications. In this process, using machine learning algorithms and predicting the value/price of the used cars would be of great advantage for the customer to pay a reasonable price according to its current market value. The important factors that impact the resale price of used cars are Brand, Mileage, Fuel Type, year of purchase, etc. Multiple algorithms like Linear Regression, Decision Tree, Random Forest analysis and X-boost were used against the data to analyze and conclude the best fit algorithm for the given sales data.

**Basic analysis of the used cars raw data**

1. Dataset contains 9 variables/factors in total and 301 rows to perform the prediction analysis

|  |  |
| --- | --- |
|  |  |

Fig. 1. Dataset Description

1. The entire dataset doesn’t contain any missing value.
2. The car model “City” was listed many times and “Corolla Altis” to the latter.

**Correlation Matrix**

|  |  |
| --- | --- |
| Fig. 2. Correlation Matrix | 1. Selling\_Price parameter is highly inversely related to Fuel\_Type\_Petrol, Seller\_Type having values of correlation coefficients as -0.54, -0.55 respectively. 2. Selling\_Price is highly related to Present\_Price with the correlation coefficient 0.88. |

**Histograms plot against the raw data.**

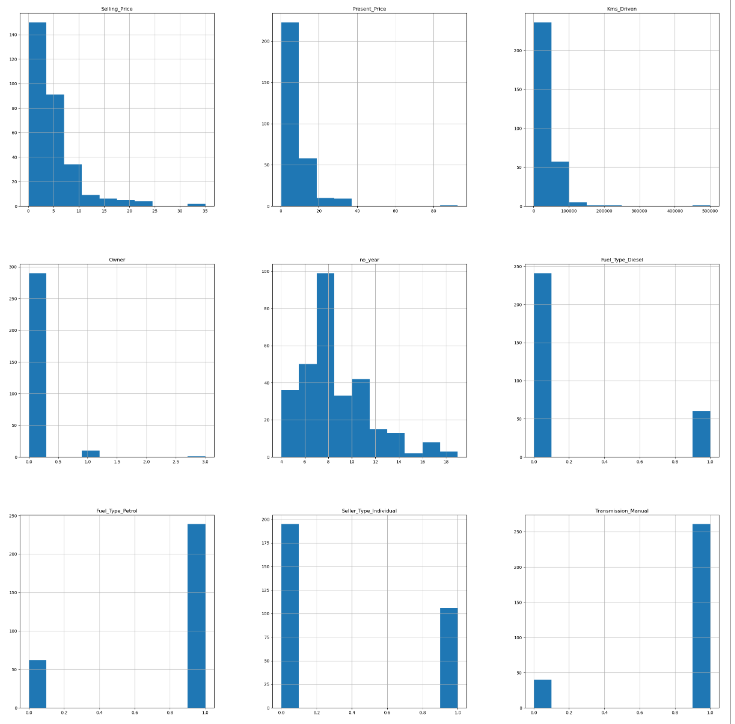
****

Fig. 3. Histograms for all the Predictor and Target variables

1. Most of the cars have selling prices around 3.5 lakhs.
2. Around 75% of the cars have travelled distance between 0 – 50,000kms.
3. Histogram of No of Years variable represents bi-modal distribution.
4. Around 85% of the cars are transmission manual.
5. Around 80% of cars use Fuel\_Petrol.
6. The months November-February recorded the highest sales.

**Observations from the Box Plots**

1. Most of the resale cars belong to the Petrol Type.
2. A negligible number of outliers were observed for Petrol and Diesel cars.

Chart, box and whisker chart

Description automatically generated

Fig. 4. Box Plots for few of the parameters

1. Diesel prices are very high when compared to Petrol and CNG.

**DATA PREPROCESSING**

1. Checking whether in case of any missing values and replacing them.
2. Removing some of the unrequired parameters (Car\_Name)
3. A new parameter named no\_year, Current Year has been introduced which stores total age and current year of the car.

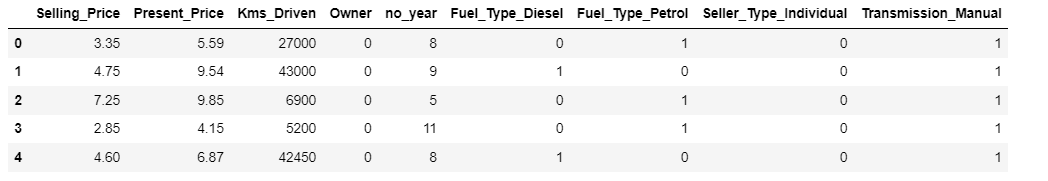


Fig. 5. Pre-Processed Table

1. Reducing Multicollinearity by converting variables into Boolean values (for eg: Petrol=1, Disel=0, Cng = 0)

**Model Development**

Linear Regressor

Text

Description automatically generated

Fig. 6. Fitting the data with Linear Regression Model

1. Parameters like Present\_price, Kms\_Driven, CPI and Fuel\_Price are measured in continuous values and the target variable Selling\_Price is also a continuous variable which led us to implement the Linear Regression approach.

Decision Tree Regressor

Text

Description automatically generated

Fig. 7. Fitting the data with Decision Tree Model

Random Forest Regressor

Text

Description automatically generated

Fig. 8. Fitting the data with Random Forest Model

1. As this is a regression problem, Random Forest algorithm can be used in order to generate the trees with some value of the sales.
2. Can handle both small and large sets of data and maintain accuracy.
3. As the dataset is huge generation of trees will be in huge numbers, which usually won’t allow overfitting trees in the model.

Voting Regressor(XG BOOST)

Text

Description automatically generated

Fig. 9. Fitting the data with Voting Regressor

**Conclusions**

1. Out of the executed four different algorithms **Voting Regressor** gives the best accuracy with 89.84 and Linear Regression stands least with 80.40.
2. And **Voting Regressor** has the least RMSE of 1.74 for the used cars dataset.

Table

Description automatically generated

Fig. 10. Description of the evaluated Machine Learning Models

1. For the given dataset we can conclude that “Voting Regressor” would be the best fit to predict the price of used cars.
2. This analysis will give some new findings for businesses as they can derive some decisions about the predictor variables like Present\_Price, Kms\_Driven, Fuel\_Type, Seller\_Type and Transmission. By analyzing the parameters that drive prices and using a reliable model (Voting Regressor for our dataset) to predict the selling price, which helps customers to make their profitable purchases.

**Model Deployment and Testing:**

1. As the **Voting Regressor** is the best fit model for this dataset, we have deployed the model and developed a few test cases to predict the price value.

Results of Model Deployment

Graphical user interface, text, application, email

Description automatically generated

Fig. 11. Results related to the Voting Regressor model.