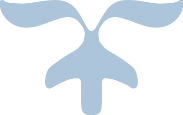


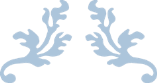
TECHNEX IIT (BHU) PROJECT REPORT

ON DATA ANALYTICS AND MACHINE LEARNING

EISYSTEMS SERVICES



**Submitted by Pavan V Chikkodikar**



Batch Code – 21DEEP - 069

Submitted to

Mr. Mayur Dev Sewak Mr. Shailendra Singh

General Manager, Operations Trainer, ICT Domain, Logicpro Infosystems

EISystems Services EISystems Services



**ACKNOWLEDGMENT**

It is a great pleasure to present this report on the topic named “Data Analytics and Machine Learning” undertaken by me as part of my B. Tech curriculum.

I am thankful to Technex’20 and IIT (BHU) Varanasi for offering me such a wonderful and challenging opportunity and I express my deepest thanks to HOD of Technex’20 for providing all the possible help and assistance and constant encouragement.

It is a pleasure that I find myself penning down these lines to express my sincere thanks to those people who helped me along the way in completing my project. I find inadequate words to express my sincere gratitude towards them.

First and foremost I would like to express my gratitude towards our training guide Mr. Shailendra Singh for placing complete faith and confidence in my ability to carry out this project and for providing me his time, inspiration, encouragement, help, and constant interest.

The internship on “Data Analytics and Machine Learning” was very helpful to me in giving the necessary background information and inspiration about the emerging field of machine learning and data science.

Their contribution and technical support in preparing this report are greatly acknowledged.

Last but not least, I wish to thanks my parents for financing my studies for Technex’20 as well as constantly encouraging me to learn engineering. Their personal sacrifice in providing this opportunity to learn engineering is gratefully acknowledged.

**TABLE OF CONTENTS**

1. Abstract……………………………………………………………05
2. Summary…………………………………………………………..06
3. Objectives………………………………………………………….7-12
4. Introduction
5. Motivation
6. Methodology
7. Software and Hardware Requirements…………………………….13
8. Data flow Diagram…………………………………………………14
9. Information about Dataset………………………………………….15
10. Source code…………………………………………………………16-27
11. Output Screenshots………………………………………………….28-29
12. Result……………………………………………………………….30
13. Future Work………………………………………………………...31
14. Conclusion………………………………………………………….32
15. References…………………………………………………………..33

**List of Figures**

1. Figure 1 Car selling image page no - 06
2. Figure 2 Linear Regression page no - 08
3. Figure 3 Random Forest page no - 09
4. Figure 4 Gradient Boost page no - 10
5. Figure 5 XGBoost page no - 11
6. Figure 6 LightGBM page no - 11
7. Figure 7 KMeans page no - 11
8. Figure 8 Deep Neural Network page no - 12
9. Figure 9 Data Flow Diagram page no - 14

**Abstract**

The demand for used cars has increased significantly in the past decade and it is prognosticated that with Covid-19 outbreak this requirement will augment considerably. Hence to enhance the reliability, with the expansion of the used car market, a model that can forecast the current market price of a used automobile on the basis of a variety of criteria. This analysis can be used to study the trends in the industry, offer better insight into the market, and aid the community in its smooth workflow. The aim of this research paper is to predict the car price as per the data set (previous consumer data like engine capacity, distance traveled, year of manufacture, etc.). The result of these algorithms will be analyzed and based on the efficiency and accuracy of these algorithms, the best one of them can be used for the said purpose.

**Summary**

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 United States. 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

Figure 1

**Objectives**

To develop a efficient and effective model which predicts the price of a used car according to user’s inputs. To achieve good accuracy. To develop a User Interface( UI ) which is user-friendly and takes input from the user and predicts the price.

1. Introduction

The used automobile market is a growing business with a market value that has nearly doubled itself in previous years. The rise of online websites and other tools like it have made it easier for both buyers and sellers to get a better understanding of the factors that determine the market value of a used car. Based on a set of factors, Machine Learning algorithms may be used to forecast the price of any automobile. The data set will include information on a variety of automobiles. There will be information regarding the vehicle's technical elements, such as the engine type, fuel type, the kilometers per liter, and more, for each car. There is no universal mechanism for establishing the retail price of used automobiles because different websites employ different methods to create it. By using statistical models to anticipate pricing, it is possible to obtain a preliminary price estimate without having to enter all of the details into the desired website. The main purpose of this study is to compare the accuracy of two different prediction models for estimating a used car's retail price. As a result, we offer a Machine Learning-based methodology for predicting the prices of secondhand cars based on their characteristics. The cost is calculated using the amount of characteristics. Then, to illustrate our findings, we construct a responsive website that includes all of the countless used car listings. Our efforts culminated in this deployed service, which integrates data, machine learning, and features. This methodology can assist consumers looking to purchase a used car in making more informed judgments. Customers can now look for all automobiles in a region without physical efforts, anytime and from any location. In this research, we used linear regression and lasso regression to develop a price model for used automobiles in a comparative research. Data was gathered from Kaggle for each algorithm. The main goal of this study is to discover the best predictive model for estimating the price of a used car.

1. Motivation

Deciding whether a used car is worth the posted price when you see listings online can be difficult. Several factors, including mileage, make, model, year, etc. can influence the actual worth of a car. From the perspective of a seller, it is also a dilemma to price a used car appropriately[2-3]. Based on existing data, the aim is to use machine learning algorithms to develop models for predicting used car prices.

1. Methodology

We utilized several classic and state-of-the-art methods, including ensemble learning techniques, with a 90% - 10% split for the training and test data. To reduce the time required for training, we used 500 thousand examples from our dataset. Linear Regression, Random Forest and Gradient Boost were our baseline methods. For most of the model implementations, the open-source Scikit-Learn package [7] was used.

1. Linear Regression was chosen as the first model due to its simplicity and comparatively small training time. The features, without any feature mapping, were used directly as the feature vectors. No regularization was used since the results clearly showed low variance.



Figure 2

1. Random Forest is an ensemble learning based regression model. It uses a model called decision tree, specifically as the name suggests, multiple decision trees to generate the ensemble model which collectively produces a prediction. The benefit of this model is that the trees are produced in parallel and are relatively uncorrelated, thus producing good results as each tree is not prone to individual errors of other trees. This uncorrelated behavior is partly ensured by the use of Bootstrap Aggregation or bagging providing the randomness required to produce robust and uncorrelated trees. This model was hence chosen to account for the large number of features in the dataset and compare a bagging technique with the following gradient boosting methods.



Figure 3

1. Gradient Boost is another decision tree based method that is generally described as “a method of transforming weak learners into strong learners”. This means that like a typical boosting method, observations are assigned different weights and based on certain metrics, the weights of difficult to predict observations are increased and then fed into another tree to be trained. In this case the metric is the gradient of the loss function. This model was chosen to account for non-linear relationships between the features and predicted price, by splitting the data into 100 regions.

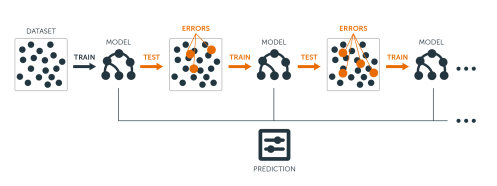


Figure 4

1. XGBoost Extreme Gradient Boosting or XGBoost is one of the most popular machine learning models in current times. XGBoost is quite similar at the core to the original gradient boosting algorithm but features many additive features that significantly improve its performance such as built in support for regularization, parallel processing as well as giving additional hyperparameters to tune such as tree pruning, sub sampling and number of decision trees. A maximum depth of 16 was used and the algorithm was run on all cores in parallel.

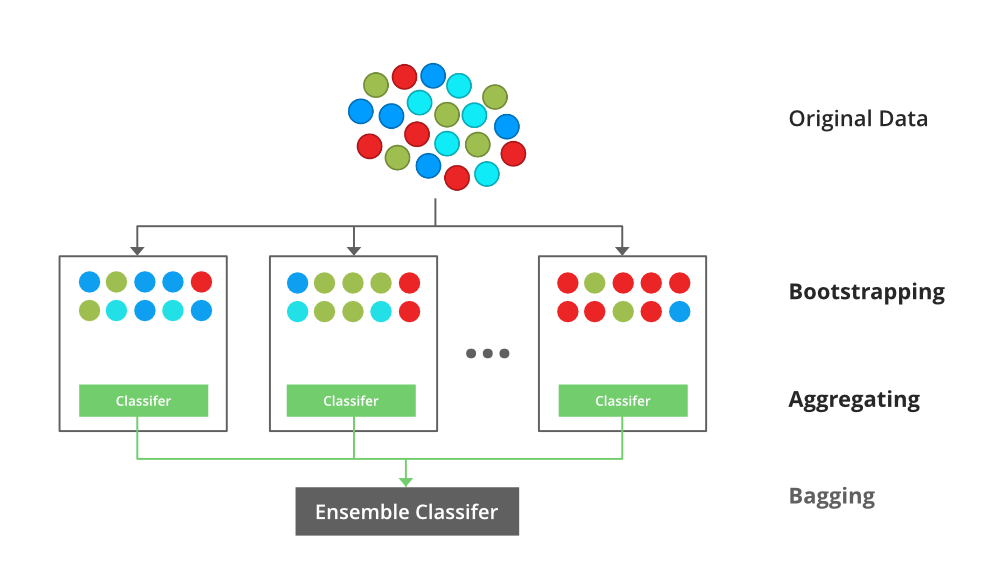


Figure 5

5. LightGBM is another gradient boosting based framework which is gaining popularity due it higher speed and accuracy compared to XGBoost or the original gradient boosting method. Similar to CS 229 Project Report Kshitij Kumbar, Pranav Gadre and Varun Nayak XGBoost, this LightGBM has a leaf-wise tree growth instead of a level-wise approach resulting in higher loss reduction. This framework can also handle categorical features [6], thus eliminating the need to one hot vectorize them and in turn, reducing memory usage. Make, Model and State and cities were declared as categorical features. The algorithm was run at tree depths in multiples of 12 and was run on all cores in parallel.

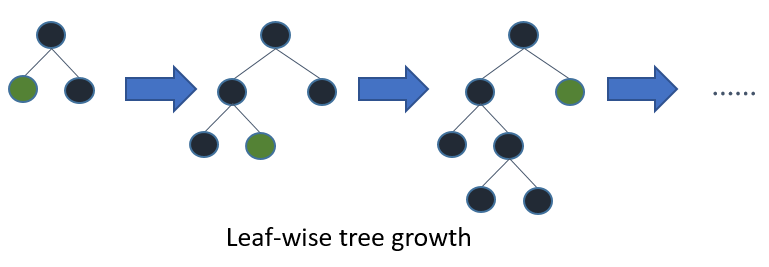


Figure 6

1. KMeans + Linear Regression In order to capitalize on the linear regression results and the apparent categorical linearity in the data as indicated in Fig. 2, an ensemble method which used KMeans clustering of the features and linear regression on each cluster was used. Due to large training time, a three-cluster model was used. Then, the dataset was classified into these three clusters and passed through a linear regressor trained on each of the three training sets.



Figure 7

1. Deep Neural Network (MLP Regressor) To introduce mode complexities in the model, the MLP regressor [7], which uses a deep neural net perceptron regressor model, was used. This model optimizes the squared-loss using LBFGS or stochastic gradient descent. Two hidden layers of width 200 and 20 were used. The learning rate was set at 0.001 and batch size at 200. ReLu was used as the activation function.

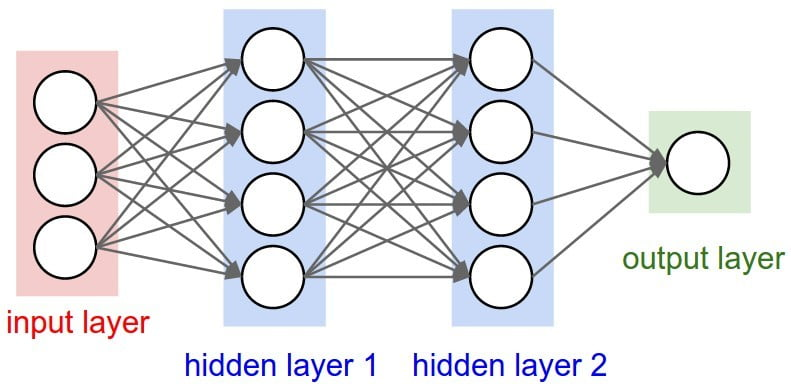


Figure 8

**Software and Hardware Requirements**

Software Requirements

* Operating System: Windows7,8,10 Processor- dual core 2.4 GHz (i5 or i7 series Intel processor or equivalent AMD) RAM-4GB

Hardware Requirements

* Hard disk: 40GB
* RAM: 128GB
* Python Pycharm PIP 2.7 Jupyter Notebook Chrome

**Data flow Diagram**

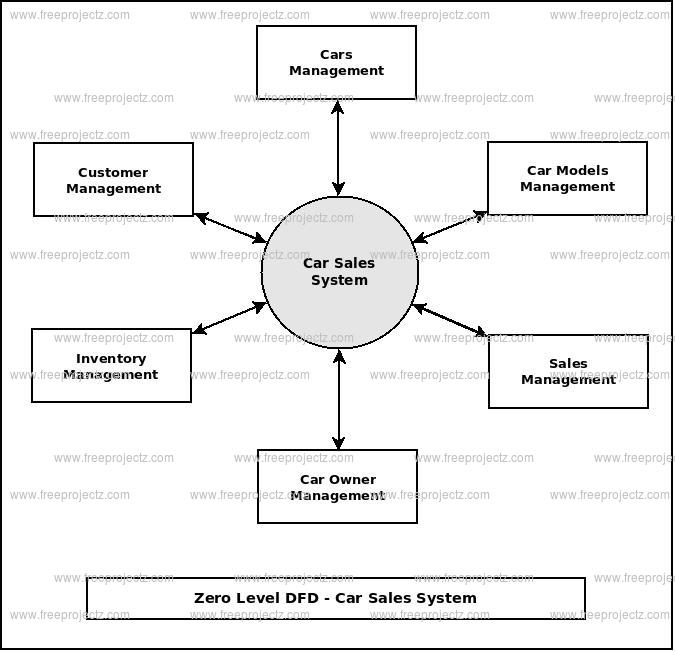
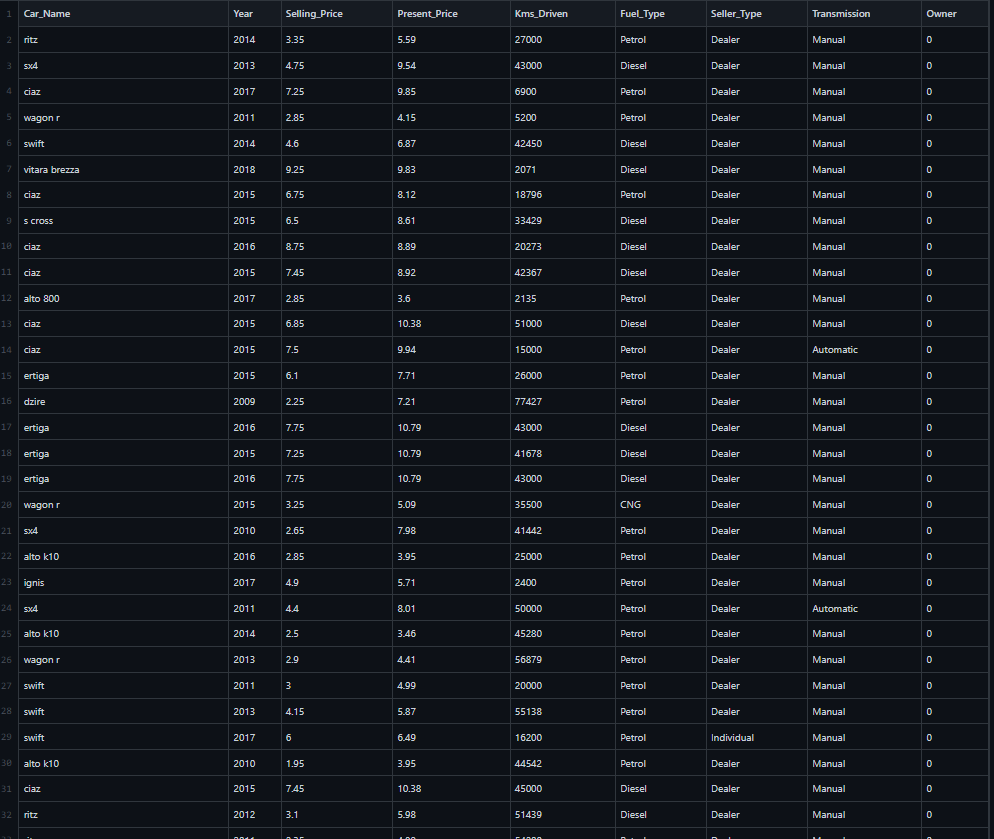


Figure 9

**Information about Dataset**

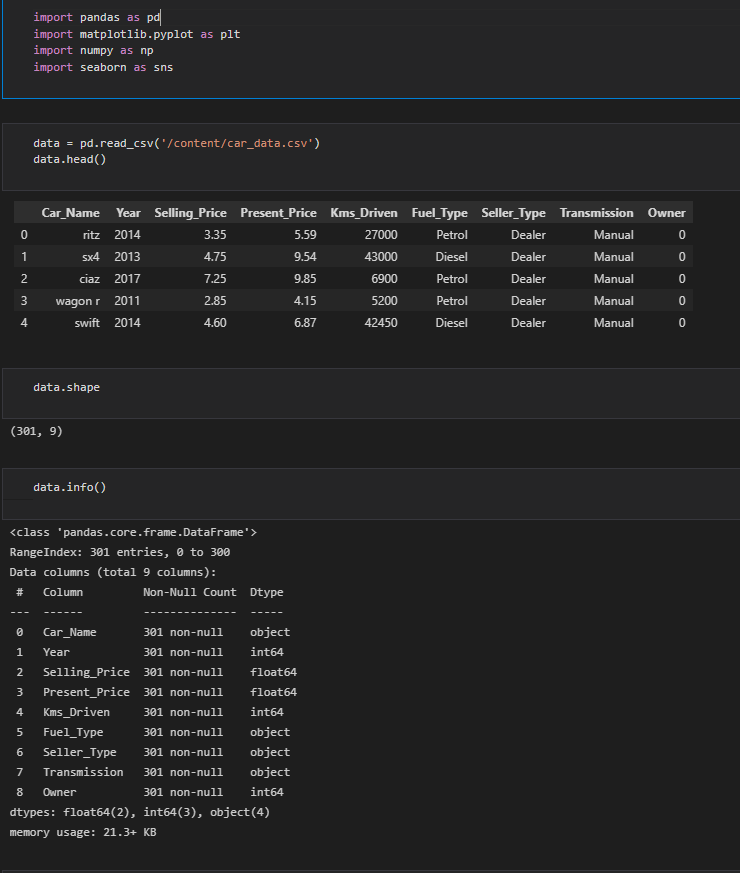
A data set is a collection of data. ... In Machine Learning projects, we need a training data set. It is the **actual data set used to train the model for performing various actions**.

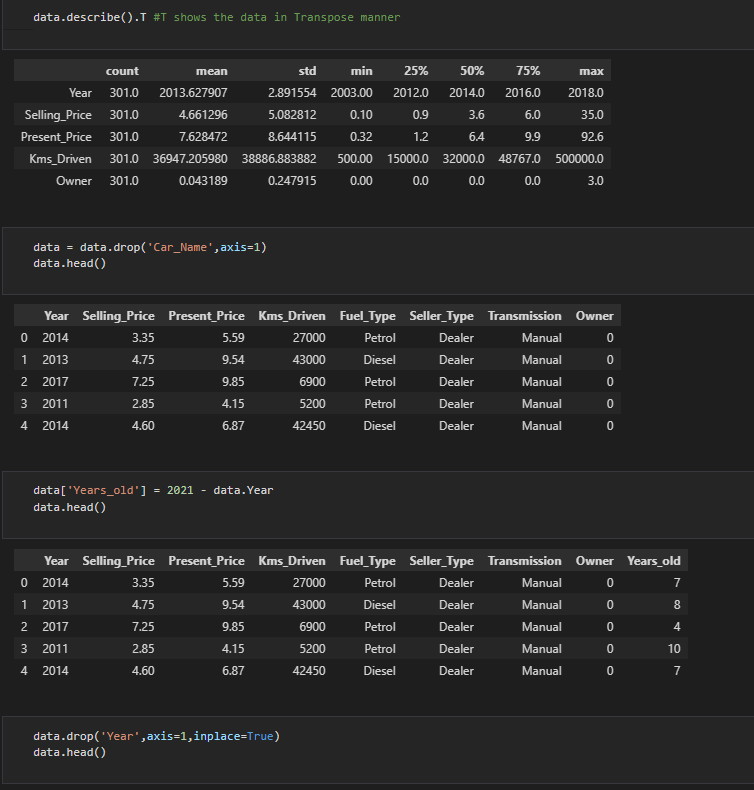
In order to make the most out of the data you have, it is **necessary to coordinate your whole data lifecycle to be usable**. For example, petabytes of data will do you no good if there is not enough computing power or network throughput to effectively work with it.

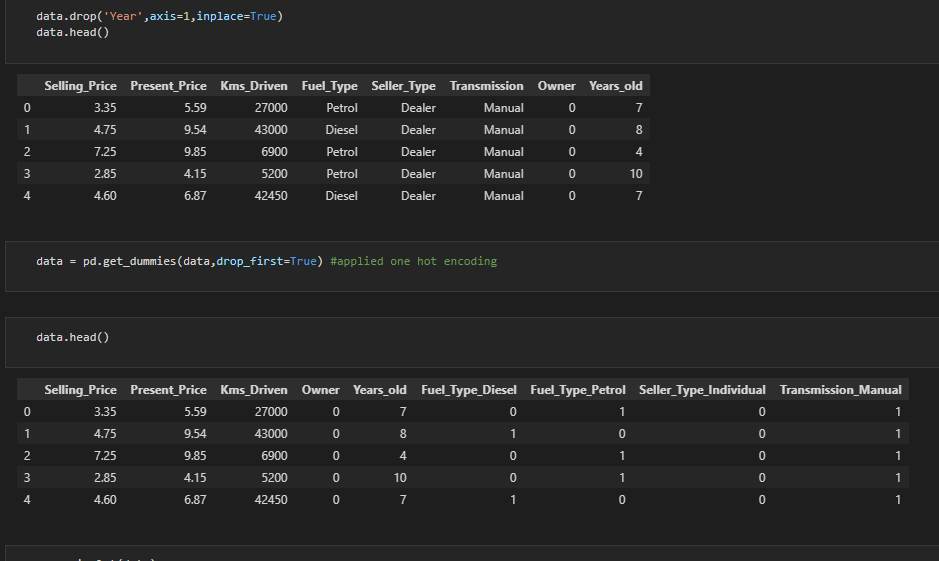


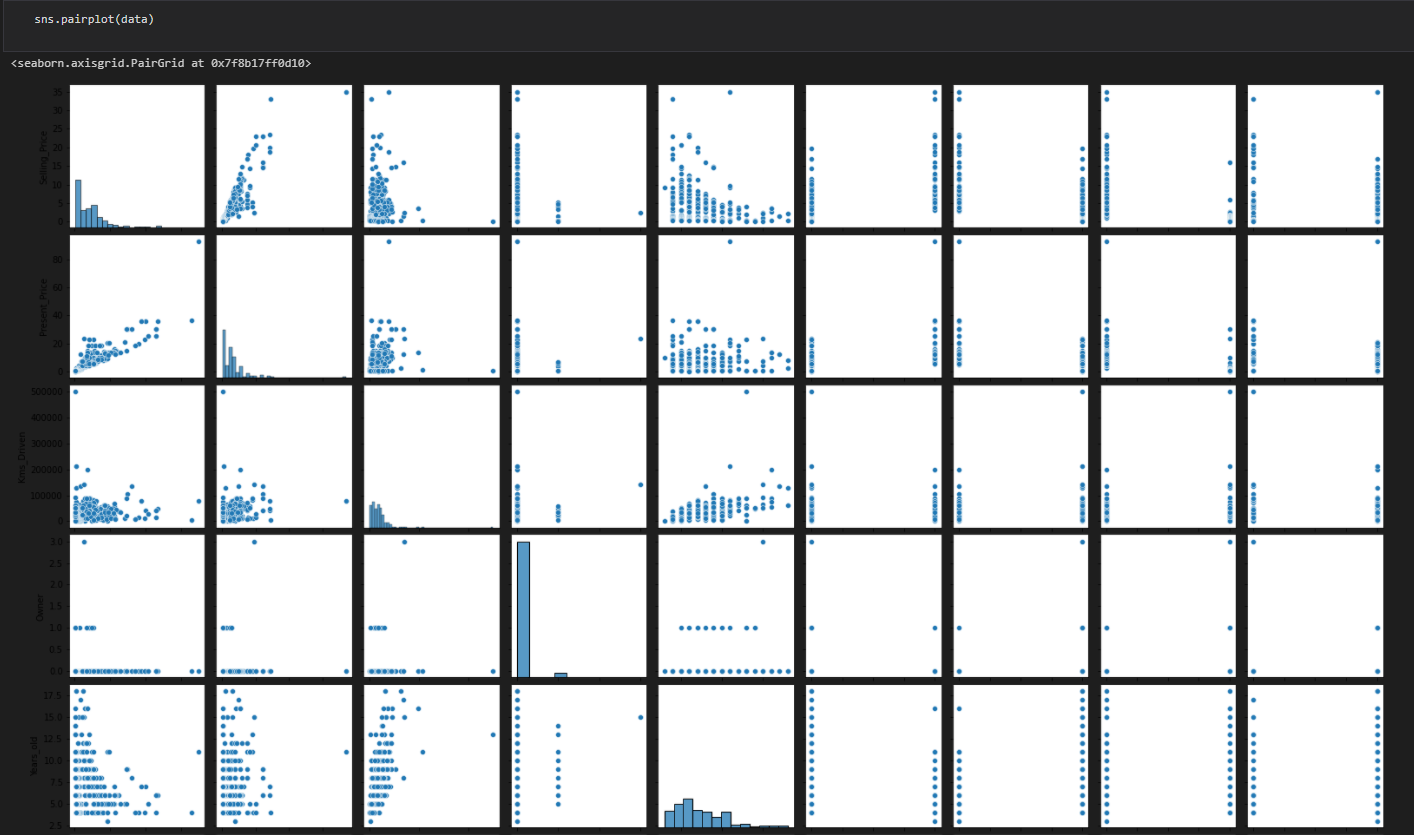
And soon…

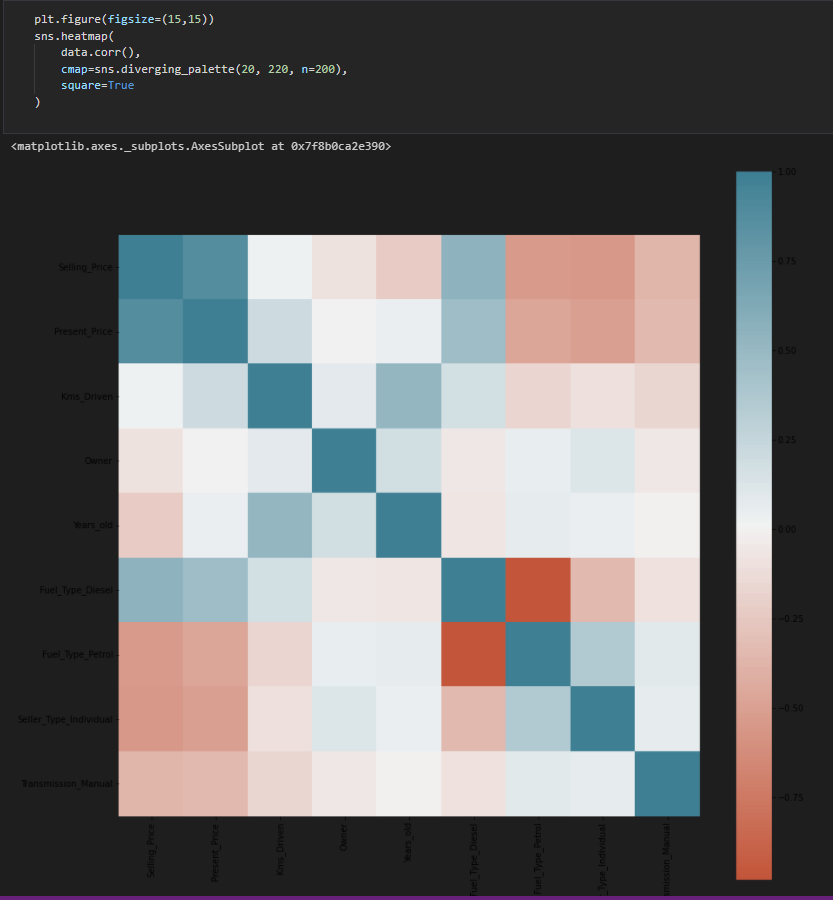
**Source code**

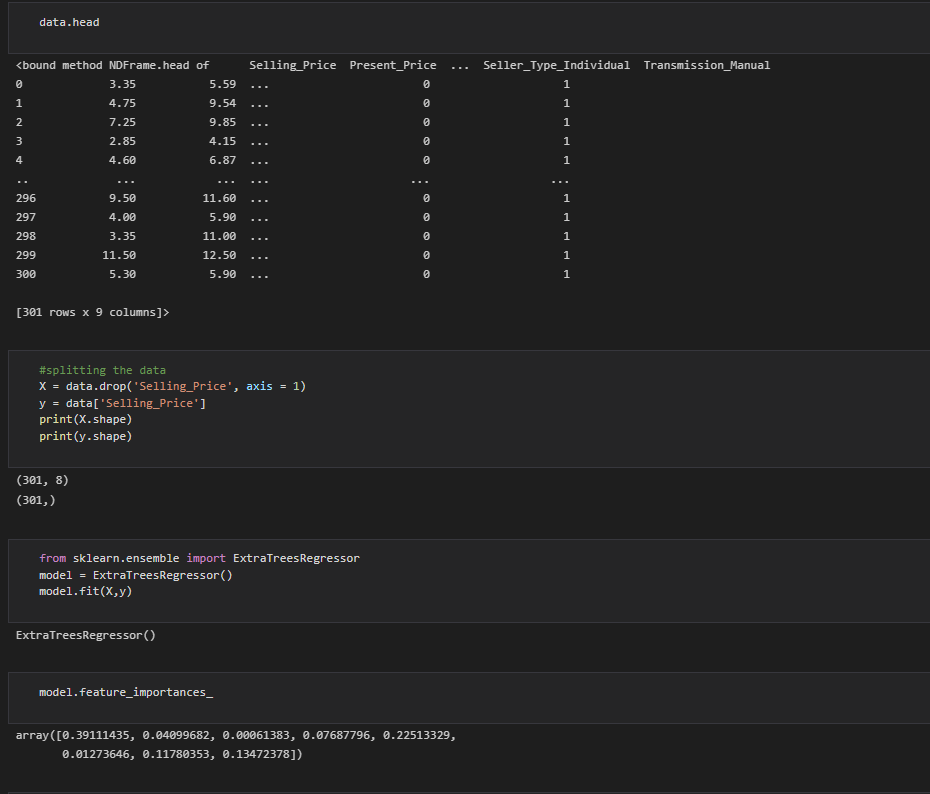


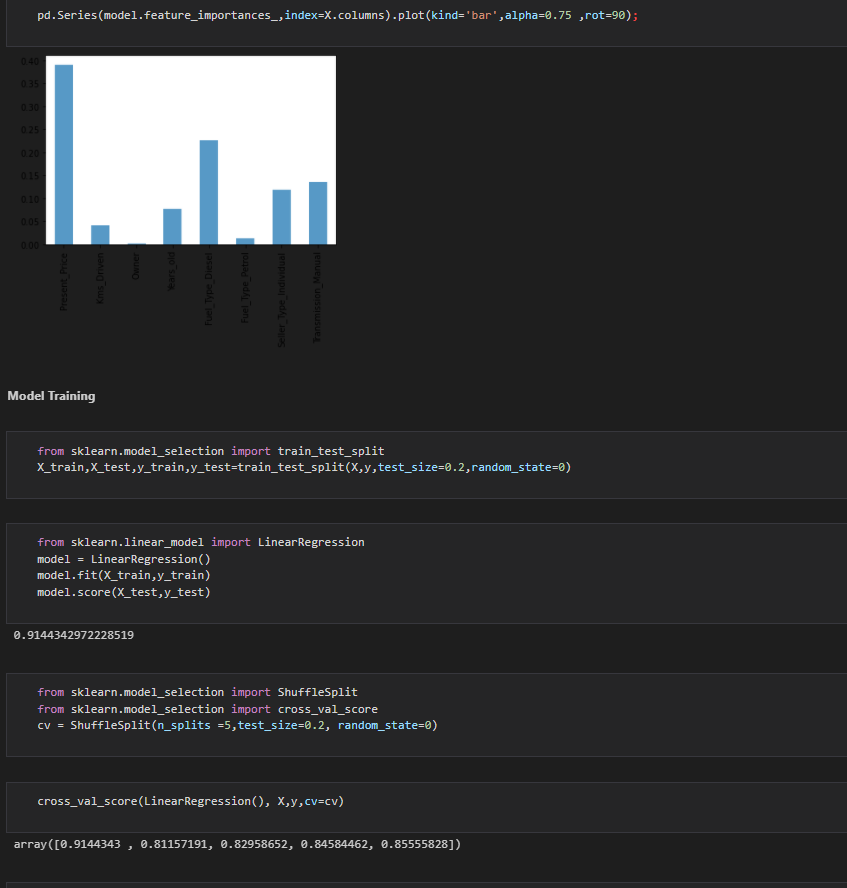


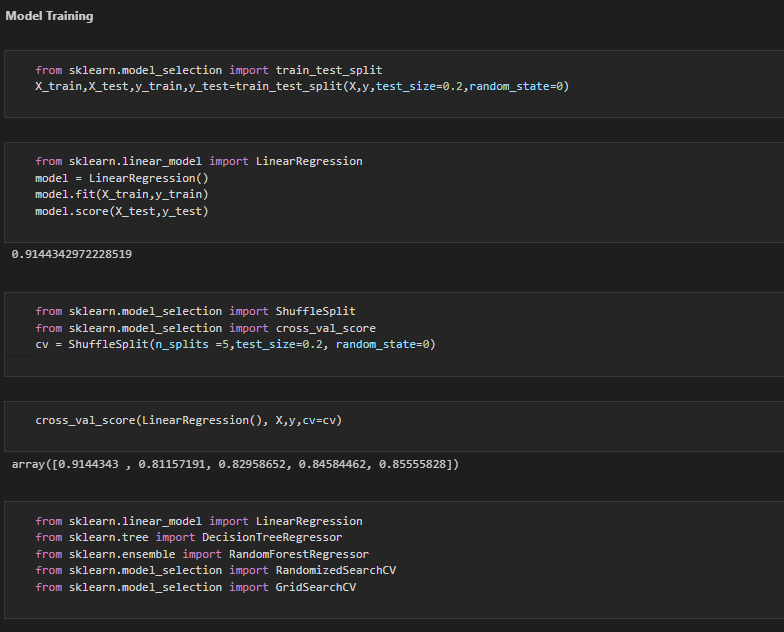


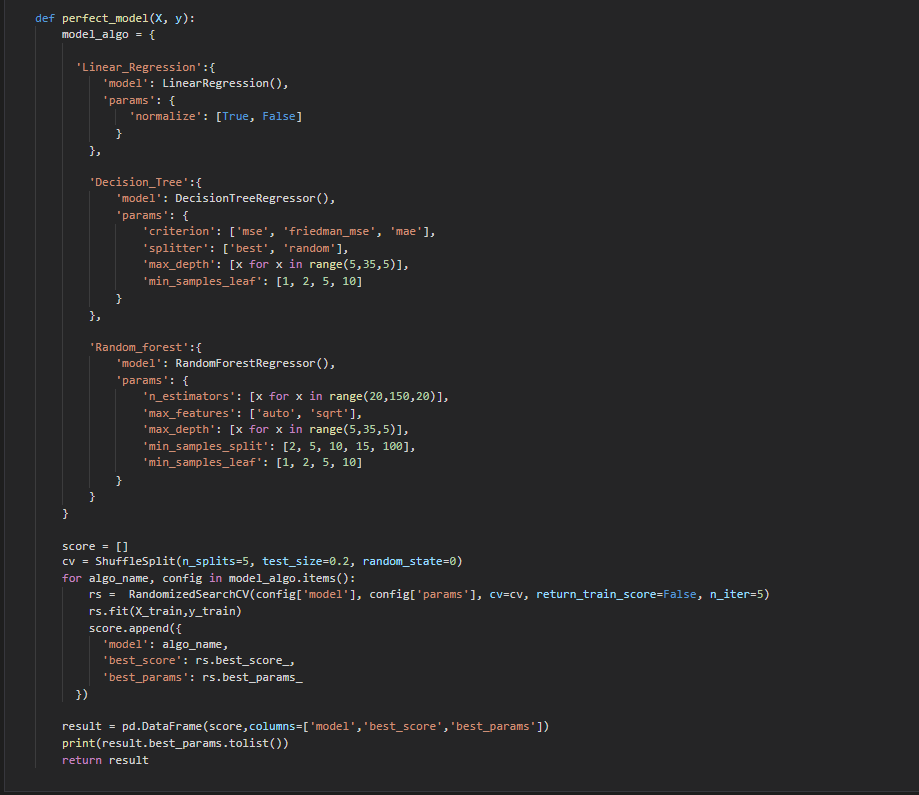


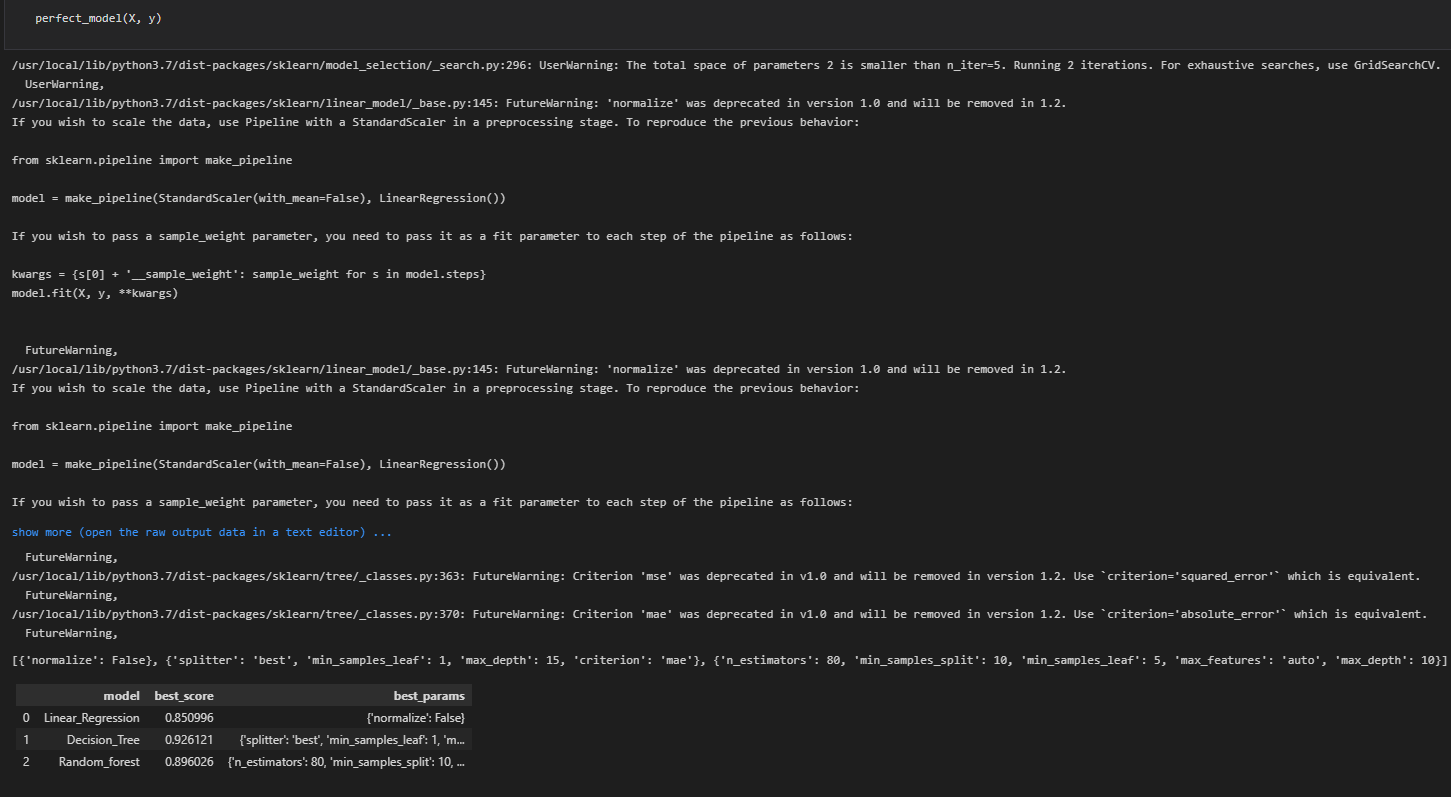


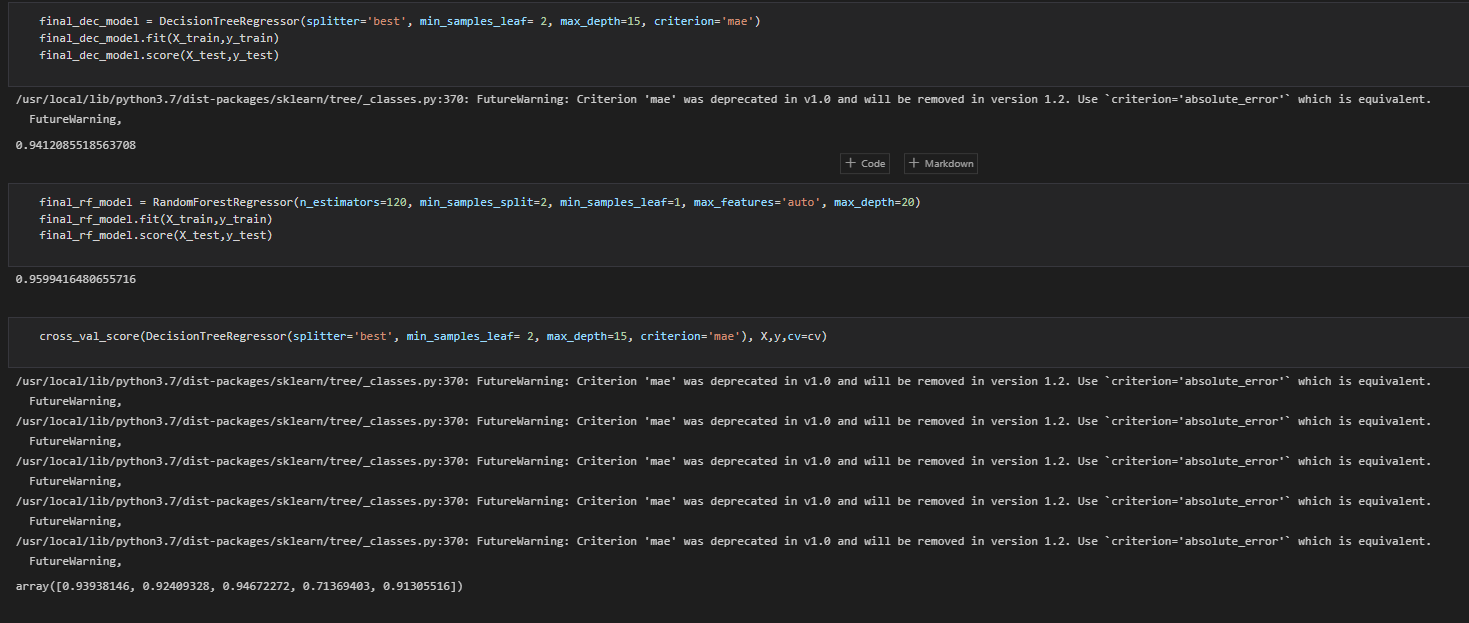


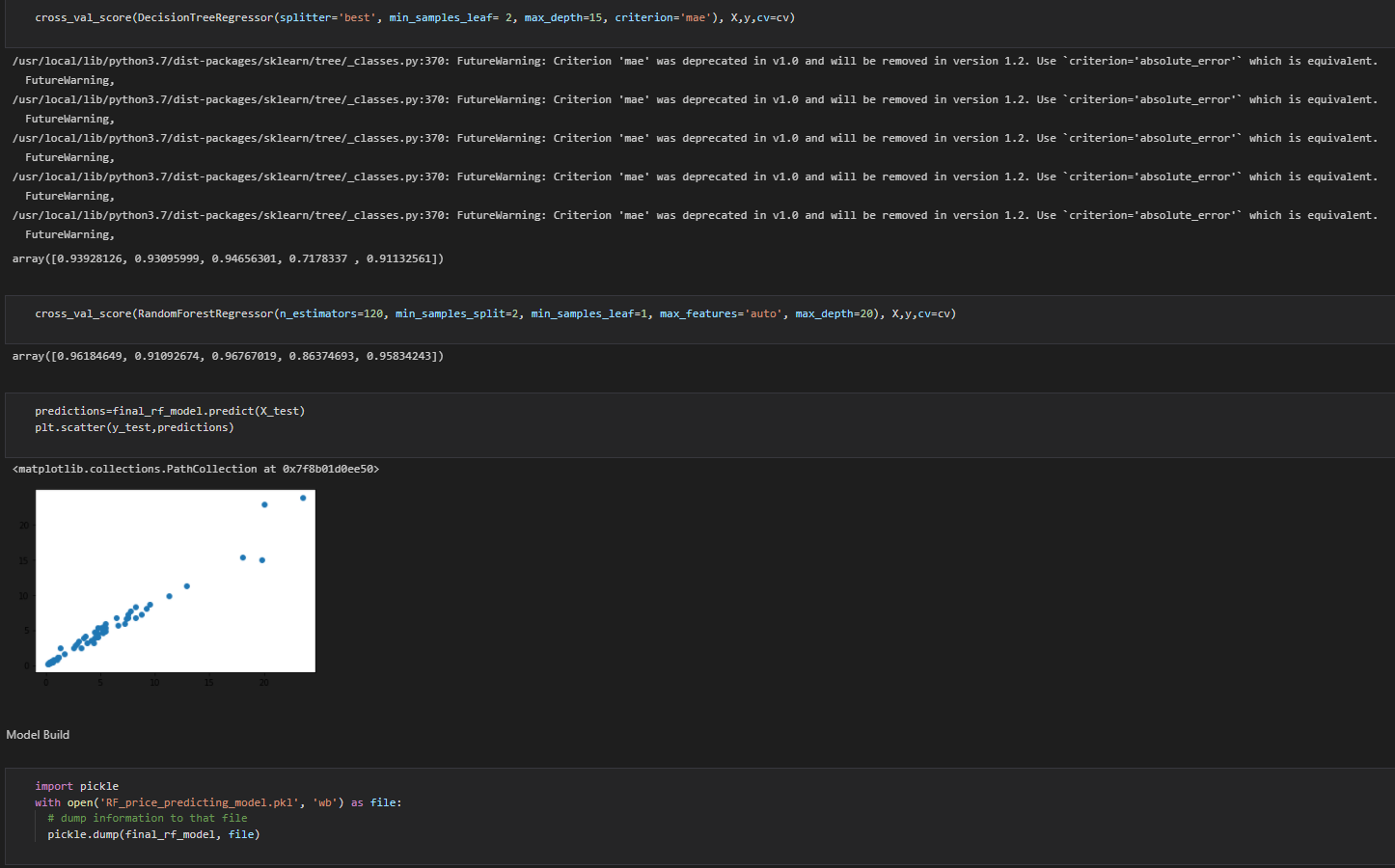




A



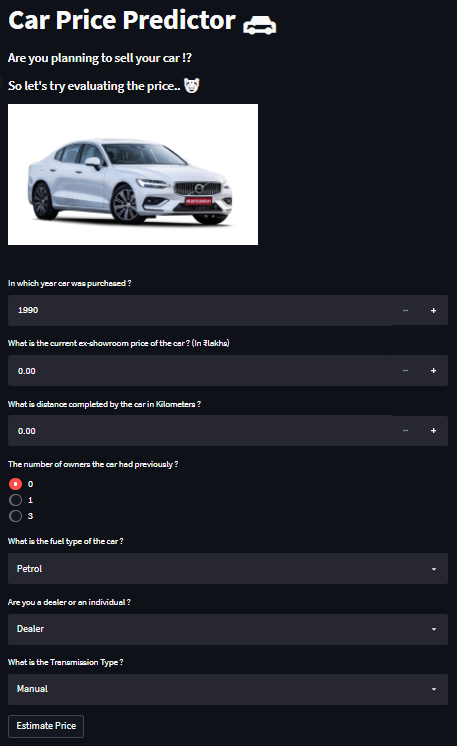


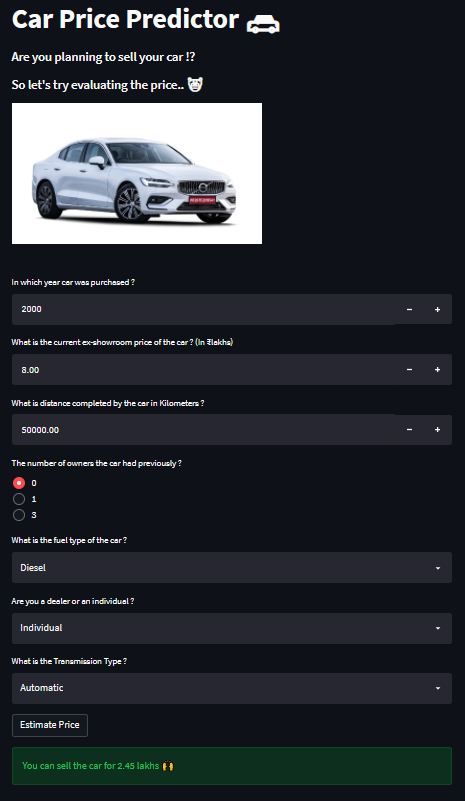


**Output Screenshots**

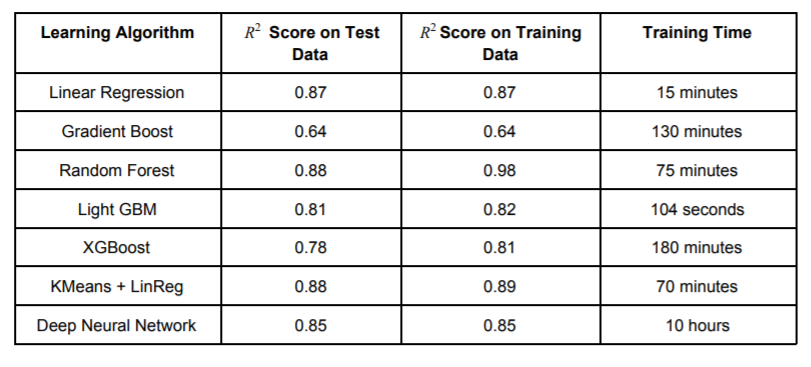
GitHub 🡪 https://github.com/PavanChikkodikar/Car-Selling-Price-Predictor

Project Demo 🡪 <https://carsellingprice-pavan.herokuapp.com/>





**Result**



**Future Work**

For better performance, we plan to judiciously design deep learning network structures, use adaptive learning rates and train on clusters of data rather than the whole dataset. To correct for overfitting in Random Forest, different selections of features and number of trees will be tested to check for change in performance.

And

In future this machine learning model may bind with various website which can provide real time data for price prediction. Also we may add large historical data of car price which can help to improve accuracy of the machine learning model. We can build an android app as user interface for interacting with user. For better performance, we plan to judiciously design deep learning network structures, use adaptive learning rates and train on clusters of data rather than the whole dataset.

**Conclusion**

The increased prices of new cars and the financial incapability of the customers to buy them, Used Car sales are on a global increase. Therefore, there is an urgent need for a Used Car Price Prediction system which effectively determines the worthiness of the car using a variety of features. The proposed system will help to determine the accurate price of used car price prediction. This paper compares 3 different algorithms for machine learning : Linear Regression, Lasso Regression and Ridge Regression.

**References**

1. Doan Van Thai, “Prediction car prices using quantify qualitative data and knowledge-based system.”

2. Pattabiraman Venkatasubbu, “Used Cars Price Prediction using Supervised Learning Techniques.”

3. Nitis Monburinon, “Prediction of Prices for Used Car by Using Regression Models”

4. https://towardsdatascience.com/used-car-priceprediction-using-machine-learninge3be02d977b2