Used Car Price Prediction

Jialong Zhang
Zehao Liu
Instructor: Dr. Sukhjit Singh Sehra
July 19, 2022

Abstract. A lot of people buy used cars these days because the price of new cars is often too expensive. However, there are many risks associated with buying a used car. Using our model to predict a reasonable price before buying a used car can help buyers clarify their goals.

1. Introduction

A huge amount of people desired to own a car to make their life convenience and the price of car is big concern of those afraid to be cheated and a lot of people waste a plenty of money on not bargain deals on used car. To address the issue, project aims to design a two-stage machine learning engine to predict the price of car by inputting the core property of the car. The price is predominantly influenced by main 14 property in table 1.

Section 2 explains how a serial of property were processed to construct the dataset. Section 3 and Section 4 deal with how different classifiers and regressors were trained and analyzed on the dataset respectively. Finally, Section 5 details the random forest model to predict car's prices

2. Project Description

2.1 Data Collection

It contains test-data.csv and train-data.csv, our project is just using train-data. data contains 14 attributes from table 1, we will do data cleaning on these raw datasets to deal with the missing data and redundant attributes.

Feature vector:

Name is the brand and model of the car.

Location is location in which the car is being sold or is available for purchase.

Year is the edition of the car.

Kilometers_Driven is The total kilometres driven in the car by the previous owner(s) in KM

Fuel_Type is The type of fuel used by the car. (Petrol / Diesel / Electric / CNG / LPG)

Transmission is The type of transmission used by the car. (Automatic / Manual)

Owner_Type determine car's ownership: Firsthand, Second hand or other.

Mileage is the fuel consumption offered by the car company in km/kg

Engine contains the displacement volume of the engine in cc.

Power describe the maximum power of the engine in bhp.

Seats is the number of seats in the car.

New_Price is the price of a new car of the same model.

Price is the price of the used car in INR Lakhs.

Table 1: Data Attributes

index	Name	Location	Year
Kilometers_Driven	Fuel_Type	Transmission	Owner_Type
Mileage	Engine	Power	Seats
New_Price	Price		

```
# Importing necessary libraries
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
sns.set()
```

2.2 Exploratory data analysis

First, we imported the necessary libraries (e.g. pandas, numpy, matplotlib and seaborn) and loaded the dataset.

We used the ".head()" and ".tail()" functions from the pandas library to return the first and last 5 observations of the dataset allowing a closer look at the data.

	Unnamed: 0	Name	Location	Year	Kilometers_Driven	Fuel_Type	...	Mileage	Engine	Power	Seats	New_Price	Price
0	0	Maruti Wagon R LXI CNG	Mumbai	2010	72000	CNG	...	26.6 km/kg	998 CC	58.16 bhp	5.0	NaN	1.75
1	1	Hyundai Creta 1.6 CRDi SX Option	Pune	2015	41000	Diesel	...	19.67 kmpl	1582 CC	126.2 bhp	5.0	NaN	12.50
2	2	Honda Jazz V	Chennai	2011	46000	Petrol	...	18.2 kmpl	1199 CC	88.7 bhp	5.0	8.61 Lakh	4.50
3	3	Maruti Ertiga VDI	Chennai	2012	87000	Diesel	...	20.77 kmpl	1248 CC	88.76 bhp	7.0	NaN	6.00
4	4	Audi A4 New 2.0 TDI Multitronic	Coimbatore	2013	40670	Diesel	...	15.2 kmpl	1968 CC	140.8 bhp	5.0	NaN	17.74

	Unnamed: 0	Name	Location	Year	Kilometers_Driven	Fuel_Type	...	Mileage	Engine	Power	Seats	New_Price	Price
6014	6014	Maruti Swift VDI	Delhi	2014	27365	Diesel	...	28.4 kmpl	1248 CC	74 bhp	5.0	7.88 Lakh	4.75
6015	6015	Hyundai Xcent 1.1 CRDi S	Jaipur	2015	100000	Diesel	...	24.4 kmpl	1120 CC	71 bhp	5.0	NaN	4.00
6016	6016	Mahindra Xylo D4 BSIV	Jaipur	2012	55000	Diesel	...	14.0 kmpl	2498 CC	112 bhp	8.0	NaN	2.90
6017	6017	Maruti Wagon R VXI	Kolkata	2013	46000	Petrol	...	18.9 kmpl	998 CC	67.1 bhp	5.0	NaN	2.65
6018	6018	Chevrolet Beat Diesel	Hyderabad	2011	47000	Diesel	...	25.44 kmpl	936 CC	57.6 bhp	5.0	NaN	2.50

We use “.info()" to know the columns and their corresponding data types.

	Year	Kilometers_Driven	Seats	Price
count	6019.000000	6.019000e+03	5977.000000	6019.000000
mean	2013.358199	5.873838e+04	5.278735	9.479468
std	3.269742	9.126884e+04	0.808840	11.187917
min	1998.000000	1.710000e+02	0.000000	0.440000
25%	2011.000000	3.400000e+04	5.000000	3.500000
50%	2014.000000	5.300000e+04	5.000000	5.640000
75%	2016.000000	7.300000e+04	5.000000	9.950000
max	2019.000000	6.500000e+06	10.000000	160.000000

Use describe() function returns the count, average, standard deviation, minimum, maximum and quantity of the data.

```
Data columns (total 14 columns):
 #   Column             Non-Null Count  Dtype
---  ------             --------------  -----
 0   Unnamed: 0         6019 non-null   int64
 1   Name               6019 non-null   object
 2   Location           6019 non-null   object
 3   Year               6019 non-null   int64
 4   Kilometers_Driven  6019 non-null   int64
 5   Fuel_Type          6019 non-null   object
 6   Transmission       6019 non-null   object
 7   Owner_Type         6019 non-null   object
 8   Mileage            6017 non-null   object
 9   Engine             5983 non-null   object
 10  Power              5983 non-null   object
 11  Seats              5977 non-null   float64
 12  New_Price          824 non-null    object
 13  Price              6019 non-null   float64
dtypes: float64(2), int64(3), object(9)
```

Now, we have a good basic understanding of the data. Then we perform data cleanup, Use ".isnull().sum()" to find the missing data, then use ".notna()" to delete the records with NULL values, and finally use "reset_index" to reset the index.

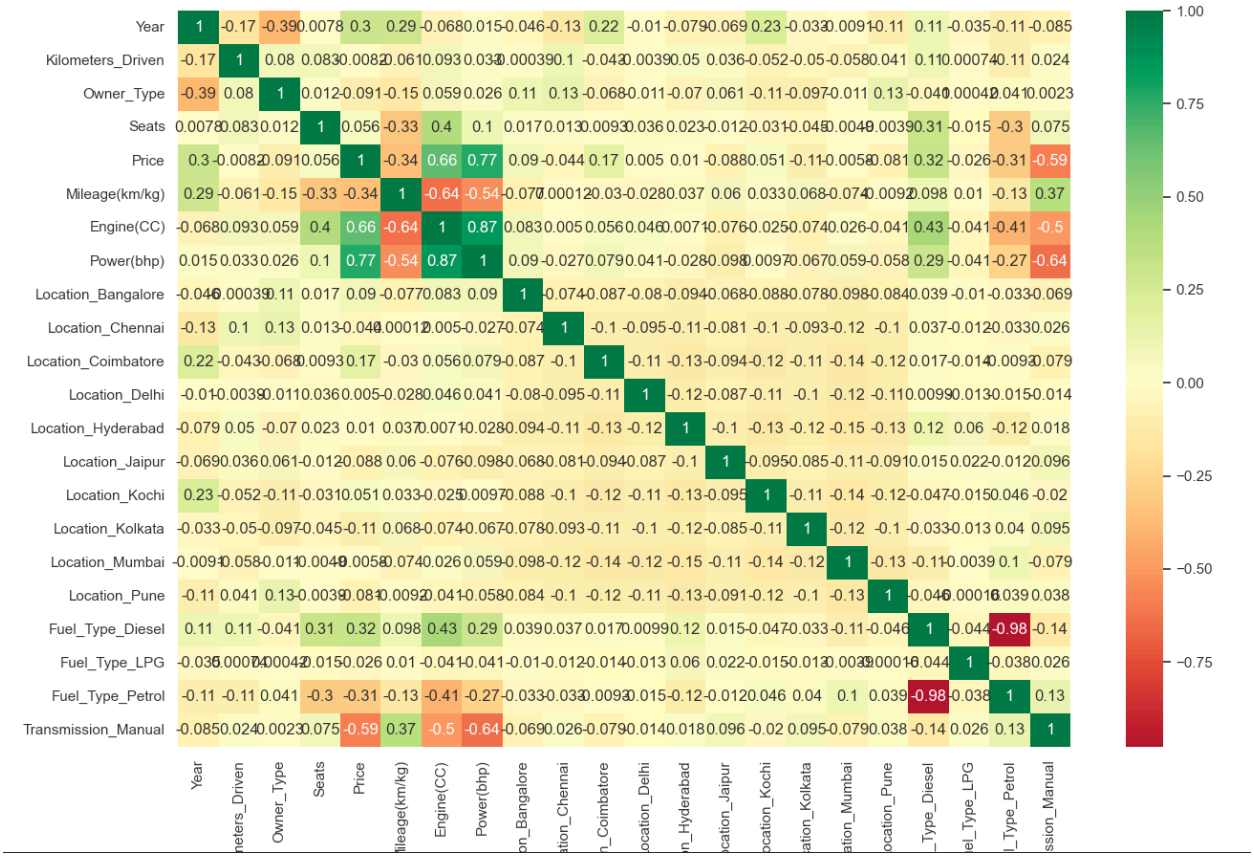
```
#delete the records with NULL values
train_data.isnull().sum()
train_data = train_data[train_data['Mileage'].notna()]
train_data = train_data[train_data['Engine'].notna()]
train_data = train_data[train_data['Power'].notna()]
train_data = train_data[train_data['Seats'].notna()]
train_data = train_data.reset_index(drop=True)
```

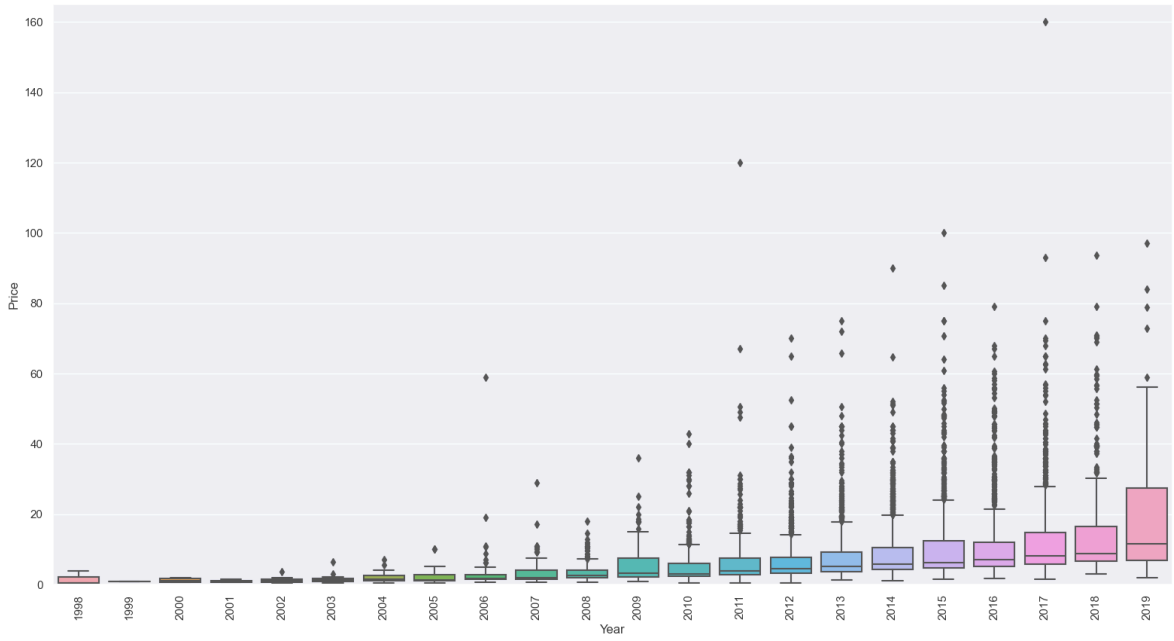
We remove some units from the data and make the data as numeric as possible. delete all useless features such as "name", “New_Price”. For non-numeric data, we use OneHotEncoder and LabelEncoder to process it, depending on whether the data is ordered or not.

```
#Working with Categorical Data

#OneHotEncoder
var = 'Location'
train_data[var].value_counts()
Location = train_data[[var]]
Location = pd.get_dummies(Location,drop_first=True)

var = 'Fuel_Type'
train_data[var].value_counts()
Fuel_t = train_data[[var]]
Fuel_t = pd.get_dummies(Fuel_t,drop_first=True)

var = 'Transmission'
train_data[var].value_counts()
Transmission = train_data[[var]]
Transmission = pd.get_dummies(Transmission,drop_first=True)

#Label Encoding
var = 'Owner_Type'
train_data[var].value_counts()
train_data.replace({"First":1,"Second":2,"Third": 3,"Fourth & Above":4},inplace=True)
```

Using the visualization library Seaborn, you can create statistical plots for univariate and multivariate analysis. To use the linear regression model, we use the ".corr()" function of pandas to find correlations and use the heat map in seaborn to visualize the correlation matrix.

```
plt.figure(figsize=(10,10))
sns.heatmap(final_train.corr(),annot=True,cmap='RdYlGn')
plt.show()
```

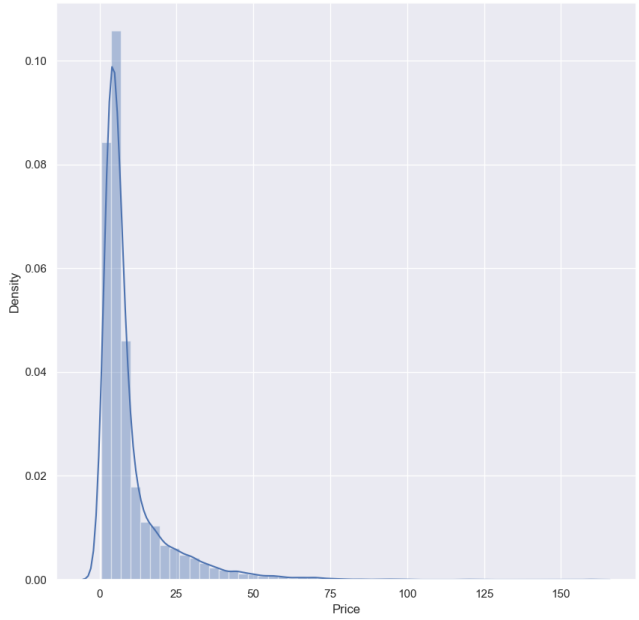




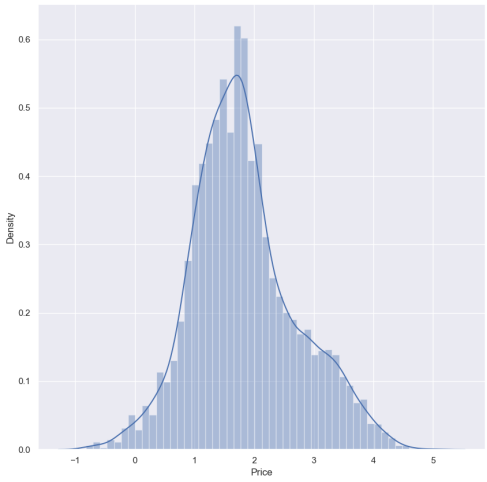
Here we can infer that "Price" has a strong positive correlation with "Power" and a strong negative correlation with "Transmission Manual".

Box plots can help us visualize outliers. We found outliers for all years after 2001.

We can observe that the distribution of prices shows a high positive skewness to the right (skew > 1). A kurtosis value of 0.1 is very high, meaning that there is a profusion of outliers in the dataset.

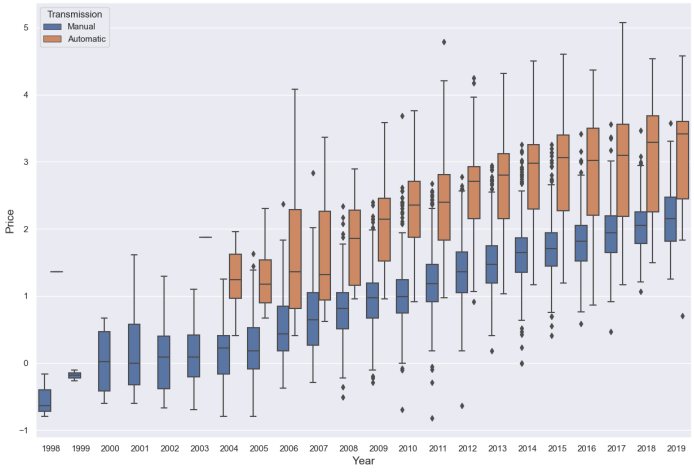


To the the graph clear, a scaling method used to convert the value of Price by Log(Price). It might be a good solution to have a more normal visualization of the distribution of the Price

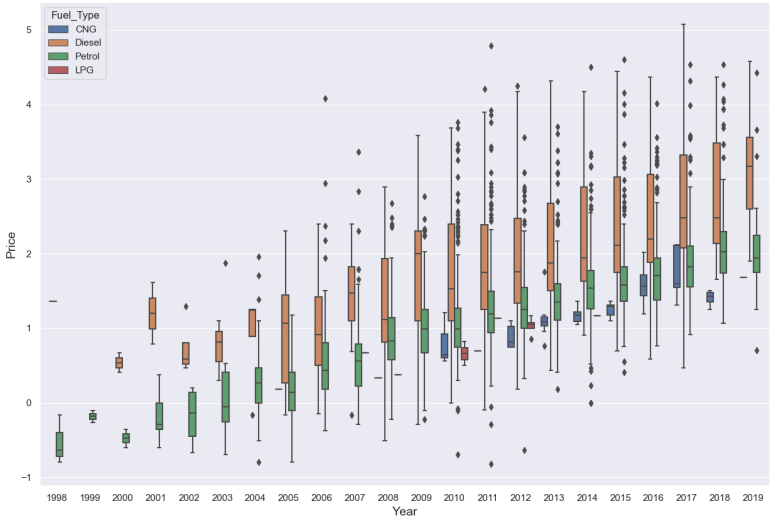


The heat map render the relationship between each parameter. It is clear to see price have strong positive correlation with car's power and car's engine.

According to the graph, the price of automatic car is always higher than manual car thus we can sum up that automatic cars are more costly



According to the graph, the price of relatively new car is always higher than relatively old car thus we can sum up that "New" cars are more expensive than "Old" cars and this is common sense

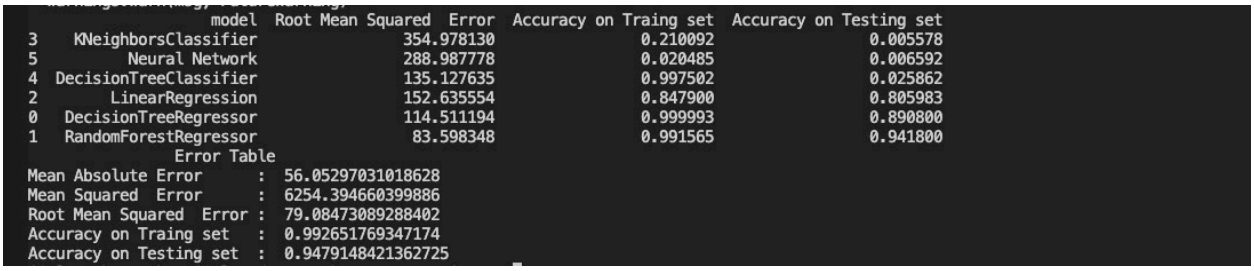


3.methodology

The project execute six difference model to fitting the data, including linear regression , neural network, KNeighbors classifier, decision tree classifier, decision tree regressor and

random forest regressor in our example. Linear regression decision tree regressor and random forest regressor is successes to predict all the content, since their accuracy is over 0.8. Random forest have the highest accuracy that is around 0.94 so that it becomes our best choice of model.

Following is our graph:



	model	Root Mean Squared Error	Accuracy on Traing set	Accuracy on Testing set
3	KNeighborsClassifier	354.978130	0.210092	0.005578
5	Neural Network	288.987778	0.020485	0.006592
4	DecisionTreeClassifier	135.127635	0.997502	0.025862
2	LinearRegression	152.635554	0.847900	0.805983
0	DecisionTreeRegressor	114.511194	0.999993	0.890800
1	RandomForestRegressor	83.598348	0.991565	0.941800

Error Table

Mean Absolute Error	:	56.05297031018628
Mean Squared Error	:	6254.394660399886
Root Mean Squared Error	:	79.08473089288402
Accuracy on Traing set	:	0.992651769347174
Accuracy on Testing set	:	0.9479148421362725

Linear regression is one of the supervised Machine learning algorithms in Python that observes continuous features and predicts an outcome. It assigns optimal weights to variables to create a line ax+b to predict the output.

Logistic regression is a supervised classification is unique Machine Learning algorithms in Python that finds its use in estimating discrete values like 0/1, yes/no, and true/false based on a given set of independent variables.

Kneighbors is a supervised learning algorithm that considers different centroids and uses a usually Euclidean function to compare distance. Then, it analyzes the results and classifies each point to the group to optimize it to place with all closest points to it. It classifies new cases using a majority vote of k of its neighbours. The case it assigns to a class is the one most common among its K nearest neighbours.

A decision tree falls under supervised Machine Learning Algorithms in Python and comes of use for both classification and regression. It takes an instance, traverses the tree, and compares important features with a determined conditional statement.

A random forest is an ensemble of decision trees. In order to classify every new object based on its attributes, trees vote for class- each tree provides a classification. The classification with the most votes wins in the forest

4.conclusion

This is a data analysis project for machine learning. we did clean the raw data by null vale. After data Preprocessing, six model are used to fitting our prediction. According to the analysis, random forest is best model in our prediction and this project is useful to predict the price of used car.

5. Reference: