

**CAR PRICE PREDICTOR**

Submitted by:

*SAYAN PAUL*

**ACKNOWLEDGMENT**

I would like to thank **cardekho.com** from where I scraped the dataset and I would also like to thank our MSE , MS Khusboo Garg for her constant help during the project.

Also I would like to thank websites like stackoverflow, geeksforgeeks and sklearn documentary for helping me when I got stuck somewhere.

**INTRODUCTION**

* Business Problem Framing

We were required to model the price of used cars with the available independent variables. This model will then be used by our client to understand how exactly the prices vary with the variables. They can accordingly manipulate their business strategy and concentrate on areas that will yield high returns. Further, the model will be a good way for the management to understand the pricing dynamics of the post-COVID market.

* Review of Literature

Cars are one of the necessary need of every middle class person around the globe and therefore used cars market is one of the markets which is one of the major contributors in a country’s economy. It is a very large market and there are various companies working in the domain. Data science comes as a very important tool to solve problems in the domain to help the companies increase their overall revenue, profits, improving their marketing strategies and focusing on changing trends in cars sales and purchases. Predictive modelling, Market mix modelling, recommendation systems are some of the machine learning techniques used for achieving the business goals for companies reselling used cars.

* Motivation for the Problem Undertaken

1. Developing an understanding of the used cars market.
2. Improving my machine learning skills.

**Analytical Problem Framing**

* Data Sources and their formats

We were provided the training data from cardekho.com by web scraping using selenium webdriver.

* Data Preprocessing Done

1. Label encoding
2. Skewness revomal of two columns using power transformer.
3. Outlier removal

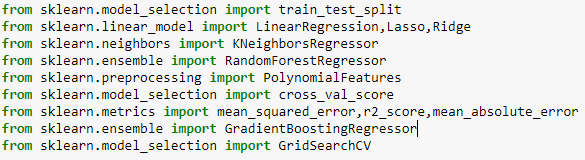
* Data Inputs- Logic- Output Relationships

GradientBoosting Regressor was used here because it was giving the best results.

* State the set of assumptions (if any) related to the problem under consideration

1. Since the number of columns were very less, we did not drop any column based on low correlation with target.
2. Models like LinearRegression, LassoRegression, RidgeRegression, ElasticNet,SGDRegressor is not going to preform well here.
3. Models like KNNRegressor, RandomForestRegressor, GradientBoostingRegressor might perform well.

* Hardware and Software Requirements and Tools Used

****

**Model/s Development and Evaluation**

* Identification of possible problem-solving approaches (methods)

The algorithms tried for solving this problem are:-

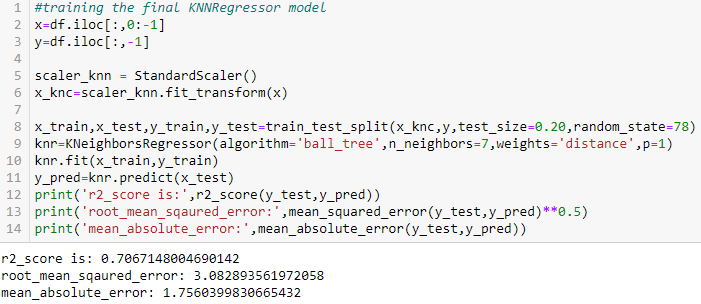
1. Linear Regression
2. KNeighbors Regression
3. Random Forest Regression
4. Gradient Boosting Regression.

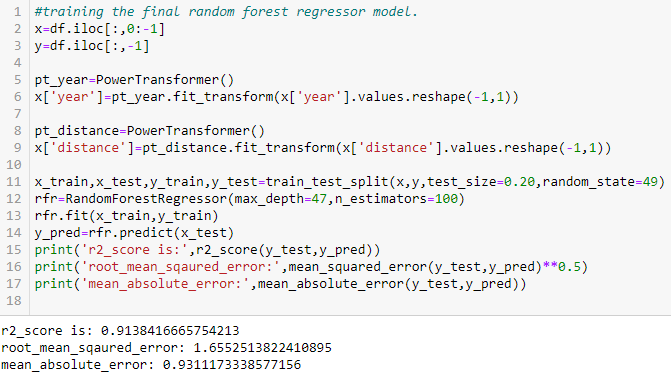
* Testing of Identified Approaches (Algorithms)

Hyperparameter tuning was performed on the above above algorithms and the best results are as shown below.

|  |  |
| --- | --- |
| **Algorithms** | **Best r2\_score** |
| KNN Regression | 0.5977441490180386 |
| Random Forest Regression | 0.8469893718646343 |
| Gradient Boosting Regression | 0.8565735439728831 |
|  |  |
|  |  |

* Run and Evaluate selected models



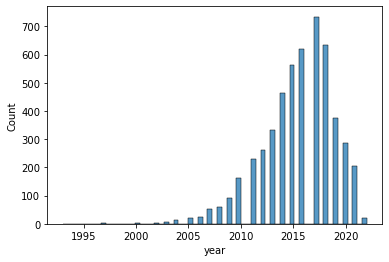


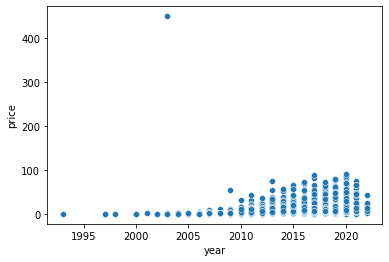


* Key Metrics for success in solving problem under consideration

1. r2\_score
2. root\_mean\_squared\_error
3. mean\_absolute\_error

* Visualizations

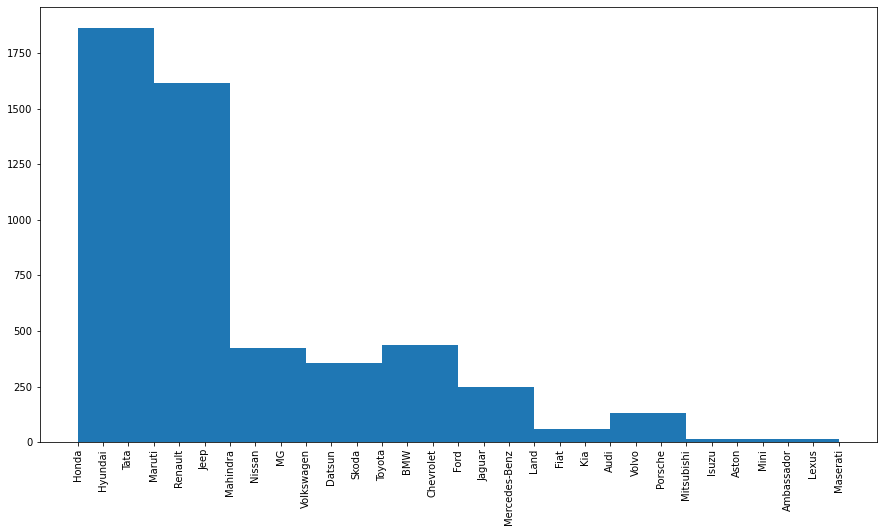




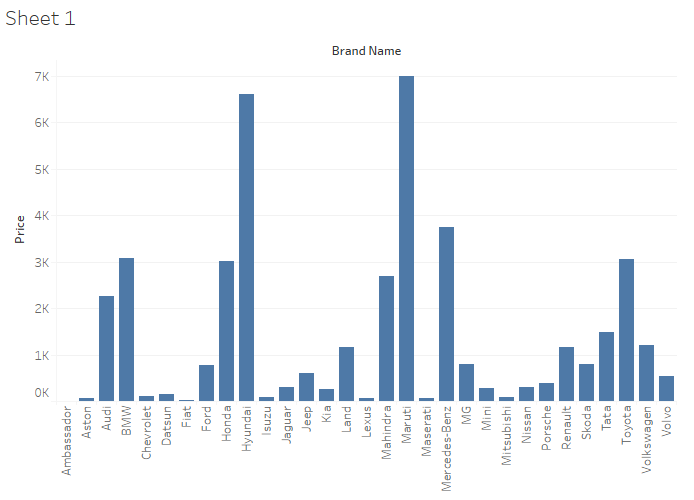
Relatively newer cars are having higher price.



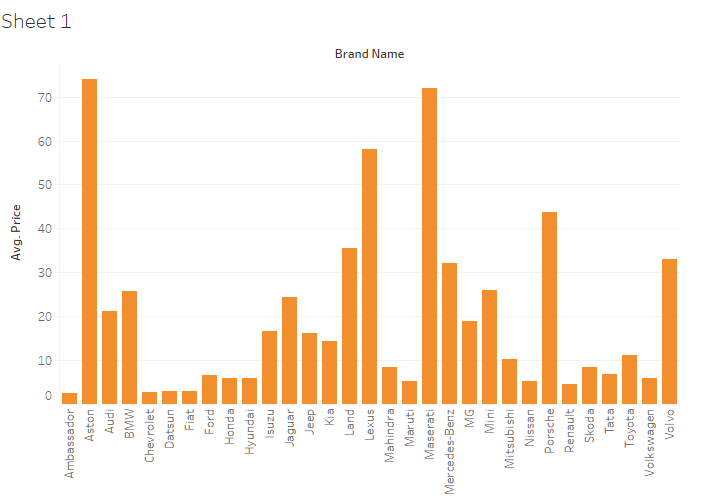
Most of the cars had only one owner.



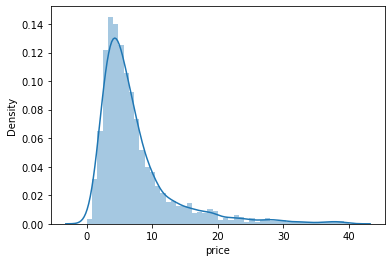
This graph shows the number of cars of each brand.



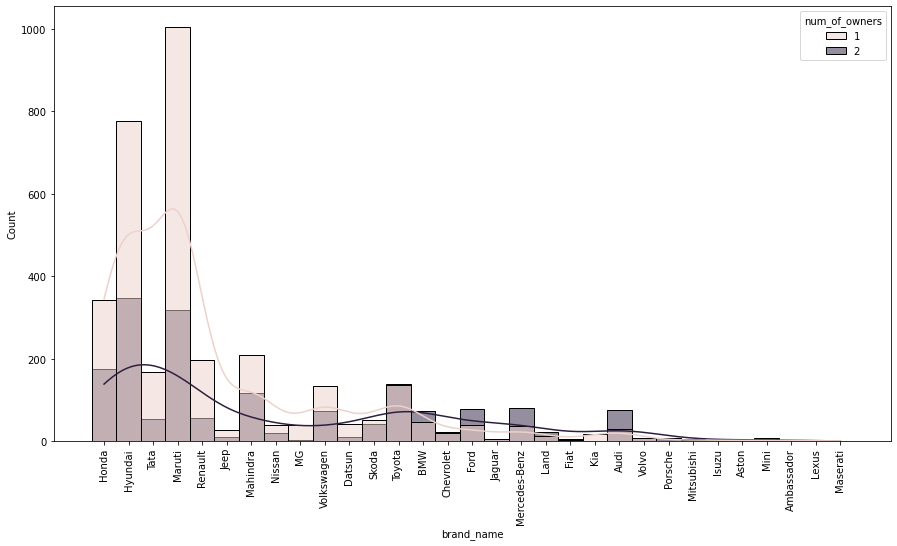
Brand name vs total price of all sold cars of that brand.



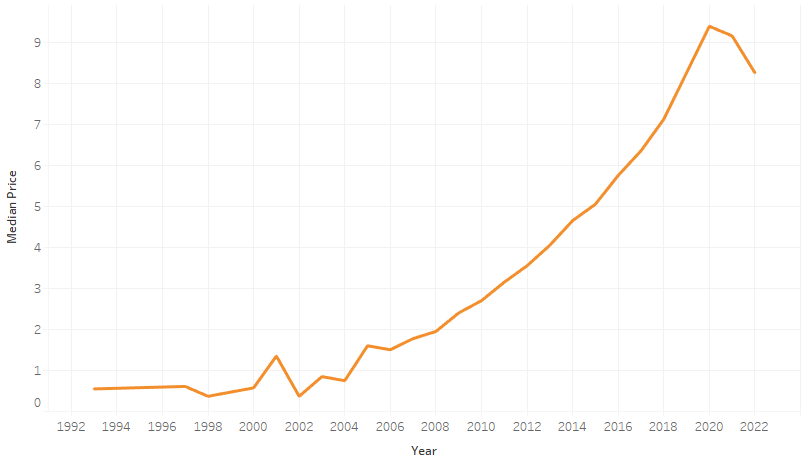
Brand name vs average price of all reselled cars of that brand.



Price distribution of used cars.



Brand name vs count of cars(for both 1 & 2 onwers)



Price is higher for newer cars.

**CONCLUSION**

* Learning Outcomes of the Study in respect of Data Science

Dropping columns having low correlation with the target is not always the only way. Sometimes those correlation attributes are necessary to give good results with KNNRegressor and DecisionTree regressor, RandomForest Regressor.

* Limitations of this work and Scope for Future Work

More hyper parameter tunings can be tried with the algorithms.

More data can be collected in order to improve model accuracy.