

Used Car Price Prediction

Submitted by:

Mohit Kumar Tomar

**ACKNOWLEDGMENT**

I take great pleasure to thank and acknowledge the help provided by **Flip Robo Technologies**. I extend whole hearted thanks to Mrs. Sapna Verma who become my Mentor and with whom I worked and learned a lot and for enlightening me with her knowledge and experience to grow with the corporate working. Her guidance at every stage of the Project enabled me to successfully complete this Project which otherwise would not have been possible without her consent encouragement and motivation. Without the support it was not possible for me to complete the report with fullest endeavour.

**Abstract**

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 India. Our results show that Random Forest model and K-Means clustering with linear regression 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.

**INTRODUCTION**

* Business Problem Framing

Today, the transportation industry is considered to be one of the backbones of the economy. Automobiles are referred to as the "Industry of Industries" in developed nations. According to industry professionals, the INDIA's automotive industry has seen remarkable growth. Besides being the fastest-growing nation in the automobile industry, it represents its global presence. In India, like most other countries, cars are gaining a great deal of popularity among the local population and the ex-pat community who work in the country. There are used cars for sale in the India of all makes.

Almost everyone wants their own car these days, but because of factors like affordability or economic conditions, many prefer to opt for pre-owned cars. Accurately predicting used car prices requires expert knowledge due to the nature of their dependence on a variety of factors and features. Used car prices are not constant in the market, both buyers and sellers need an intelligent system that will allow them to predict the correct price efficiently. In this intelligent system, the most difficult problem is the collection of the dataset which contains all important elements like the manufacturing year of the car, its gas type, its condition, miles driven, horsepower, doors, number of times a car has been painted, customer reviews, the weight of the car, etc. It is clear that the price of the product is affected by many factors, but unfortunately, information about these features is not always readily available. Since this project primarily focuses on the Indian market, the benchmark dataset containing most key features is scraped from the well known websites.

It is necessary to pre-process and transform collected data in the proper format prior to feeding it directly to the data mining model. As a first step, the dataset was statistically analyzed and plotted. Missing, duplicated, and null values were identified and dealt with. Features were chosen and extracted using correlation matrices. To build an efficient model, the most correlated features were retained, and others were discarded. This prediction problem can be considered a regression problem since it belongs to the supervised learning domain. Many Regressor known as random forest, linear regression, and bagging regression were trained and compared. A random forest Regressor outperformed all others in this project, so it was chosen as the main algorithm model.

* Statement of problem

The research objective of this study is to predict used cars prices in India using data mining techniques, by scraping data from websites that sell used cars, and analysing the different aspects and factors that lead to the actual used car price valuation. To enable consumers to know the actual worth of their car or desired car, by simply providing the program with a set of attributes from the desired car to predict the car price.

The purpose of this study is to understand and evaluate used car prices in the India, and to develop a strategy that utilizes data mining techniques to predict used car prices.

* Project goals

This project aims to deliver price prediction models to the public, to help guide the individuals looking to buy or sell cars and to give them a better insight into the automotive sector. Buying a used car from a dealer can be a frustrating and an unsatisfying experience as some dealers are known to deploy deceitful sale tactics to close a deal. Therefore, to help consumers avoid falling victims to such tactics, this study hopes to equip consumers with right tools to guide them in their shopping experience.

Another goal of the project is to explore new methods to evaluate used cars prices and to compare their accuracies. Considering this is an interesting research topic in the research community, and in continuing their footsteps, we hope to achieve significant results using more advanced methods of previous work.

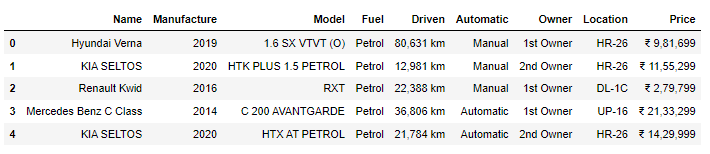
**Analytical Problem Framing**

* Mathematical/ Analytical Modeling of the Problem

Data Understanding and preparation is an essential part of building a model as it gives the insight into the data and what corrections or modifications shall be done before designing and executing the model, preliminary analysis of the data must be done to have deeper understanding into the quality of the data, in terms of outliers and the skewedness of the figures, descriptive Statistics of categorical and numerical variables was done for that to be achieved. As well as the ability to understand the main attributes that affect the results of the price. That was done through a correlation matrix for every attribute to understand the relations between the different factors

* Data Sources and their formats

The project deals with India used cars. Using Selenium, the benchmark dataset from OLX.com, cardekho.com and car24.com was scraped in order to build the effective intelligent model.



9 Features have been scrapped.

1. Name: Car complete name
2. Manufacture: Car manufacturer company name
3. Model: Car Model/ variant
4. Fuel: Which fuel is being used in the car (Petrol/Diesel/CNG etc)
5. Driven: Total driven km by car
6. Automatic: if the car is Manual or Automatic
7. Owner: How many owners have been changed of the car
8. Location: which state of India
9. Price: Selling price of the car

* Data Preprocessing Done

After data collection the dataset was pre-processed to remove samples that have missing value, and remove non-numerical part from numerical attributes, converting categorical values into numerical (if needed), fix any discrepancies in the units, as well as removing attributes that doesn’t affect the price evaluations if needed to reduce the complexity of the model

Data Understanding and preparation is an essential part of building a model as it gives the insight into the data and what corrections or modifications shall be done before designing and executing the model, preliminary analysis of the data must be done to have deeper understanding into the quality of the data, in terms of outliers and the skewedness of the figures, descriptive Statistics of categorical and numerical variables was done for that to be achieved. As well as the ability to understand the main attributes that affect the results of the price. That was done through a correlation matrix for every attribute to understand the relations between the different factors.

* Data Inputs- Logic- Output Relationships

Afterwards when the data is organized and transformed into a form that could be processed by the data mining technique. Different data mining models were designed to predict prices and values of used cars. In this study three models are proposed to be built using Logistic Regression model technique, Random Forest Regressor and Bagging Regressor. Firstly, the data was portioned into section for training and the other part for testing, portioning percentage can be tested with different ratios to analyse different results. All three models were evaluated on four evaluation matrices known as model score, Mean Square Error (MSE), Mean Absolute Error (MAE) and Root Mean Square Error (RMSE). From all, the Random Forest Regressor outperformed.

Describe the relationship behind the data input, its format, the logic in between and the output. Describe how the input affects the output.

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

In the past year the world of automobiles has seen a drastic change with the semiconductor shortages after the pandemic, which led to spike in used car prices. Hence, there was fast change in car prices during this study which will affect the actual car pricing prediction future. As the current dataset will undervalue the cars in the market. Therefore, a model that is built on real time data can be best integrated into a mobile app for public use would be the idea solution.

* Hardware and Software Requirements and Tools Used

Hardware:

CPU: Intel(R) Core(TM) i5-4460 CPU @ 3.20GHz

Ram: 20GB

Monitor= 32”inch Samsung

Software: Latest Anaconda for Jupyter

Python Libraries:

Pandas , Numpy, seaborn, matplotlib, scikit-learn,

**Model/s Development and Evaluation**

* Identification of possible problem-solving approaches (methods)

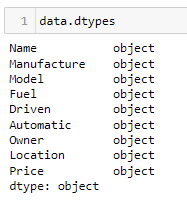
Pre-processing is a Data Mining technique that involves converting raw data into a comprehensible format. There is often a lack of specific activity or trend data, and many inaccurate facts are included in real-world data. Consequently, this may result in poor-quality data collection, and, in turn, poor-quality models constructed from the data. Such problems can be resolved by pre-processing the data.

Pre-processing in Machine Learning is the process of modifying, or encoding, data so that the machine can parse it more easily. Thus, the algorithm can now properly interpret the data.

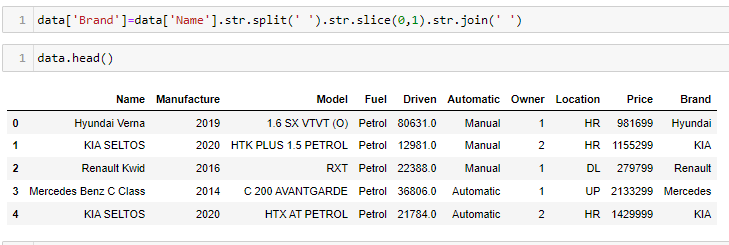
In this project, following steps are preformed to pre-process the dataset.

1. Dataset collection: we have collected the dataset from 3 leading websites which deals with used car sell-buy.
2. Pre-Processing: scrapped data was very messy. We have to perform many pre-processing on that dataset.

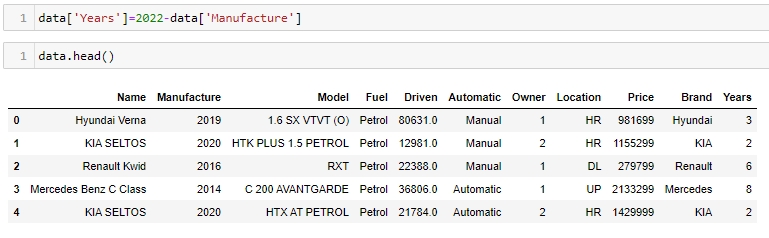
As the data is scrapped from the websites, all features were in Object form, even the integer features also were in object form.



**Name** feature was having the complete name of the car, while we required the car manufacture company name so we extract this from the car name.



**Manufacture** feature provided the year of car manufacture which is required basically to know the age of that particular car at instance. So we calculated the Age of car at this moment



**Fuel** feature were having many garbage inputs so we have replaced the garbage inputs with the median of feature

**Driven** feature were having small null values, we have decided to fill those null values with median of Driven feature

**Location** features were having the registration code for the car, so we have to scrap the state of the India. Because due to taxation, every state have different on-road price of car.

**Price** feature were in object form, we removed unwanted rupee sign and convert into integer datatype. Because this is our target feature here.

We have used StandardScaler to standardize the continuous features and OneHotEncoding to encode categorical features into integer feature.

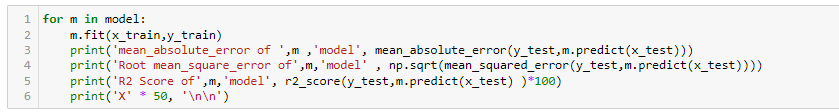
* Testing of Identified Approaches (Algorithms)

We have used several available Regressor Algorithms



We have calculated RMSA for all the available Algos then decide to which one to proceed for model building

* Run and Evaluate selected models



We have calculated MAE, RMSE and R2 score for all ML Algorithms.

* Key Metrics for success in solving problem under consideration

The regression model can be evaluated on following parameters:

1. Mean Square Error (MSE):

MSE is the single value that provides information about goodness of regression line. Smaller the MSE value, better the fit because smaller value implies smaller magnitude of errors. 𝑴𝑺𝑬= 𝟏𝑵Σ|𝒚𝒊−𝒚|𝟐𝑵𝒊=𝟏

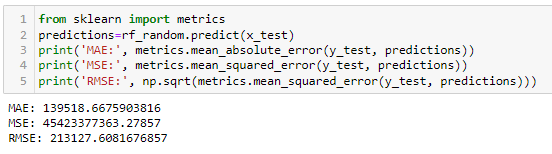
Equation 3 MSE equation

2. Root Mean Square Error (RMSE):

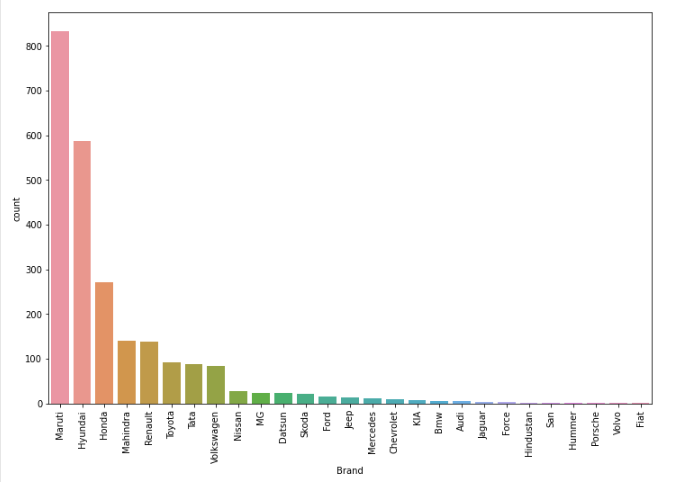
RMSE is the quadratic scoring rule that also measures the average magnitude of the error. It is the square root of average squared difference between prediction and actual observation.

3. Mean Absolute Error (MAE):

This measure represents the average absolute difference between the actual and predicted values in the dataset. It represents the average residual from the dataset. 𝑴𝑨𝑬= 𝟏𝑵Σ|𝒚𝒊−𝒚|

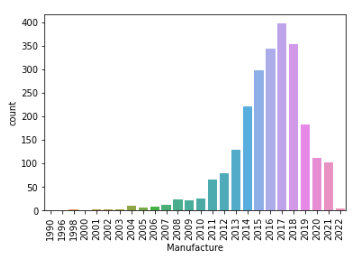


* Visualizations



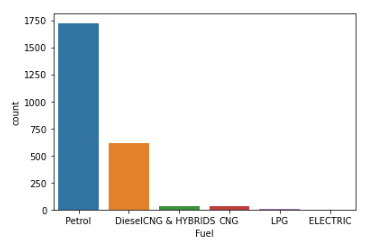
In our dataset: 34.6% cars are of Maruti and 24% are of Hyundai

11% are of Honda



Most of the Cars are available for resale are from the year 2014 to 2021. Absolutely, No one would buy very old car.

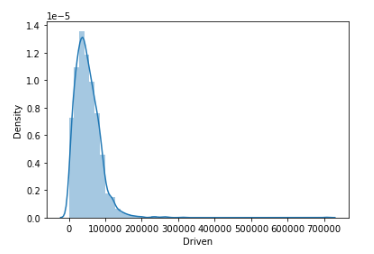
India also have Scrap law of 10years of Diesel car and 15 yrs for Petrol Car.



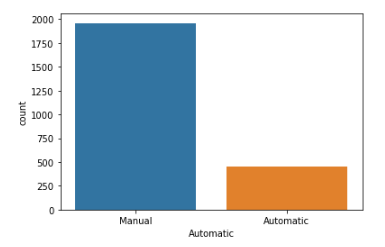
71.56% cars are on Petrol

25.49% cars are on Diesel

Majority of Cars in India are on Petrol and Diesel. However, India has started to manufacture Electric cars. So In coming years, we would see large number of electric cars in Indian market.

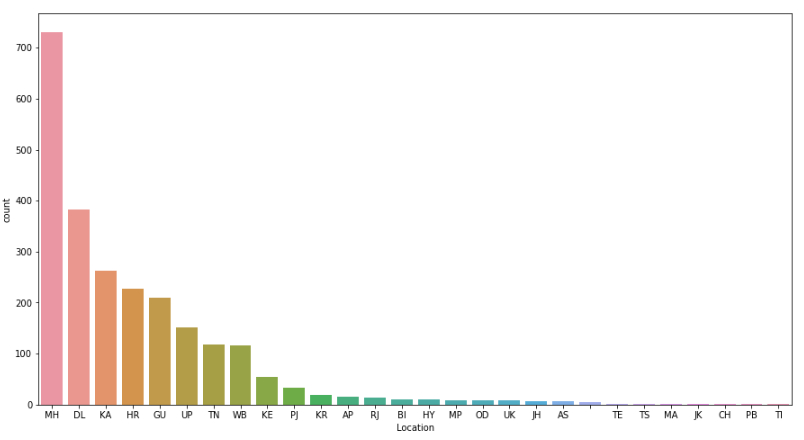


Distribution of KM driven by the old cars are right skewed but having a bell-curve shape

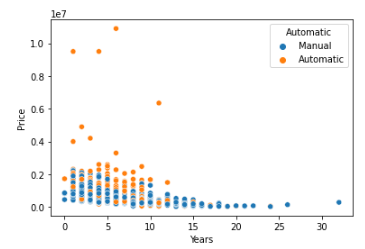


81.32% cars are manually operated

And remaining are Automatic.



Most of the Cars are from Maharastra and then Delhi



Before 15 years, India hardly had Automated car while in recents years we had many Automated Car’s.

This graphs also tells that Automated car price is relatively high than manual car.

**CONCLUSION**

* Key Findings and Conclusions of the Study

Using data mining and machine learning approaches, this project proposed a scalable framework for India based used cars price prediction. Car24.com, olx.com and cardekho websitees was scraped using the Selenium scraping tool to collect the benchmark data. An efficient machine learning model is built by training, testing, and evaluating three machine learning regressors named Random Forest Regressor, Linear Regression, and Bagging Regressor. As a result of pre-processing and transformation, Random Forest Regressor came out on top with 64% accuracy.

Each experiment was performed in real-time Jupyter notebook.

* Limitations of this work and Scope for Future Work

In the future, more data will be collected using different web-scraping techniques, and deep learning classifiers will be tested. Algorithms like Quantile Regression, ANN and SVM will be tested.

Afterwards, the intelligent model will be integrated with web and mobile-based applications for public use. Moreover, after the data collection phase Semiconductor shortages have incurred after the pandemic which led to an increase in car prices, and greatly affected the secondhand market. Hence having a regular Data collection and analysis is required periodically, ideally, we would be having a real time processing program