Used Cars Price Prediction

Submitted by:

LAKSHITA KAIN
DATA SCIENCE INTERN

ACKNOWLEDGMENT

I would like to express my deepest appreciation to all those who provided me the possibility to complete this report. A special gratitude I give DataTrained and FlipRobo, whose contribution in stimulating suggestions and encouragement, helped me to coordinate my project. Furthermore I would also like to acknowledge with much appreciation the crucial role of the our SME of Mr Sajid Chaudhary.

1. INTRODUCTION

A car is not only a symbol of social status but also a valuable asset. In 2018-19, while new car sales were recorded at 3.6 million units, 4 million second-hand cars were bought and sold, according to a recent report on India's pre-owned car market by IndianBlueBook, a used car pricing guide by Mahindra First Choice Wheels.

The growth rate of new car sales has slowed owing to a variety of reasons, including cyclical slowdown in auto sales in election years and an overall consumption slowdown in the economy. New car sales grew 2.70% in 2018-19, the slowest growth rate for the industry in four years. In April, passenger vehicle sales saw a sharp decline compared with the same month last year, and domestic sales saw a contraction of 17.07%.

People buy used cars for various reasons like to learn how to drive, to have a business car, to have a seperate car for work, financial crunch, selling the old car to get a new one and the reasons can be endless.

The value of a car depreciates significantly in the first year itself, up to 50% for some models. From a buyer's perspective, the aspiration of owning a car is getting fulfilled at a lower price point. According to the IndianBlueBook report, 45% of the buyers want a car that is four to five years old. However, 46% of the sellers want to sell their vehicle when it is six to eight years old. Getting your used car insured will cost you a lot less as compared to a new car.

1.2 PROBLEM DEFINITION:

Predicting the resale value of a car is not a simple task. It is trite knowledge that the value of used cars depends on a number of factors. The most important ones are usually the age of the car, its make (and model), the origin of the car (the original country of the manufacturer), its mileage (the number of kilometers it has run) and it's fuel type.

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. Due to rising fuel prices, fuel economy is also of prime importance.

1.3 OBJECTIVE :

This study focuses on predicting a used car's selling price all over India using various supervised machine learning models based on various quantitative and qualitative features of a car.

2. Literature Review

2.1. Machine Learning Algorithms

Machine learning is a subfield of Artificial Intelligence (AI) that works with algorithms and technologies to extract useful information from data. Machine learning methods are appropriate in big data since attempting to manually process vast volumes of data would be impossible without the support of machines. Machine learning in computer science attempts to solve problems algorithmically rather than purely mathematically. Therefore, it is based on creating algorithms that permit the machine to learn. However, there are two general groups in machine learning which are supervised and unsupervised. Supervised is where the program gets trained on pre-determined set to be able to predict when a new data is given. Unsupervised is where the program tries to find the relationship and the hidden pattern between the data.
Machine learning-based systems are growing in popularity in research applications in most disciplines.

However, some of them give better performance in certain circumstances. Thus, this thesis attempts to use regression algorithms to compare their performance when it comes to predicting values of a given dataset. The performance will be measured upon predicting house prices since the prediction in many regression algorithms relies not only on a specific feature but on an unknown number of attributes that result in the value to be predicted. House prices depend on an individual house specification. Houses have a variant number of features that may not have the same cost due to its location. For instance, a big house may have a higher price if it is located in desirable rich area than being placed in a poor neighbourhood.

2.2. Multiple Linear Regression

Multiple Linear Regression (MLR) is a supervised technique used to estimate the relationship between one dependent variable and more than one independent variables. Identifying the correlation and its cause-effect helps to make predictions by using these relations. To estimate these relationships, the prediction accuracy of the model is essential; the complexity of the model is of more interest. However, Multiple Linear Regression is prone to many problems such as multicollinearity, noises, and overfitting, which effect on the prediction accuracy.
Regularised regression plays a significant part in Multiple Linear Regression because it helps to reduce variance at the cost of introducing some bias, avoid the overfitting problem and solve ordinary least squares (OLS) problems. There are two types of regularisation techniques L1 norm (least absolute deviations) and L2 norm (least squares). L1 and L2 have different cost functions regarding model complexity.

2.3. Lasso Regression

Least Absolute Shrinkage and Selection Operator (Lasso) is an L1-norm regularised regression technique that was formulated by Robert Tibshirani in 1996 [6]. Lasso is a powerful technique that performs regularisation and feature selection. Lasso introduces a bias term, but instead of squaring the slope like Ridge regression, the absolute value of the slope is added as a penalty term. Lasso is defined as

$$L=Min(sum\ of\ squared\ residuals + \alpha*|slope|)$$

Where $Mi(sum\ of\ squared\ residuals)$ is the Least Squared Error, and $\alpha*|s\ ope|$ is the penalty term. However, alpha α is the tuning parameter which controls the strength of the penalty term. In other words, the tuning parameter is the value of shrinkage. $|slope|$ is the sum of the absolute value of the coefficients .
Cross-validation is a technique that is used to compare different machine learning algorithms in order to observe how these methods will perform in practice. Cross-validation method divides the data into blocks. Each block at a time will be used for testing by the algorithm, and the other blocks will be used for training the model. In the end, the results will be summarised, and the block that performs best will be chosen as a testing block . However, α is determined by using cross-validation. When $\alpha = 0$, Lasso becomes Least Squared Error, and when $\alpha \neq 0$, the magnitudes are considered, and that leads to zero coefficients. However, there is a reverse relationship between alpha α and the upper bound of the sum of the coefficients t. When $t \to \infty$, the tuning parameter $\alpha = 0$. Vice versa when $\alpha = 0$ the coefficients shrink to zero and $t \to \infty$. Therefore, Lasso helps to assign zero weights to most redundant or irrelevant features in order to enhance the prediction accuracy and interpretability of the regression model.

Throughout the process of features selection, the variables that still have non-zero coefficients after the shrinking process are selected to be part of the regression model. Therefore, Lasso is powerful when it comes to feature selection and reducing the overfitting.

2.4. Ridge Regression

The Ridge Regression is an L2-norm regularised regression technique that was introduced by Hoerl in 1962 [9]. It is an estimation procedure to manage collinearity without removing variables from the regression model. In multiple linear regression, the multicollinearity is a common problem that leads least square estimation to be unbiased, and its variances are far from the correct value. Therefore, by adding a degree of bias to the regression model, Ridge Regression reduces the standard errors, and it shrinks the least square coefficients towards the origin of the parameter space [10]. Ridge formula is:

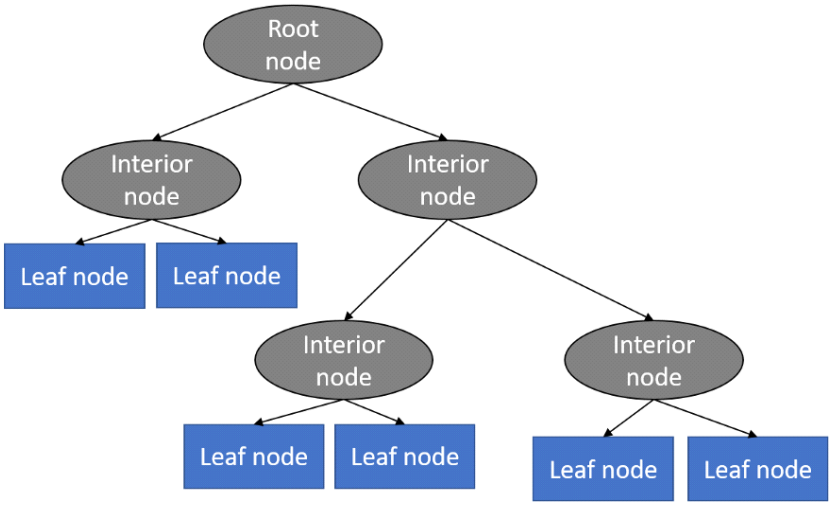
$$R=Mi(sum\ of\ squared\ residuals + \alpha*slope_2)$$

Where $Mi(sum\ of\ squared\ residuals)$ is the Least Squared Error, and $\alpha * slope^2$ is the penalty term that Ridge adds to the Least Squared Error.
When Least Squared Error determines the values of parameters, it minimises the sum of squared residuals. However, when Ridge determines the values of parameters, it reduces the sum of squared residuals. It adds a penalty term, where α determines the severity of the penalty and the length of the slope. In addition, increasing the α makes the slope asymptotically close to zero. Like Lasso, α is determined by applying the Cross-validation method. Therefore, Ridge helps to reduce variance by shrinking parameters and make the prediction less sensitive.

2.5 Decision Tree Regressor

Decision Tree is one of the most commonly used, practical approaches for supervised learning. It can be used to solve both Regression and Classification tasks with the latter being put more into practical application.

It is a tree-structured classifier with three types of nodes. The Root Node is the initial node which represents the entire sample and may get split further into further nodes. The Interior Nodes represent the features of a data set and the branches represent the decision rules. Finally, the Leaf Nodes represent the outcome. This algorithm is very useful for solving decision-related problems.



A Decision Tree can be defined as a model : $\varphi = X \longmapsto Y$

Where any node t represents a subspace $X_t \subseteq X$ of the input space and internal nodes t are labelled with a split s_t taken from a set of questions Q. However, to determine the best separation in Decision Trees, the Impurity equation of dividing the nodes should be taken into consideration, which is defined as:

$$\Delta i(s,t) = i(t) - pLi(t_L) - pRi(t_R)$$

2.6 Random Forest Regressor

Random Forest is a model that constructs an ensemble predictor by averaging over a collection of decision trees. Therefore, it is called a forest, and there are two reasons for calling it random. The first reason is growing trees with a random independent bootstrap sample of the data. The second reason is splitting the nodes with arbitrary subsets of features . However, using the bootstrapped sample and considering only a subset of the variables at each step results in a wide variety of trees. The variety is what makes Random Forest more effective than individual Decision Tree.

To improve prediction performance, Random Forest acquires out-of-bag (OOB) estimates, which is based on the fact that, for every tree, approximately $1/e \approx 0.367$ *or* 36% of cases are not in the bootstrap sample [15]. There are several advantages to using OOB. One advantage is that the complete original example is used both for constructing the Random Forest classifier and for error estimation. Another advantage is its computational speed, especially when dealing with large data dimensions .

2.7 Adaptive Boosting Regressor

Adaptive boosting (AdaBoost) is an ensemble algorithm incorporated by Freund and Schapire (1997), which trains and deploys trees in time series. Since then, it evolved as a popular boosting technique introduced in various research disciplines. It merges a set of weak classifiers to build and boost a robust classifier that will improve the decision tree's performance and improve accuracy.

The mathematical presentation of the AdaBoost classifier at a high level is described of the following:

Consider that training data (x1 and y1), . . . , (xm and ym) are the xi 2 X, yi 2 {−1, +1}. Then the parameters of AdaBoost classifier are initialized: D1(i) = 1/m for i = 1, . . . , m. For t = 1, . . . , T:

• Train weak learner using distribution Dt.

• Aim: select ht with low weighted error:

$$\epsilon = \Pr_i \sim D_t[h_t(x_i) \neq y_i]$$

Select at = 1

$$H(x) = sign\left(\sum_{t=1}^{T} \alpha_t h_t(x)\right).$$

$$D_{t+1}(i) = \frac{D_t(i)\exp(-\alpha_t y_i h_t(x_i))}{Z_t}$$

Hence, for i = 1, . . . ,m:

where Zt is a normalization factor (chosen so that Dt + 1 will be a distribution). Which generates an output of the final hypothesis as following:

H(x) = si.

3. Methodology

3.1 Data Collection

Data has been scrapped using Selenium webdriver with python. Selenium is a powerful tool for controlling web browsers through programs and performing browser automation.

Our data has been retrieved from https://www.car24.com which consists of used car information such as model, brand, location, kilometers driven, fueltype, transmission, engine, Price, Monthly EMI price and direct URL. The data collected is from Delhi – NCR, Chennai, Maharashtra, Pune, Hyderabad, Gujarat, Jaipur Chandigarh and Kolkata.

3.2. About Dataset

- Shape of our datset : records - 9100, features - 13

```
0   Unnamed: 0        9100 non-null   int64
1   Name              9100 non-null   object
2   Brand             9100 non-null   object
3   Price             9100 non-null   object
4   Kilometers        9100 non-null   object
5   Transmission      9100 non-null   object
6   Fuel              9100 non-null   object
7   Monthly_EMI       9100 non-null   object
8   Downpayment       9100 non-null   object
9   Ownership         9100 non-null   object
10  Engine            9100 non-null   object
11  Year_of_Purchase  9100 non-null   int64
12  URL               9100 non-null   object
```

We have two int data type features and 11 object data types although some our features have misinterpreted data types such as Kilometers, Price and Monthly_EMI should be of numeric data type not categorical. "Unnamed : 0" is an necessary column for our analysis therefore we drop it.

We don't have any null values or duplicated values in our obtained dataset.

3.3. Data Preprocessing

3.3.1 Feature Engineering

From our URL column states from which the car is sold has been extracted and added to a new column df['State'].

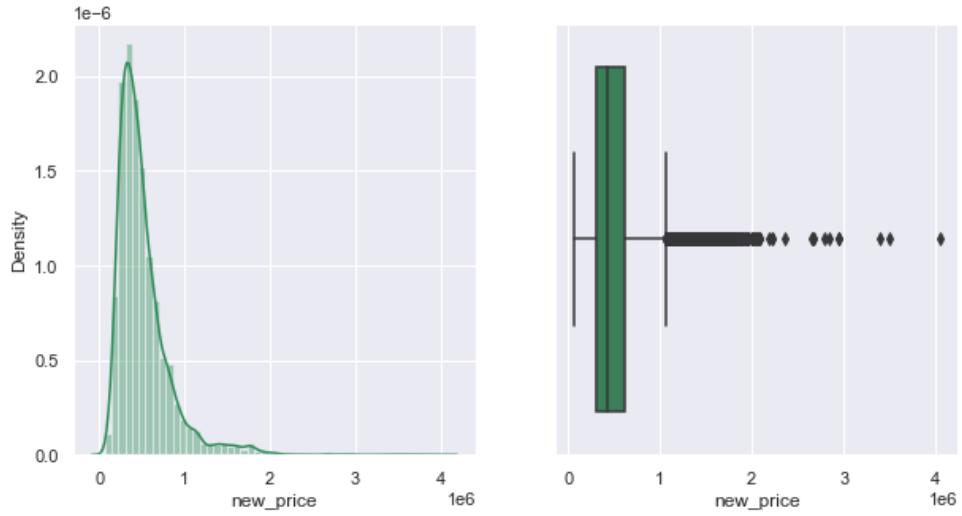
In many of our features such as Kilometers, Price, Monthly_EMI, Name, Brand have

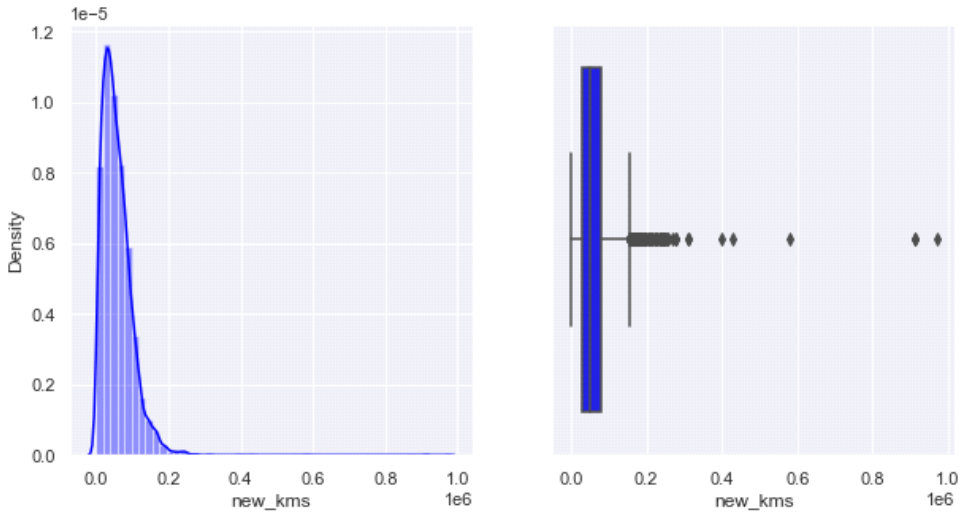
unneccessary delimeters which needed to be extracted and removed to have cleaner data entities.

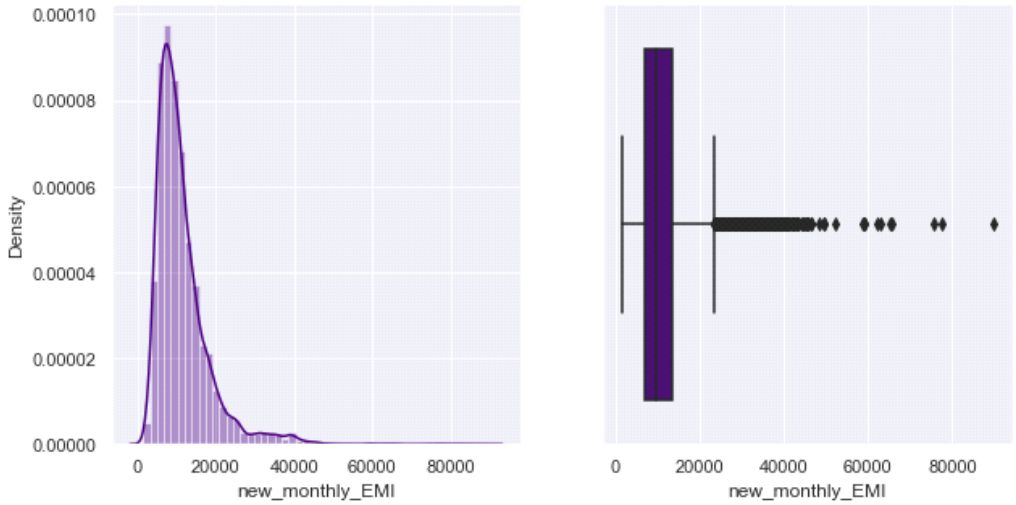
- From Kilometers delimeters like commas, white spaces, and numeric were extracted.
- From Name and brand delimeters like : "[,'']' and white spaces were removed.
- From Price and Monthly_EMI special character "₹" adnd comma were removed.

3.4 Outliers

Outliers are noisy data that they do have abnormal behaviour comparing with the rest of the data in the same dataset. Outliers can influence the prediction model and performance due to its oddity. There are three types of outliers, which are point, contextual, and collective outliers.
3 of our features are right skewed : price, kilometers, monthly_EMI.







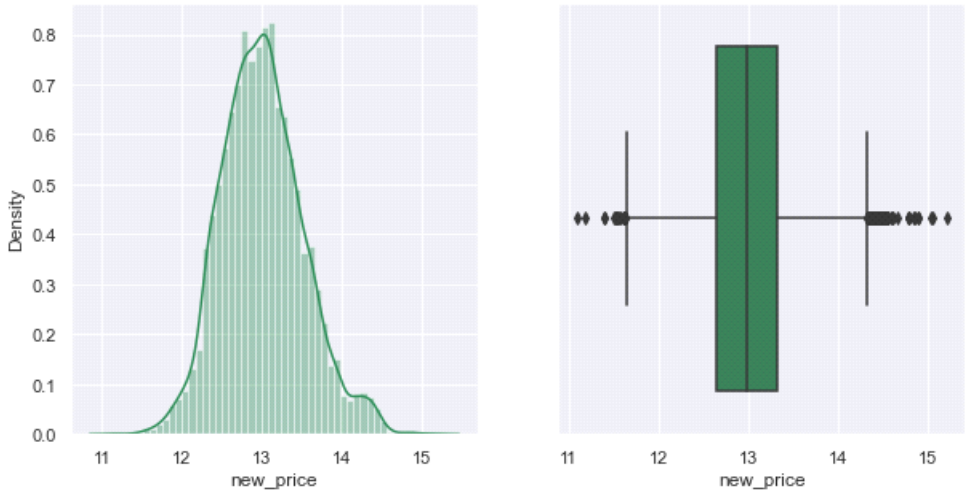
Visualizing outliers using histogram and box - plot

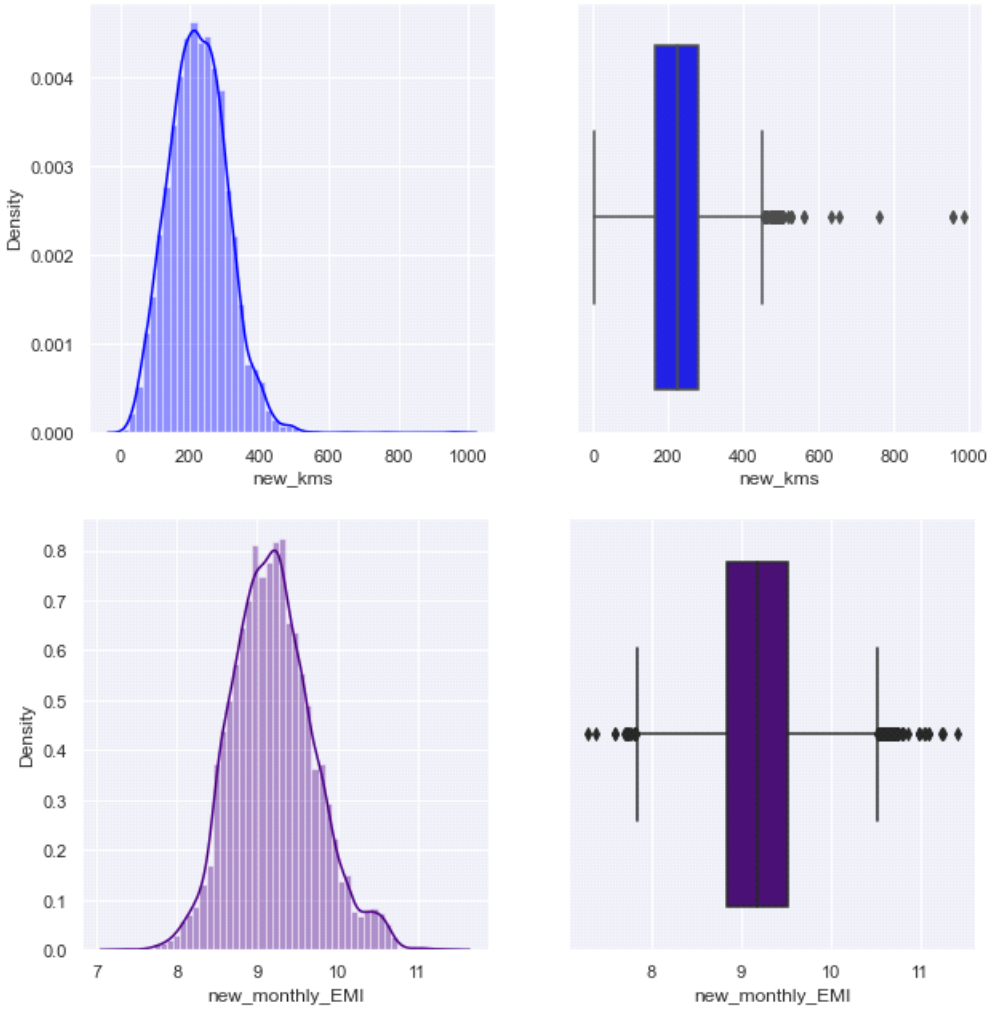
3.5 Square - Root/ Cube - Root Transformation

The square root, x to x^(1/2) = sqrt(x), is a transformation with a moderate effect on distribution shape: it is weaker than the logarithm and the cube root. It is also used for reducing right skewness, and also has the advantage that it can be applied to zero values. Note that the square root of an area has the units of a length. It is commonly applied to counted data, especially if the values are mostly rather small.

The cube root, x to x^(1/3). This is a fairly strong transformation with a substantial effect on distribution shape: it is weaker than the logarithm. It is also used for reducing right skewness, and has the advantage that it can be applied to zero and negative values. Note that the cube root of a volume has the units of a length.

After performing transformations to reduce skewness.





3.6 Encoding

Label Encoding refers to converting the labels into a numeric form so as to convert them into the machine-readable form. Machine learning algorithms can then decide in a better way how those labels must be operated. It is an important pre-processing step for the structured dataset in supervised learning. This may lead to the generation of priority issues in the training of data sets. A label with a high value may be considered to have high priority than a label having a lower value.

3.7. Training–Test Set Split

For all the models fitted in this study, we split the balanced data into 75% for the training set and 25% for the testing set (validation).

4. Hardware and Software Requirements and Tools Used

4.1. Software Requirements :

- Python
- Anaconda (Jupyter Notebook)
- Python Libraries (Scikit - Learn, Imblearn, Pandas, Numpy , etc)

4.2. Hardware Requirements :

- Minimum 4 GB RAM
- Intel Core-i3 or above processor

5. Results

In this section, I present analytic results of the various machine learning models adopted in this research. All model diagnostic metrics in this paper are based on the validation/test set. In order to compare the machine learning models they have been tested on data where the price of the properties are known. Therefore, the data set is split into training data and testing data. Training data is, as the name suggests, used to train the machine learning model and constitutes the larger share of the split data set. When the model is fit with the training data the sales price of the cars are known to the model in order for it to learn from the data. The testing data is used to get a measurement of how well the machine learning model predicts house prices. The machine learning model is given the test data but without the price of the car in order to predict the price for them given the various features for the used-car. The predicted price is then compared to the actual price in the test data.

5.1 Error metrics

For measuring how good predictions the model makes, four error metrics have been used. Mean absolute error (MAE), Mean squared error (MSE), Median absolute error (MedAE) and Coefficient of determination (R2).

5.1.1 R-Squared Score

In statistics, the coefficient of determination, denoted R2 or r2 and pronounced "R squared", is the proportion of the variation in the dependent variable that is predictable from the independent variable(s).

It is a statistic used in the context of statistical models whose main purpose is either the prediction of future outcomes or the testing of hypotheses, on the basis of other related information. It provides a measure of how well observed outcomes are replicated by the model, based on the proportion of total variation of outcomes explained by the model.

$$R^2 = 1 - \frac{\text{Unexplained Variation}}{\text{Total Variation}}$$

5.1.2 Mean absolute error (MAE)

Mean absolute error measures the prediction error by taking the mean of all

absolute values of all errors, that is:

$$MAE = \frac{\sum_{i=0}^{n} |y_i - \hat{y}_i|}{n}$$

Where n is the number of samples, y are the target values and ^y are the predicted values. A MAE closer to 0 means that the model predicts with lower error and that the prediction is better the closer the MAE is to 0.

5.1.3 Mean squared error (MSE)

Mean squared error is similar to MAE, but the impact of a termis quadratically proportional to its size. It measures the prediction error by taking the mean of all squared absolute values of all errors, that is:

$$MSE = \frac{\sum_{i=0}^{n} (y_i - \hat{y}_i)^2}{n}$$

As for MAE, values close to 0 are better.

5.1.4 Root Mean squared error (RMSE)

Root Mean Square Error (RMSE) is the standard deviation of the residuals (prediction errors). Residuals are a measure of how far from the regression line data points are; RMSE is a measure of how spread out these residuals are. In other words, it tells you how concentrated the data is around the line of best fit.

$$\mathbf{RMSE}_{fo} = [\sum_{i=1}^{N} (z_{f_i} - z_{o_i})^2 / N]^{1/2}$$

	r2 Score (%)	Mean Absolute Error(e-05)	Mean Squared Error (e-09)	Root Mean Squared Error(e-05)
Linear Regression	99.99	2.8340	1.3370	3.6571
Lasso Regularization	99.99	4.8300	3.5660	5.9721
Decision Tree Regressor	99.99	0.0007	8.0760	0.0089
Random Forest Regressor	99.96	0.0006	8.8880	0.0094
Ridge Regularization	99.99	2.8344	1.3370	3.6578
K-Nearest Neighbors	7.08	0.3788	0.2430	0.4932
Gradient Boost Regressor	99.96	0.0030	9.0090	0.0094
Adaptive Boosting Regressor	98.88	0.0410	0.0029	0.0540
Bagging Regressor	99.99	0.0007	8.7580	0.0093

fig : Models used and their evaluation metrics

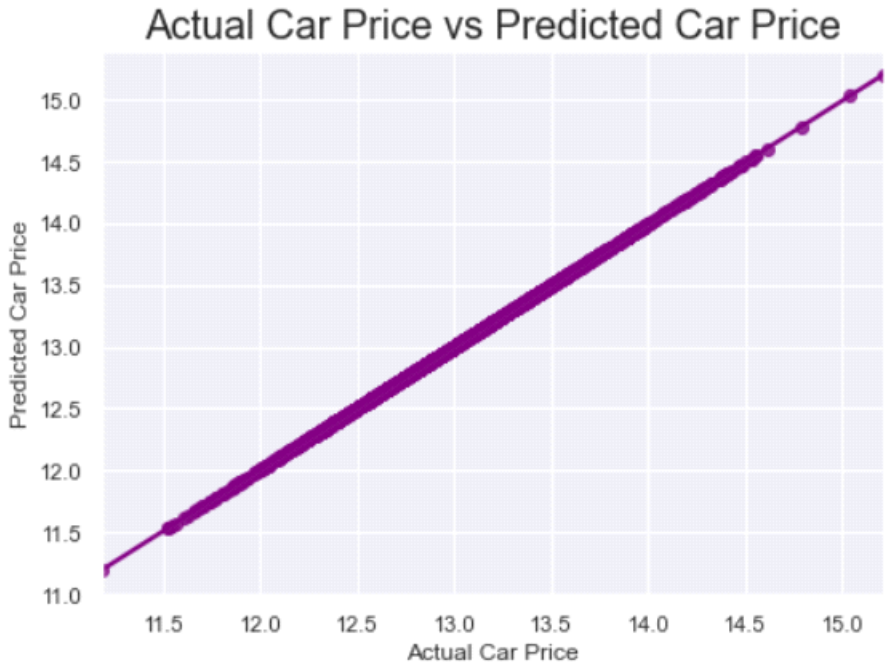
5.6. Tuning of Hyperparameters

In this subsection, we present the optimal hyperparameters to tune our best model i.e. Linear Regression model with the following hyperparameters:

```
#Using Randomized Search
param_grid = {'alpha': uniform()}

model = Lasso()
rand_search = RandomizedSearchCV(estimator=model,
                                 param_distributions=param_grid,
                                      n_iter=100)

rand_search.fit(X_train,y_train)

print(rand_search.best_estimator_.alpha)
print(rand_search.best_score_)
```

```
0.019989745963960304
0.9939549005186666
```

We've deployed Randomized Search CV for Hyperparameter Tuning of our model with 100 iterations.



We concluded the following error metrics for our final tuned model :

- R2 Score for Tuned Lasso Regression Model: 0.9999
- Mean Absolute Error for our Tuned Lasso Regression Model: 0.00039
- Mean Squared Error for our Tuned Lasso Regression Model: 2.5785e-07
- Root Mean Squared Error for our Tuned Lasso Regression Model: 0.00050

6. Conclusions :

This study focuses on how well used-car prices can be predicted by different regression models. In this study we have found that Multiple Linear Regression algorithm not only performed better at predicting used-car but also was the fastest among all algorithms. The lowest error achieved was for Linear Regression model with an MAE of 0.00039(which is almost negligible).

7. Future Scope

The auto industry is changing rapidly and car prices are only going up. So to speak, new cars are getting costlier each year, making them a very high value purchase for the common man and quite ironically, the average life span of a car is going down despite the steady rise in prices.

In india used-car market is bigger than the new car market. There can be endless amount of data that can be collected which can further improve our machine learning models. Larger the data better the model.