***CAR RESALE VALUE PREDICTOR***

TEAM:NICHE CODERS

TEAM MEMBERS

1:Ravi Teja Karanam

2:Sanapala Aravind

3:N.S Pranay Manas

4:C.T.R Tejodhar

CONTENTS

1. Title of the project
   1. Introduction
   2. Objectives of research
   3. Problem Statement
   4. Industry Profile

1. Review of literature
2. Data Collection
3. Methodology
   1. Data processing

4.1.1 Statistical Techniques and Data Visualization

* 1. Exploratory Data Analysis
  2. Data Modeling using Supervised ML techniques

1. Findings and Suggestion
2. Conclusion
3. References

***1.1 INTRODUCTION :***

Predicting the price of used cars is both an important and interesting problem. With difficult economic conditions, it is likely that sales of second-hand imported (reconditioned) cars and used cars will increase. In many developed countries, it is common to lease a car rather than buying it outright. A lease is a binding contract between a buyer and a seller (or a third party – usually a bank, insurance firm or other financial institutions) in which the buyer must pay fixed instalments for a pre-defined number of months/years to the seller/financer. After the lease period is over, the buyer has the possibility to buy the car at its residual value, i.e. its expected resale value. Thus, it is of commercial interest to seller/financers to be able to predict the salvage value (residual value) of cars with accuracy.

***1.2 OBJECTIVE:***

Objective of the project is to find the nearest possible resale value of any car depending on the brand, year of purchase, model, no of kilometres travelled etc.

***1.3 PROBLEM STATEMENT:***

Due to the advent of companies such as ola and uber the people who are interested in buying cars on resale have increased exponentially. This project is expected to help such people

***1.4 INDUSTRY PROFILE:***

With a scintillating 2.3 million units produced in 2008 the Indian automobile industry bagged the position of being the ninth largest in the world. Following economic liberalization, Indian domestic automobile companies like Tata Motors Maruti Suzuki and Mahindra and Mahindra expanded their production and export operations in and across the country and since then the industry has only shown signs of growth.

The Indian automobile industry seems to come a long way since the first car that was manufactured in Mumbai in 1898. The automobile sector today is one of the key sectors of the country contributing majorly to the economy of India. It directly and indirectly provides employment to over 10 million people in the country. The Indian automobile industry has a well-established name globally being the second largest two wheeler market in the world, fourth largest commercial vehicle market in the world, and eleventh largest passenger car market in the world and expected to become the third largest automobile market in the world only behind USA and China.

This development in the automobile industry marks the increase in the resale industry as well thus giving the project a good use.

***2 REVIEW OF LITERATURE***

One major improvement that we did in our research work is that we didn’t drop all the missing values instead what we did is replace all the missing data with the mode of the respective categorical column by that we were able to save more that 1 lakh entries and achieve 89.7% accuracy.

***3 DATA COLLECTION:***

The current data set provides us with a lot of data about the used cars like the brand, date of purchase, type of gear, seller, power, model ,etc.



***4 METHODOLOGY:***

***4.1 DATA PREPROCESSING***

First the given data is encoded cause the column “name” is not in correct format the result is shown below

A screenshot of a cell phone

Description automatically generated

Instead of using name as a string value we use name as (length of name) an integer variable and add this as a column and drop ‘name’ column for easy analysis.

As there are only 2 types of sellers and most of the given data belongs to only one type of seller (i.e. privat) this column can be dropped and so it has been dropped.





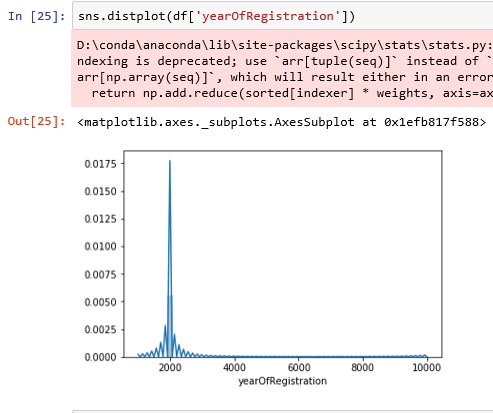
And similarly there are only two types of offers and only one of them is used in majority so this column has also been dropped.



There are a lot of null values in the type of car column so instead of deleting all the rows we define a dictionary with the information that was provided from the same data set with the models as keys and if there are any same model names as keys with model names as null they will be replaced with the data values .



The year of registration is considered only from 1980 to 2019 as the data is highly distributed in this range.

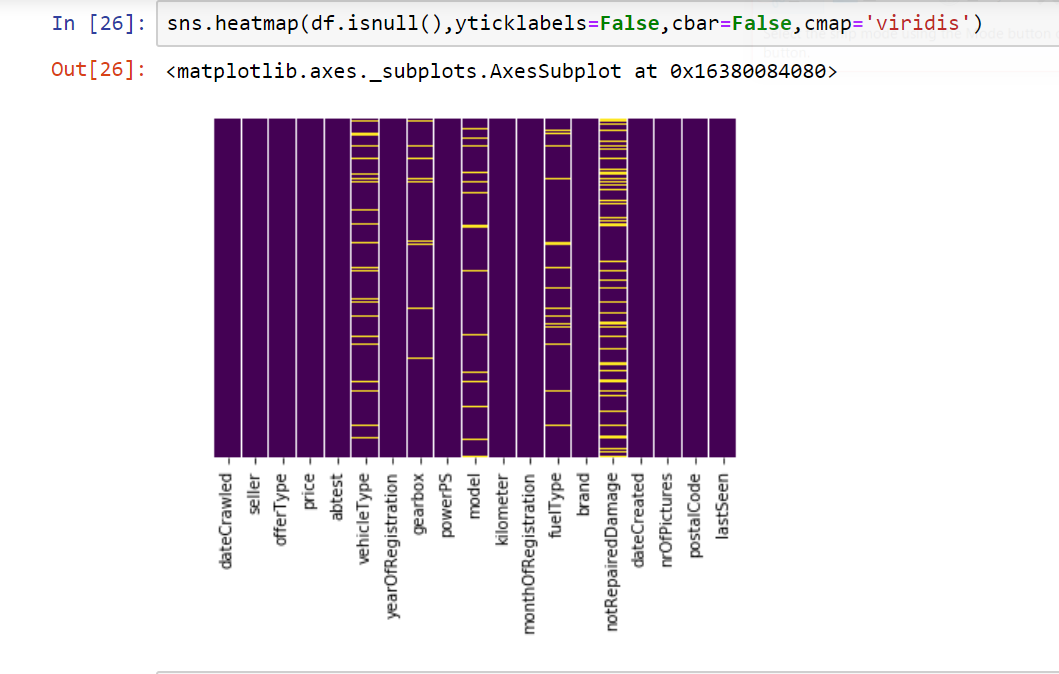


We now fill all the NaN values with a string “Not-specified” instead of dropping the missing data so that later we can replace the not specified with mode of the following categorical column .

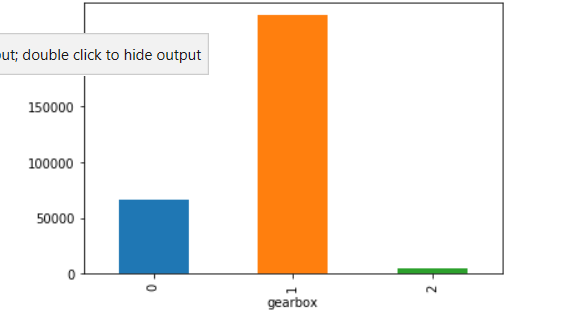


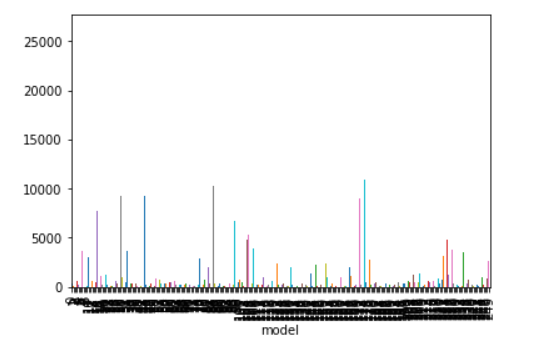
***4.1.1 Statistical Analysis and Data Visualization***

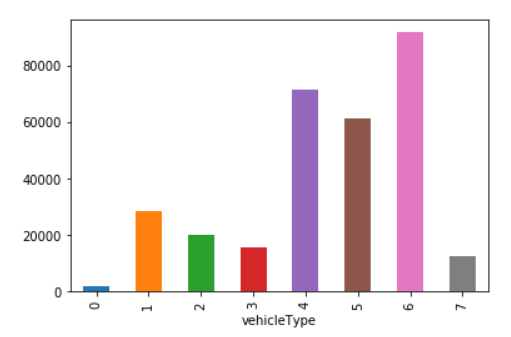
Let’s visualize the number of null values in the given data set using a heatmap.

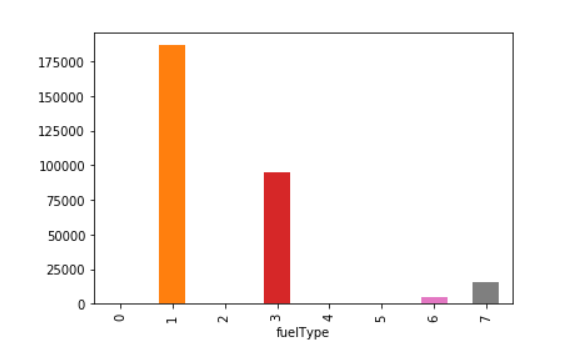


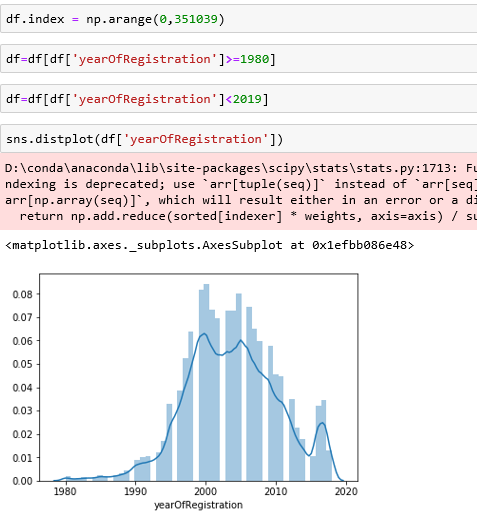
We now check the count of different categories in the respective columns to see which categories influence the prediction the most.







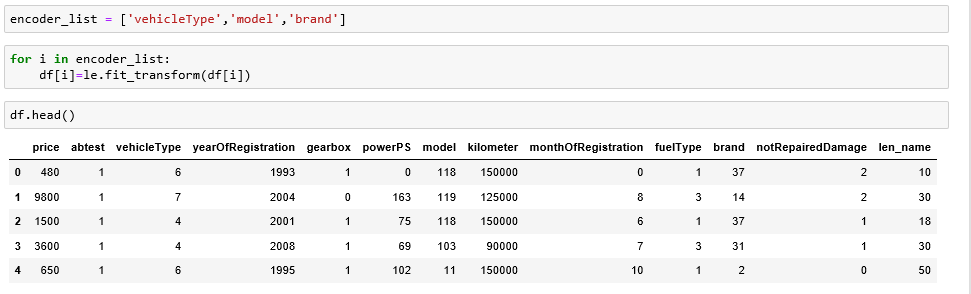




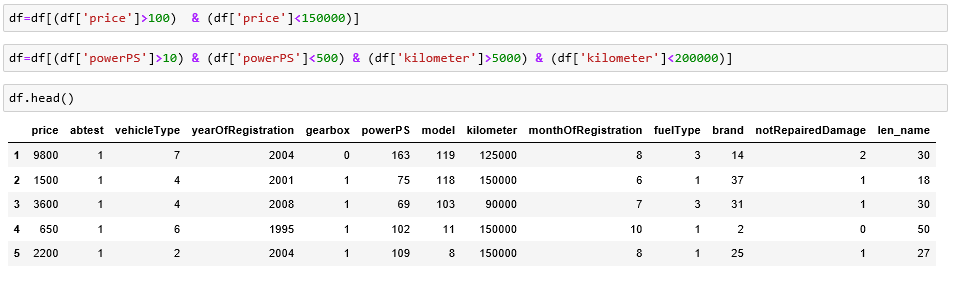
Then gearbox, fuel type, not repaired damage are encoded and unused data columns are dropped i.e. number of pictures posted, postal code, last seen ,date crawled.



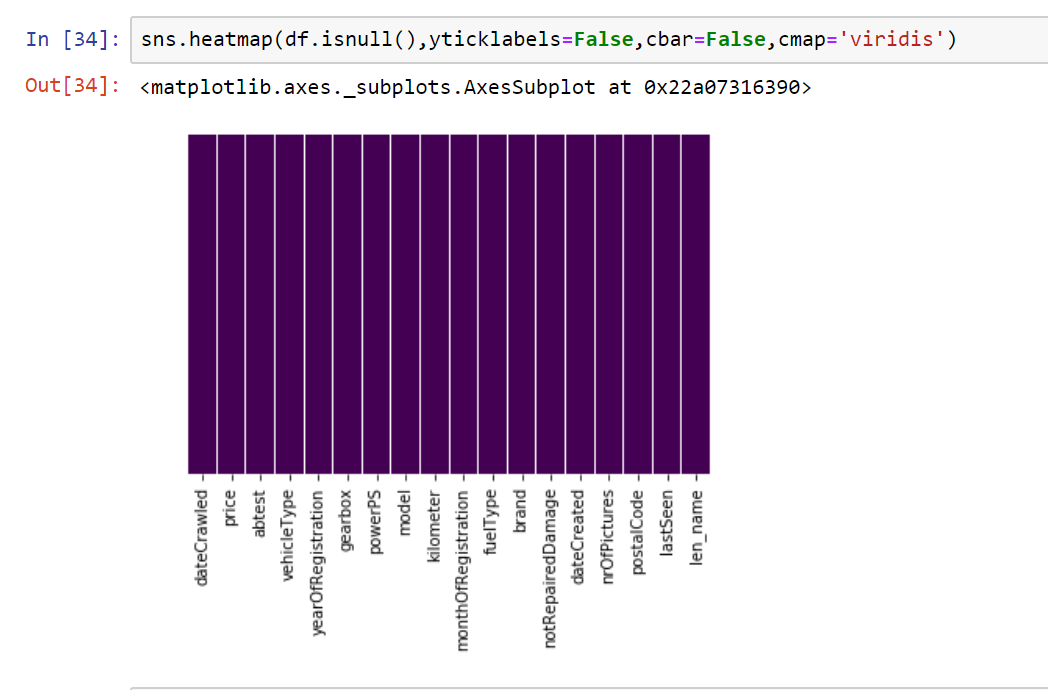
Vehicle type, model , brand are also encoded



A few cars as per the data are free of cost such values are removed and also there are cars whose power is less than 10 they are removed as well. The data after processing looks like this

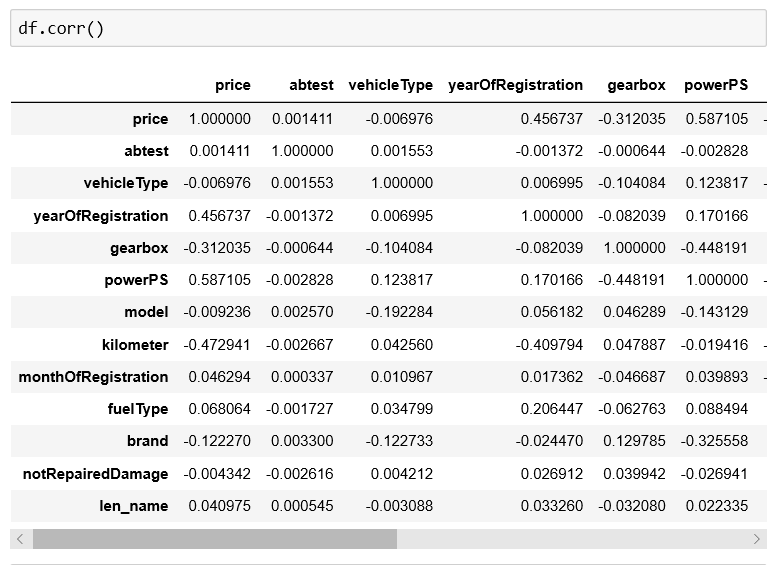


After data pre-processing , the heatmap looks like :



***4.2 EXPLORATORY DATA ANALYSIS***

After completing all the pre-processing we finally see the correlation between different variables



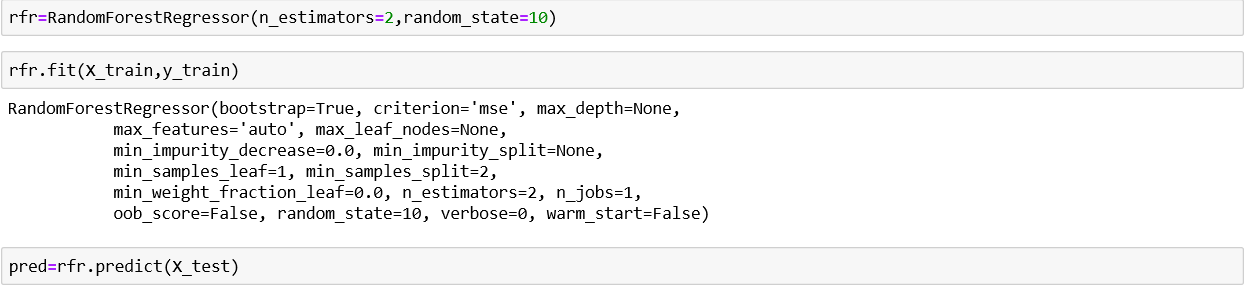
4.3 DATA MODELING AND SUPERVISED LEARNING

Since now the dataset is ready for modelling we shall divide our dataset into train, test sets.

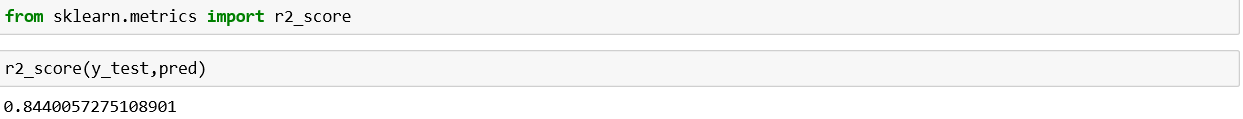




The model that we are using for our dataset is Random forest Regressor with the (n\_estimators =2)

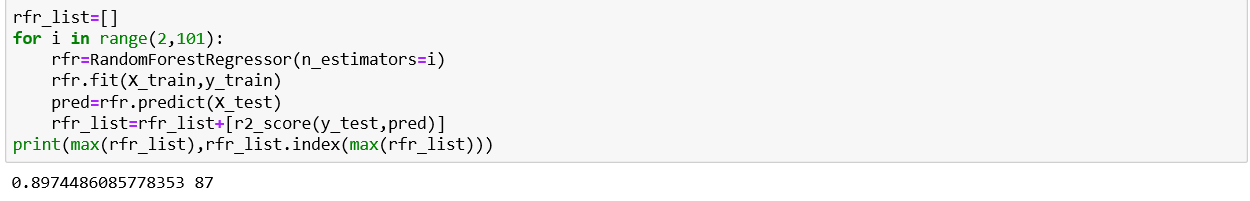


After fitting the data it is really important to know how our model responds to the new data ,for that we are using r2\_score model which gives us a measure as to how our model predicts.



***5. FINDINGS AND SUGGESTIONS***

We now check whether if any further alterations(Such as changing the n\_estimators value) would improve our accuracy.fawf



As we can see we get better accuracy when the n\_estimator value is 87 , with the accuracy of 89.7 .

***6 CONCLUSION***

Based on our Data Visualization, Exploratory Data Analysis,

We conclude that our analysis went smoothly, except for the fact that time consumed(complexity) is high as we have dealt over 3,00,000 entries.

An interesting future study might involve testing the model using a range of values for n\_estimators and random\_state.

Any help or advice is welcome!!!

***7 REFERENCES***

1:https://skills.thesmartbridge.com/courses/course-v1:CognitiveClass+ML0101ENv3+v4

2:https://stackoverflow.com/questions/54618253/how-to-deal-with-missing-values-in-pandas-dataframe