**VEHICLE RESALE VALUE PREDICTION**

**Using Random Forest Regression**

Developed by: Ashik Hatter, Anusha A R, Arpitha MC,

Navami B G, Rashika N

**Smart Bridge-Remote Summer Internship Program**

**1. INTRODUCTION:**

A **used vehicle**, a **pre-owned vehicle**, or a **secondhand vehicle**, is a vehicle that has previously had one or more retail owners. Used vehicles are sold through a variety of outlets, including franchise and independent vehicle dealers, rental vehicle companies, buy here pay here dealerships, leasing offices, auctions, and private party sales. Some vehicle retailers offer "no-haggle prices," "certified" used vehicles, and extended service plans or warranties.

**History**: Established in 1898, the Empire State Motor Wagon Company in Catskill, New York was one of the very first American used vehicle lots.

The used vehicle market is substantially larger than other large retail sectors, such as the school and office products market.

**Used vehicle pricing:**

Used vehicle pricing reports typically produce three forms of pricing information.

* Dealer or retail price is the price expected to pay if buying from a licensed new-vehicle or used-vehicle dealer.
* Dealer trade-in price or wholesale price is the price a shopper should expect to receive from a dealer if trading in a vehicle. This is also the price that a dealer will typically pay for a vehicle at a dealer wholesale auction.
* Private-party price is the price expected to pay if buying from an individual. A private-party seller is hoping to get more money than they would with a trade-in to a dealer. A private-party buyer is hoping to pay less than the dealer retail price.

Pricing of used vehicles can be affected by geography. For example, [convertibles](https://en.wikipedia.org/wiki/Convertible) have a higher demand in warmer climates than in cooler areas. Similarly, [pickup trucks](https://en.wikipedia.org/wiki/Pickup_truck) may be more in demand in [rural](https://en.wikipedia.org/wiki/Rural_area) than [urban](https://en.wikipedia.org/wiki/Urban_area) settings. The overall condition of the vehicle has a major impact on pricing. Condition is based on appearances, vehicle history, mechanical condition, and mileage. There is much subjectivity in how the condition of a vehicle is evaluated.

There are various theories as to how the market determines the prices of used vehicles sold by private parties, especially relative to new vehicles. One theory suggests that new vehicle dealers are able to put more effort into selling a vehicle, and can therefore stimulate stronger demand. Another theory suggests that owners of problematic vehicles ("lemons") are more likely to want to sell their vehicles than owners of perfectly functioning vehicles. Therefore, someone buying a used vehicle bears a higher risk of buying a lemon, and the market price tends to adjust downwards to reflect that.

**1.1 Overview:**

Determining whether the listed price of a used vehicle 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 vehicle 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, but is compute heavy. Conventional linear regression also yielded satisfactory results, with the advantage of a significantly lower training time in comparison to the aforementioned methods.

**1.2 Purpose:**

Deciding whether a used vehicle 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 vehicle. From the perspective of a seller, it is also a dilemma to price a used vehicle appropriately. Based on existing data, the aim is to use machine learning algorithms to develop models for predicting used vehicle prices.

**1.3 Data Description**

Dataset For this project, we are using the dataset on used vehicle sales from all over the United States, available on Kaggle. Dataset contains information about used vehicles name, brand, vehicle type, abtest, powerPS, model, gear type, kilometers driven, year of registration, month of registration,

**2. Existing problem:**

The prices of new vehicles in the industry is fixed by the manufacturer with some additional costs incurred by the Government in the form of taxes. So, customers buying a new vehicle can be assured of the money they invest to be worthy. But due to the increased price of new vehicles and the incapability of customers to buy new vehicles due to the lack of funds, used vehicles sales are on a global increase (Pal, Arora and Palakurthy, 2018). There is a need for a used vehicle price prediction system to effectively determine the worthiness of the vehicle using a variety of features. Even though there are websites that offers this service, their prediction method may not be the best. Besides, different models and systems may contribute on predicting power for a used vehicle’s actual market value. It is important to know their actual market value while both buying and selling.

3. T**he Client:**

To be able to predict used vehicles market value can help both buyers and sellers.

**Used vehicle sellers (dealers):** They are one of the biggest target group that can be interested in results of this study. If used vehicle sellers better understand what makes a vehicle desirable, what the important features are for a used vehicle, then they may consider this knowledge and offer a better service.

**Online pricing services:** There are websites that offers an estimate value of a vehicle. They may have a good prediction model. However, having a second model may help them to give a better prediction to their users. Therefore, the model developed in this study may help online web services that tells a used vehicle’s market value.

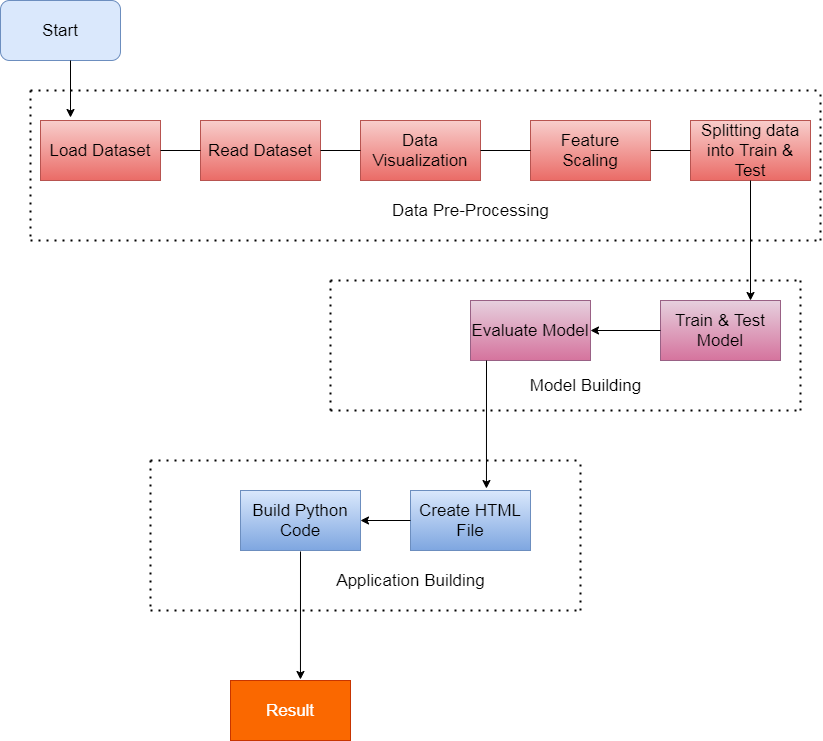
**Individuals:** There are lots of individuals who are interested in the used vehicle market at some points in their life because they wanted to sell their vehicle or buy a used vehicle. In this process, it’s a big corner to pay too much or sell less then it’s market value.

**4. Machine-Learning model with high accuracy:**

**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. This model gave an accuracy of **81%.**

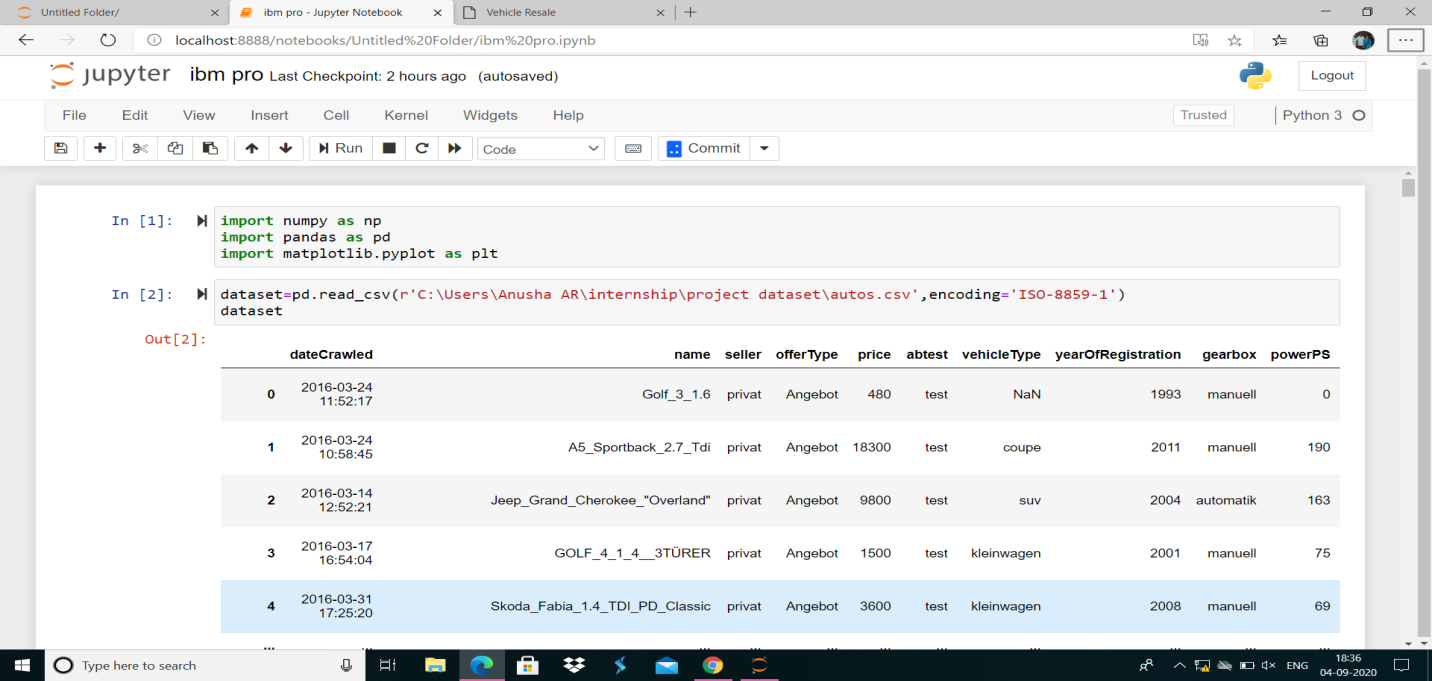
Random forest is a set of multiple decision trees. Deep decision trees may suffer from overfitting, but random forest prevents overfitting by creating trees on random subsets. That’s why, it’s a good model to in the analysis.

**5. Flow Chart:**

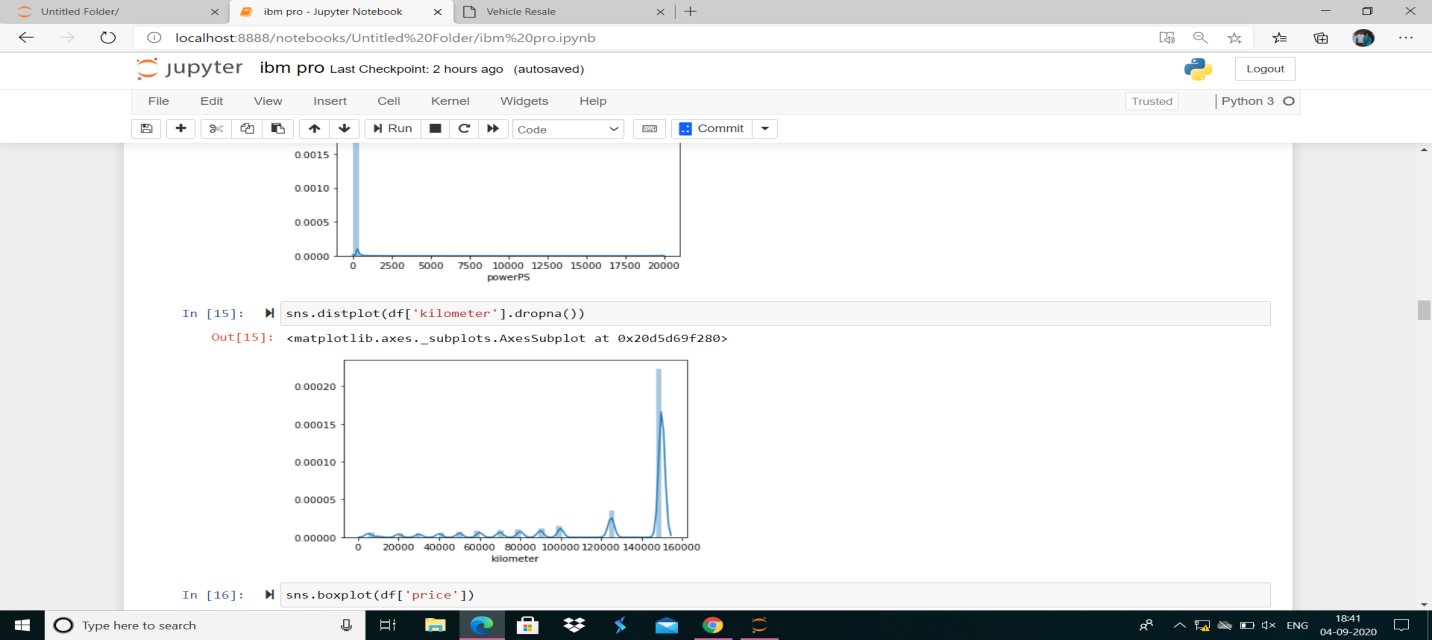
. Fig 1.1 :Flow Chart

**6. Screenshot:**

6.1Importing Libraries and Dataset:

****Fig 1.2 : Screenshot of Libraries and Dataset

6.2Data Visualisation by Seaborn:

****

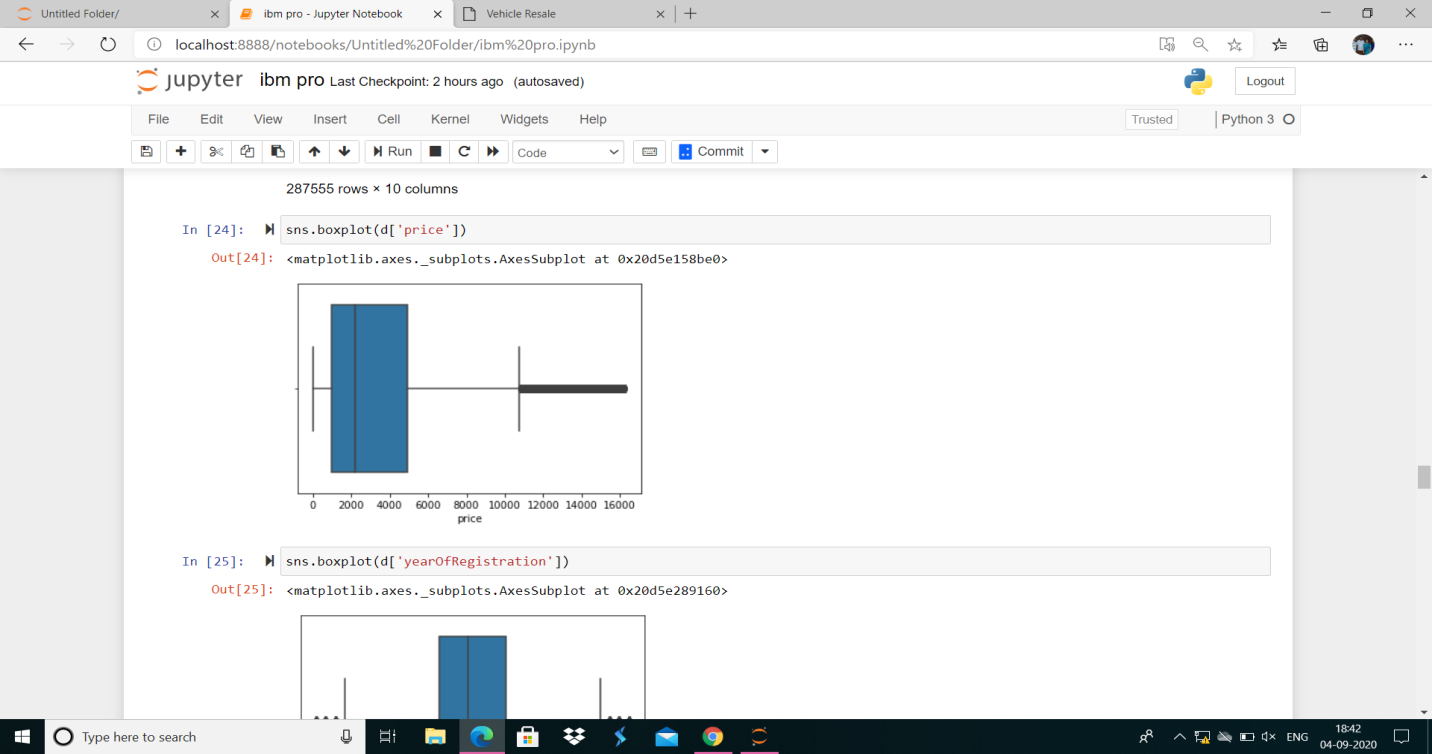


Fig 1.3 : Screenshot of Data Visualisation

6.3Taking Care Of Missing Data:

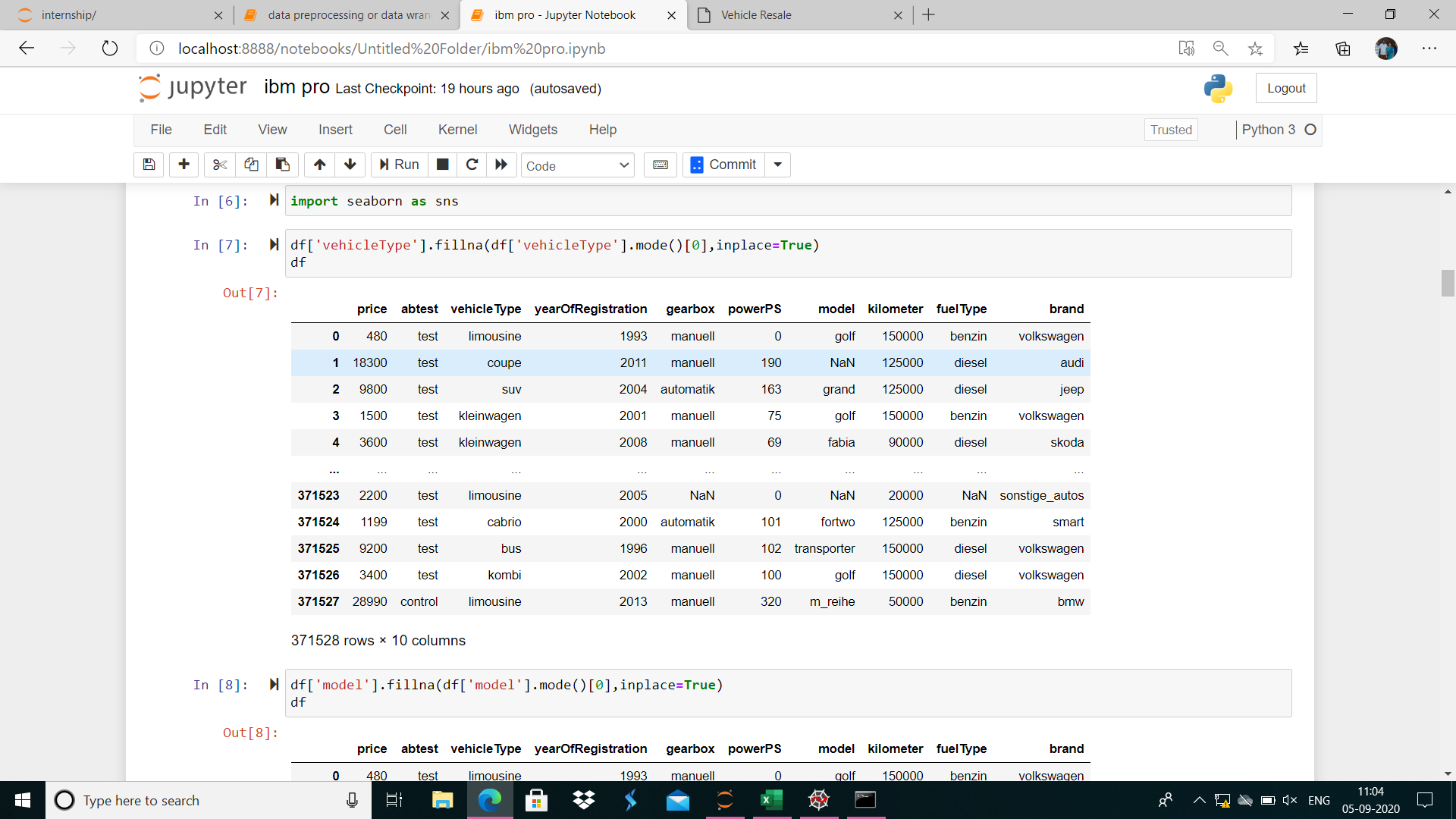
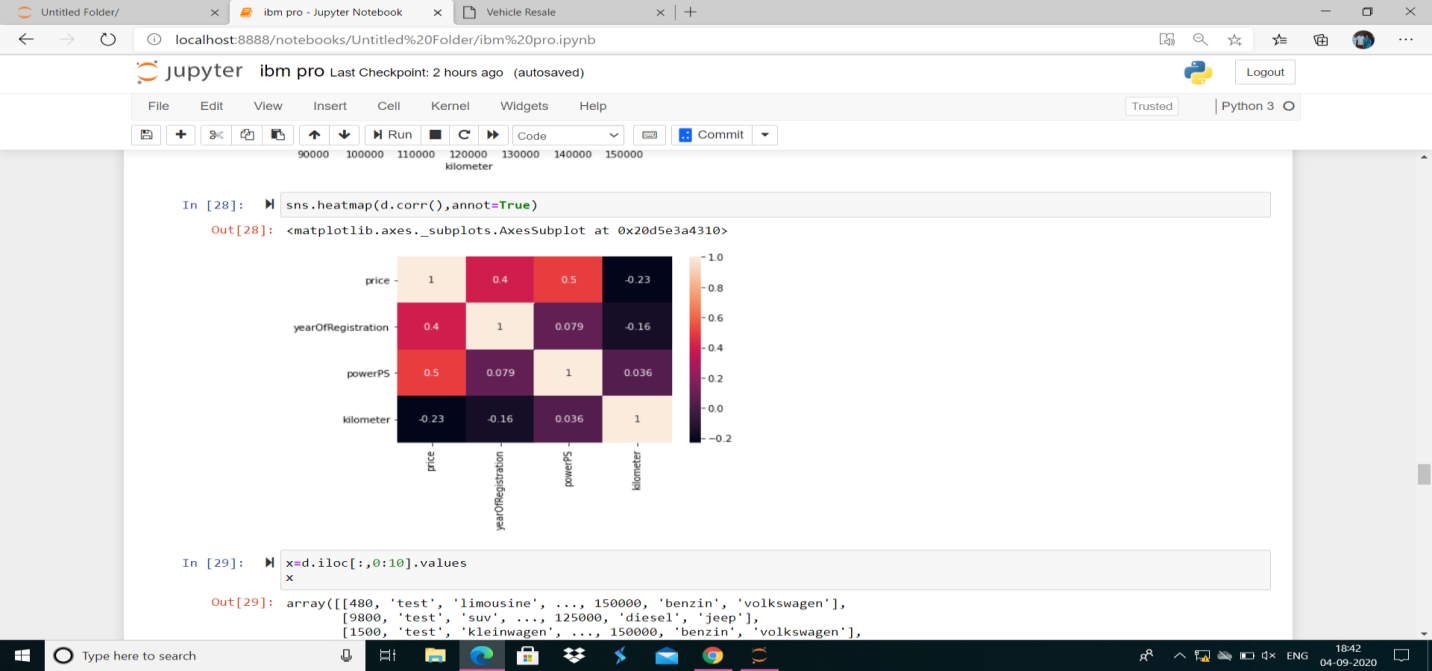
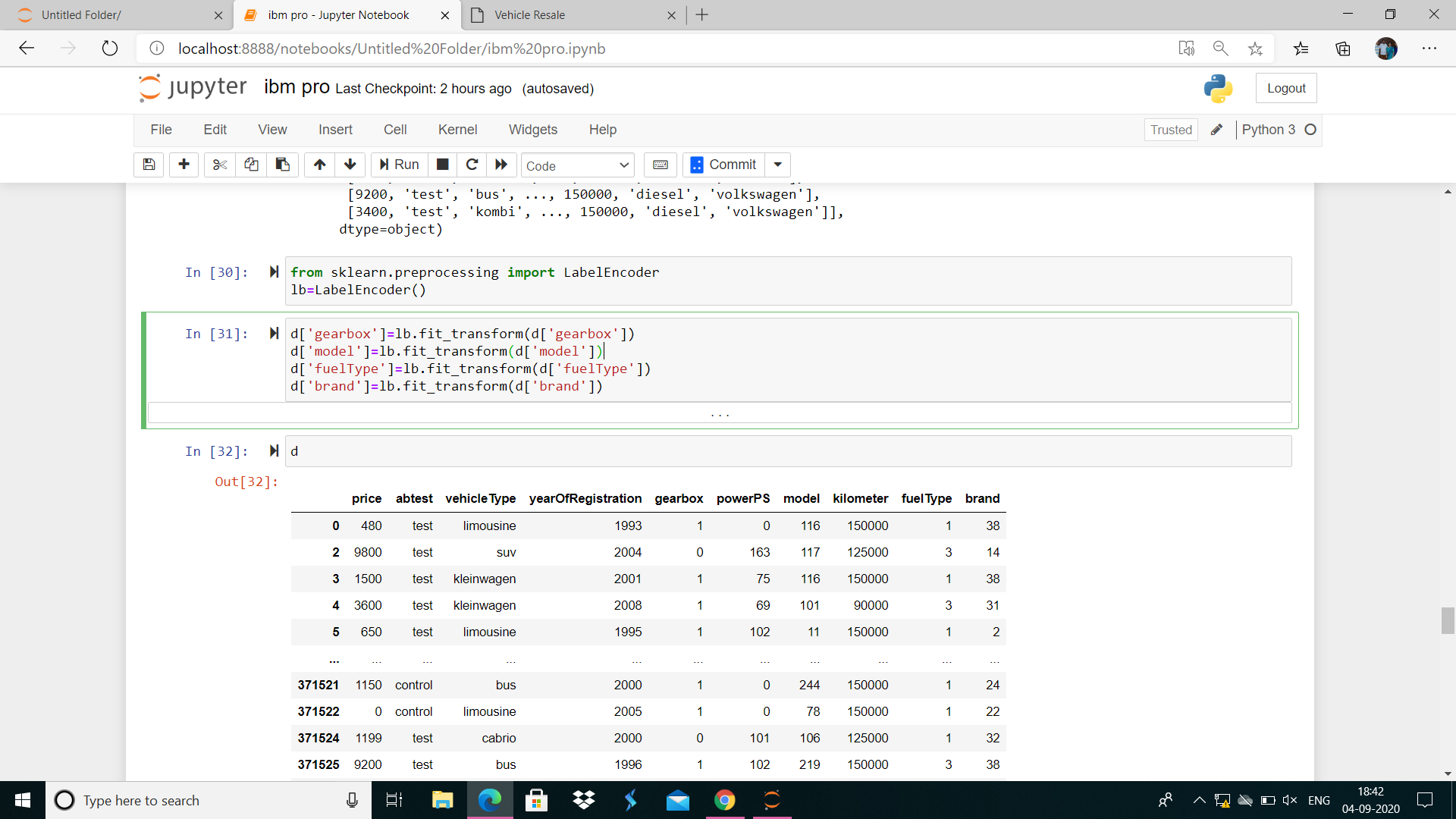


Fig 1.4 : Screenshot of handling of missing values

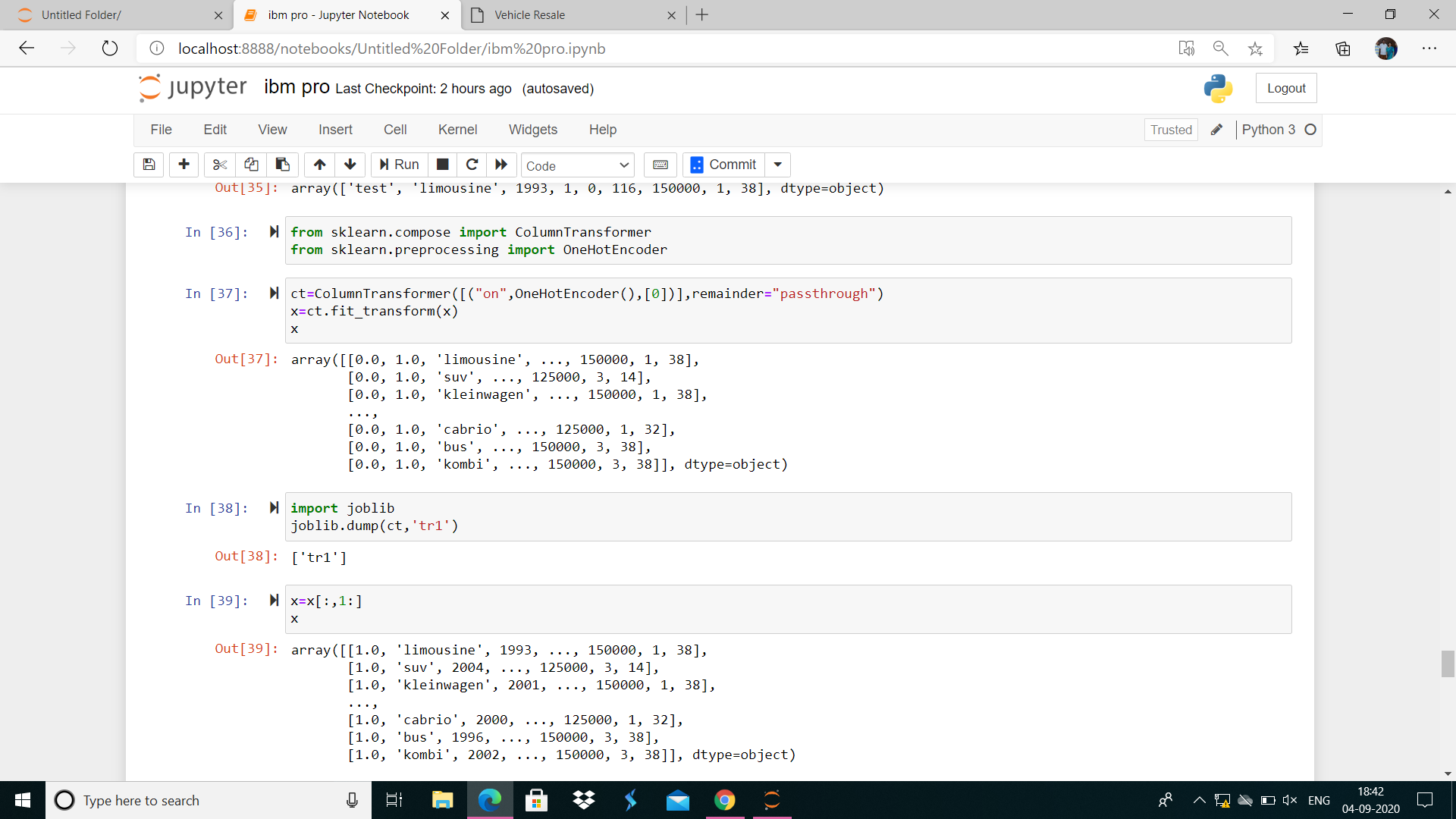
6.4 Heatmap:

Fig 1.5 : Screenshot of heatmap

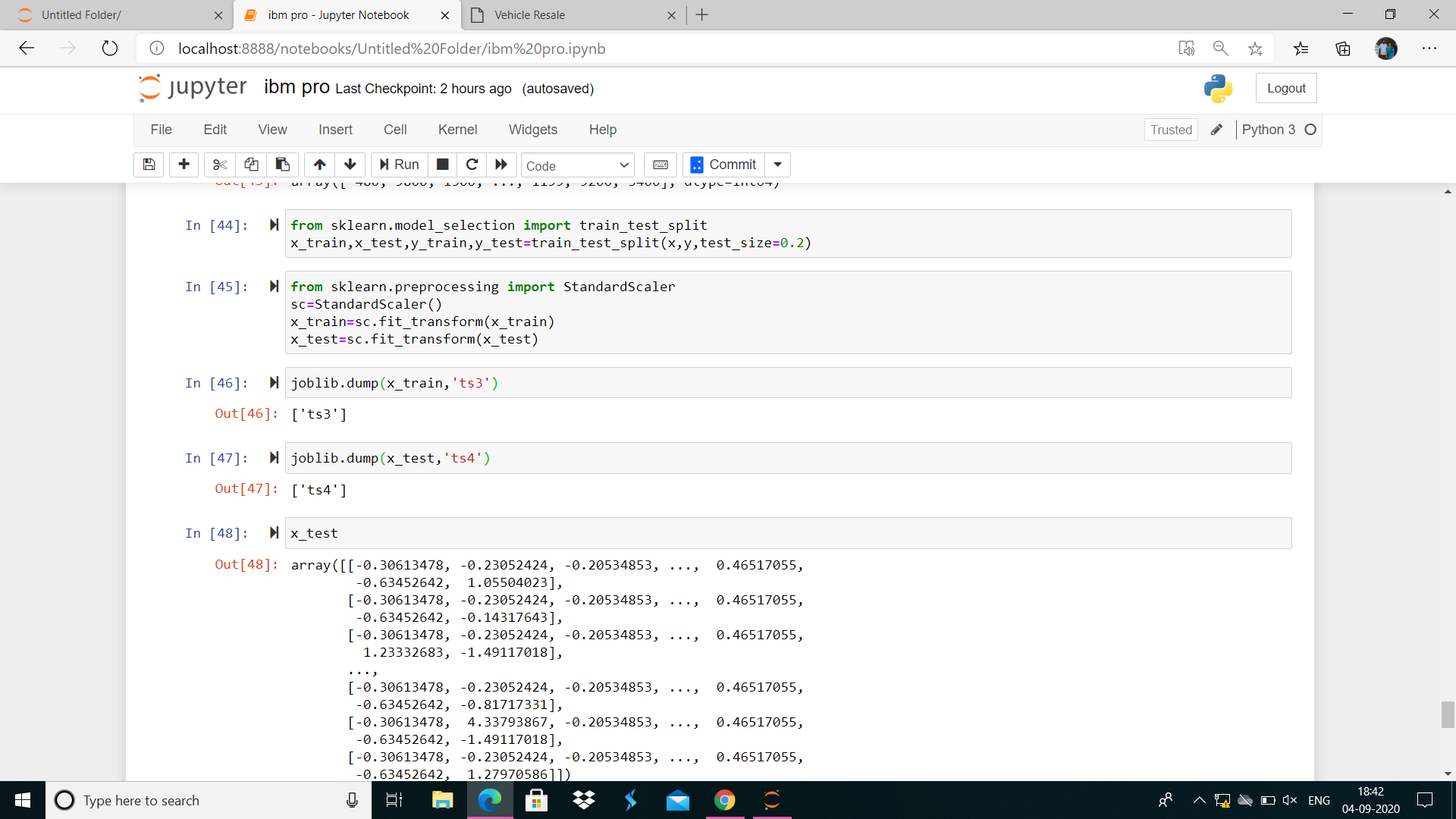
6.5 Label Encoding**:**

****Fig 1.6 : Screenshot of label encoding

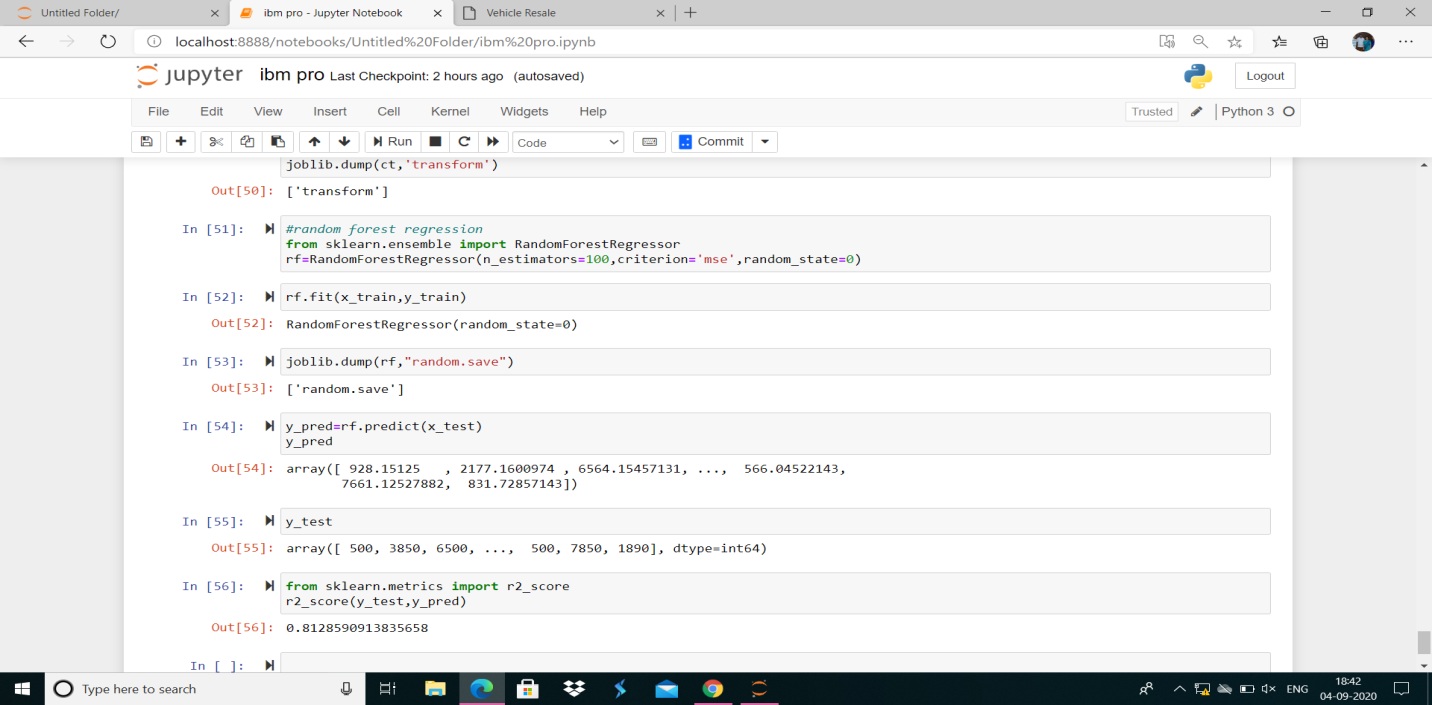
6.6 One Hot Encoding:

Fig 1.7 : Screenshot of one hot encoding

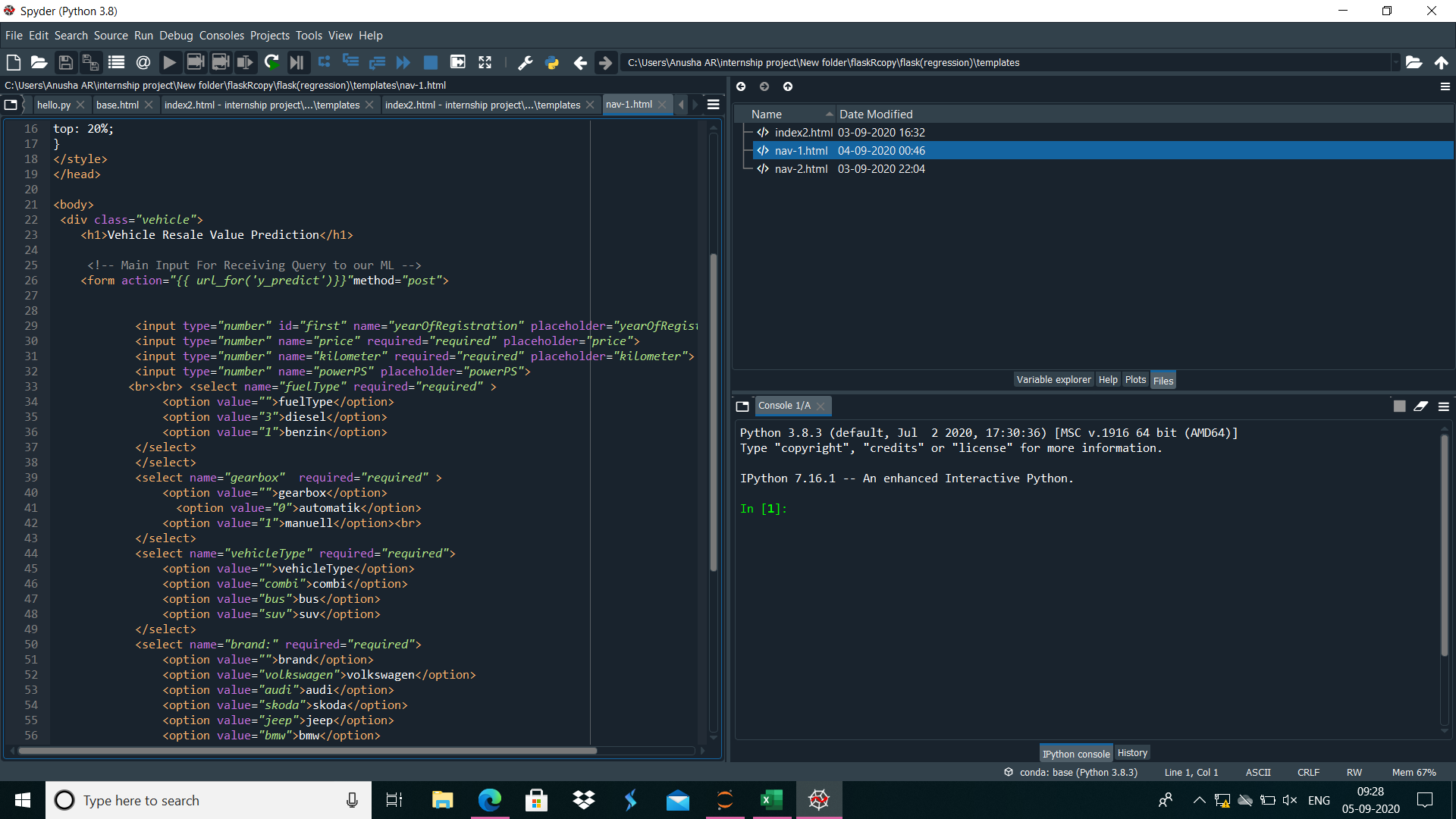
6.7 Feature Scaling and Splitting Data Into Test and Train:

****Fig 1.8 : Screenshot of splitting data

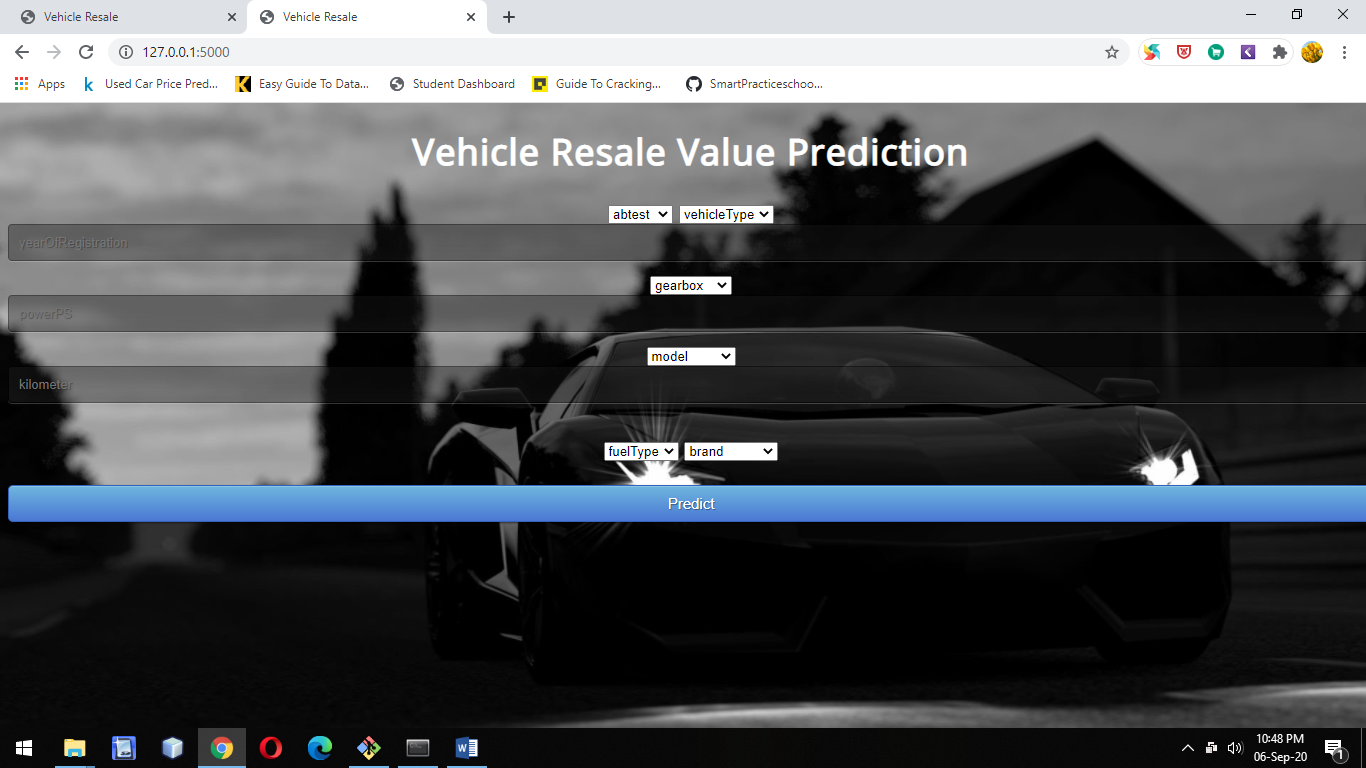
6.8 Model Building:

****Fig 1.9 : Screenshot of model building

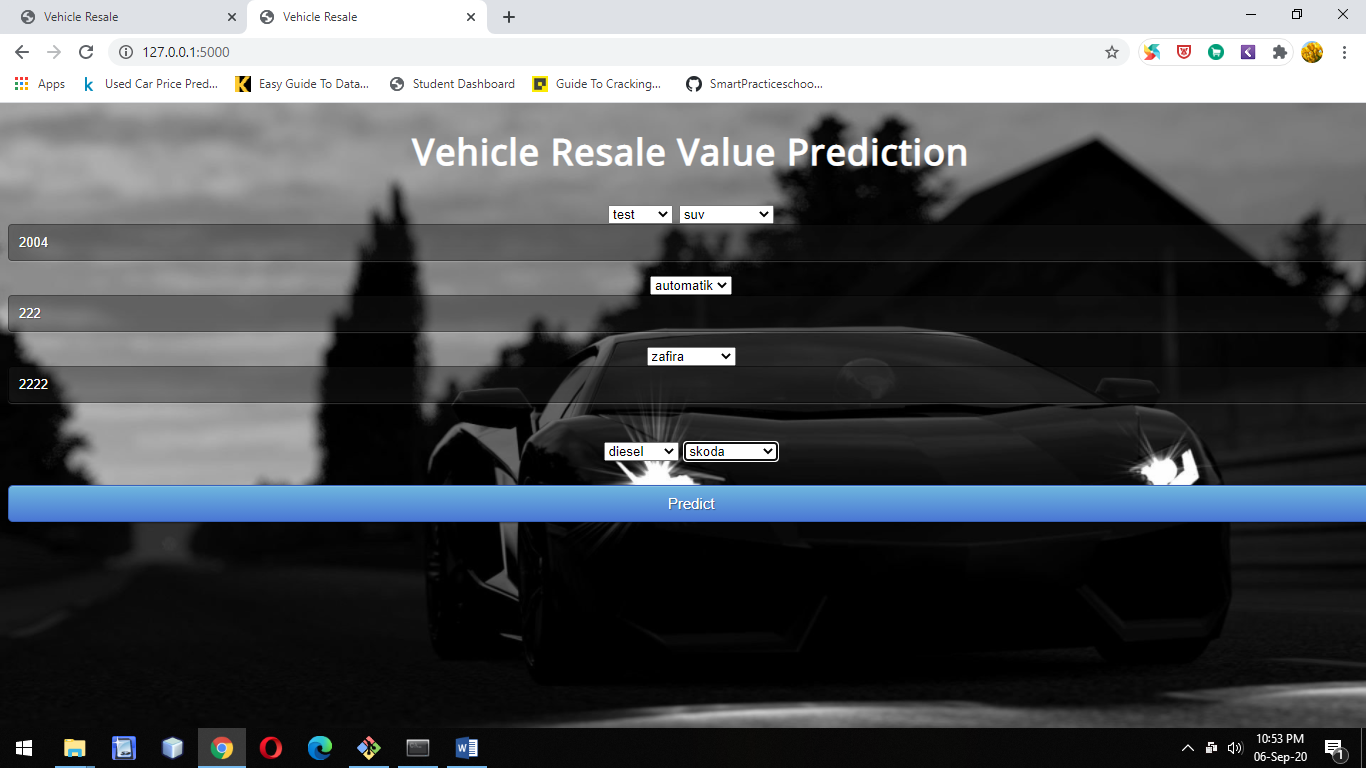
6.9 HTML File in Spyder:

Fig 1.10 : Screenshot of html file

6.10 Web Page Before Entering The Inputs:

Fig 1.11 : Screenshot of output

6.11 Web Page After Entering The Input:



6.12 Output:

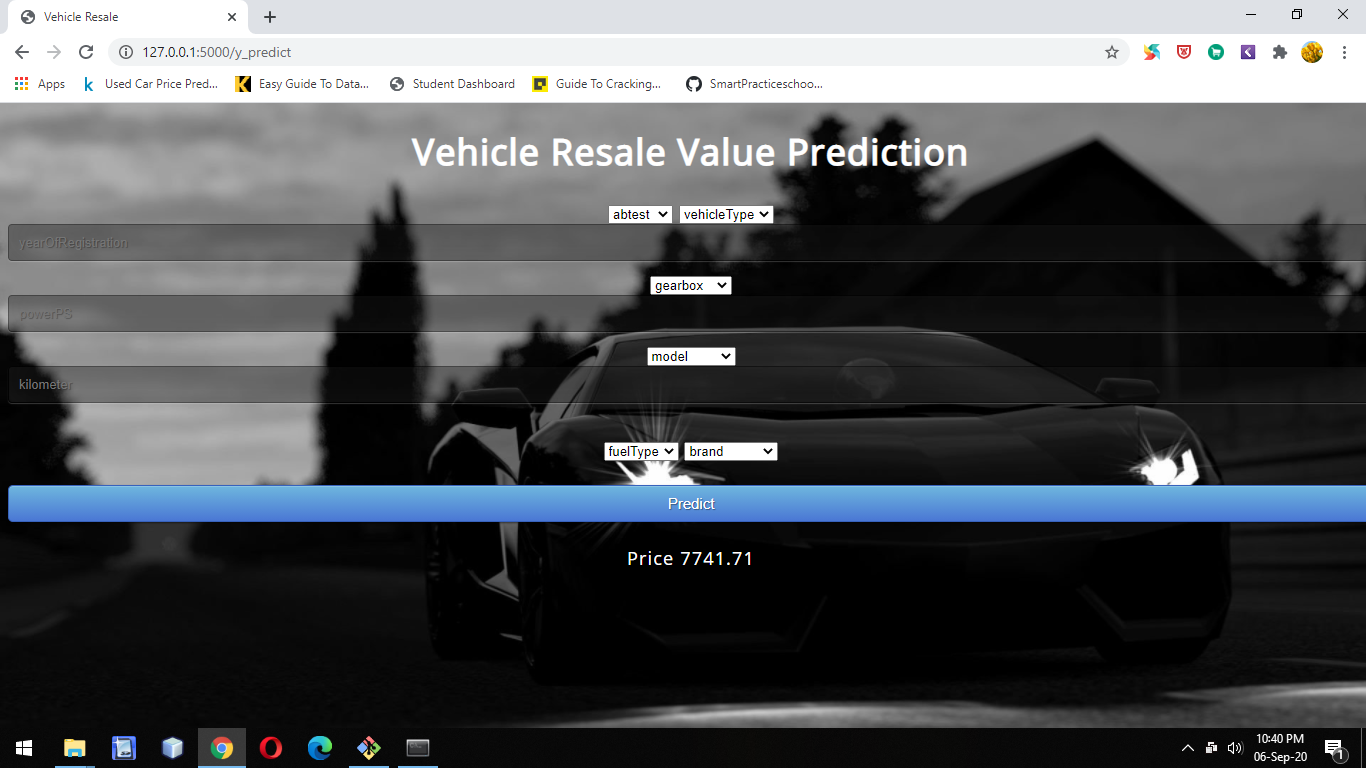


Fig 1.13 : Screenshot of output

**7. Software Designing:**

● Jupyter Notebook Environment

● Spyder Ide

● Machine Learning Algorithms

● Python (pandas, numpy, matplotlib, seaborn, sklearn)

● HTML

● Flask

We developed this resale value prediction by using the Python language which is a interpreted and high level programming language and usng the Machine Learning algorithms. for coding we used the Jupyter Notebook environment of the Anaconda distributions and the Spyder, it is an integrated scientific programming in the python language. For creating an user interface for the prediction we used the Flask. It is a micro web framework written in Python. It is classified as a microframework because it does not require particular tools or libraries. It has no database abstraction layer, form validation, or any other components where pre-existing third-party libraries provide common functions, and a scripting language to create a webpage is HTML by creating the templates to use in the functions of the Flask and HTML.

**8. Results and conclusion:**

The results of our tests were quantified in terms of the R score of our predictions. score is a statistical 2 R2 measure of how close the data are to the fitted regression line. Random Forest method which marginally outperforms Linear Regression. However Random Forests tend to overfit the dataset due to the tendency of growing longer trees. This was worked upon by restricting the depth of trees to different values.

**9. Limitations:**

This study used different models in order to predict used vehicle prices. However, there was a relatively dataset for making a strong inference because number of observations was only 6 lakh. Gathering more data can yield more robust predictions. Secondly, there could be more features that can be good predictors. For example, here are some variables that might improve the model: number of doors, number of seats, gas/mile (per gallon), color, mechanical and cosmetic reconditioning time, used-to-new ratio, appraisal-to-trade ratio.

**11. Bibliography:**

1. https://www.kaggle.com/orgesleka/used-cars-dataset

2. N. Monburinon, P. Chertchom, T. Kaewkiriya, S. Rungpheung, S. Buya and P. Boonpou, "Prediction of prices for used vehicle by using regression models," 2018 5th International Conference on Business and Industrial Research (ICBIR), Bangkok, 2018, pp. 115-119.

3. Listiani M. 2009. Support Vector Regression Analysis for Price Prediction in a Vehicle Leasing Application. Master Thesis. Hamburg University of Technology

4. Chen, Tianqi, and Vehiclelos Guestrin. "Xgboost: A scalable tree boosting system." Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining. ACM, 2016.

5. Ke, Guolin, et al. "Lightgbm: A highly efficient gradient boosting decision tree." Advances in Neural Information Processing Systems. 2017.

6. Fisher, Walter D. "On grouping for maximum homogeneity." Journal of the American statistical Association 53.284 (1958): 789-798. 7. https://scikit-learn.org/stable/modules/classes.html: Scikit-learn: Machine Learning in Python, Pedregosa et al., JMLR 12, pp. 2825-2830, 2011.

**Appendix:**

**HTML:**

<!DOCTYPE html>

<html >

<!--From https://codepen.io/frytyler/pen/EGdtg-->

<head>

<meta charset="UTF-8">

<title>Vehicle Resale</title>

<link href='https://fonts.googleapis.com/css?family=Pacifico' rel='stylesheet' type='text/css'>

<link href='https://fonts.googleapis.com/css?family=Arimo' rel='stylesheet' type='text/css'>

<link href='https://fonts.googleapis.com/css?family=Hind:300' rel='stylesheet' type='text/css'>

<link href='https://fonts.googleapis.com/css?family=Open+Sans+Condensed:300' rel='stylesheet' type='text/css'>

<link rel="stylesheet" href="{{ url\_for('static', filename='css/style.css') }}">

<style>

.login{

top: 20%;

}

</style>

</head>

<body>

<div class="vehicle">

<h1>Vehicle Resale Value Prediction</h1>

<!-- Main Input For Receiving Query to our ML -->

<form action="{{ url\_for('y\_predict')}}"method="post">

<select name="abtest:" required="required">

<option value="">abtest</option>

<option value="1">test</option>

<option value="0">control</option><br>

</select>

<select name="vehicleType" required="required">

<option value="">vehicleType</option>

<option value="limousine">limousine</option>

<option value="kombi">kombi</option>

<option value="bus">bus</option>

<option value="suv">suv</option>

</select>

<input type="number" id="first" name="yearOfRegistration" placeholder="yearOfRegistration" >

<select name="gearbox" required="required" >

<option value="">gearbox</option>

<option value="0">automatik</option>

<option value="1">manuell</option><br>

</select>

<input type="number" name="powerPS" placeholder="powerPS">

<select name="model:" required="required">

<option value="">model</option>

<option value="116">golf</option>

<option value="117">grand</option>

<option value="101">fabia</option>

<option value="11">3er</option>

<option value="244">zafira</option>

<option value="78">colt</option>

<option value="106">fortow</option>

<option value="219">transporter</option><br>

</select>

<input type="number" name="kilometer" required="required" placeholder="kilometer">

<br><br> <select name="fuelType" required="required" >

<option value="">fuelType</option>

<option value="3">diesel</option>

<option value="1">benzin</option>

</select>

<select name="brand:" required="required">

<option value="">brand</option>

<option value="38">volkswagen</option>

<option value="32">smart</option>

<option value="31">skoda</option>

<option value="14">jeep</option>

<option value="2">bmw</option>

<option value="22">mitsubishi</option><br>

</select>

<br><br><button type="submit" class="btn btn-primary btn-block btn-large">Predict</button>

</form><br>

{{ prediction\_text }}

</div>

</body>

</html>

**App.py**

import numpy as np  
from flask import Flask, request, jsonify, render\_template  
from joblib import load  
app = Flask(\_\_name\_\_)  
model= load('random1.save')  
trans1=load('ts1')  
trans2=load('ts2')  
  
  
@app.route('/')  
def home():  
 return render\_template('index.html')  
  
@app.route('/y\_predict',methods=['POST'])  
def y\_predict():  
 *'''  
 For rendering results on HTML GUI  
 '''* x\_test = [[x for x in request.form.values()]]  
 print(x\_test)  
   
 test=trans1.transform(x\_test)  
 test=test[:,1:]  
   
 test=trans2.transform(x\_test)  
 test=test[:,1:]  
   
   
   
 print(test)  
 prediction = model.predict(test)  
 print(prediction)  
 output=prediction[0]  
   
 return render\_template('index.html', prediction\_text='Price {}'.format(output))  
  
'''@app.route('/predict\_api',methods=['POST'])  
def predict\_api():  
   
 #For direct API calls trought request  
   
 data = request.get\_json(force=True)  
 prediction = model.y\_predict([np.array(list(data.values()))])  
  
 output = prediction[0]  
 return jsonify(output)'''  
  
if \_\_name\_\_ == "\_\_main\_\_":  
 app.run(debug=True)