

**USED CARS PRICE PREDICTION**

Submitted by:



DEEPTHI MALLU

**ACKNOWLEDGMENT**

I take this opportunity to acknowledge everyone who have helped me in every stage of this project.

Firstly, I am indebtedly grateful to my SME MR. Sajid Chowdary sir, who helped me from beginning of my Projects. Am also thankful to my Mentor Shankar Gowda Sir and my whole Data Trained team, where I have learnt Analysing the datasets and building the models using Machine learning and making the projects. Finally, am so thankful to my Flip Robo Technologies team, as they provided me the opportunity to work as intern in their company.

I feel pleasure, to make project report on “Used Cars Price Prediction”. It has been my privilege to have a team of project guide who have assisted me from the commencement of this project. The project is a result of my hard work, and determination put on by me with the help of web-scarping techniques, car selling websites, skikit-learn.org, Wikipedia and referred some old projects for completion of project.

**INTRODUCTION**

**Business Problem Framing**

In current generation car is being considered as status, whoever own’s it and also being considered as their essential part of their life. Cars have controls for driving, parking, passenger comfort, and a variety of lights. Over the decades, additional features and controls have been added to vehicles, making them progressively more complex, but also more reliable and easier to operate.

The price of a new car in the industry is fixed by the manufacturer with some additional costs incurred by the Government in the form of taxes. So, customers buying a new car can be assured of the money they invest to be worthy. But, due to the increased prices of new cars and the financial incapability of the customers to buy them, Used Car sales are on a global increase. [buying a pre-owned car](https://www.bankrate.com/loans/auto-loans/) rather than a new one can save you big bucks. new cars typically depreciate about 20 percent when they are driven off the lot. [Most cars will lose](https://www.bankrate.com/loans/auto-loans/) another 10 percent in value during the first year. That’s a 30-percent loss in value during the initial year of ownership. A 22,14,390 worth car loses roughly 6,64,317in value during that period. You can avoid that hit by buying a one-year-old used car. He cautions that one-year-old used cars are hard to come by, but it’s becoming easier to find two- and three-year-old used cars.

Therefore, there is an urgent need for a Used Car Price Prediction system which effectively determines the worthiness of the car using a variety of features. Existing System includes a process where a seller decides a price randomly and buyer has no idea about the car and its value in the present-day scenario. In fact, seller also has no idea about the car’s existing value or the price he should be selling the car at.

With the impact of covid19 in the market, lot of changes took place in the car market. Now some cars are in demand hence making them costly and some are not in demand hence cheaper. With the change in market, our clients are facing problems with their old data. So, in order to get new data, I have used Web scraping for scarping the details and applied machine learning models on new data.

To overcome this problem, we have developed a model which will be highly effective. Regression Algorithms are used because they provide us with continuous value as an output and not a categorized value. Because of which it will be possible to predict the actual price a car rather than the price range of a car. User Interface has also been developed which acquires input from any user and displays the Price of a car according to user’s inputs.

**Conceptual Background of the Domain Problem**

A car (or automobile) is a wheeled [motor vehicle](https://en.wikipedia.org/wiki/Motor_vehicle) used for [transportation](https://en.wikipedia.org/wiki/Transportation). Most definitions of cars say that they run primarily on roads, seat one-to-eight people, have four [wheels](https://en.wikipedia.org/wiki/Wheels) and mainly transport people rather than goods.

There are costs and benefits to car use. The costs to the individual include acquiring the vehicle, interest payments (if the car is financed), repairs and [maintenance](https://en.wikipedia.org/wiki/Auto_maintenance), fuel, [depreciation](https://en.wikipedia.org/wiki/Depreciation), driving time, [parking fees](https://en.wikipedia.org/wiki/Parking_fee), taxes, and insurance. The costs to society include [maintaining roads](https://en.wikipedia.org/wiki/Maintaining_road), [land use](https://en.wikipedia.org/wiki/Land_use), [road congestion](https://en.wikipedia.org/wiki/Road_congestion), [air pollution](https://en.wikipedia.org/wiki/Air_pollution), [public health](https://en.wikipedia.org/wiki/Public_health), healthcare, and [disposing of the vehicle at the end of its life](https://en.wikipedia.org/wiki/Vehicle_recycling). [Traffic collisions](https://en.wikipedia.org/wiki/Traffic_collisions) are the largest cause of injury-related deaths worldwide.

Personal benefits include on-demand transportation, mobility, independence, and convenience. Societal benefits include economic benefits, such as job and wealth creation from the [automotive industry](https://en.wikipedia.org/wiki/Automotive_industry), transportation provision, societal well-being from leisure and travel opportunities, and revenue generation from [taxes](https://en.wikipedia.org/wiki/Category:Vehicle_taxes). People's ability to move flexibly from place to place has far-reaching implications for the nature of societies. There are around 1 billion cars in use worldwide. Car usage is increasing rapidly, especially in [China](https://en.wikipedia.org/wiki/Automotive_industry_in_China), [India](https://en.wikipedia.org/wiki/Automotive_industry_in_India) and other [newly industrialized countries](https://en.wikipedia.org/wiki/Newly_industrialized_country).

For the last 80 plus years the car industry has changed the way we manoeuvre around the world. From the first sets of horseless carriages to cars that can reach top speeds that were previously never thought possible, automobiles have come a long way from when they were first introduced and mass-produced. Automobiles come in all shapes and sizes, personalized to fit every need of the driver now. The invention of the first automobile opened up a whole new way of living for so many people, making things more convenient and places easily accessible that would have taken a lot longer to get to in a carriage or on foot.

Car Price Prediction is a really an interesting machine learning problem as there are many factors that influence the price of a car in the second-hand market. In this competition, we will be looking at a dataset based on sale/purchase of cars where our end goal will be to predict the price of the car given its features to maximize the profit.

**Review of Literature**

Cars came into global use during the 20th century, and [developed economies](https://en.wikipedia.org/wiki/Developed_country) depend on them. The year 1886 is regarded as the birth year of the car when German inventor [Karl Benz](https://en.wikipedia.org/wiki/Karl_Benz) patented his [Benz Patent-Motorwagen](https://en.wikipedia.org/wiki/Benz_Patent-Motorwagen) Cars became widely available in the early 20th century. One of the first cars accessible to the masses was the 1908 [Model T](https://en.wikipedia.org/wiki/Ford_Model_T), an American car manufactured by the [Ford Motor Company](https://en.wikipedia.org/wiki/Ford_Motor_Company). Cars were rapidly adopted in the US, where they replaced [animal-drawn](https://en.wikipedia.org/wiki/Draft_animal) [carriages](https://en.wikipedia.org/wiki/Carriage) and carts. In Europe and other parts of the world, demand for automobiles did not increase until after [World War II](https://en.wikipedia.org/wiki/World_War_II).

Cars have controls for driving, parking, passenger comfort, and a variety of lights. Over the decades, additional features and controls have been added to vehicles, making them progressively more complex, but also more reliable and easier to operate. These include rear-reversing cameras, [air conditioning](https://en.wikipedia.org/wiki/Automobile_air_conditioning), [navigation systems](https://en.wikipedia.org/wiki/Automotive_navigation_system), and [in-car entertainment](https://en.wikipedia.org/wiki/In-car_entertainment). Most cars in use in the early 2020s are propelled by an [internal combustion engine](https://en.wikipedia.org/wiki/Internal_combustion_engine), fuelled by the [combustion](https://en.wikipedia.org/wiki/Combustion) of [fossil fuels](https://en.wikipedia.org/wiki/Fossil_fuel). [Electric cars](https://en.wikipedia.org/wiki/Electric_car), which were invented early in the [history of the car](https://en.wikipedia.org/wiki/History_of_the_automobile), became commercially available in the 2000s and are predicted to cost less to buy than gasoline cars before 2025. The transition from fossil fuels to electric cars features prominently in most [climate change mitigation scenarios](https://en.wikipedia.org/wiki/Climate_change_mitigation_scenarios), such as [Project Drawdown](https://en.wikipedia.org/wiki/Project_Drawdown)'s 100 actionable solutions for climate change.

**Motivation for the Problem Undertaken**

The pre-owned car market is absolutely booming. Recent statistics have it that the total annual revenue for used car dealership in the United States of America is US0 billion with an annual growth of 2.4 percent. With massive annual sales of nearly US$370 billion, one can confidently say that the used vehicle dealership industry represents almost half of the United States’ auto retail market and it is of course the largest retail segment of the U.S. economy.

India’s used car market is projected to reach over $ 66 billion by 2022, on the back of growing population and rising urbanization in the country. Increasing focus of automakers towards setting up used car networks in different parts of the country and growing inclination of consumers towards used cars owing to their affordability and improved after sales services are some of the other major factors expected to boost demand for used cars in India in the coming years.

Moreover, market growth is anticipated to be driven by rising penetration of online platforms such as OLX, Car Dekho, Cars24 etc., that enable used car dealers to boost their reach to a larger audience. So, as it is on growing business, I want to predict the used car prices and give correct information for the people in order to sell or buy a used car from which we can save huge bugs. Based on existing data, the aim is to use machine learning algorithms to develop models for predicting used car prices.

**Analytical Problem Framing**

**Mathematical/ Analytical Modelling of the Problem**

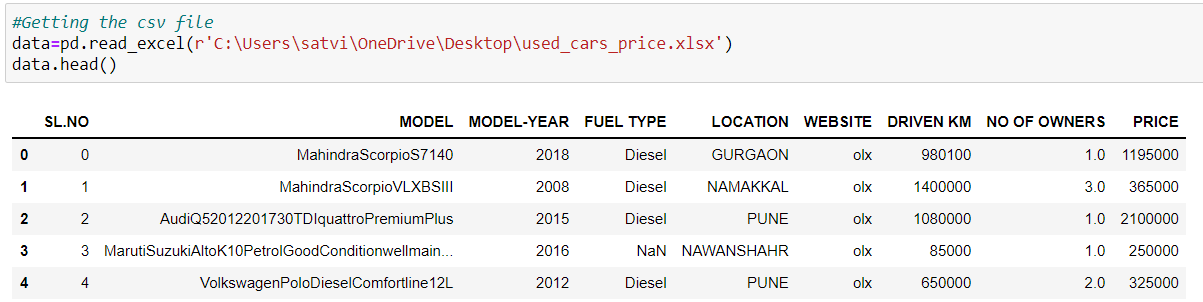
In this project, I have collected the details of used cars from different websites like olx, cardekho, cars24 so for collecting the data I have used web-scraping, selenium for scraping the details. Then I have collected all the data from different locations in India and collected the details like car model, Location, kilometres driven, type of fuel used, price etc. So, by using all the information of the data collected we need to predict our label price. After collecting the data, I have put all together in a data frame and saved the data as excel file.

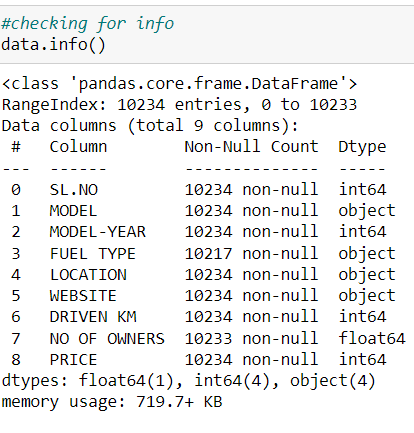
Then by using my Jupiter notebook I have imported required libraries like pandas, numpy, matplotlib, seaborn where we use them for our problem using then I imported my collected excel file data and checked for top 5 rows using head method and then I have checked for the shape of the data. From shape method I got to know that there are 10234 rows and 9 columns. Then I checked for is null method to find the NaN’s then there are very few null values in the columns and then I tried filling them and checked for type of the data in the columns using info method. Then I checked for the describe method and then plotted some graphs and visualized and tried removing outliers from the data and then scaled the data. As the label is a continuous variable, I have used regression models for predicting our label. I have used various algorithms for checking the patterns and concluded a final model on the basis of performance and evaluated the model.

**Data Sources and their formats**

By using pandas, I have first imported the Excel file and it consists of different columns which includes data in it. Our dataset consists of Features and label. After importing I have checked for shape of the dataset and which consists of rows and columns. Then I checked for null values and need to be treated and then I checked for info () method for knowing the type of the data then I checked for stats using describe method.

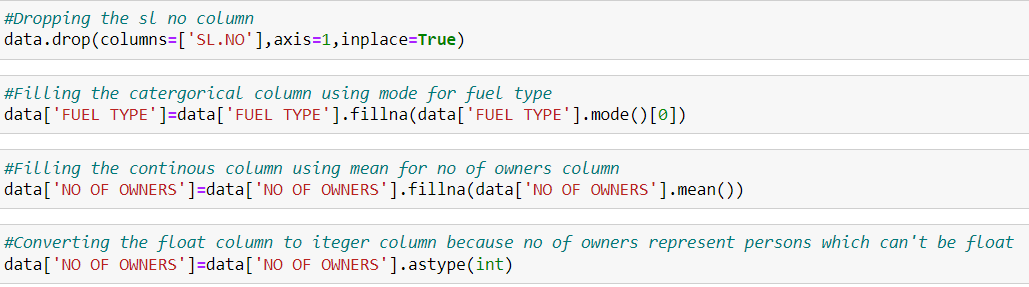
Our label is price prediction which is a continuous variable based on the values of independent variables our dependent variable depends.





**Data Pre-processing Done**

There are few null values present in the dataset by using some of the imputing techniques, we need to fill the NaN values using mean or mode method on the type of the data in the columns. For the categorical variables I have use mode method to fill and for continuous columns I used mean to fill the NaN’s.



**Data Inputs- Logic- Output Relationships**

For checking the relation between the columns, I have used correlation matrix to find the relation and plotted heat map to visualise the percentage of the correlation. The below are the observations from the heatmap.

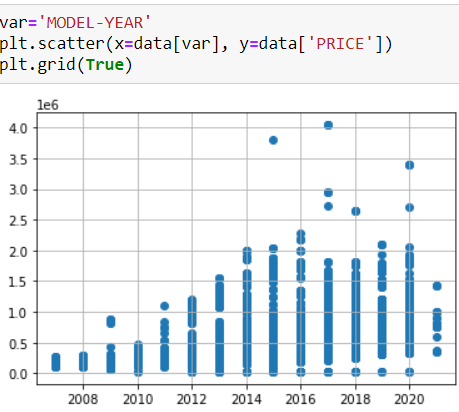
1.The Dark blue indicates high correlation and light blue indicates less correlation.

2.Our label is highly correlated with model-year, and very less correlated with other cities

3.Even some of the features are correlated with the other features like driven km is correlated with model year, website is correlated with jaipur, etc.

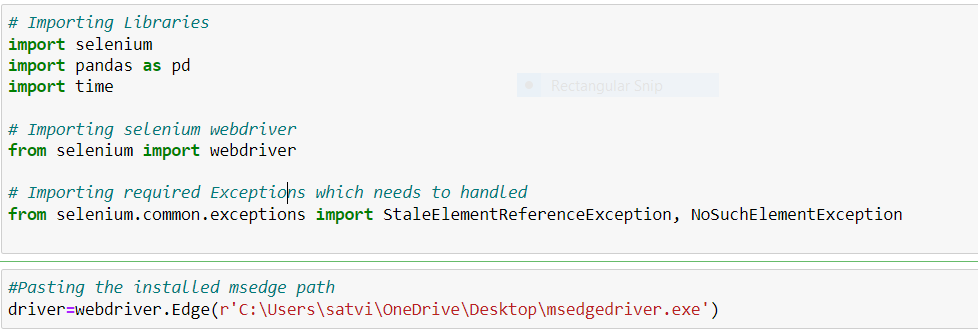


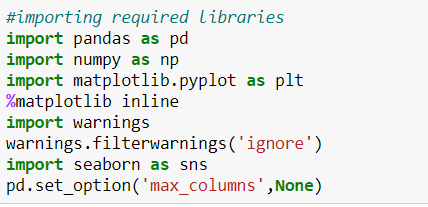
I have plotted scatter plots and checked the correlation between the features and the label.

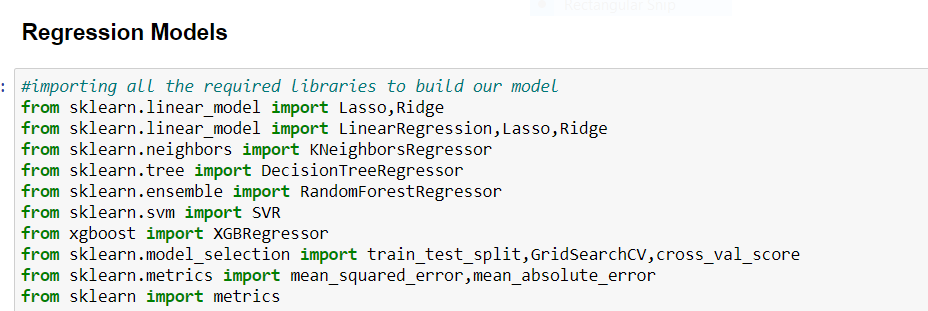


**Hardware and Software Requirements and Tools Used**

I have used my laptop, wed server, micro-soft edge, Jupiter Notebook which is having GUI interface. Imported necessary libraries from python such as pandas, NumPy, seaborn, matplotlib, then imported the required model libraries from Scikit learn to import our algorithms.



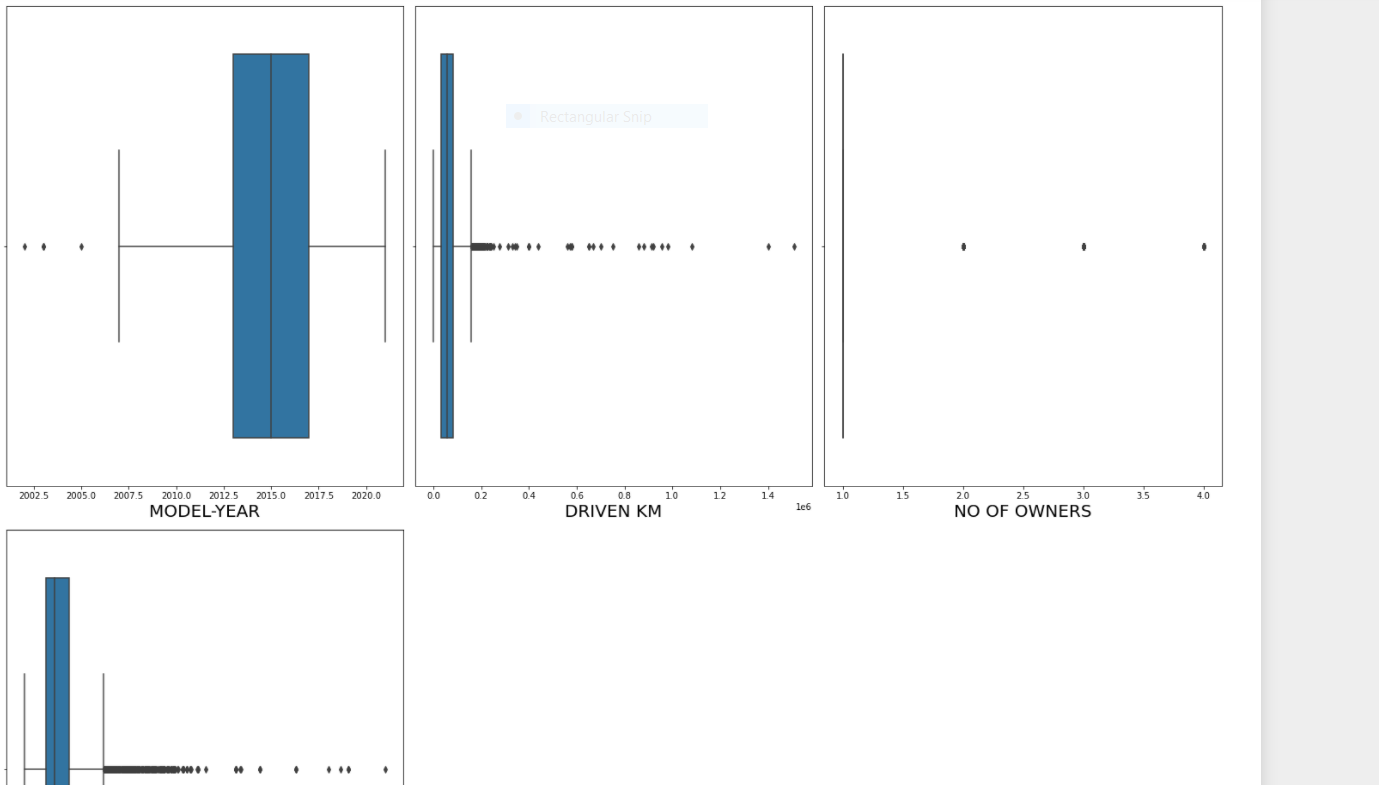




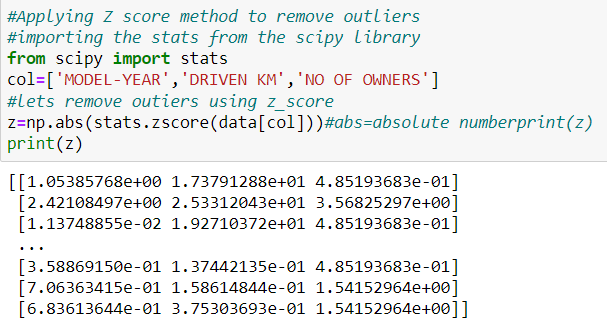
**Model/s Development and Evaluation**

I have plotted box plots to check for outliers and distribution plots to check the skewness so I found the presence of outliers and skewness in the continuous data. So, I have used Z-Score method for removing the outliers and log transformation method the remove the skewness.

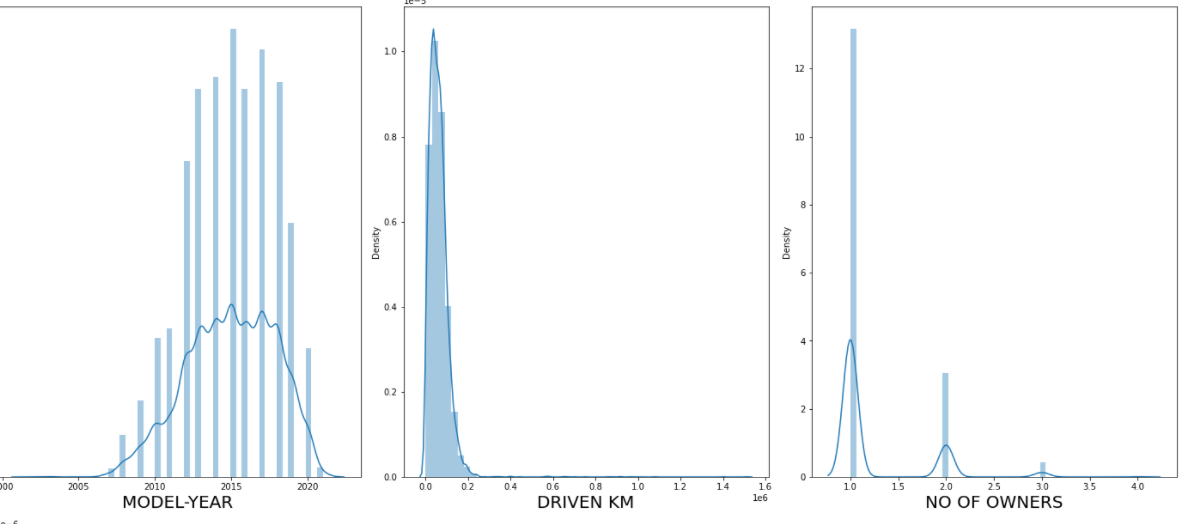
**Box Plots**:



**Z-Score**



**Distribution Plots** are used to check the flow of the data in the columns.

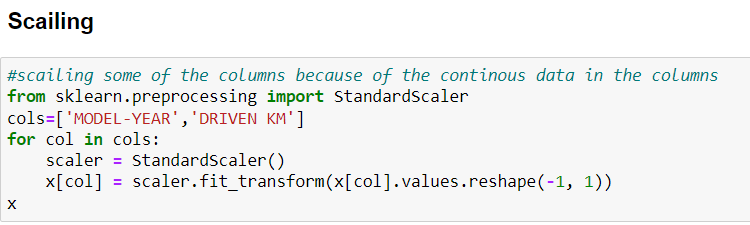


**Testing of Identified Approaches (Algorithms)**

The price of the used cars prediction is a numerical variable so it comes under regression problem, So I have used 8 different algorithms to check the model patterns. In order to final a model we have to check on different evaluation metrics like finding the score of the training data and testing data and finding the errors like mean absolute error (MAE), mean squared error (MSE) and Root mean squared error (RMSE). In order to tell a model is good their RMSE value should be as less as possible then we can say the model is efficiently working on the given data.

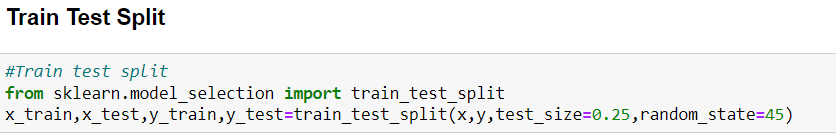
**Scaling the data**

I have used standard Scaler to scale the data**.**



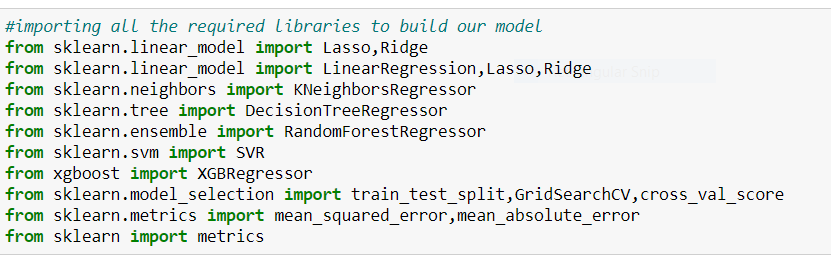
**Train Test Split**

I have imported the train\_test\_split from the module sklearn from model\_selection. And used 75% of the data for training and 25% of the data for testing and splitted the data into x-train, x-test, y-train, y-test.



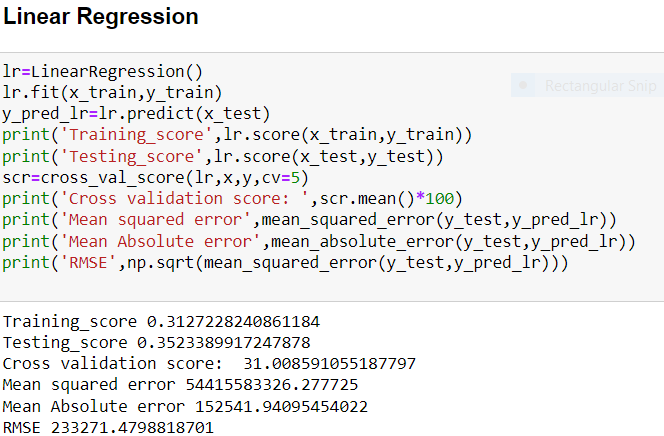
**Regression Algorithms used for our prediction**

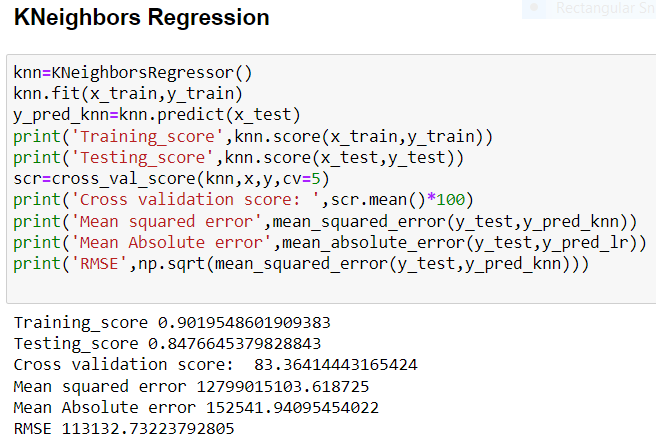
Since our price prediction is a continuous variable then this comes under Regression problem so I have used different Regression algorithms for predicting our label.

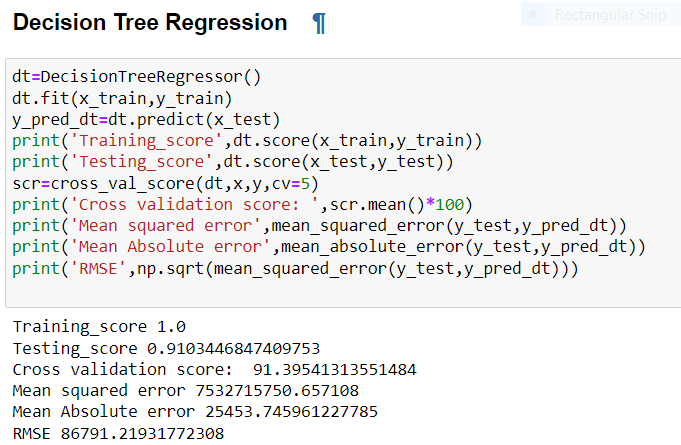


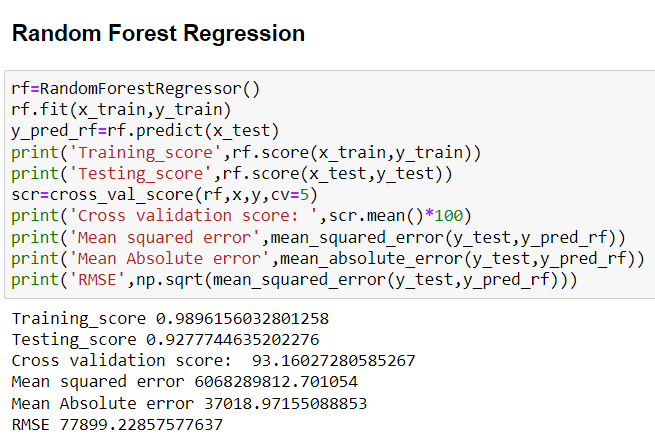
**Run and Evaluate selected models**

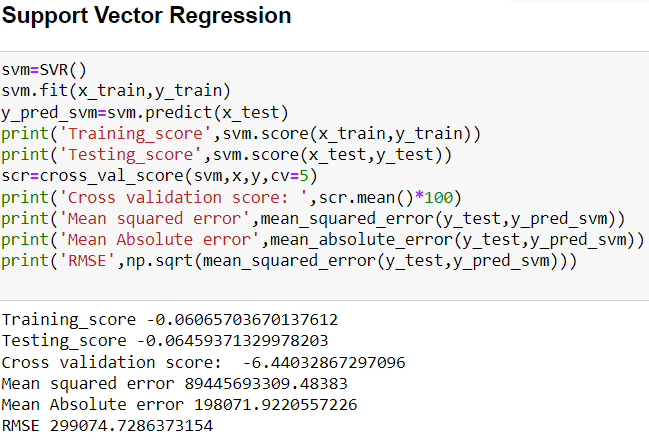
I have used various algorithms for predicting our label like Linear Regression, KNeighbors Regression, Decision Tree Regression, Random Forest Regression, Support Vector Regression, Ada Boost Regression, Lasso Regression. For evaluating the model, I have used Mean Squared Error (MSE), training score, testing score and root mean squared root (RMSE).

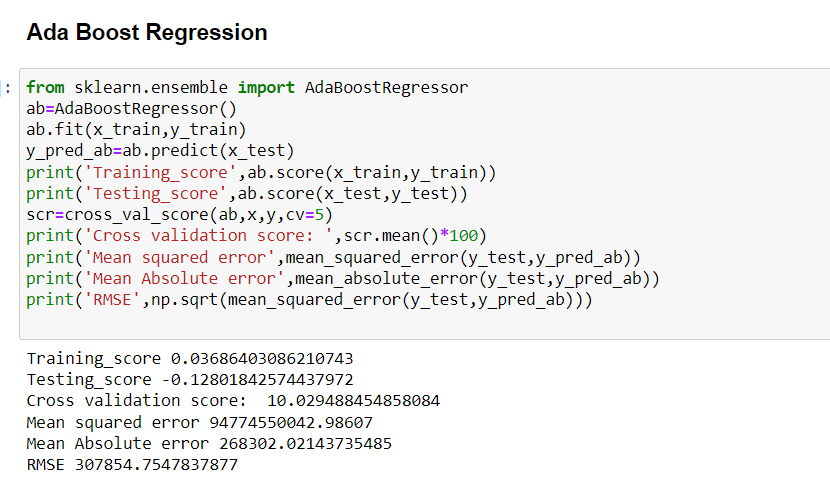












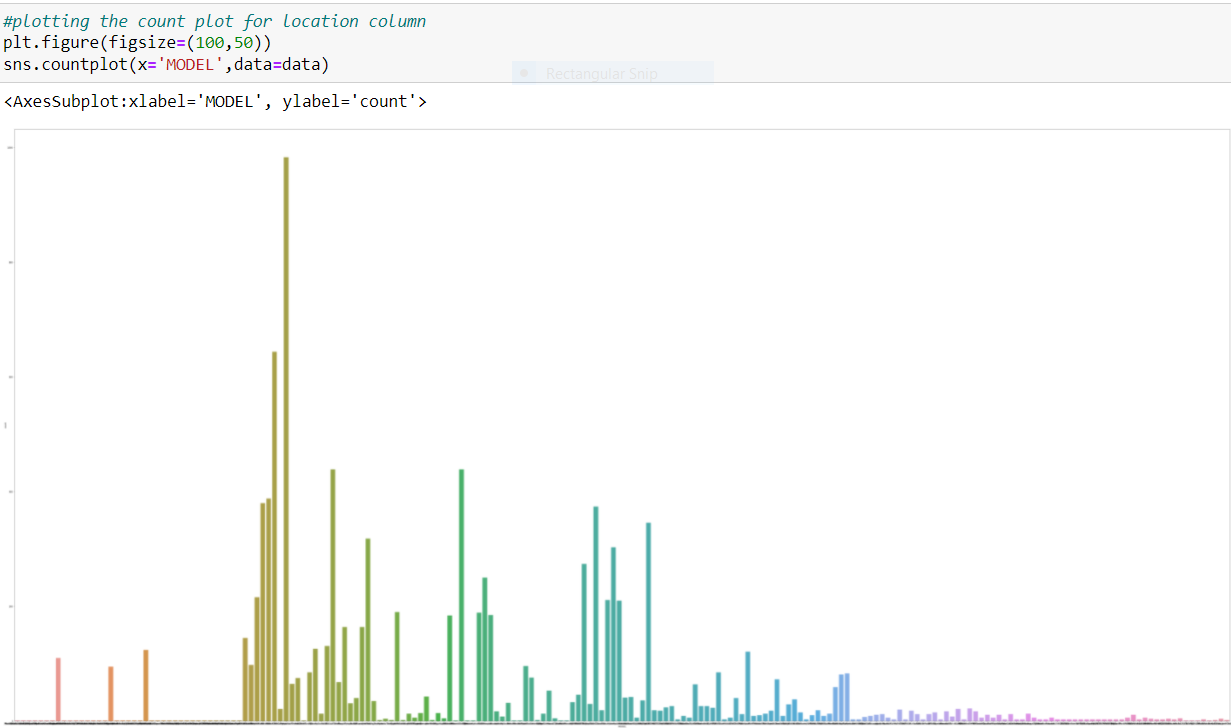
**Key Metrics for success in solving problem under consideration**

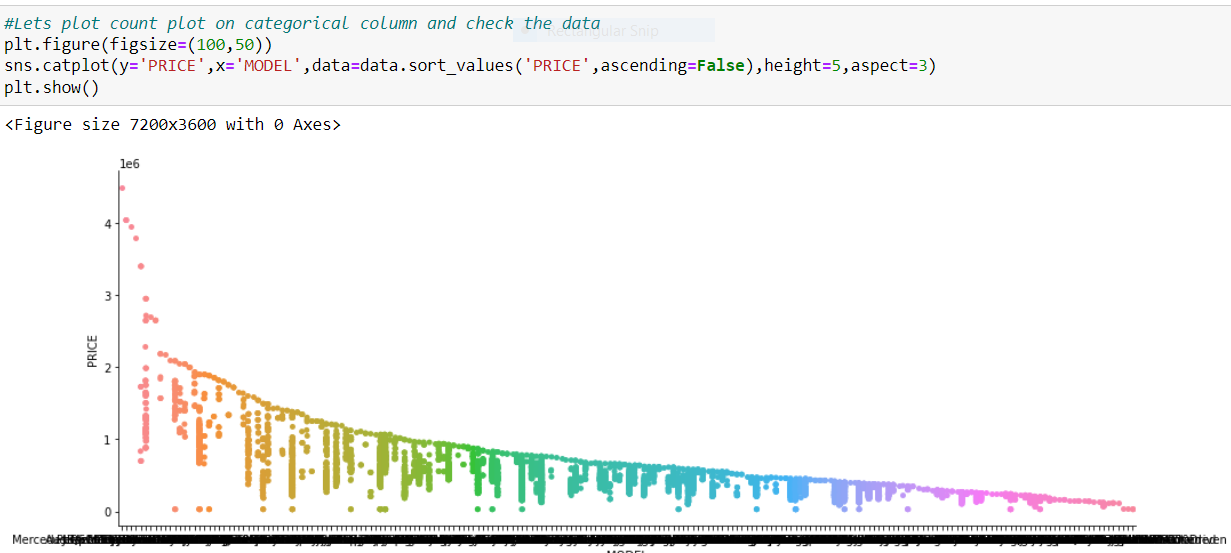
After Checking the model evaluation, I have considered Random Forest Regression as my final model and I have applied hyperparameter tuning to improve the score of the model. By using Grid search cv we are going to pass different parameters for the final algorithm which improves the score of the model and reduction in errors. But after applying grid search cv there is no change in the score and error. So, I have considered old values as for checking the model performance.

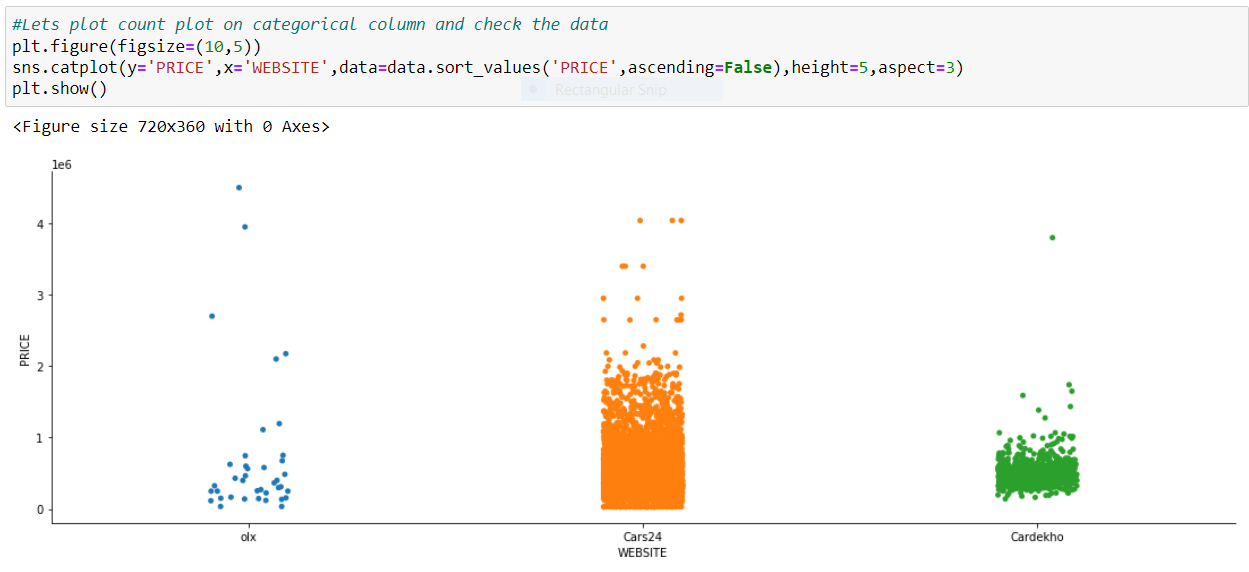


**Visualizations**

I have plotted count plots on categorical columns and checked the type of classes present in the columns and even plotted cat plots between the features and the label and found the relation and tried to know how label is varying with the features.

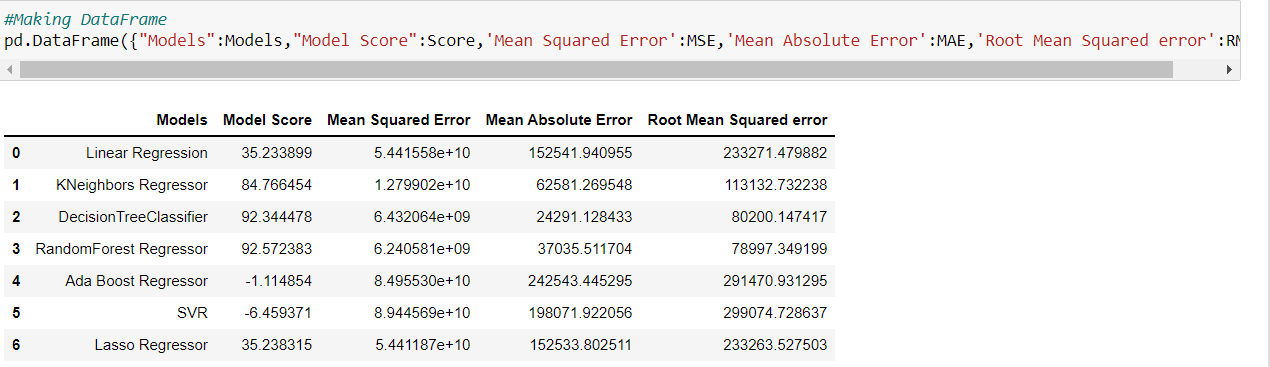






**Interpretation of the Results**

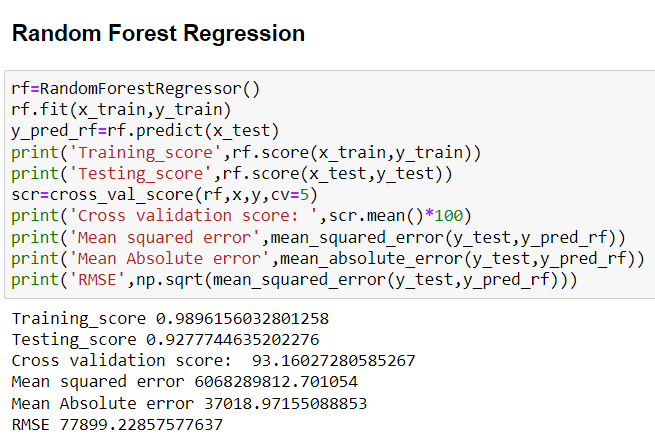
I have created a data frame using all the models used for prediction, their scores and errors given by the models as shown in the below screenshot.

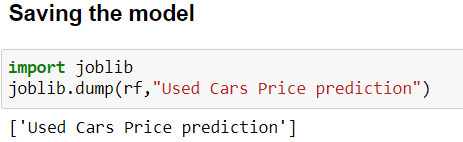


**CONCLUSION**

**Key Findings and Conclusions of the Study**

I have used various models for predicting the price of used cars and used various evaluation metrics for evaluating the model like finding the training score, testing score, Mean Squared error (MSE), Mean Absolute error (MAE), Root Mean squared error (RMSE). So, after evaluation metrics on different models, Random Forest regressor giving high accuracy and low RMSE Value. So I finalised the model and saved the model using joblib library.



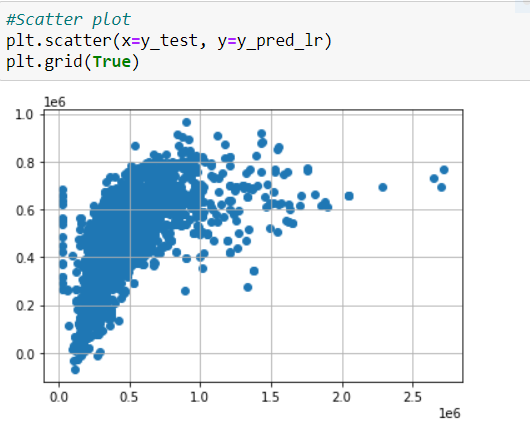


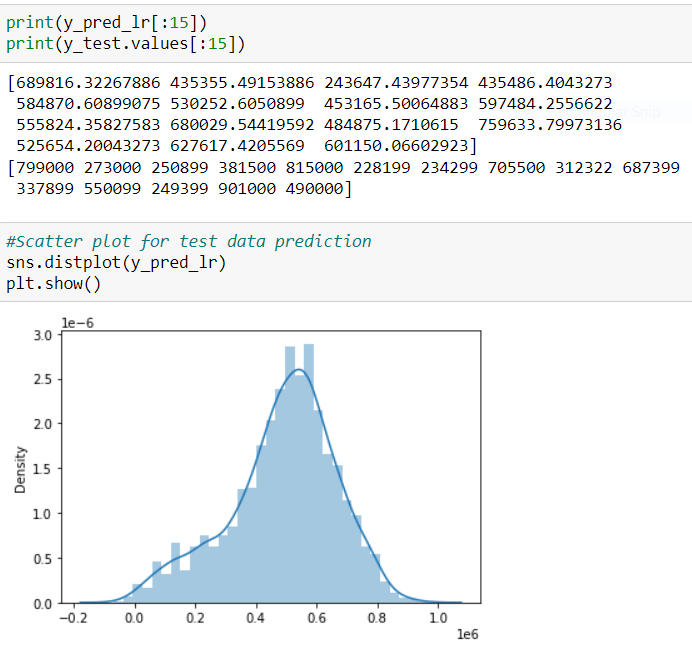
**Conclusions on our model building**

We got our best model as Random Forest Regressor with the score of 93% and both the training and testing scores are balanced and the RMSE value is also less compared to all other models. So, we can go further build our model using Random Forest Regressor.

**Learning Outcomes of the Study in respect of Data Science**

After finalising the model Random Forest Regressor, I have taken the values of prices which are predicted by the model and compared with the actual Price values and checked the relation between them by plotting the scatter plot and plotted distribution plot for predicted value and checked the data distribution and visualised the predicted price values are normally distributed.





**Limitations of this work and Scope for Future Work**

In the pre-owned car price prediction, I have performed many Regression algorithms to predict the used cars price. The Root Mean squared error (RMSE) errors calculated for all the algorithms are very high. Statistical methods work better, on large set data. But the length of the dataset is very less so using different methods that match the time-series data will be used in the future research to obtain smaller error prediction values (RMSE) and using more data to get the better result. In future this machine learning model may bind with various website which can provide real time data for price prediction. Also, we may add large historical data of car price which can help to reduce the RMSE error.