# Car Price Prediction using Linear Regression And Recursive Feature Of Elimination

Satender Kumar Tiwari
*Department of Information Technology*
*ABES Engineering College*
Ghaziabad, India
satender.19b131191@abes.ac.in

Saurabh Singh
*Department Of Information Technology*
*ABES Engineering College*
Ghaziabad, India
saurabh.19b131083@abes.ac.in

Satyajeet Dhama
*Department Of Information Technology*
*ABES Engineering College*
Ghaziabad, India
satyajeet.19b131104@abes.ac.in

***Abstract*— The production of cars has been steadily increasing in the past decade, with over 70 million passenger cars being produced in the year 2016. This has given rise to the used car market, which on its own has become a booming industry. The recent advent of online portals has facilitated the need for both the customer and the seller to be better informed about the trends and patterns that determine the value of a used car in the market. Using Machine Learning Algorithms such as Linear Regression, Multiple Regression and Regression trees, Recursive Feature Of Elimination. we will try to develop a statistical model which will be able to predict the price of a used car, based on previous consumer data and a given set of features. We will also be comparing the prediction accuracy of these models to determine the optimal one.**

***Keywords*— *ANOVA, Liner Regression, Regression Tree, Multiple Regression, RFE***

## I. INTRODUCTION

The used car market is an ever-rising industry, which has almost doubled its market value in the last few years. The emergence of online portals such as CarDheko, Quikr, Carwale, Cars24, and many others has facilitated the need forboth the customer and the seller to be better informed about the trends and patterns that determine the value of the used car in the market. Machine Learning algorithms can be used to predict the retail value of a car, based on a certain set of features.

Different websites have different algorithms to generate the retail price of the used cars, and hence there isn't a unified algorithm for determining the price. By training statistical models for predicting the prices, one can easily get a rough estimate of the price without actually entering the details intothe desired website. The main objective of this paper is to use three different prediction models to predict the retail price of a used car and compare their levels of accuracy.

The data set used for the prediction models was created by Shonda Kuiper[1]. The data was collected from the 2005 Central Edition of the Kelly Blue Book and has 804 records of2005 GM cars, whose retail prices have been calculated. The data set primarily comprises of categorical attributes along with two quantitative attributes.

The following are the variables used:

**Price:** The calculated retail price of GM cars. The cars which were selected for this data set were all less than a year old andwere considered to be in good condition.

**Mileage:** The total number of miles the car has been driven

**Make:** The manufacturer of the car

**Model:** The specific models for each car

**Trim:** The type of car model

**Type:** The car's body type

**Cylinder:** The number of cylinders present in the Engine

**Liter:** The fuel capacity of the Engine

**Doors:** The number of doors in the car

**cruise:** A categorical variable (binary), which represents whether cruise control is present in the car (coded 1 if present) **sound:** A categorical variable (binary), that represents whether upgraded speakers are present in the car (coded 1 if present)

**Leather:** A categorical variable (binary), that represents whether the car has leather interiors (coded 1 if present)

Using these attributes, we will try to predict the price by using the Statistical Analysis System (SAS) for exploratory data analysis.

## II. LITERATURE SURVEY

Overfitting and underfitting come into picture when we create our statistical models. The models might be too biasedto the training data and might not perform well on the test data set. This is called overfitting. Likewise, the models might nottake into consideration all the variance present in the population and perform poorly on a test data set. This is called underfitting. A perfect balance needs to be achieved between these two, which leads to the concept of Bias-Variance tradeoff. Pierre Geurts [2] has introduced and explained how bias-variance tradeoff is achieved in both regression and classification. The selection of variables/attribute plays a vital role in influencing both the bias and variance of the statisticalmodel. Robert Tibshirani [3] proposed a new method called Lasso, which minimizes the residual sum of squares. This returns a subset of attributes which need to be included in multiple regression to get the minimal error rate. Similarly, decision trees suffer from overfitting if they are not pruned/shrunk. Trevor Hastie and Daryl Pregibon [4] have explained the concept of pruning in their research paper. Moreover, hypothesis testing using ANOVA is needed to verify whether the different groups of errors really differ from each other. This is explained by TK Kim and Tae Kyun in their paper [5]. A Post-Hoc test needs to be performed along with ANOVA if the number ofgroups exceeds two.

Tukey's Test has been explored by Haynes W. in his research paper [6]. Using these techniques, we will create, train and test the effectiveness of our statistical models.

### Problem Statement

A Chinese automobile company Geely Auto aspires to enter the US market by setting up their manufacturing unit there and producing cars locally to give competition to their US and European counterparts.

They have contracted an automobile consulting company to understand the factors on which the pricing of cars depends. Specifically, they want to understand the factors affecting the pricing of cars in the American market, since those may be very different from the Chinese market. The company wants to know:

- Which variables are significant in predicting the price of a car

- How well those variables describe the price of a car

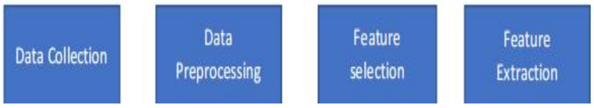
Based on various market surveys, the consulting firm has gathered a large dataset of different types of cars across the Americal market.

### Business Goal

You are required to model the price of cars with the available independent variables. It will be used by the management to understand how exactly the prices vary with the independent variables. They can accordingly manipulate the design of the cars, the business strategy etc. to meet certain price levels. Further, the model will be a good way for management to understand the pricing dynamics of a new market.

## III. PROPOSED MODEL

### 3. Proposed Methodology

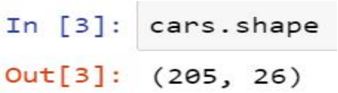


**• Data Collection**

For this project the data has been downloaded from Kaggle .
It consist of
26 classes namely

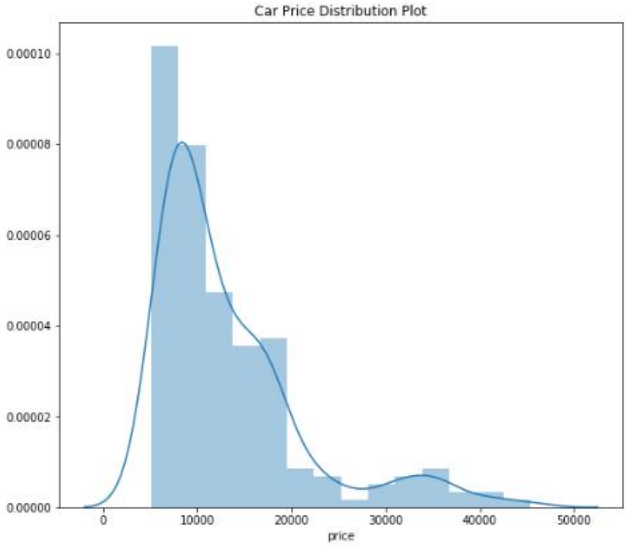
| car_ID | symboling | CarName | fueltype | aspiration |
|---|---|---|---|---|

etc .There are total of 205 number of samples classes as shown in figure 1.



```
In [3]: cars.shape
Out[3]: (205, 26)
```

### Visualizing the data

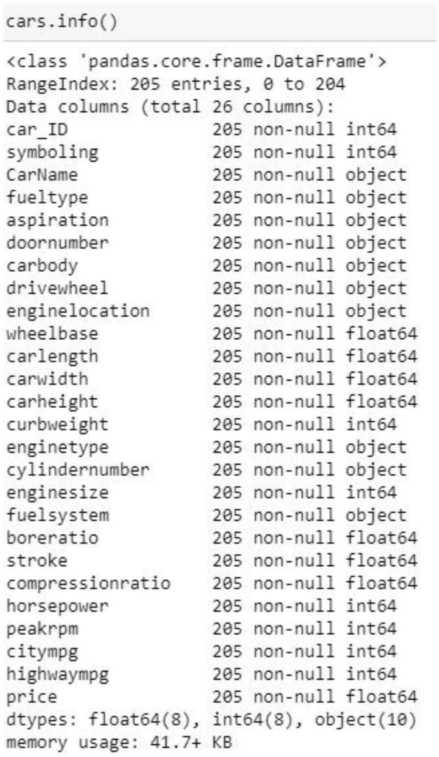


### Inference

1. The plot seemed to be right-skewed, meaning that the most prices in the dataset are low(Below 15,000)
2. There is a significant difference between the mean and the median of the price distribution.
3. The data points are far spread out from the mean, which indicates a high variance in the car prices.(85% of the prices are below 18,500, whereas the remaining 15% are between 18,500 and 45,400.)

### • Data Preprocessing

The data is preprocessed it does not require to make any change in
dataset. Id is not considered as any valid parameter and diagnosis is our
final output hence we are not considering both these as input parameter



```
cars.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 205 entries, 0 to 204
Data columns (total 26 columns):
car_ID              205 non-null int64
symboling           205 non-null int64
CarName             205 non-null object
fueltype            205 non-null object
aspiration          205 non-null object
doornumber          205 non-null object
carbody             205 non-null object
drivewheel          205 non-null object
enginelocation      205 non-null object
wheelbase           205 non-null float64
carlength           205 non-null float64
carwidth            205 non-null float64
carheight           205 non-null float64
curbweight          205 non-null int64
enginetype          205 non-null object
cylindernumber      205 non-null object
enginesize          205 non-null int64
fuelsystem          205 non-null object
boreratio           205 non-null float64
stroke              205 non-null float64
compressionratio    205 non-null float64
horsepower          205 non-null int64
peakrpm             205 non-null int64
citympg             205 non-null int64
highwaympg          205 non-null int64
price               205 non-null float64
dtypes: float64(8), int64(8), object(10)
memory usage: 41.7+ KB
```

### *A. Linear Regression*

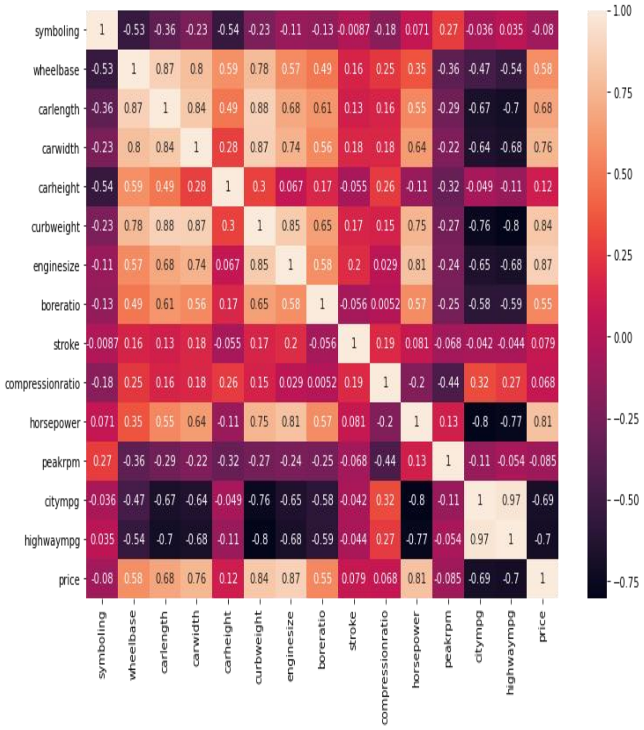
Linear regression analysis is **used to predict the value of a variable based on the value of another variable**. The variable you want to predict is called the dependent variable. The variable you are using to predict the other variable's value is called the independent variable.

### *B. Recursive Feature Elimination*

Recursive feature elimination (RFE) is **a feature selection method that fits a model and removes the weakest feature (or features) until the specified number of features is reached**.

### Feature selection

For feature selection we take parameters with mean values and draw a
heat-map between these parameters of the dataset.

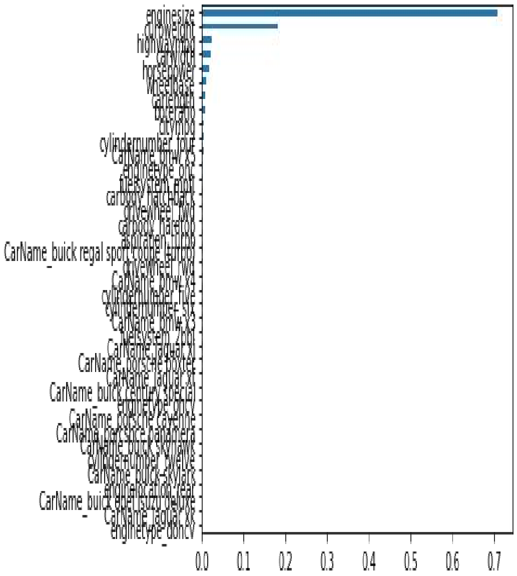


### Feature extraction

PCA (Principle Component Analysis) is used to extract features by
reducing the numbers of input variables, we use Random forest

```
Mean Absolute Error: 1535.83 degrees.
Accuracy: 87.5 %.
```

Feature importance is also used to find
importance of every parameter



### Result and Discussion

All experiment has been carried out in widows 10 with Intel(R) Core(TM) i5-
8265U CPU @ 1.60GHz 1.80 GHz, RAM 8 GB. In this analysis we worked on
two Machine Learning Models providing accuracy 87.5% By RFE and 84.6% by LR.

### Conclusion & Future Work

Conclusion
We have developed model to predict car price problem.

First, we made the detailed exploratory analysis.

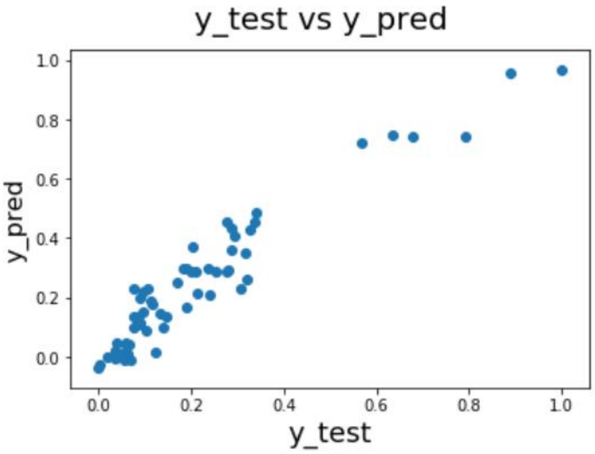
We have decided which metric to use.

We analyzed both target and features in detail.
We transform categorical variables into numeric so we can use them in the model.
We transform numerical variables to reduce skewness and get close to normal distribution.
We use pipeline to avoid data leakage.
We looked at the results of the each model and selected the best one for the problem in hand.
We made hyperparameter tuning of the best model see the improvement

We looked at the feature importance.

### Evaluation of test via comparison of y_pred and y_test

```
Text(0, 0.5, 'y_pred')
```



## References

[1] Shonda Kuiper (2008) Introduction to Multiple Regression: How Much Is Your Car Worth?, Journal of Statistics Education, 16:3, DOI:10.1080/10691898.2008.11889579

[2] Geurts P. (2009) Bias vs Variance Decomposition for Regression and Classification. In: Maimon O., Rokach L. (eds) Data Mining and Knowledge Discovery Handbook. Springer, Boston, MA

[3] Robert T. (1996) Regression Shrinkage and Selection Via the Lasso. In: Journal of the Royal Statistical Society: Series B (Methodological)Volume 58, Issue 1

[1] G. Eason, B. Noble, and I. N. Sneddon, “On certain integrals of Lipschitz-Hankel type involving products of Bessel functions,” Phil. Trans. Roy. Soc. London, vol. A247, pp. 529–551, April 1955. *(references)*

[2] J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68–73.

[3] I. S. Jacobs and C. P. Bean, “Fine particles, thin films and exchange anisotropy,” in Magnetism, vol. III, G. T. Rado and H. Suhl, Eds. New York: Academic, 1963, pp. 271–350.

[4] K. Elissa, “Title of paper if known,” unpublished.

[5] R. Nicole, “Title of paper with only first word capitalized,” J. Name Stand. Abbrev., in press.

[6] Y. Yorozu, M. Hirano, K. Oka, and Y. Tagawa, “Electron spectroscopy studies on magneto-optical media and plastic substrate interface,” IEEE Transl. J. Magn. Japan, vol. 2, pp. 740–741, August 1987 [Digests 9th Annual Conf. Magnetics Japan, p. 301, 1982].

[7] M. Young, The Technical Writer’s Handbook. Mill Valley, CA: University Science, 1989.