

Used Car Price Prediction

Submitted by:

Aamir Rahman

**ACKNOWLEDGMENT**

I would like to express my special thanks of gratitude to FlipRobo, who gave me the golden opportunity to do this wonderful project of Used Car Price Prediction.

Secondly, I would also like to thank DataTrained teachers and mentors who teach me all the basic and professional concept for building the project.

**INTRODUCTION**

* Business Problem Framing

Determining whether the listed 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 developing machine learning models that can accurately predict the price of a used car based on its features, in order to make informed purchases. We implement and evaluate various learning methods on a dataset consisting of the sale prices of different makes and models across cities in the United States. Our results show that Random Forest model yield the best results, but are compute heavy. Conventional linear regression also yielded satisfactory results, with the advantage of a significantly lower training time in comparison to the aforementioned methods.

* Conceptual Background of the Domain Problem

To understand this problem the developer should have a basic knowledge related to the working principles of the car. Like what are the most important thing that a buyer look at the time of buying a car like Km driven, Transmission, Mileage, Engine, etc. Sometimes, some buyer focus only one thing that is most important to them. Some buyer prefer mileage over anything, some prefer power, some prefer seating capacity. So before solving this problem the developer should be aware from these things.

* Motivation for the Problem Undertaken

Desire to face the challenge in solving the unsolved problems. And also the objective behind taking this project is to improve my skill in statistical as well as in analytical knowledge in machine learning and artificial intelligence.

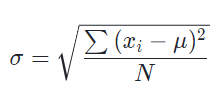
**Analytical Problem Framing**

* Mathematical/ Analytical Modeling of the Problem

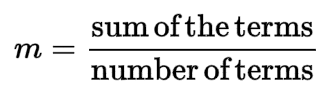
There are various mathematical modelling is carried out in this project. For example:

* finding the mu and sigma for our target variable using the scipy built-in function scipy.stats.norm. And making it normally distributed.

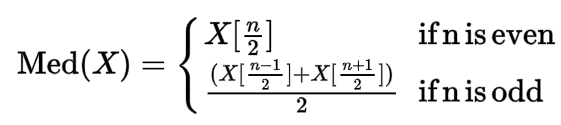
The formula for finding sigma:



The formula for finding mu:



* Treating the Null values using median and mode:



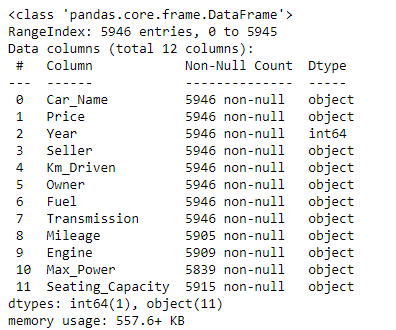


* Data Sources and their formats

I have scraped around 6000 used car data from cardekho.com.

It contains 12 columns in which 11 are independent variable and rest 1 is dependent variable.

The given data is given in form of numerical and categorical format.



* Data Preprocessing Done

In the given data there are some ambiguities present. In order to develop a machine learning model, data pre-processing is the most important step to build an effective machine learning model.

These are some data pre-processing which is done on our data:

* **Null Values:** This is the first thing that we should check in our data. Null values can be treated in two ways either replace it or remove it. If a column contains more than 75% null values then we should better remove it. And we replace the null value by using mean, median or mode of that column.
* **Outliers:** An outlier is any data point that is distinctly different from rest of your data points. When you are looking at a variable that is relatively normally distributed, you can think of outliers as anything that falls 3or more standard deviations from its mean.

There are three causes of outliers – data entry/ an experiment measurement errors, sampling problems, and natural variations.

* **Converting label column into numerical column:** Machine understand numbers, not text. We need to convert each text category to numbers in order for the machine to process them using mathematical equations. We use label encoding in our data pre-processing to convert the label data into numerical data.

Label Encoding**:** It is popular encoding technique for handling categorical variables. In this technique, each label is assigned a unique integer based on alphabetical ordering.

* Data Inputs- Logic- Output Relationships

In machine learning model the relation between input and output can be defined in 3 ways, either the relation is positive, negative or no relation. The type of relation between input and output can be find by using the correlation value. We can build the correlation matrix by using heatmap, if the value comes near to 1 then there is positive relationship, if the value comes near to -1 then there is negative relationship, and if the value comes near to 0 then there is no relationship.

* Hardware and Software Requirements and Tools Used

List of tools used in the project:

1. Scientific computing libraries: Pandas, Numpy, Scipy.
2. Visualization libraries: Matplotlib, Seaborn.
3. Algorithmic libraries: Scikit learn, Stats model.
4. Programming language: Python.
5. IDE used: Jupyter notebook.

**Model/s Development and Evaluation**

* Testing of Identified Approaches (Algorithms)

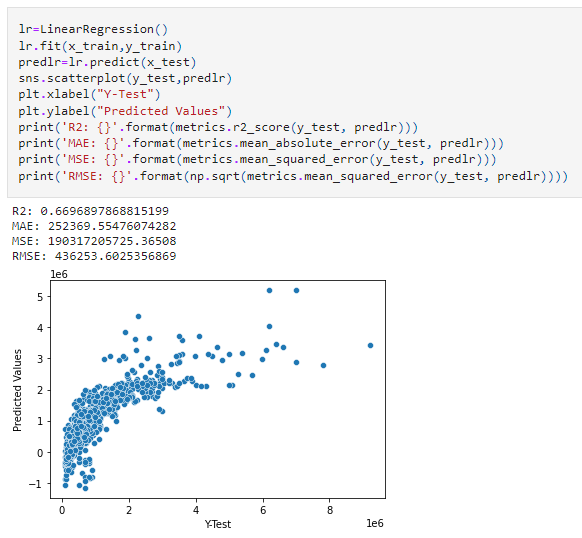
1. **Linear Regression:** Linear regression attempts to model the relationship between two variables by fitting a linear equation to observed data. One variable is considered to be an exploratory variable, and the other is considered to be a dependent variable.
2. **Elastic Net:** Elastic Net first emerged as a result of critique on lasso, whose variable selection can be too dependent on data and thus unstable. The solution is to combine the penalties of ridge regression and lasso to get the best of both worlds.
3. **Decision Tree Regression:** The decision tree is used to fit a sine curve with addition noisy observation. As a result, it learns local linear regression approximating the sine curve.

We can see that if the maximum depth of the tree (controlled by the max\_depth parameter) is set too high, the decision trees learn too find details of the training data and learn from the noise, i.e. they overfit.

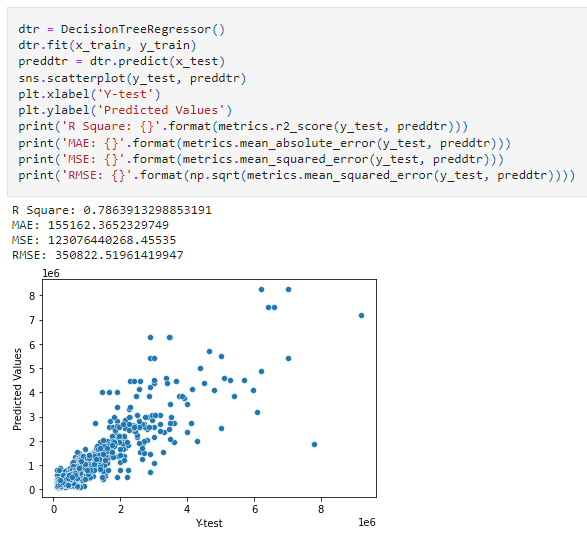
1. **Random Forest Regression:** It is a supervised learning algorithm that uses ensemble learning method for regression. Ensemble learning is a technique that combines predictions from multiple machine learning algorithm to make a more accurate prediction than a single model.
2. **AdaBoost Regression:** An adaboost regression is a meta-estimator that begins by fitting a regressor on the original and then fits additional copies of the regressor on the same dataset but where the weights of instances are adjusted according to the error of the current prediction.

* Run and Evaluate selected models

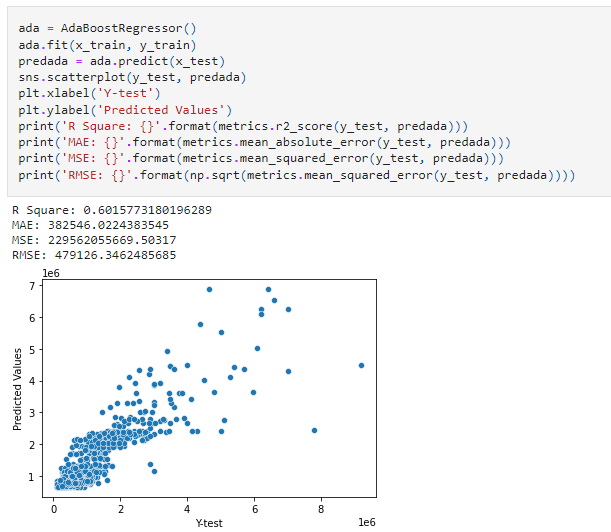
1. **Linear Regression:**



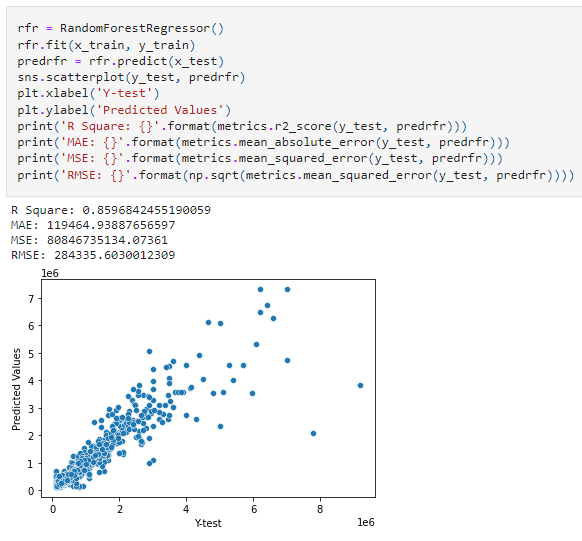
1. **Decision Tree Regression:**



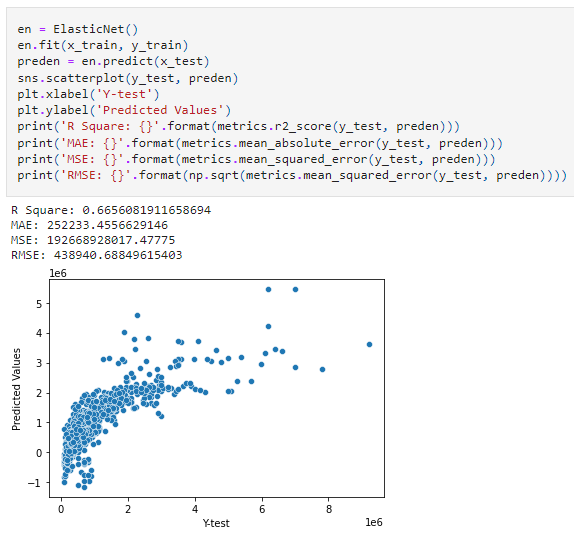
1. **AdaBoost regression:**

****

1. **Random forest regression:**



1. **Elastic Net regression:**



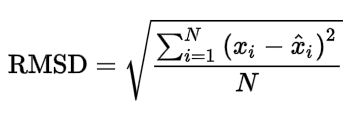
* Key Metrics for success in solving problem under consideration

This project is regression problem. These are the metrics used in project:-

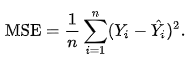
1. **Mean Absolute Error:** It is a measure of error between paired observations expressing the same phenomenon. Example of Y versus X include comparisons of predicted versus observed, subsequent time versus initial time, and one technique of measurement versus an alternative technique of measurement. MAE is calculated as:



1. **Root mean squared error:** It is frequently used measure of the differences between values (sample or population values) predicted by a model or an estimator and the values observed. RMSE is calculated as:

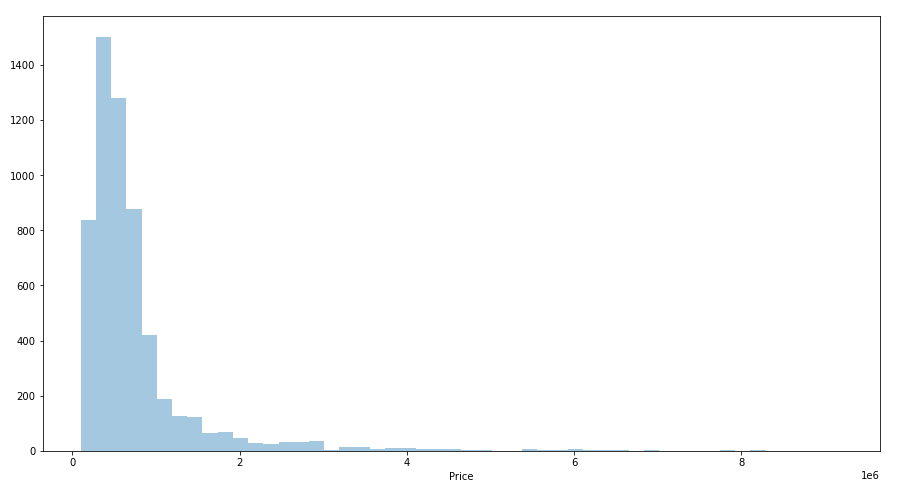
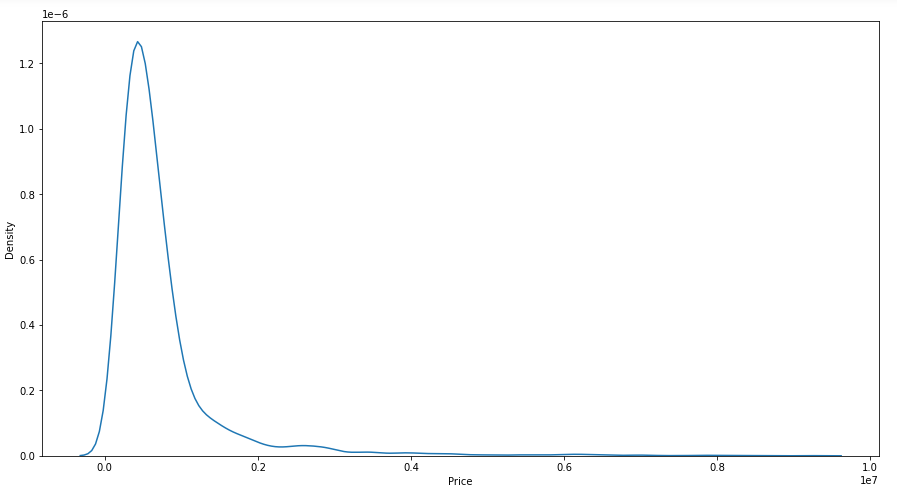


1. **Mean Squared Error:** It measures the average of the squares of the error, that is, the average squared difference between the estimated values and the actual value.MSE is a risk function, corresponding to the expected value of the squared error loss.

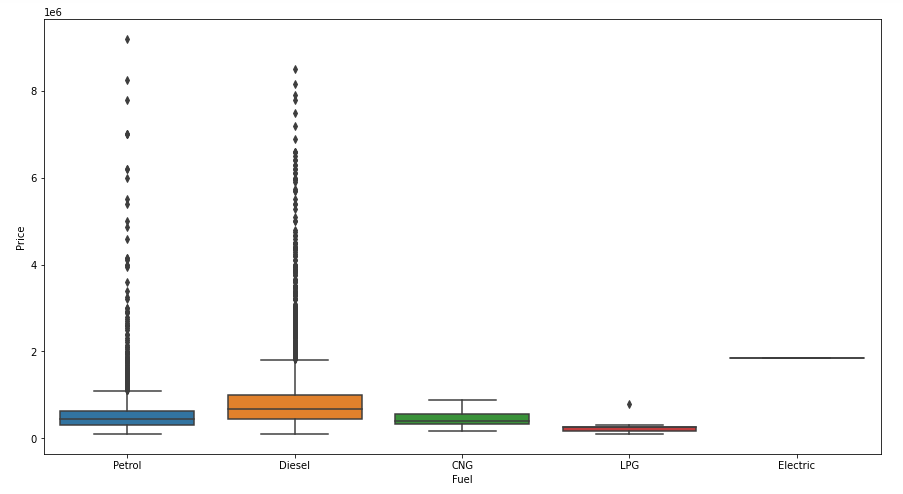


1. **R2 Score:** The main purpose of this metrics is either in prediction of future outcomes or the testing of hypotheses, on the basis of other related information. It provides a measure of how well observed outcomes are replicated by the model, based on the proportion of total variation of outcomes explained in the model.

* Visualizations

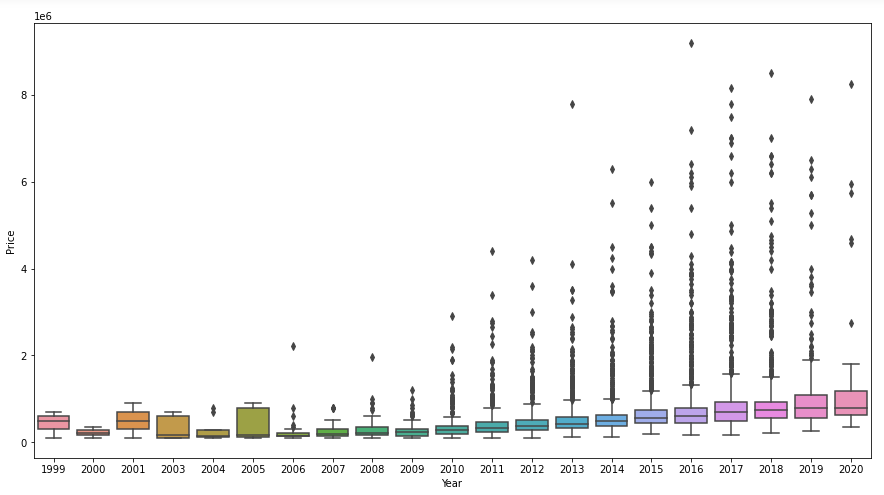
 

The two graphs mentioned above are density distribution plot (right) and count plot (left). Looking at the count plot, we can see that 1400 car has the price between 500,000.



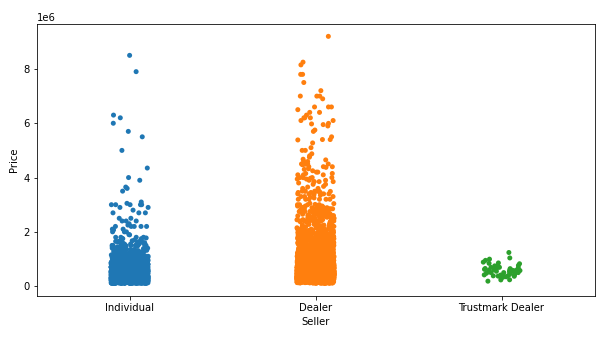
We can see in the above boxplot that diesel car would cost more followed

petrol.

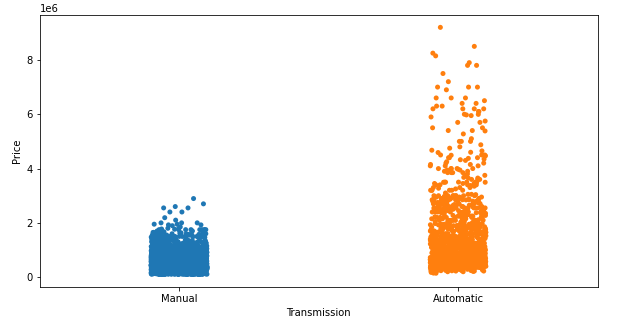


Clearly, it’s a strong tendency.

Pic of newest car is high compared to old car.



The price given by dealer seller is high as compared to others.



The automatic transmission type cars are more costly than manual transmission

type.

* Interpretation of the Results

In the given project, data pre-processing, data visualization and modelling is done.

1. We can conclude from the data pre-processing that there are many null values present in the dataset, if that null value is present with some actual true data then our model can perform even more accurate. The data contains more labelled data due to which we perform the label encoding on independent variable, but more often we use one hot encoding for the independent variable. The data could be more precise.
2. In modelling phase, we train our data with Linear Regression, Poisson Regression, Support vector regression, Ridge regression, Random forest regression, Adaboost regression. And perform a cross validation on these model. Out of these model support vector regression performs worst while Ridge regression performs very well. Then we perform the hyper parameter tuning of the best model.

**CONCLUSION**

* Key Findings and Conclusions of the Study

This study employs machine learning techniques, to develop a price prediction model for used car. It uses a rather large publicly available dataset of real used car transaction. The regression model performances of the model are compared with one another. The empirical result shows that the random forest algorithm provides superior performances for all metrics under study, the coefficient of determination of r2 score, and the computational time.

* Limitations of this work and Scope for Future Work

The Study can be enlarged in a subsequent research by increasing the dataset size so potentially uncovered details and feature of the dataset and of this study can be addressed. An increased dataset would potentially be good enough for employing deep neural networks, which can assure that more in-depth analysis on the housing price prediction can be performed. Then, the enlarged used car price prediction problem can be tackled as a classification problem.