**Abstract-** The prediction of car price has always been the area of high research interest, as it requires a large amount of effort to be made, and the knowledge of field experts. There is a large number of different attributes to be evaluated so that the prediction can be both trustworthy and accurate. In order to build a model so that it can predict the price of used cars in India, we have applied three machine learning techniques (i.e., Random Forest, Linear Regression, Lasso Regression) [17,18]. The data which we gathered from the cardekho.com website is used to predict the price of used cars. After having respective performances on all these three algorithms on the given dataset were compared so that we can find out which one suits them best on the available dataset. The predicted model was then integrated with the Javascript application. Furthermore, the model was tested on the test dataset and an accuracy of 84 percent was obtained.

**Keyword: - Car Price prediction, Random Forest, Machine learning.**

1. **Introduction**

Car price prediction is one of the popular and interesting one. As per the information that we have got from the websites of cardekho that more than 50000 cars are currently registered under cardekho. Mainly the reselling of cars can be done in two ways either offline or on internet. In offline mode there is a middle man who is present between the buyer and seller who is making a huge amount of profits over the transaction. The second alternative is through online mode where the customers can get the best price for their cars. If he decides to sell it

* **Kilometer Travelled: -** As we know that the price is also estimated on the basis of a kilometer traveled. The more the vehicle has traveled, the older that vehicle is.
* **Year of registration: -** Basically the year of registration contains the information about the time when that automobile has been registered with the Road Authority of India. Newer the registration of the automobile more will be the price of the vehicle, and by the passing year, its price gets decreased.
* **Fuel Type: -** Currently the cars come up with three types of fuel i.e., Diesel, Petrol, and CNG, which will also help to predict the price of the cars.

Due to the factor mentioned above, we have developed a machine learning system that can predict the price of old cars.

**OBJECTIVES**

* To prepare a supervised learning model, so that it can estimate the selling price of cars based on various aspects.
* The prepared model must be capable of making predictions based on various features.
* Provide the comparison on the behalf of the graph for better visualization.

**MOTIVATION**

The Automobile business has been dominated by a few multinational corporations and a number of retailers across the world. Mainly the multinational corporation are manufacturers by trade but the retailer deals with both the new and secondhand cars. The used automobile market has increased significantly, the value increased contribute by a bigger percentage of the total market. In India the secondhand vehicle in accounts for around 3.5 million of vehicle annually.

**FEATURES**

Majorly there are two features given in this project

* Re-sale platform: - A centralized platform that will predict the actual price of the cars.
* Feature Selection: - Feature-based prediction and search.

In this section, we mention the introduction of the project, the objective, project’s motivation, and features of the module. Further we will be covering literature review after that in section III we will cover the technology used which is being used in machine learning, Section IV will explain the methods we used, Section V will discuss the results, and Section VI will discuss the conclusion and future work.

**II LITERATURE REVIEW**

Here, we'll go through some of the applications and approaches that motivated us to create this project. We conducted an analysis on the core concepts of this project and utilized those notions to collect information such as the technological stack, algorithm, and flaws of this project, which lead us to develop a better working model.

**CarDekho**

CarDekho is a website where sellers may sell old vehicles or cars.

This startup has a simple UI that asks sellers for information such as model of car, distance traveled, year of registration, and type of vehicle (petrol, diesel, CNG). This enables online model to execute certain algorithms on Applied parameters in order to estimate the selling price.

**GET VEHICLE PRICE**

This is a website that predicts the same as that of the CarDekho Android app, which works on some algorithm that predicts on the basis of a kilometer traveled, fuel type, age of the car. The site uses a ML approach to estimate the price of the vehicle. It will help you to compare the price with different sellers.

**III TECHNOLOGY USED**

Here, we have used python for machine learning as a major technology for implementing the concepts because of its inbuilt function and libraries package available in python. Some of the major libraries used here are: -

**NUMPY**

This is a Python-based array-processing package. It comes with a high-performance multidimensional array object and utilities for manipulating them. It's a Python module that's required for scientific computing. NumPy may be used as a multi-dimensional container of general data in addition to its normal applications.

NumPy can define any data type, allowing it to interact with a wide range of databases fast and easily.

**SCIKIT-LEARN**

It provides a variety of supervised and unsupervised learning methods through a uniform Python interface. It is provided under various Linux distributions and is licensed under a liberal simplified BSD license, promoting academic, commercial use. The library is being constructed.

**MATPLOTLIB**

Matplotlib is a library used in a python programming language, which is being used for plotting the graph with having a mathematical numerical extension NumPy, which serves as visualization utility.

**SEABORN**

Seaborn is a Python data visualization framework built on top of matplotlib and tightly integrated with pandas data structures. Seaborn's visualization is a critical component that aids in data exploration and comprehension.

**PANDAS**

Pandas are generally based on two fundamental Python libraries: matplotlib for data visualization and NumPy for mathematical calculations. Pandas functions as a wrapper around these libraries, enabling you to use fewer lines of code to access numerous matplotlib and NumPy methods.

Pandas are also used for importing the data from the files.

**IV METHODOLOGY**

Here, we discussed the various algorithm and the required dataset were administered on our model. The training of the dataset has been done on a random basis. The dataset is having several attributes such as kilometer traveled, fuel type, transmission type, and age of the cars, which will determine the price of the cars[2]. Since we have tested our model five algorithms RandomForestRegression, Lasso Regression, Linear Regression, Ridge Regression, and Support Vector Regression (SVR), After applying all these algorithms we have obtained which fits best with the model to predict the actual price of it [5,6].

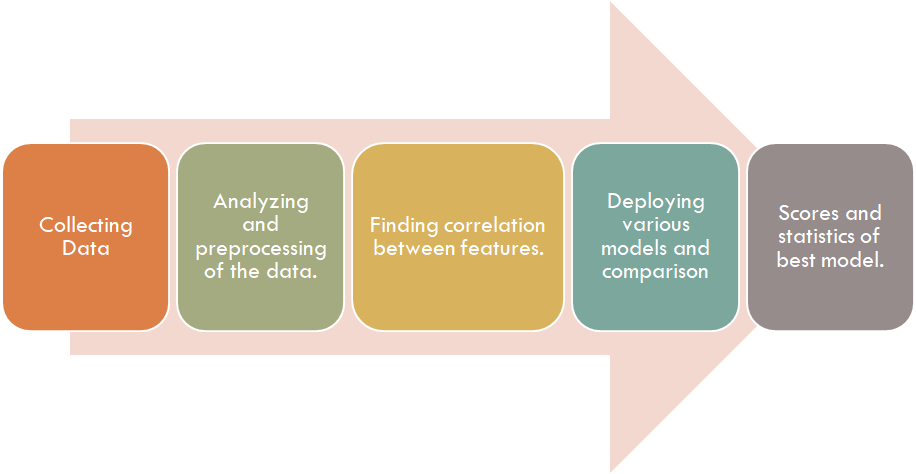


Figure 4.1 Shows the Methodology of our model

**DECISION TREE REGRESSION**

Decision tree learning utilizes a decision tree (as a predictive model) to make predictions about the target value of an object (represented in the fields) (represented in the leaves). The modeling technique in statistics, data mining, and ML [11] is one of the predicting approaches. Tree models in which the objective variable is capable of taking a discrete set of values are termed classification trees; in this type of structures, the leaven is a class label. Decision Trees usually take the continuous values so that they are called regression trees. The main objective is to predict the price of the cars on the basis of several inputs’ variable that the decision tres variable generally uses in mining the data.

**V EXECUTION**

Here, we have discussed the steps and the method that has been implemented in our model and by plotting some of the comparison graphs, and several statical graphs, to study the best algorithm which works on our model. [14]

Firstly, we checked any null entries in our dataset, which we found no null entries in our dataset. After that, we have added current year columns in our dataset so that we can find the age of the car, then in the next step, we have pre-processing and cleaning of our dataset. We have plotted the heat map to show the correlation within our dataset, owner, fuel type, kilometer traveled, present price, and the age of the car. [16,17]

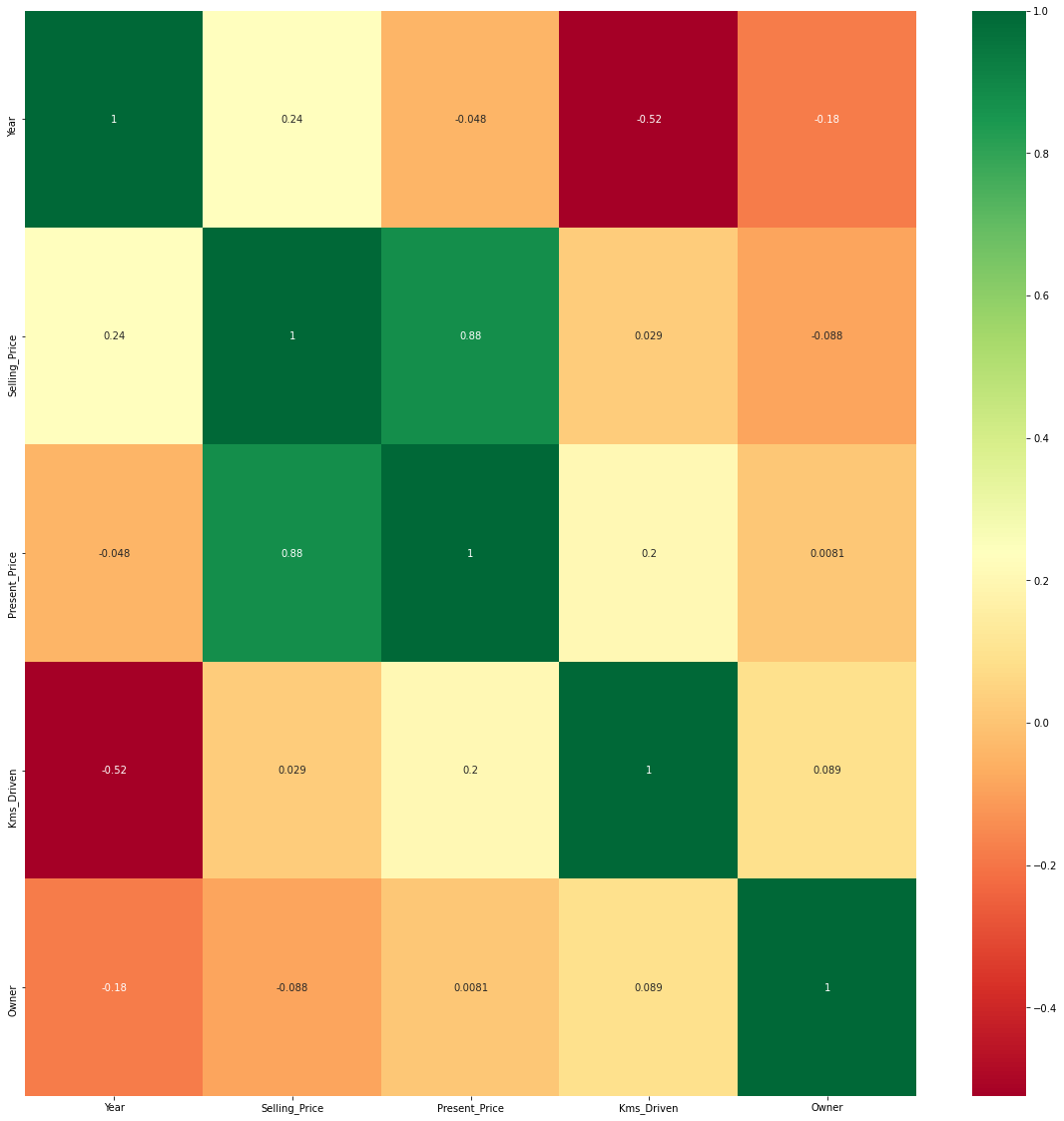


Figure 5.1 shows the heat map of the dataset, which shows the correlation between different features.

We have split the dataset in which 70% of the data has been used for training and the rest 30% of them have been used for testing.

From sklearn, we have imported RandomForestRegression and Linear Regression to predict the price of the car.

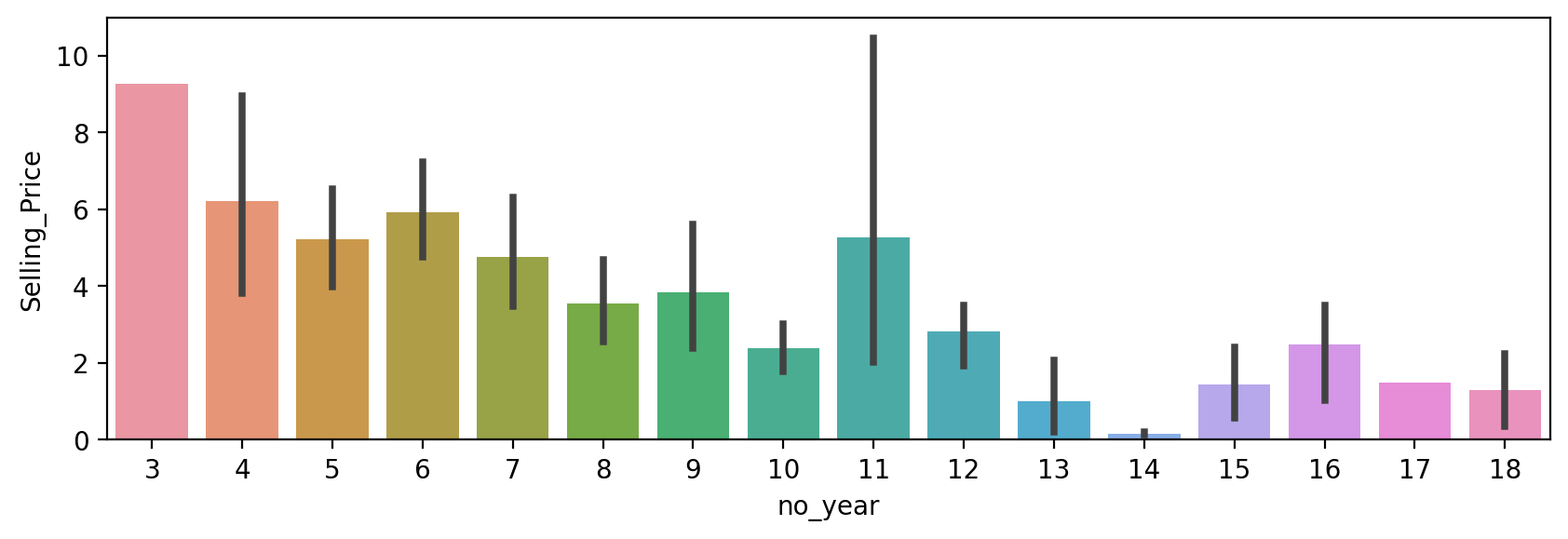


Figure 5.2 We have plotted here the bar graph between the selling price and the age of the car (no\_year).

As we have observed here that the cars having age of 3 years has a maximum selling price.

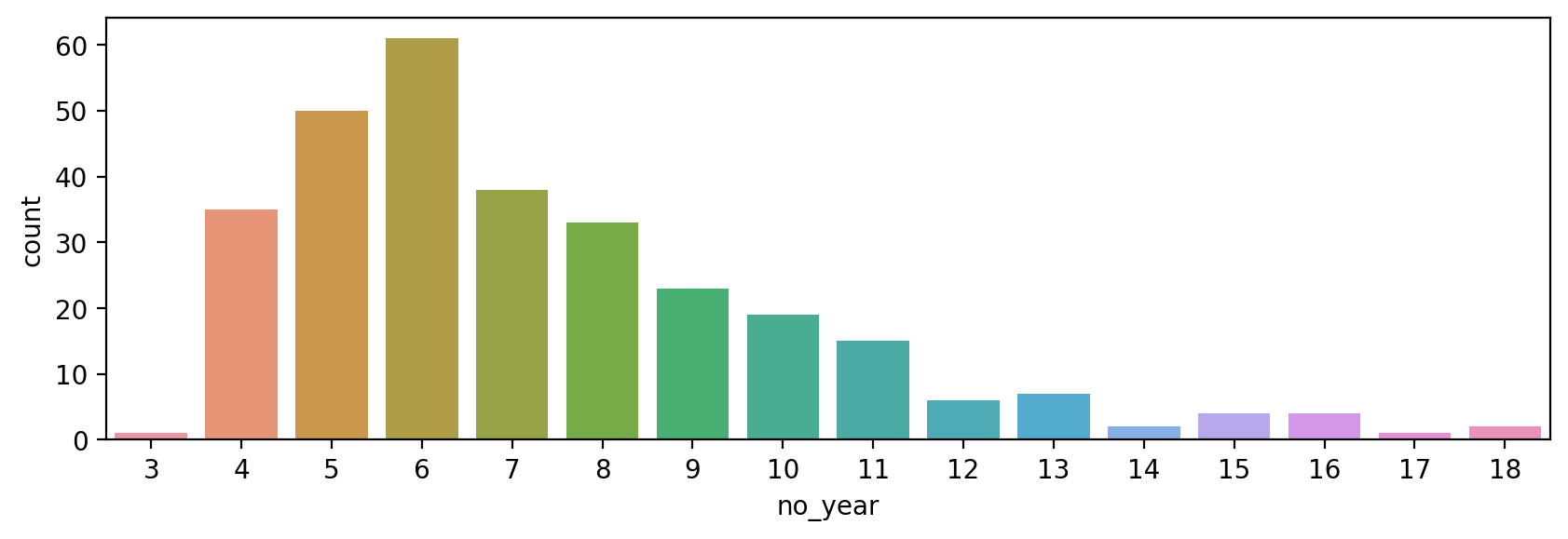


Figure 5.3 shows the graphs between the age(no\_year) and count(total\_numbers) of cars.

As we observe here that cars having an age of 6 years are of maximum numbers in our dataset.

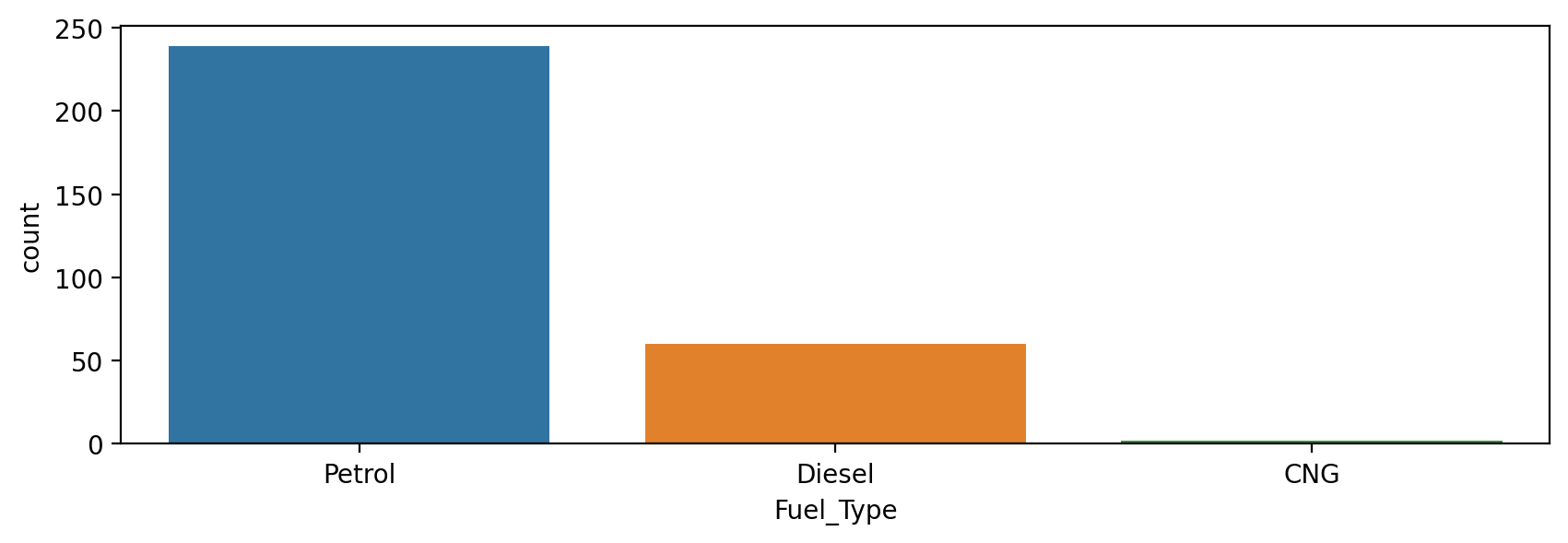


Figure 5.4 shows the graphs between fuel type and the number of cars(count).

As we observe here that the cars had petrol as fuel type are sold more than any others.

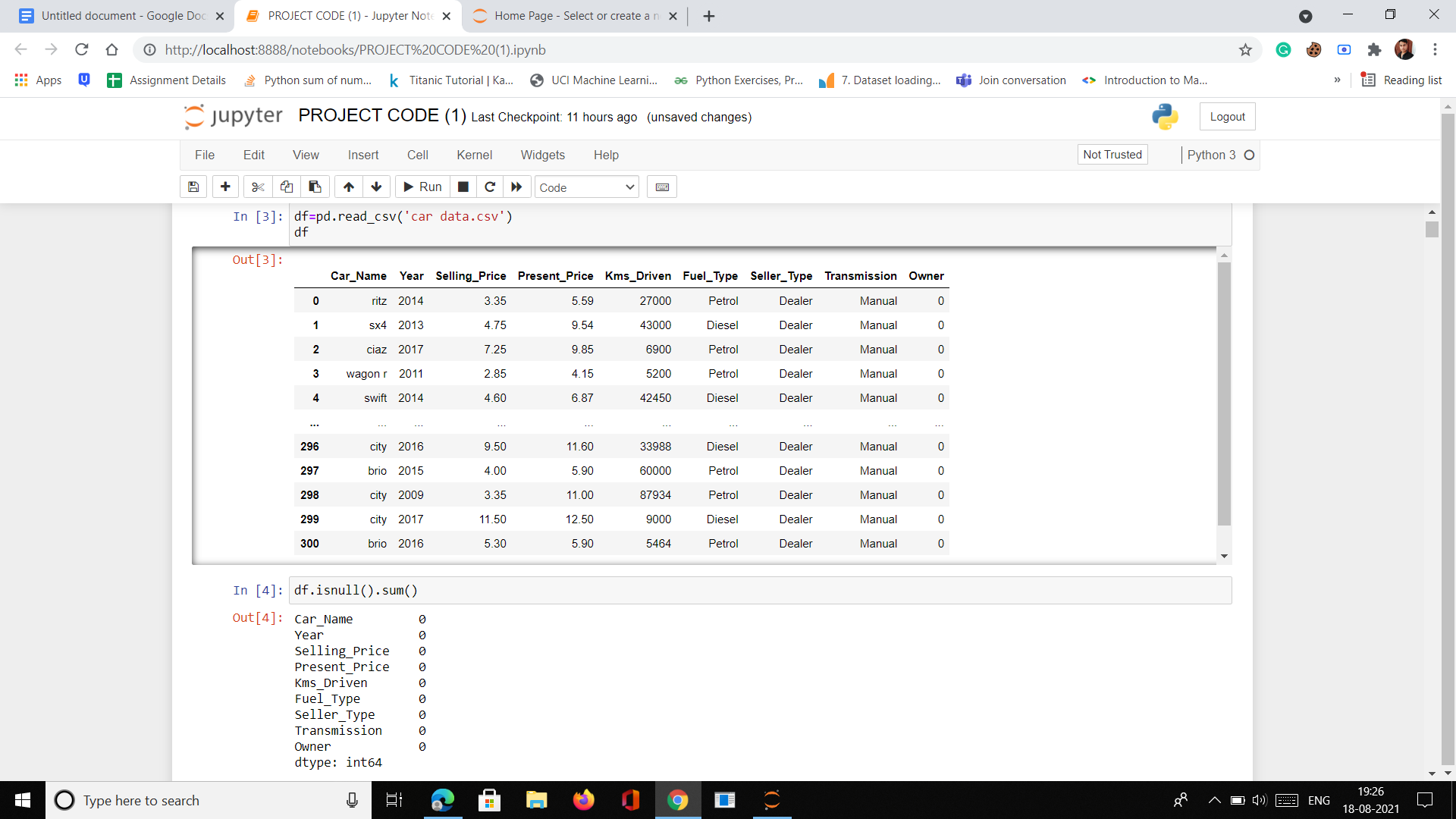


Figure 5.5 Snapshot of dataset header

Fro fig 4.5 as you can see in the header of our dataset and there are having no null values in our dataset.

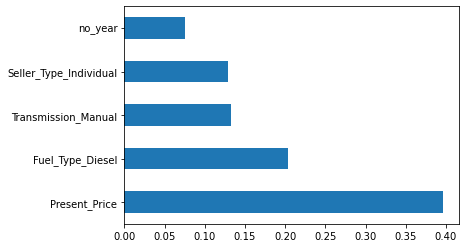


Figure 5.6 Show the graphs of feature importance

As you can see the present price has more importance than any others.

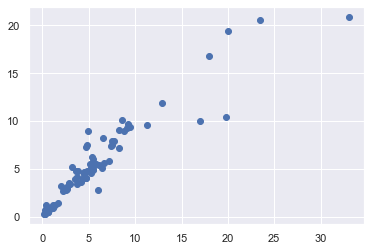


Figure 5.7 showing the scatter plot of random forest regression between testing and predictions. [7,8]



Figure 5.8 showing the scatter plot of the graph for linear regression[7,8]



Figure 5.9 shows the scatter plot for Lasso Regression [7,8]



Figure 5.10 showing the scatter plot for Ridge Regression[7,8]

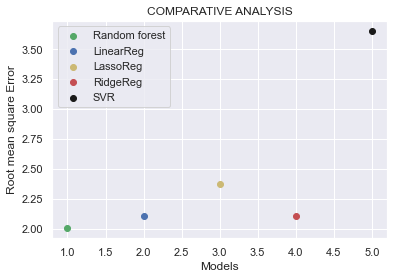


Figure 5.11 Here the scatter plot shows the comparative analysis of the Random Forest, Linear Regression, Lasso regression, and SVR. [7,8]

As we can observe here the Random Forest has more accuracy and SVR has the least accuracy as it shows the outline of the plot.

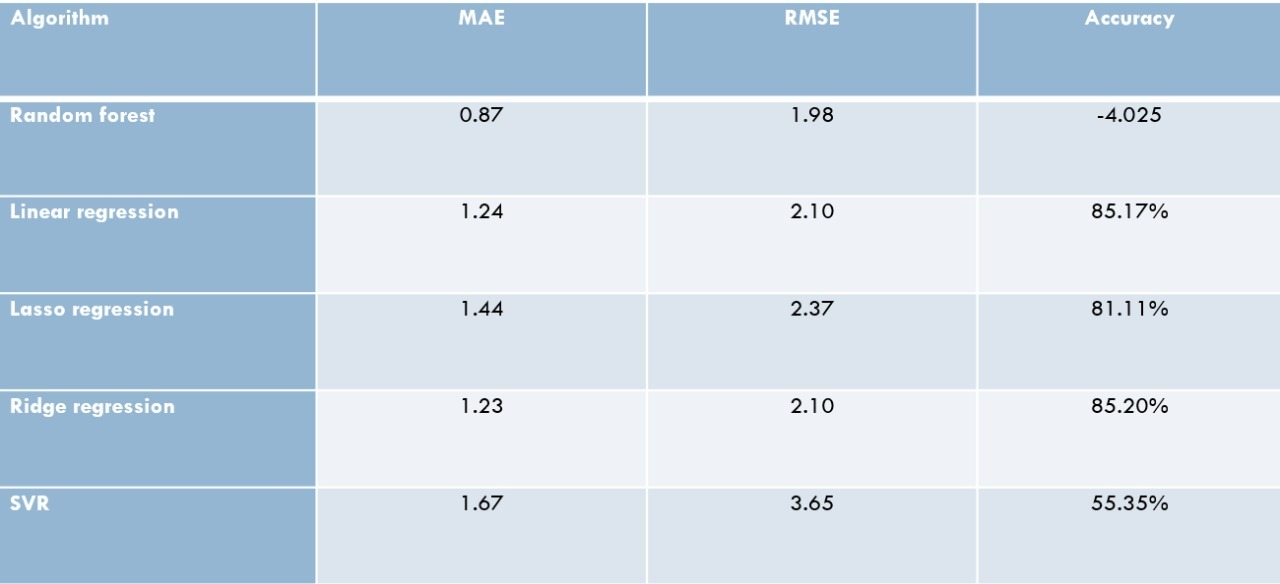


Figure 5.13 This table shows the Accuracy table of different Algorithm.

As we can see here this table shows the accuracy of a different algorithm, so that we can choose that which will shoot the best for our model.

**VI CONCLUSION**

Here, we will go through the findings and observations we made while putting this project together. We successfully implemented ML algorithmic model in Python utilizing well-known algorithms from libraries. On our dataset, we first do pre-processing and data cleaning. The findings revealed a positive association between price and kilometers traveled, a negative interrelation between price and year of registration, and a positive interrelation between price and year of registration.

After testing our dataset on a different algorithm, we found that the Random Forest algorithm has maximum accuracy of 84%. That’s why we can say Random Forest best suits our model because Decision Trees itself are a low bias and high variance model but when decision trees combine to form a random forest their accuracy increases on testing data as well.[20]

That’s why random forest has low bias and low variance thus yields better output.

**FUTURE WORK**

As part of future work, we aim to enlarge our dataset by adding datasets from Olx, Cars24, and others so that our predictions should be more accurate.

¨Doing more hyperparameter tuning (learning rate, batch size, number of layers, number of units, dropout rate, batch normalization, etc.). We can implement this on SVR and find out if the accuracy improves (i.e., use RandomisesSearchCV, GridSearchCV).

We are planning to develop a front-end for our model with the help of our python file and Html code. That we can have our model on websites for public use So that they can use it whenever they purchase the cars.

**REFERENCES**

[1].M. Antonakakis, T. April, M. Bailey, M. Bernhard, E. Bursztein, J. Cochran, Z. Durumeric, J. A. Halderman, L. Invernizzi, M. Kallitsis, D. Kumar, C. Lever, Z. Ma, J. Mason, D. Menscher, C. Seaman, N. Sullivan, K. Thomas, and Y. Zhou, "Understanding the Mirai botnet," in Proc. of USENIX Security Symposium, 2017.

[2].Iman Sharafaldin, Arash Habibi Lashkari, and Ali A. Ghorbani, ―Toward Generating a New Intrusion Detection Dataset and Intrusion Traffic Characterization‖, 4th International Conference on Information Systems Security and Privacy (ICISSP), Portugal, January 2018

[3].Hossein Hadian Jazi, Hugo Gonzalez, Natalia Stakhanova, and Ali A. Ghorbani. "Detecting HTTP-based Application Layer DoS attacks on Web Servers in the presence of sampling." Computer Networks, 2017

[4]. A. Shiravi, H. Shiravi, M. Tavallaee, A.A. Ghorbani, Toward developing a systematic approach to generate benchmark datasets for intrusion detection, Comput. Security 31 (3) (2012) 357–374.

[5].Z. He, T. Zhang, and R. B. Lee, ―Machine Learning Based DDoS Attack Detection from Source Side in Cloud,‖ in Proceedings of the 2017 IEEE 4th International Conference on Cyber Security and Cloud Computing (CSCloud), pp. 114–120, New York, NY, USA, June 2017

[6].R. Doshi, N. Apthorpe, and N. Feamster, "Machine Learning DDoS Detection for Consumer Internet of Things Devices," 2018 IEEE Security and Privacy Workshops (SPW), San Francisco, CA, 2018, pp. 29-35.

[7].Jerome H. Friedman, (2002), Stochastic gradient boosting, Computational Statistics & Data Analysis, 38, (4), 367-378

[8].Friedman, Jerome H. Greedy function approximation: A gradient boosting machine. Ann. Statist. 29 (2001), no. 5, 1189--1232.

[9].https://www.ijcseonline.org/pub\_paper/74-IJCSE-07051.pdf

[10].https://www.ceeol.com/search/article-detail?id=746689

[11] MITCHELL, T. M., 1997. Machine Learning. McGraw-Hill, Inc. New York, NY, USA.

[12] DATA MINING SOFTWARE IN JAVA. 2014. Available from: http://www.cs.waikato.ac.nz/ml/weka/index.html [Accessed 17 January 2014].

[13] LEXPRESS.MU ONLINE. 2014. Available from: http://www.lexpress.mu/ [Accessed 17 January 2014]

[14] Gareth, J., Daniela, W., Trevor, H., & Tibshirani, R. (2013). An Introduction to Statistical

[15] Hurwitz, E., & Marwala, T. (2012). Common mistakes when applying computational intelligence and machine learning to stock market modelling. arXiv preprint arXiv:1208.4429

[16] Pal, N., Arora, P., Kohli, P., Sundararaman, D., & Palakurthy, S. S. (2018, April). How Much Is My Car Worth? A Methodology for Predicting Used Cars’ Prices Using Random Forest. In Future of Information and Communication Conference (pp. 413–422). Springer, Cham.

[17] Raschka, S., & Mirjalili, V. (2017). Python machine learning. Packt Publishing Ltd.

[18] IBM Knowledge Center. (n.d). Use of KNN. Retrieved from: <https://www.ibm.com/support/knowledgecenter/SSHRBY/com.ibm.swg.im.dashdb.analytics.doc/doc/r_knn_usage.html>

[19] Agencija za statistiku BiH. (n.d.), retrieved from: http://www.bhas.ba . [accessed July 18, 2018.

[20] Pudaruth, S. (2014). Predicting the price of used cars using machine learning techniques. Int. J. Inf. Comput. Technol, 4(7), 753-764.