

CAR PRICE PREDICTION



Submitted by:

Utkarsha A Jadhav

**ACKNOWLEDGMENT**

This includes mentioning of all the references, research papers, data sources, professionals and other resources that helped you and guided you in completion of the project.

**INTRODUCTION**

* Business Problem Framing

Describe the business problem and how this problem can be related to the real world.

Answer: The business problem was to scrap the data from the olx website and then use machine learning model for prediction of the car price. In real life this may provide a close cost of the car and car type with model.

* Conceptual Background of the Domain Problem

Describe the domain related concepts that you think will be useful for better understanding of the project.

Answer: The domain related concepts like Brand, model, variant, manufacturing year, driven kilometres, fuel, number of owners etc. The car made and model will make difference as it give the different features in the car. The number of owner also influence the cost based on the maintenance and usage. Driven kilometres influence the cost as the wear and tear of the car. Again manufacturing year and spare parts also makes difference in the price.

* Review of Literature

This is a comprehensive summary of the research done on the topic. The review should enumerate, describe, summarize, evaluate and clarify the research done.

Answer: We can scrap the details from different websites and then predict the prices.

* Motivation for the Problem Undertaken

Describe your objective behind to make this project, this domain and what is the motivation behind.

Answer: The motivation for the project is to get good condition car for a reasonable price.

**Analytical Problem Framing**

* Mathematical/ Analytical Modeling of the Problem

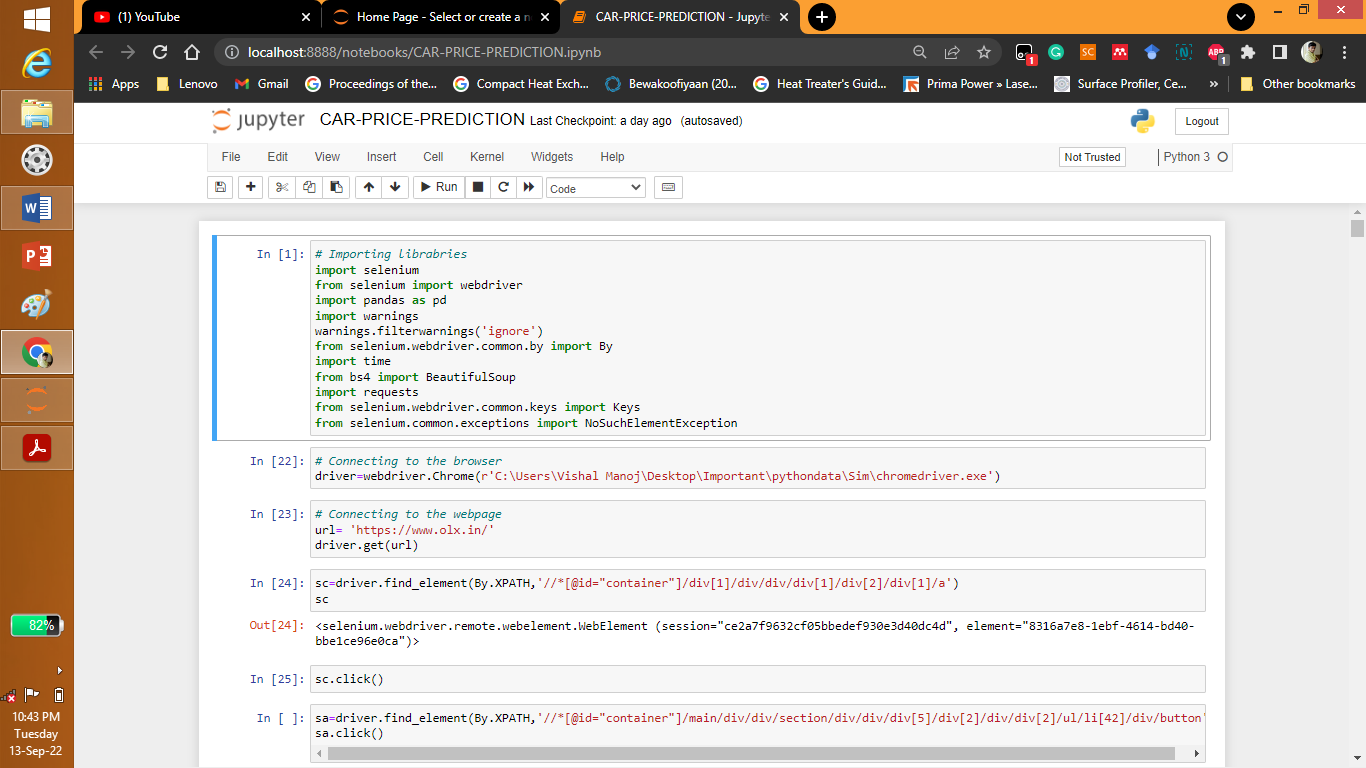
Describe the mathematical, statistical and analytics modelling done during this project along with the proper justification.

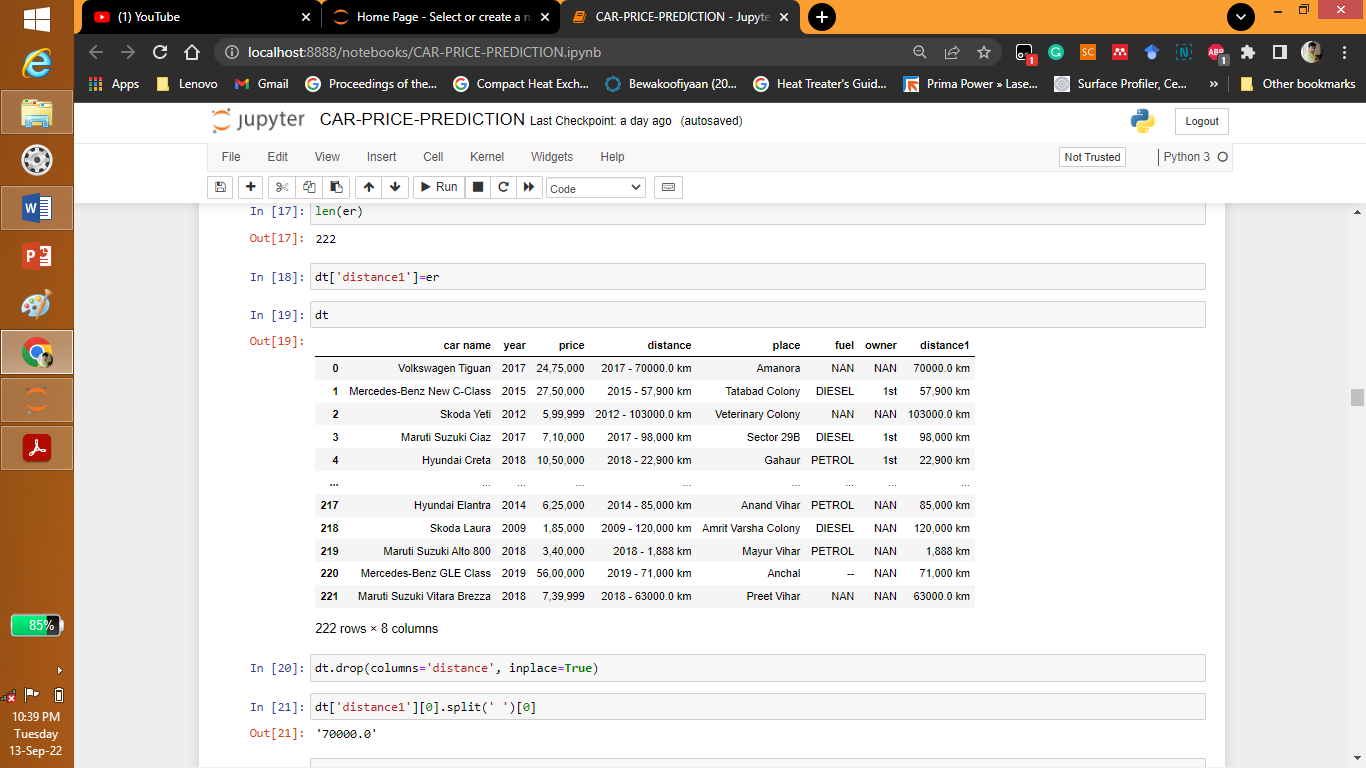
Answer: The different models used like linear regression, logarithmic regression, K-NN regression, random forest regression, ada-boost regression, EGboost regression, gradient-boost regression and SVC regression.

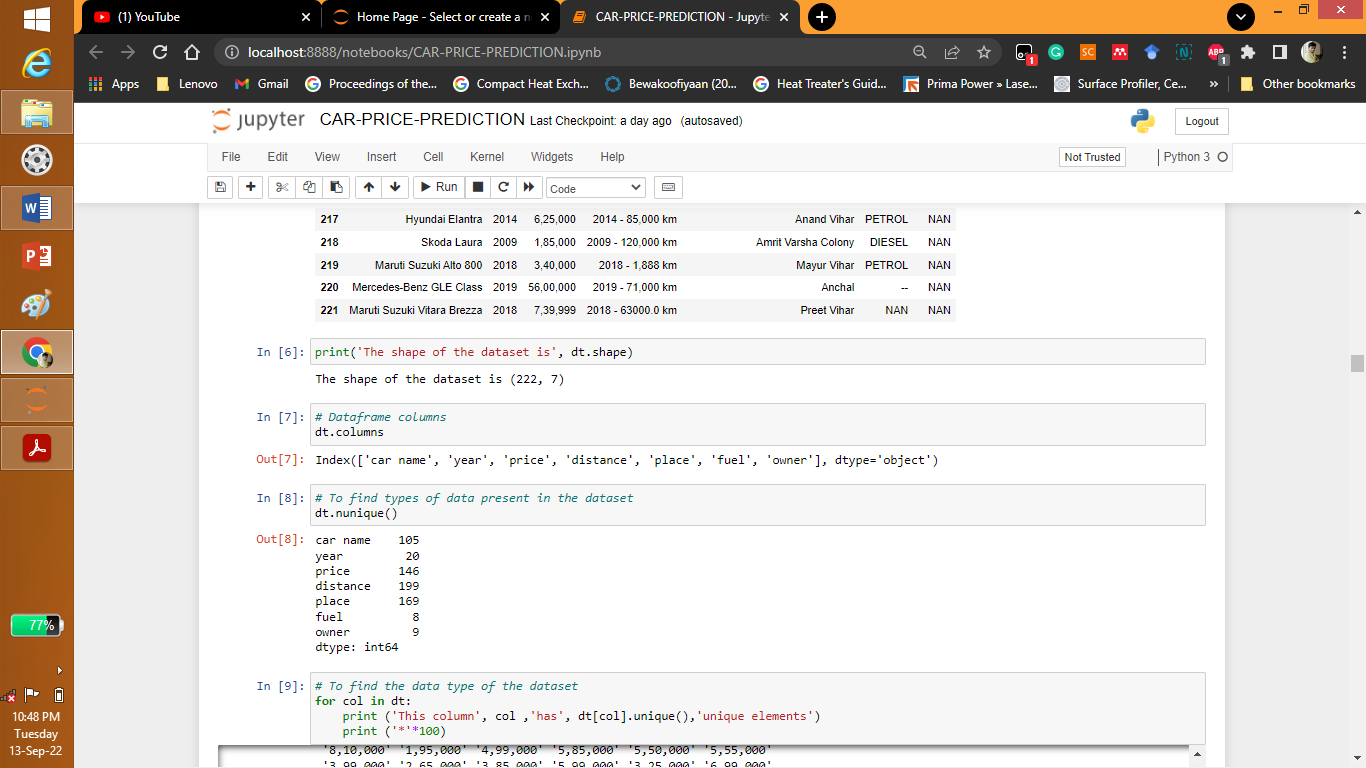
* Data Sources and their formats

What are the data sources, their origins, their formats and other details that you find necessary? They can be described here. Provide a proper data description. You can also add a snapshot of the data.

Answer: The data can be taken by a survey conducted by micro credit company, open source websites like Kaggel etc. The data is in the form of .csv file it may also be in .json or Excel files. In the present form we have taken dataset from olx website.



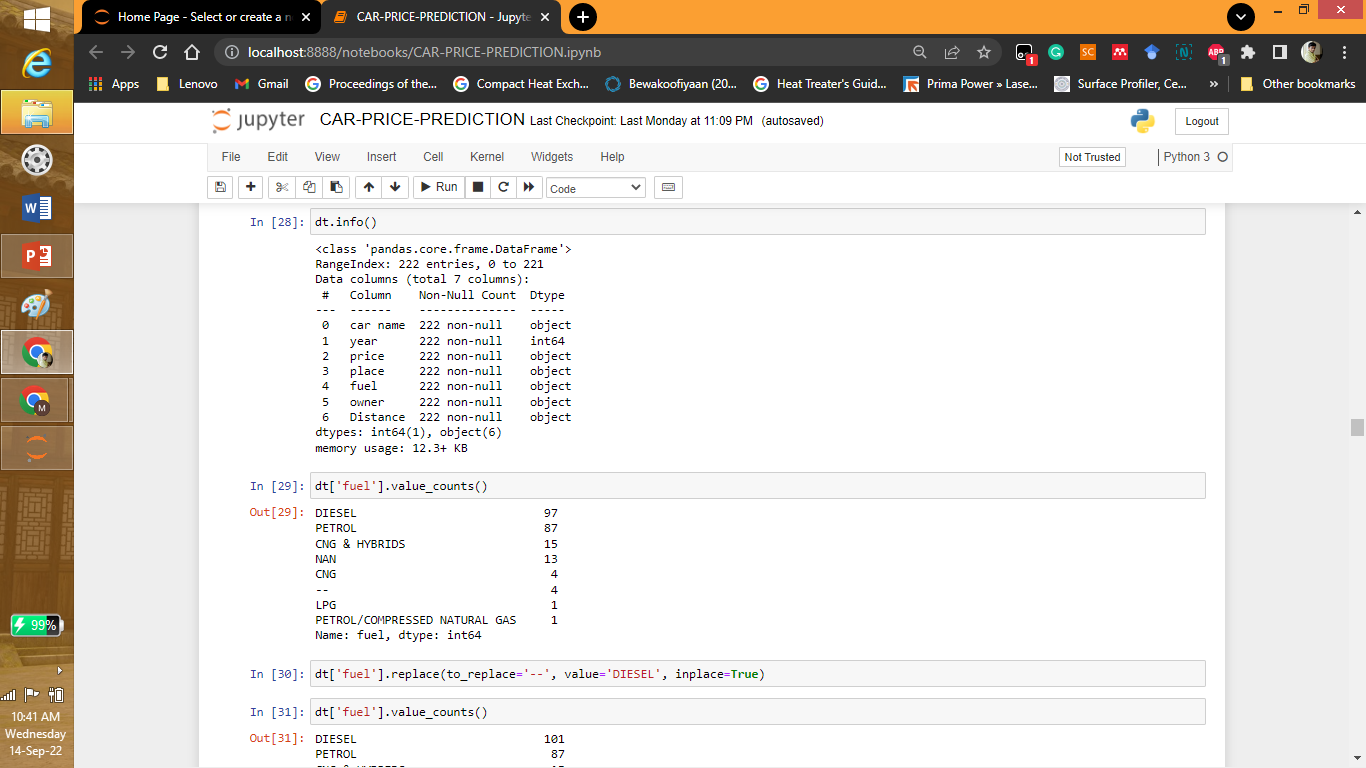


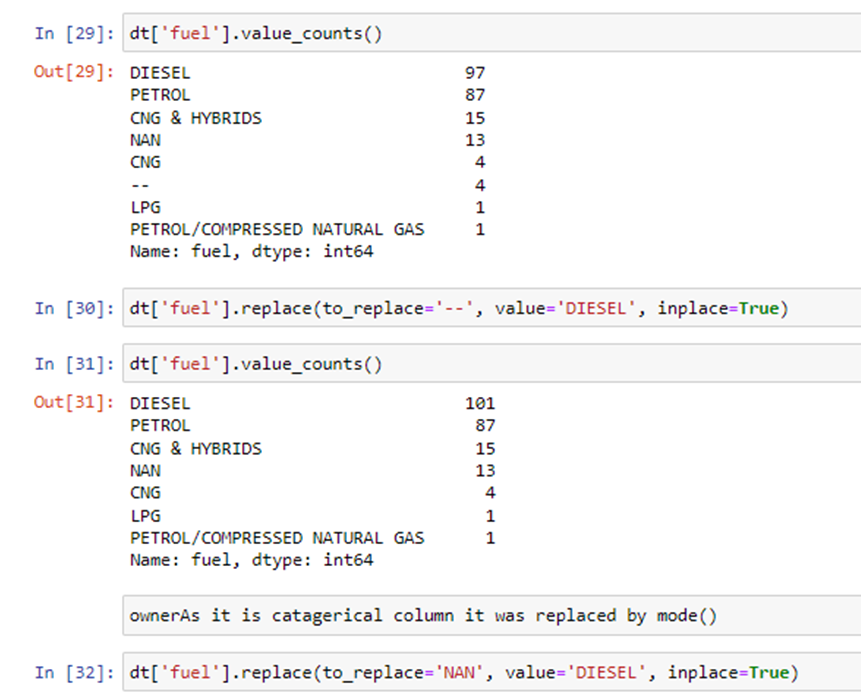


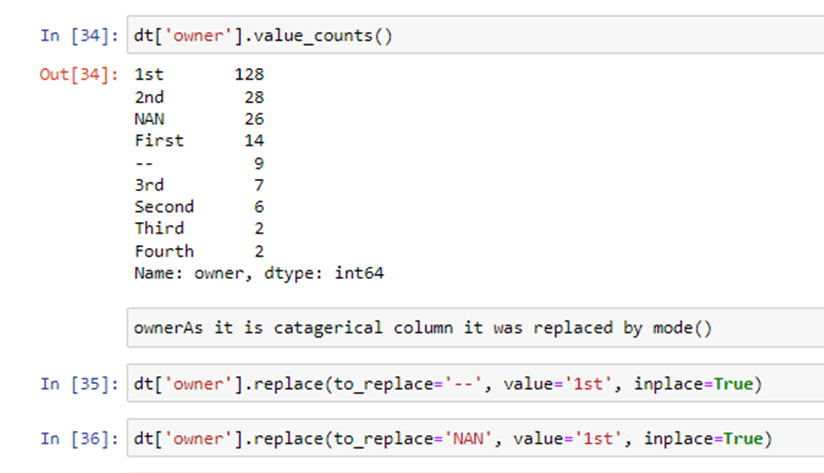
* Data Preprocessing Done

What were the steps followed for the cleaning of the data? What were the assumptions done and what were the next actions steps over that?

Answer: The place is neglected as it least influencing the car price it is neglected.







* Data Inputs- Logic- Output Relationships

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

Answer: The number of columns are car name, year, price fuel, owner, distance. The car name has characteristics like model, type etc. so this may influence the cost. Manufacturing year also influences the cost of car. Recent manufacturing years costs more and others costs lesser. Owner like 1, 2 or 3 also influences the cost. Distance travelled also influences the price. More the distance travelled lesser the price.

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

Here, you can describe any presumptions taken by you.

Answer: Only data from olx of 200 data points are web scraped.

* Hardware and Software Requirements and Tools Used

Listing down the hardware and software requirements along with the tools, libraries and packages used. Describe all the software tools used along with a detailed description of tasks done with those tools.

Answer: The different hardware and software tools, libraries and packages used are selenium, numpy, pandas, seaborn, matplotlib, sklearn.

**Model/s Development and Evaluation**

* Identification of possible problem-solving approaches (methods)

Describe the approaches you followed, both statistical and analytical, for solving of this problem.

* Testing of Identified Approaches (Algorithms)

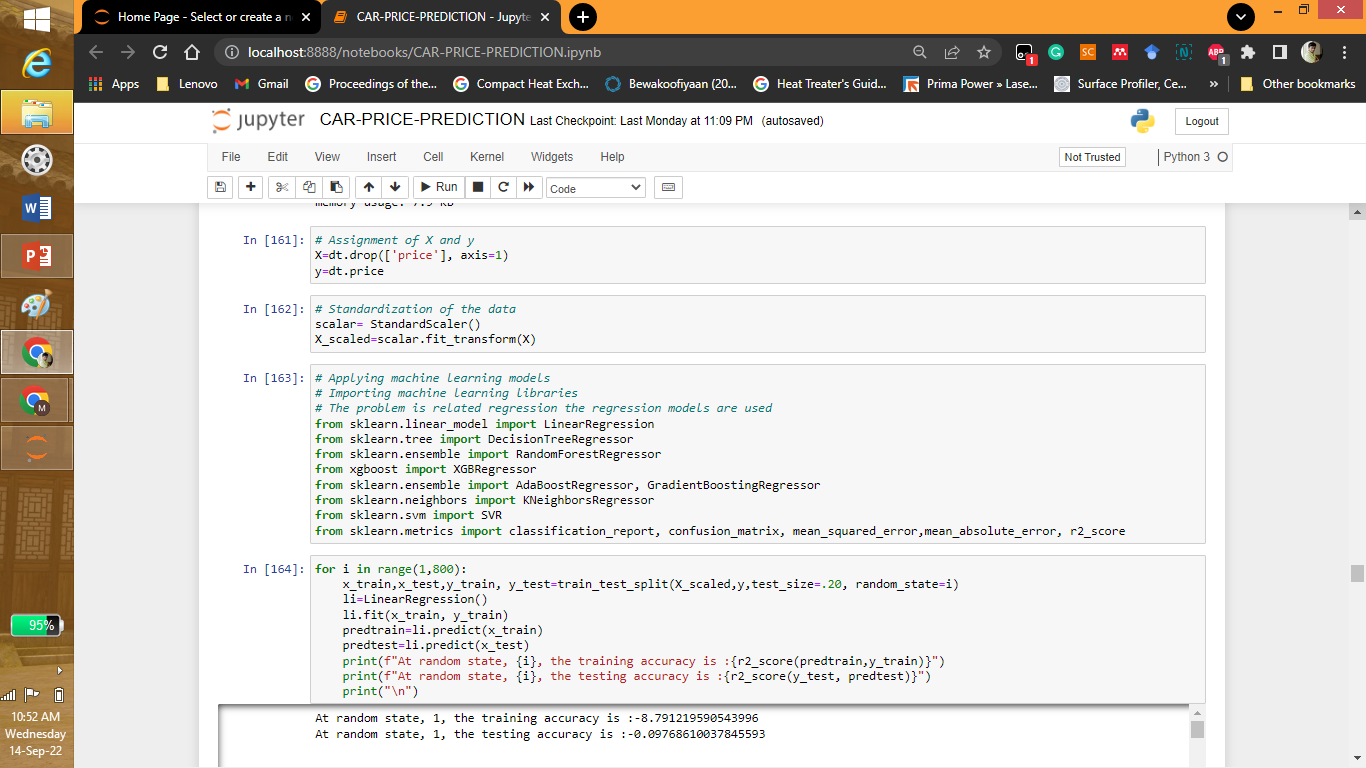
Listing down all the algorithms used for the training and testing.

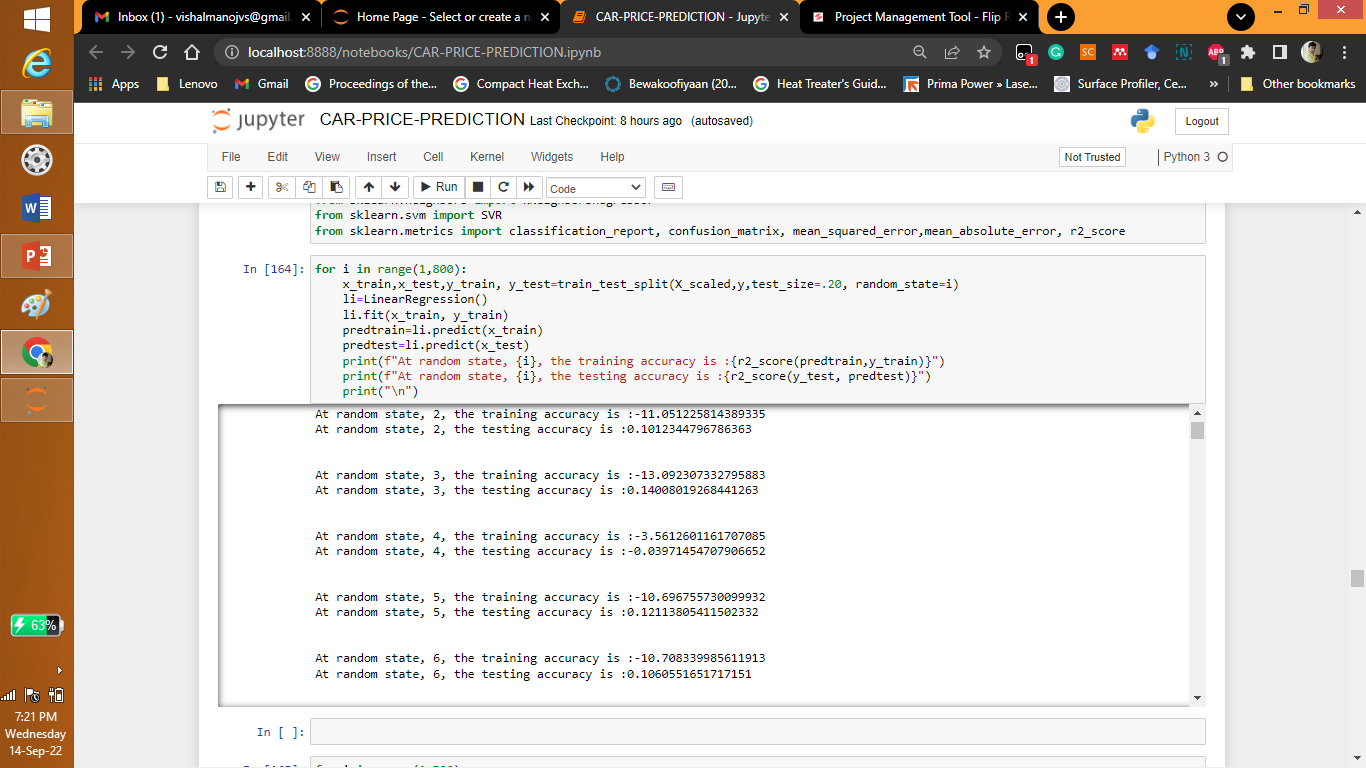
The different models used like linear regression, logarithmic regression, K-NN regression, random forest regression, ada-boost regression, EGboost regression, gradient-boost regression and SVC regression.

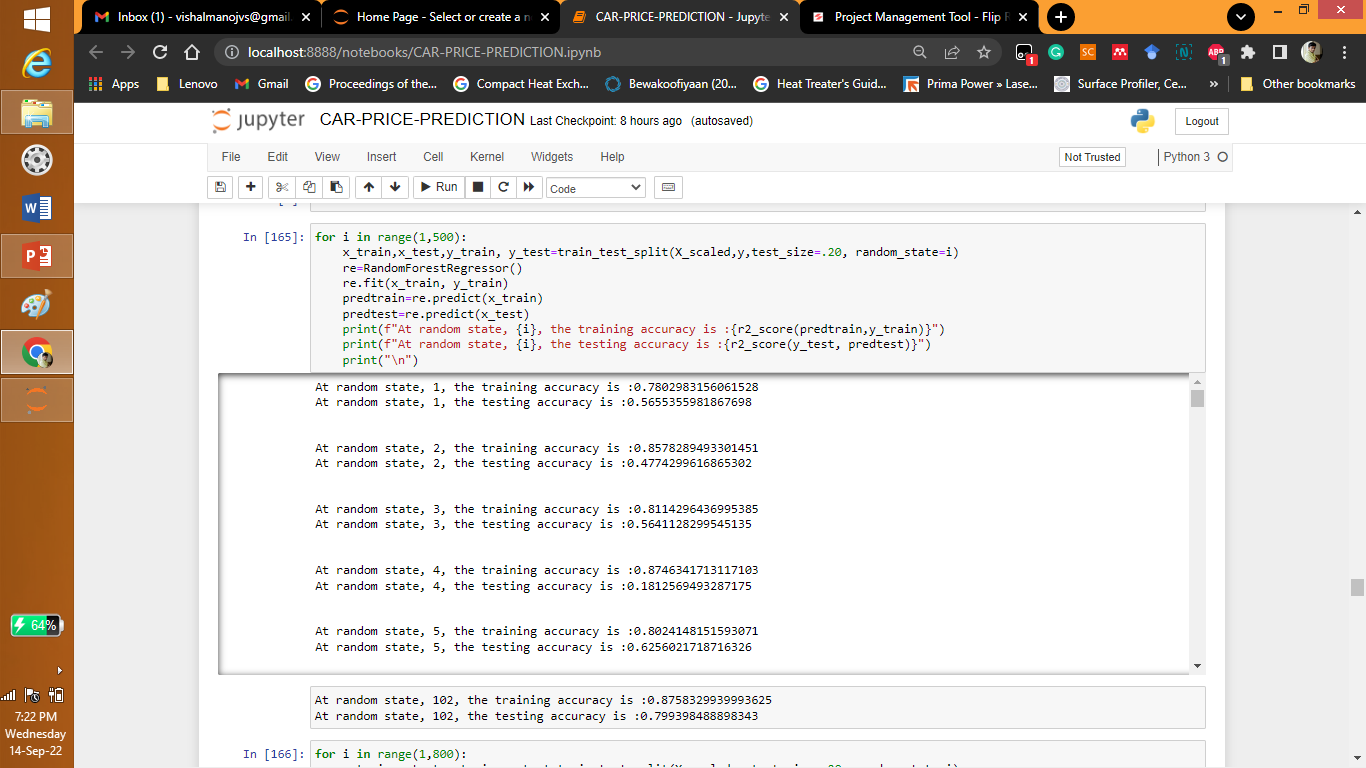
* Run and Evaluate selected models

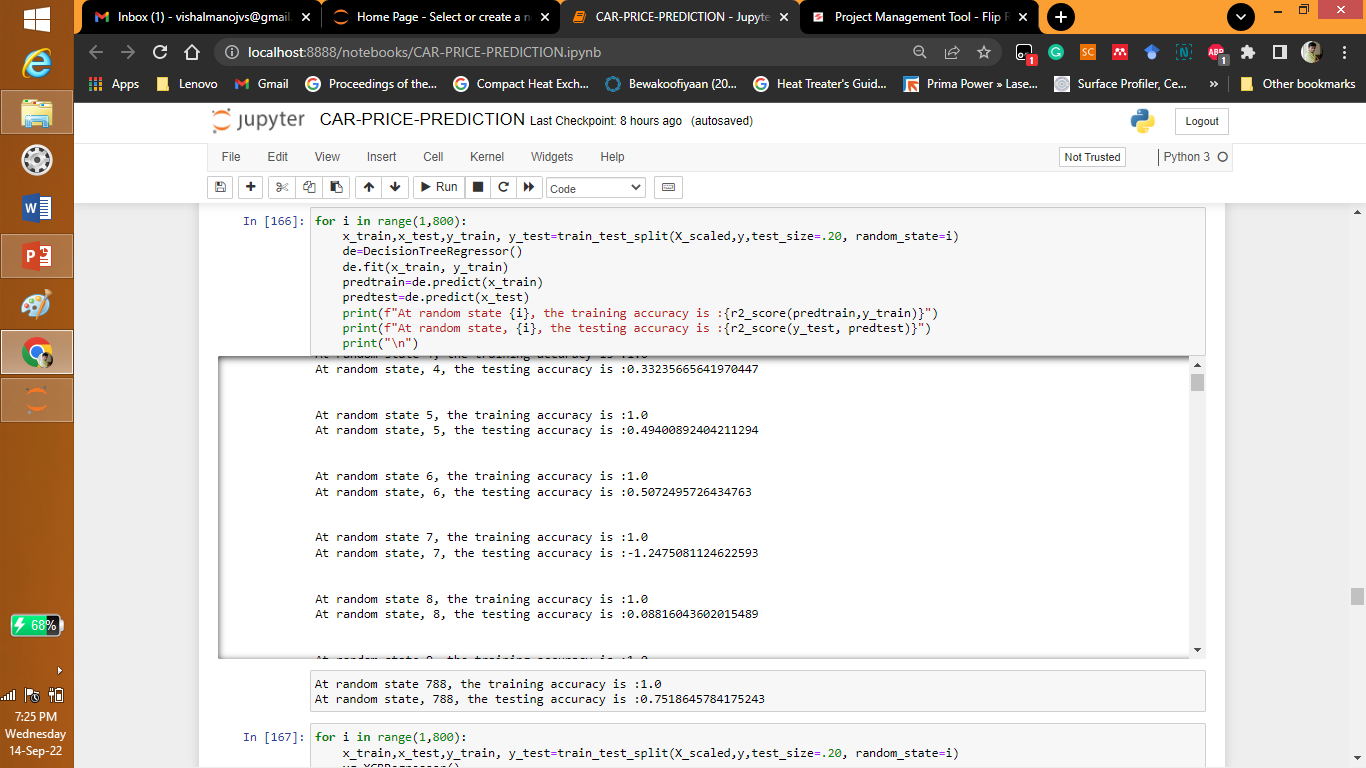
Describe all the algorithms used along with the snapshot of their code and what were the results observed over different evaluation metrics.

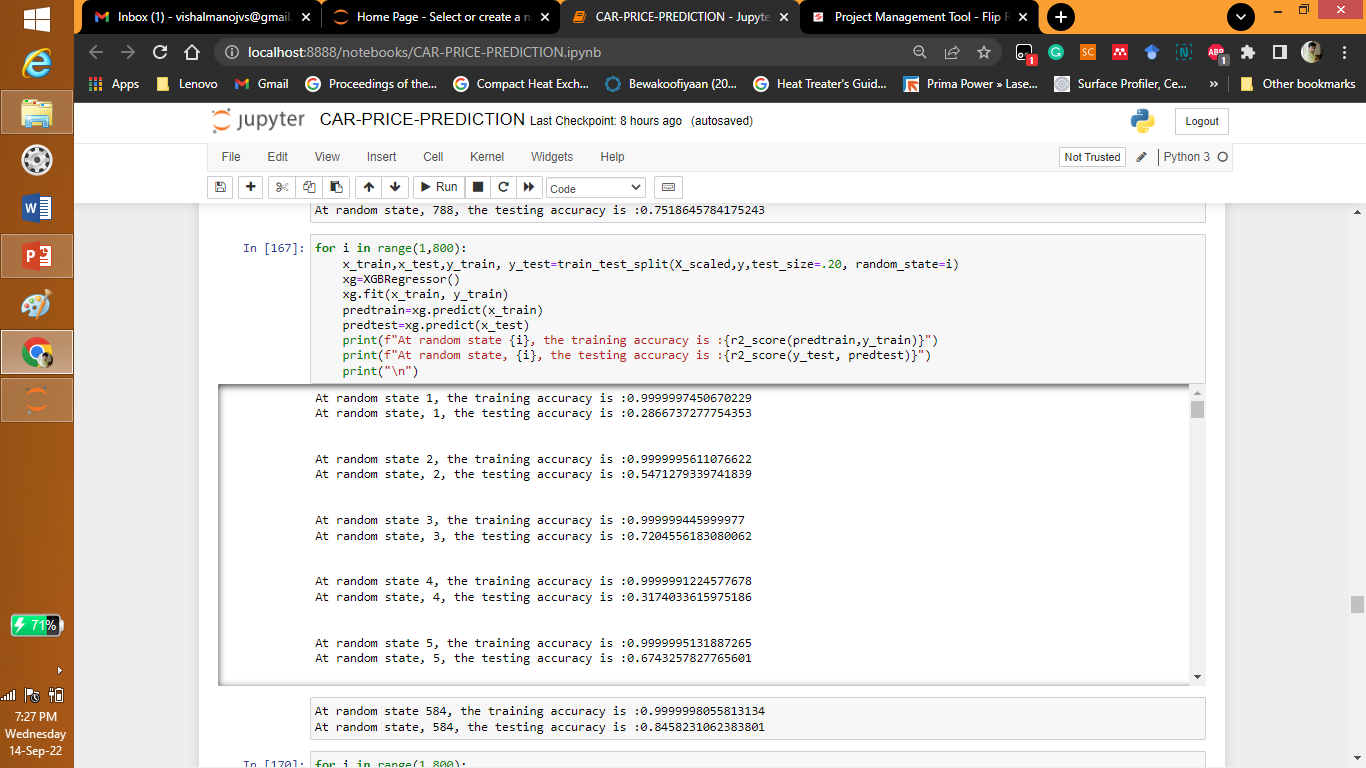
Answer:

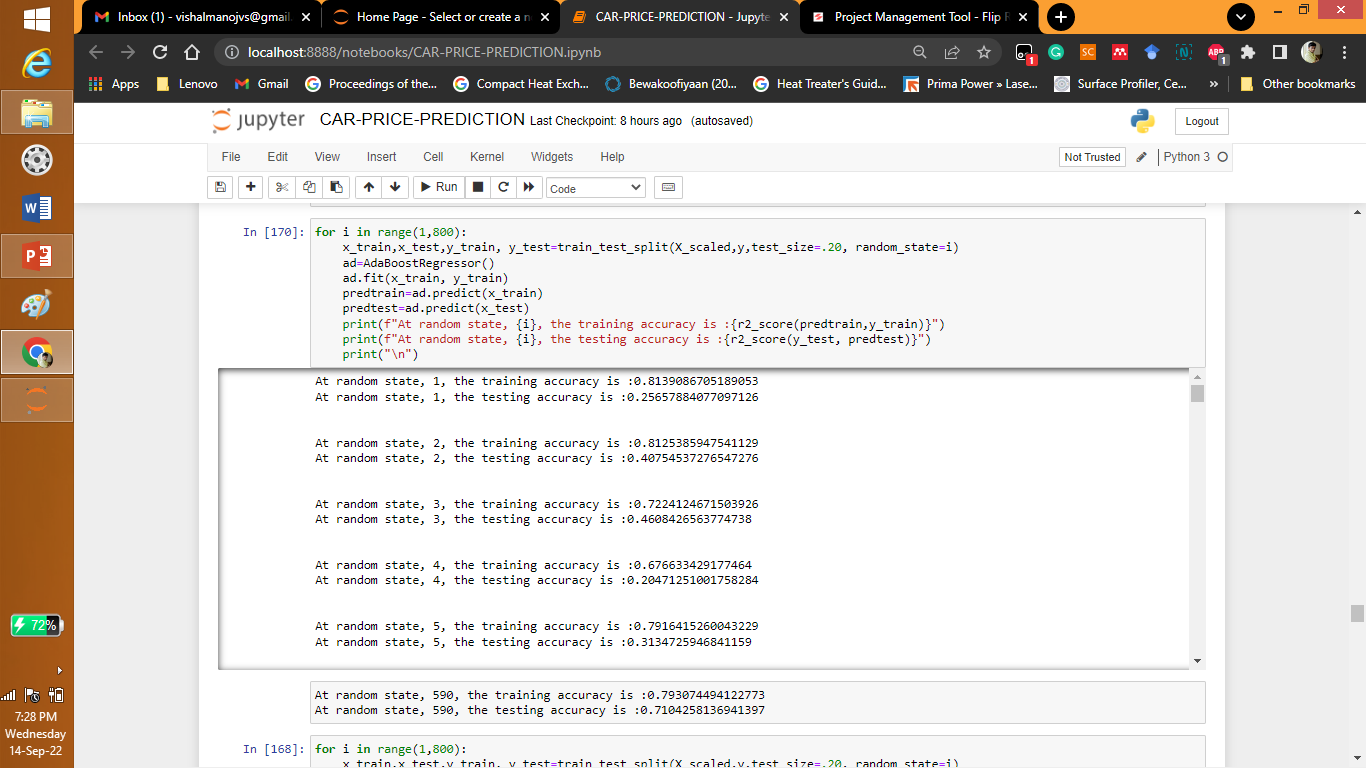


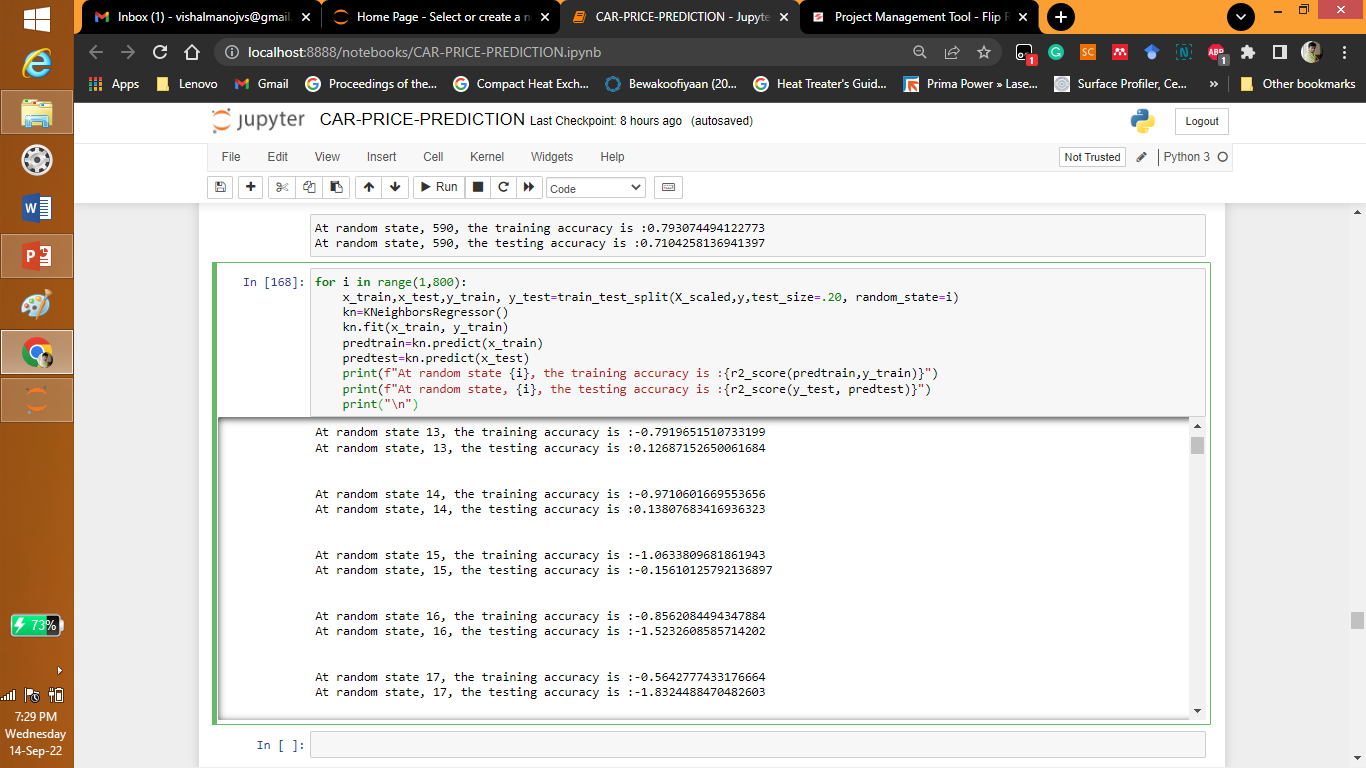


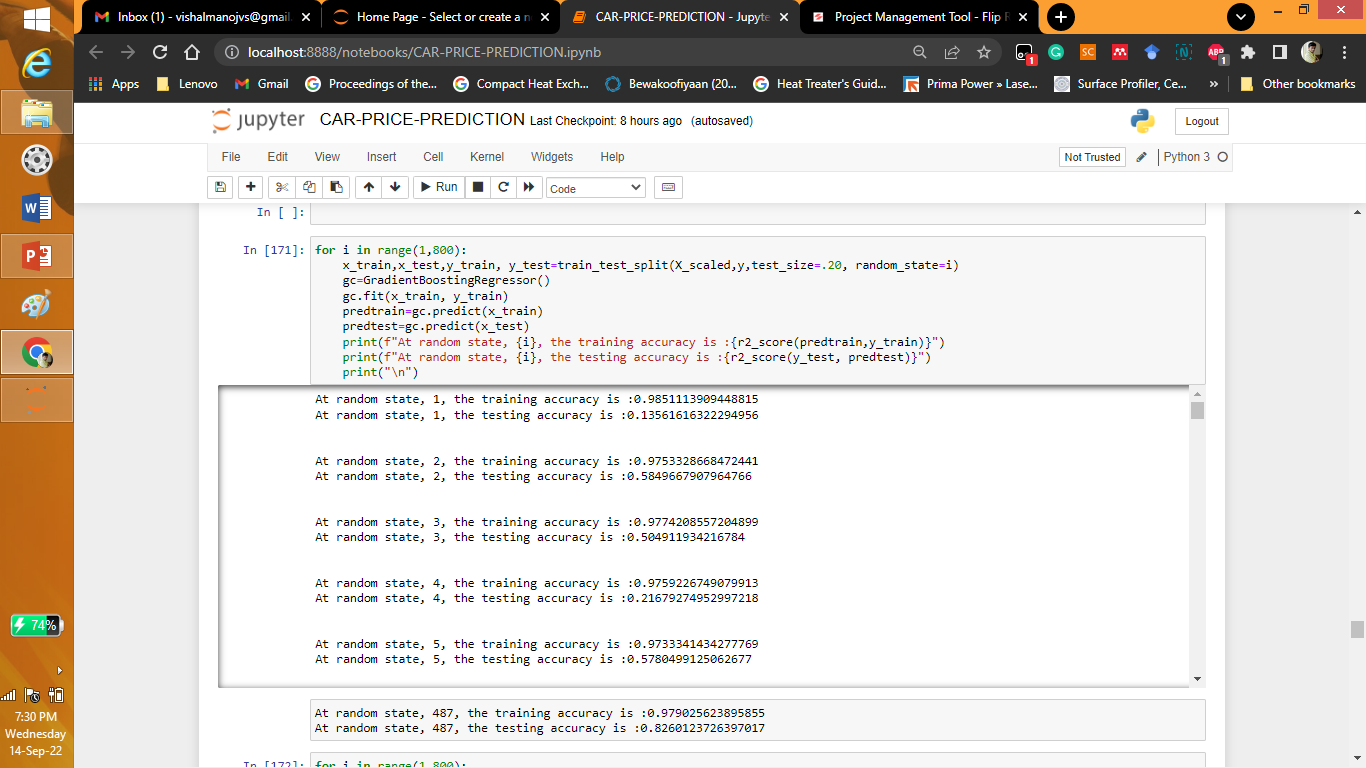


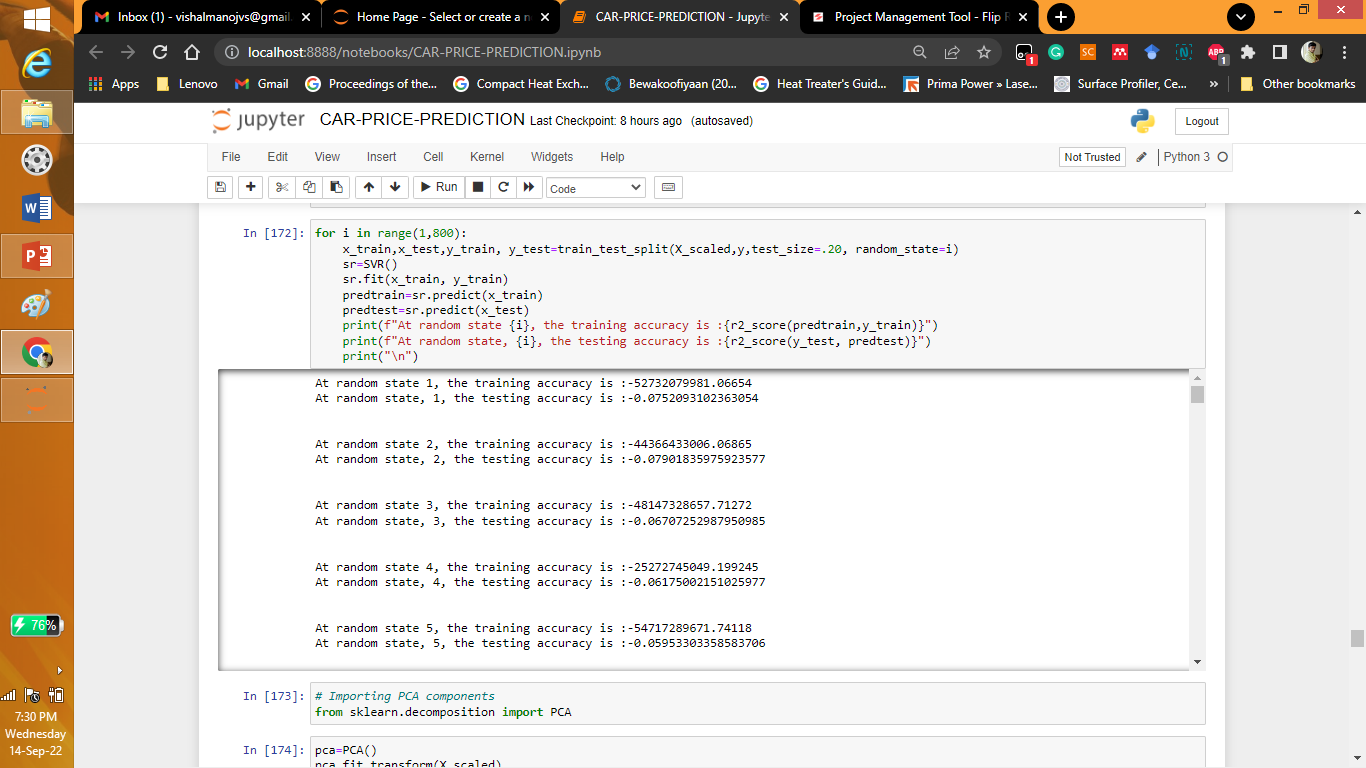


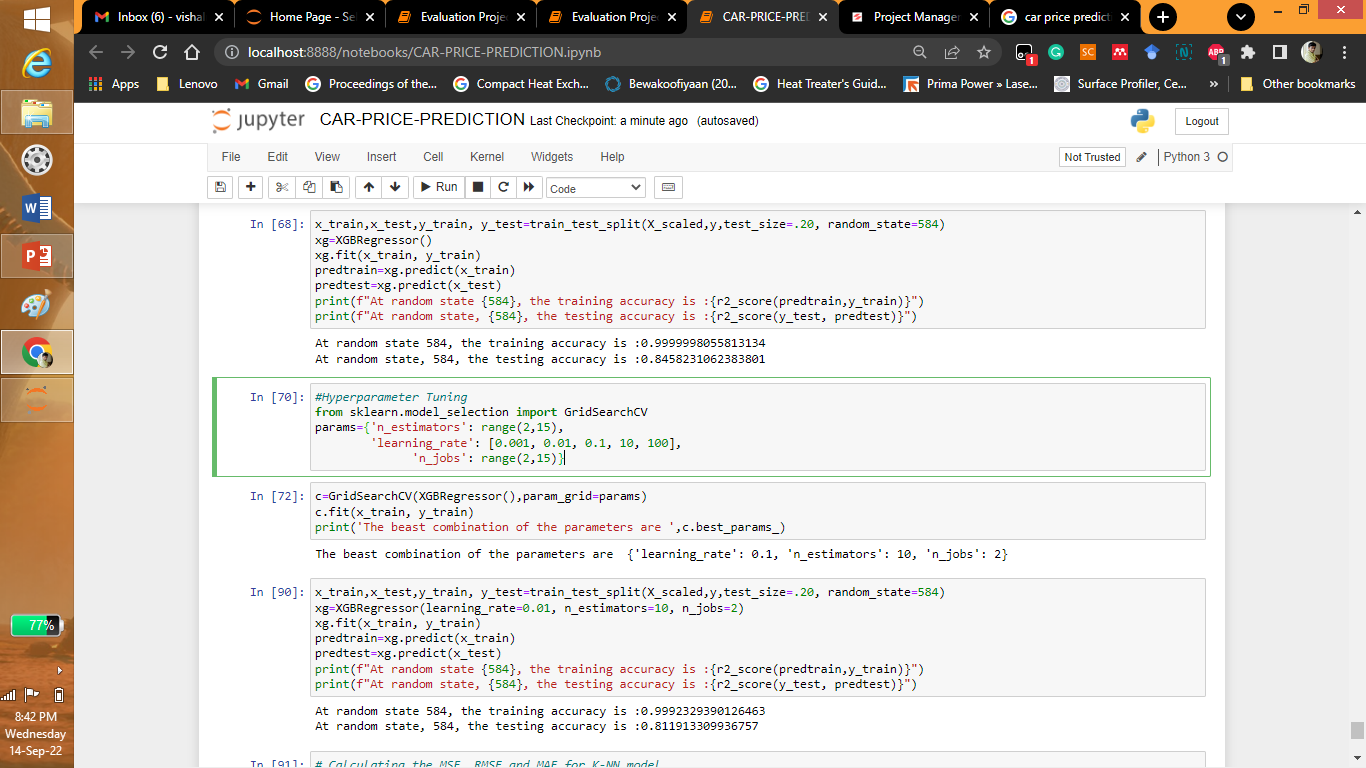


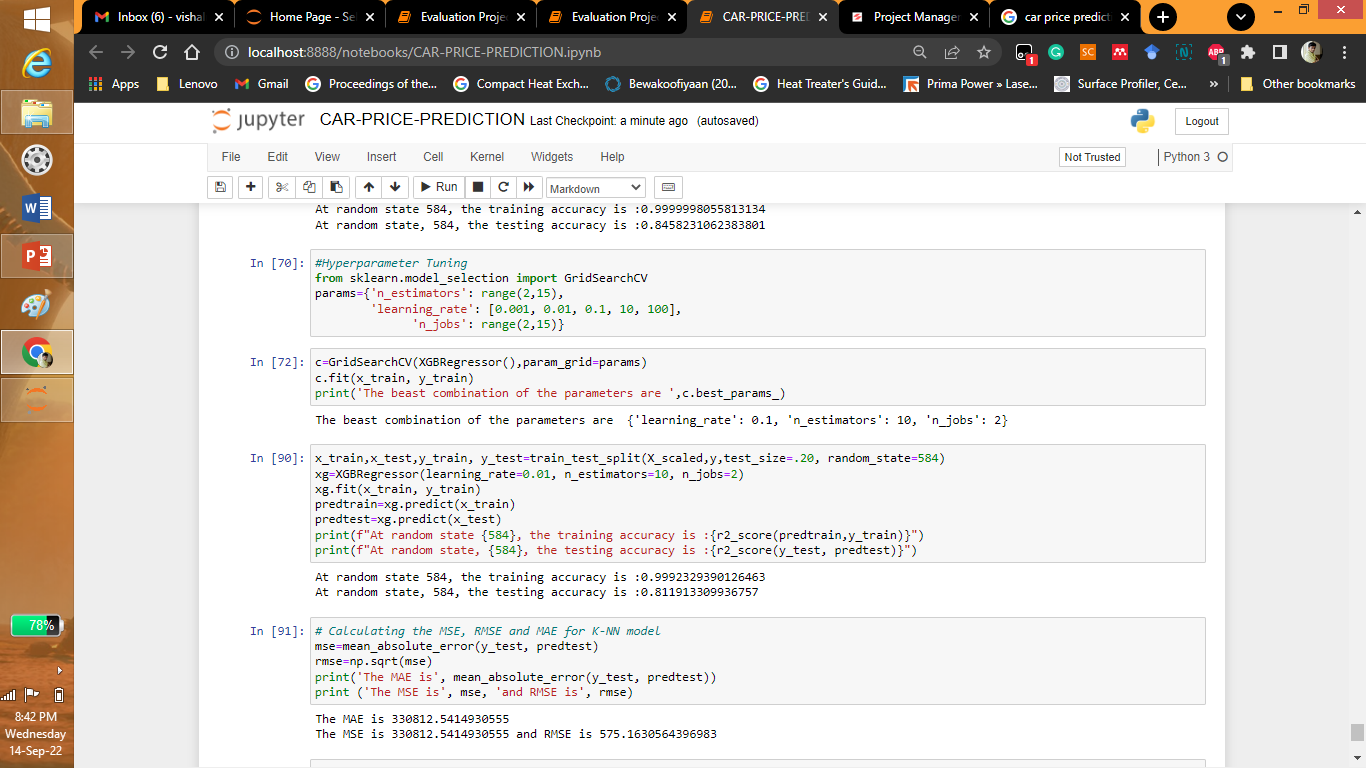












* Key Metrics for success in solving problem under consideration

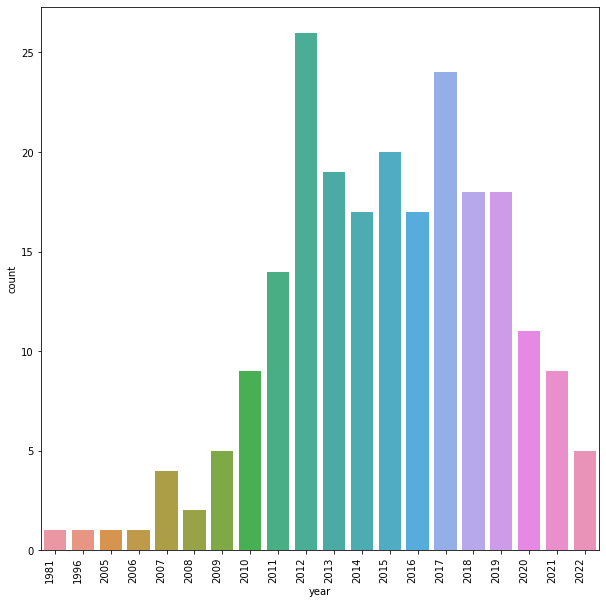
What were the key metrics used along with justification for using it? You may also include statistical metrics used if any.

* Visualizations

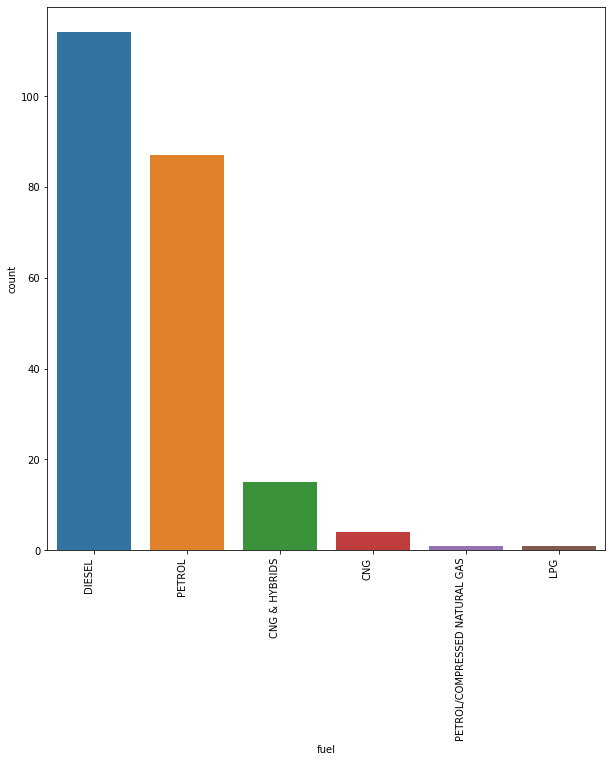
Mention all the plots made along with their pictures and what were the inferences and observations obtained from those. Describe them in detail.

If different platforms were used, mention that as well.

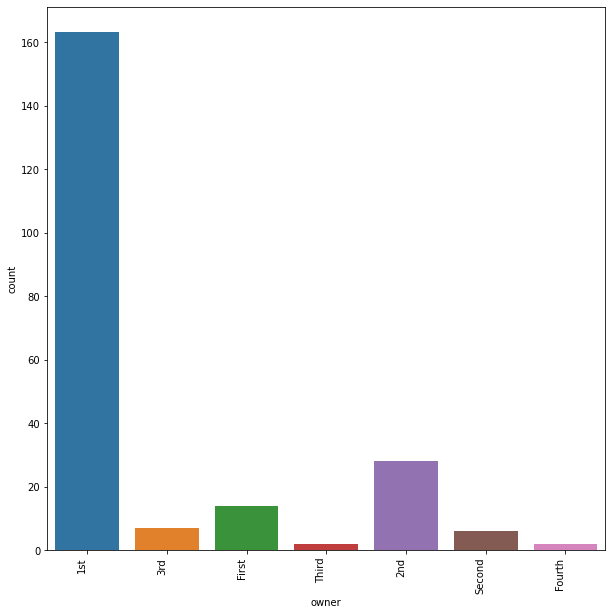
Answer:



From the graph we can see that 2012 and 2017 manufactured vehicles are for sales in the market.



The diesel fuel is the most selling vehicles in the olx website followed by petrol and CNG and hybrids.



The 1st owner vehicles are the most selling followed by 2nd owner vehicles.

* Interpretation of the Results

Give a summary of what results were interpreted from the visualizations, pre-processing and modelling.

Answer: We can see that 2012 and 2017 manufactured vehicles, 1st owner vehicles and diesel fuel are for sales.

**CONCLUSION**

* Key Findings and Conclusions of the Study

Describe the key findings, inferences, observations from the whole problem.

Answer: We can see that 2012 and 2017 manufactured vehicles, 1st owner vehicles and diesel fuel are for sales.

* Learning Outcomes of the Study in respect of Data Science

List down your learnings obtained about the power of visualization, data cleaning and various algorithms used. You can describe which algorithm works best in which situation and what challenges you faced while working on this project and how did you overcome that.

Answer: It can be seen the model was web scrapped. Then it was made as a dataset which was used for modelling. The data is cleaned and the different graphs were plotted for explorative data analysis. Then the label encoder is used and categorical variables were converted to numerical format. Different model were used and XgBoost model was seen to be most efficient. After hyper-parameter tuning is at random state 584, the training accuracy is 0.9992329390126463. At random state, 584, the testing accuracy is 0.811913309936757

* Limitations of this work and Scope for Future Work

What are the limitations of this solution provided, the future scope? What all steps/techniques can be followed to further extend this study and improve the results?

Answer: More data can be scrapped and applied for prediction. The solution provided only has the accuracy of 81% approximately by using neural networks this accuracy can be increased.