

**Project on**

**Car Price Prediction**

Submitted by:

Arpita Rai

Int\_33

**ACKNOWLEDGMENT**

I would like to thank Flip Robo Technologies, for giving me this opportunity to work on this project. I got to learn more from this project about Data Scraping, and practical implementations of using machine learning modules.

I take this opportunity to express my gratitude and regards to my mentor Mr. Shwetank Mishra for his guidance, monitoring and constant encouragement by giving new projects. The help and guidance given by him time to time shall carry me a long way in the journey of life on which I am about to embark.

Lastly, I thank almighty, my parents, brother, sister and friends for their constant encouragement without which this assignment would not be possible.

**INTRODUCTION**

* Business Problem Framing

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. One of our clients works with small traders, who sell used cars. With the change in market due to covid 19 impact, our client is facing problems with their previous car price valuation machine learning models. So, they are looking for new machine learning models from new data. We have to make car price valuation model.

The primary point of this venture is to create a dataset with the help of web scraping and anticipate the cost of trade-in vehicle in view of different elements.

* Conceptual Background of the Domain Problem

The increased preference for personal mobility due to safety concerns amid the Covid-19 pandemic has led to supply constraints in the used car market, as people are holding on to their cars longer.

Meanwhile, demand for used cars has increased rapidly compared to pre-Covid times, and this mismatch between demand and supply has led to a 2-7 per cent increase in the price of used cars, said experts.

While the pandemic has led to an increased requirement of personal mobility, this has been met with lower expected income amid the economic distress, said Shashank Srivastava, Executive Director (Marketing and Sales), Maruti Suzuki India. This has led to a downward telescoping of demand amid the pandemic, resulting in greater demand for used cars, whilst also leading people to not sell their old cars in the market, which is the source of supply in the pre-owned car market, he explained.

* Review of Literature

With the recent arrival of internet portals, buyer and seller may obtain an appropriate status of the factors that ascertain the market price of a used automobile.

We have tried to develop statistical models that can forecast the value of a pre-owned automobile based on prior customer details and different parameters of the vehicle. This projects aims to compare the efficiency of different models’ prediction to find the appropriate one. On the subject of automobile price prediction, several previous studies have been conducted.

I have collected the data from different automobile websites which are-

1. Car Dekho
2. Cars24
3. Olx

* Motivation for the Problem Undertaken

**Objective of the Project**

1. **Data Collection-** To scrape the data of at least 5000 cars from various websites like Olx, cardekho, Cars24 etc.
2. **Model Building-** Need to build a machine learning model for forecasting value of vehicle based on multiple attributes.

**Analytical Problem Framing**

* Data Sources and their formats

We have scraped 4340 data of cars from websites like Olx, Car Dekho, Cars 24,etc., and we have saved it for Machine Learning Model.

We have extracted attributes like Brand name and model along with its manufacturing year, seller type, and number of owners, fuel type, transmission, kilometres driven and lastly price.

Price is our target column. Year, price and kilometre driven are int type and rest are object data type.

* Mathematical/Analytical Modeling of the Problem

We find out the attributes like brand, model, manufacturing year, varient, total driven kilometers, fuel type and seller type have very wide range of variables and it’s not very smart to study their plots. We do not need to perform bivariate analysis on these.

* Data Preprocessing Done

We observe that there are no missing values in any column. We didn’t check for skewness as most of the columns are of object type.

* Hardware and Software Requirements and Tools Used

**Python** is the most popular technology for implementing machine learning ideas, owing to the fact that it has a large number of built-in algorithms in the form of bundled libraries.

The following are some of the most important libraries and tools we used in our project:

1. **Numpy:** NumPy is a Python module for array processing. It is the most important Python module for scientific computing. NumPy may be used as a multi-dimensional container of general data in addition to its apparent scientific applications.

It allows any data types to be created, allowing NumPy to connect with a broad range of databases cleanly and quickly.

1. **Jupyter Notebook**: Jupyter Notebook is an open-source online software that lets you create and share documents with live code, equations, visualisations, and narrative prose. Data cleansing and transformation, numerical simulation, statistical modelling, data visualisation, machine learning, and more are all included.
2. **Sklearn**- Power transform, label encoder, standard scaler, linear, random forest, decision tree, Gradient boosting Regressor, k-nearest neighbours, r2 score, mean absolute error, mean squared error, train test split, grid search cv and ensemble technique.
3. **Chromedriver**: Chromedriver is used to web scrape the data and automate the process. I have scraped the data using Selenium python.

**Model/s Development and Evaluation**

* Identification of possible problem-solving approaches (methods)

We go over many techniques and datasets that was used to create this module.

This model is trained by using a dataset comprising over 4500 tuples. The value of a car is determined by factors such as the number of kilometres driven, the year of registration, the kind of gasoline used, and the financial strength of the owner. We created regressor methods and compared the two on different car models because this is a regression problem.

Anaconda seeks to address Python's dependency hell, where distinct projects have various dependency versions, so that project dependencies do not require separate versions, which might conflict.

* Testing of Identified Approaches (Algorithms)

The models used training and testing datasets are as followed:

1. Linear Regression

2. SGD Regressor

3. KNeighbors Regressor

4. Decision Tree Regressor

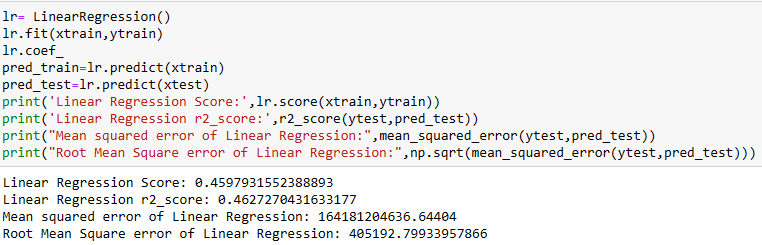
5. Random Forest Regressor

* Run and Evaluate selected models

1. **Liner Regression**

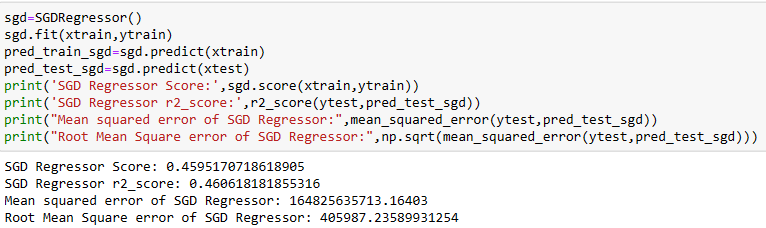
Linear regression is a type of regression analysis in which the independent(x) and dependent(y) variables can be constrained in a linear relationship.

The method is commonly used to predict and calculate correlations between independent and dependent variables. The regression model establishes a linear or exponential connection between independent and dependent variables.

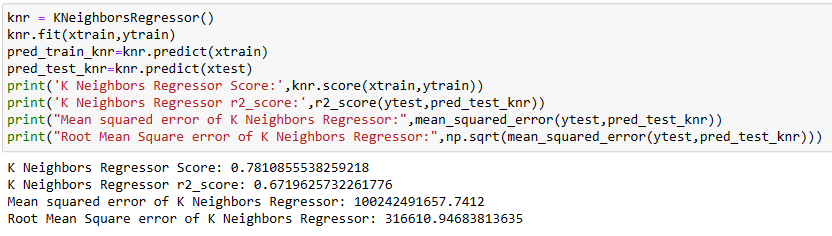
****

1. **SGD Regressor**: The loss gradient is calculated each sample at a time, and the model is updated along the way using a decreasing strength schedule. SGD stands for Stochastic Gradient Descent (aka learning rate).

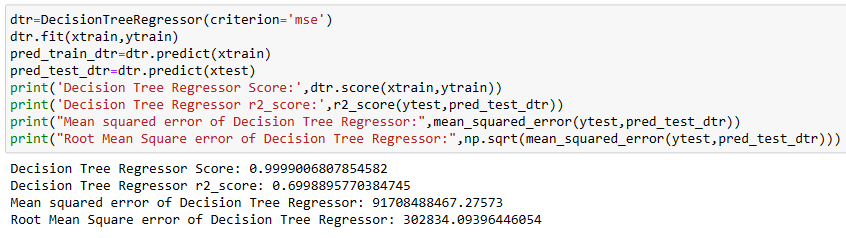
The regularizer is a penalty applied to the loss function that decreases model parameters towards zero using either the squared euclidean norm L2 or the absolute norm L1 or a mix of the two (Elastic Net). The update is trimmed to 0.0 whenever the parameter update passes the 0.0 value due to the regularizer, allowing for the learning of sparse models and online feature selection.



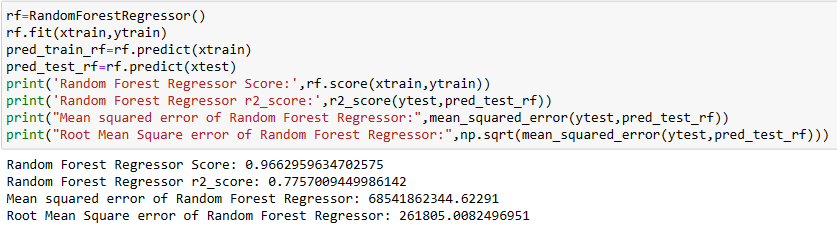
1. **K Neighbors Regressor**: Algorithm Calculating the average of the numerical goal of the K nearest neighbours is a straightforward implementation of KNN regression. An inverse distance weighted average of the K closest neighbours is another method. The distance functions used in KNN regression are the same as those used in KNN classification.



1. **Decision Tree Regressor**: Classification trees are tree models in which the goal variable can take a discrete set of values; in these tree structures, leaves indicate class labels and branches represent feature combinations that lead to those class labels. Regression trees are decision trees in which the target variable can take continuous values (usually real numbers). The objective is to build a model that predicts the value of a target variable from a set of input variables.

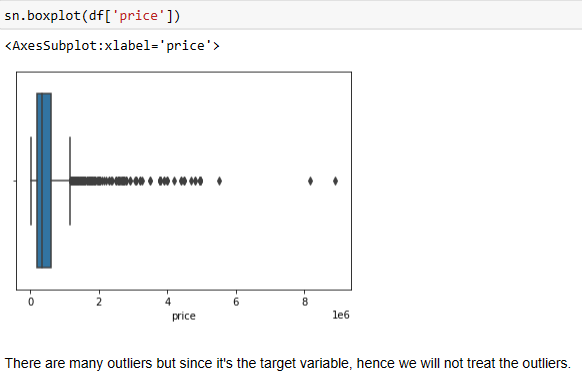


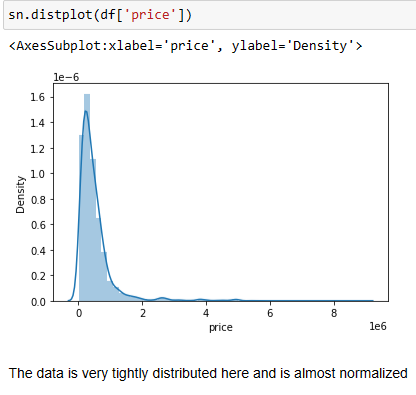
1. **Random Forest Regressor**: A regressor with a random forest. A random forest is a meta estimator that employs averaging to increase predicted accuracy and control over-fitting by fitting a number of classification decision trees on various sub-samples of the dataset.



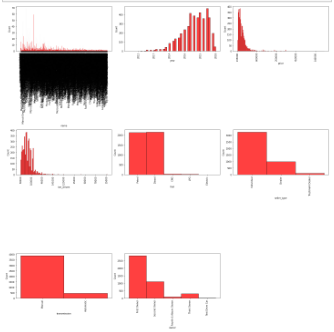
* Visualizations

We have plotted histograms and distribution plot in univariate analysis, which interpreted that all the columns are equally important but the columns like brand, variant, location, date and total driven kilometres have a wide range of data spread hence we will not perform it’s bivariate analysis.

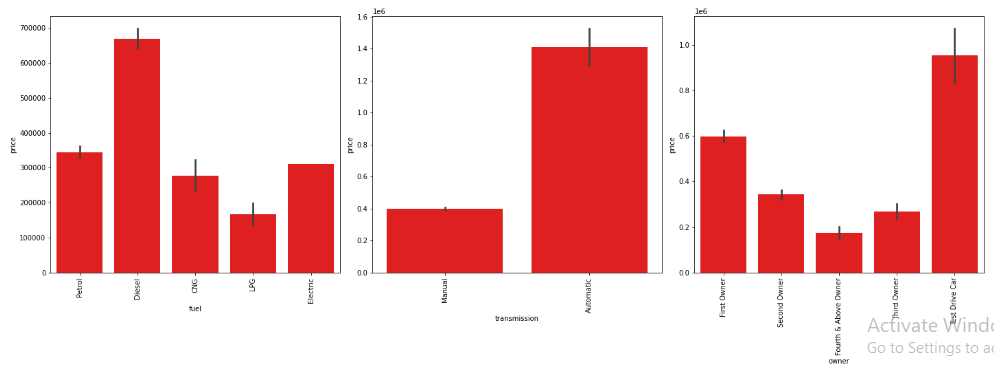


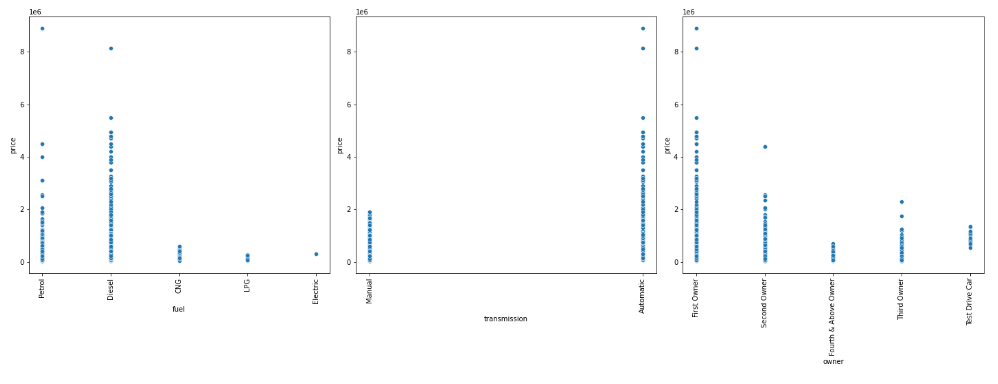


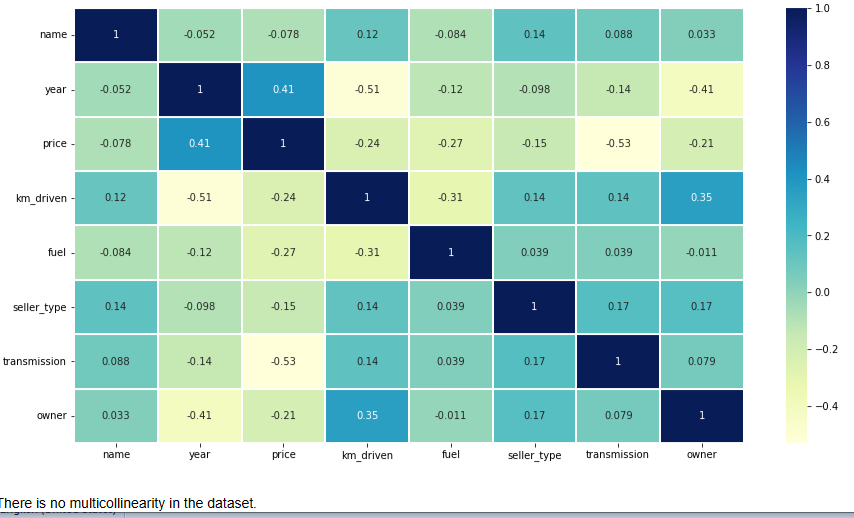
**Histogram**



From bivariate analysis we conclude that, Since Brands, Varients and Driven Kilometres have a wide range of values in them, we will not perform bivariate analysis for them as they will not give us any specific details. Now by plotting graph of Fuel type, Transmission and Owner against Price, we conclude that Car that uses Diesel, have automatic Transmission and Has only 1 owner is more likely to have a high price.







**CONCLUSION**

* Key Findings and Conclusions of the Study

Car price prediction has picked researchers' interest since it takes a significant amount of work and expertise on the part of the field expert. For a dependable and accurate forecast, a large number of unique attributes are analysed. We employed 5 different machine learning approaches to develop a model for forecasting the price of used automobiles. The respective performances of different algorithms were then compared to discover the one that best suited the existing data set. The final prediction model was implemented in a python programme. Furthermore, the model was tested with test data, yielding an accuracy of 75.61 percent.

* Learning Outcomes of the Study in respect of Data Science

Using well-known algorithms from Python libraries, we were able to successfully construct machine learning algorithmic paradigms. On our dataset, we first do pre-processing and data cleaning. We trimmed the tuples that contained null values, which accounted for less than 1% of the total. The findings revealed a positive relationship between price and kilometres travelled, as well as year of registration and kilometres travelled. Negative correlation is related to the notion of inverse proportion, whereas positive correlation is related to the concept of direct proportion. The model was trained using over 4340 tuples.

* Limitations of this work and Scope for Future Work

As a part of future work, we aim at the variable choices over the algorithms that were used in the project. We could only explore two algorithms whereas many other algorithms which exist and might be more accurate. More specifications will be added in a system or providing more accuracy in terms of price in the system i.e.

1) Horsepower

2) Battery power

3) Suspension

4) Cylinder

5) Torque

As we know technologies are improving day by day and there is also advancement in car technology also, so our next upgrade will include hybrid cars, electric cars, and Driverless cars.

s