CS 552 – Data Science with Python Assignment-1
Cumhuriye Tülüğ Küçüköğüt
S025791

Abstract

Cars are the dreams of many people and they are a tool that makes people's lives easier.Cars have a lot of features, and these features are big factors that affect the prices of cars. Some of these features are: year,fuel,seller type etc.Most of the people, when buying their cars, examine these features in detail and evaluate the price.

The aim of this assignment, prediction of the car prices according to the independent variables that are year,km driven,seller type etc.

Introduction

Today, most people try to establish the relationship between features and try to predict the dependent value according to these features.For this reason many algorithms are implemented. Linear regression is the one of them.

Linear regression is a technique which is generally used in statistics and machine learning.It is used to make a relationship between independent and dependent data. Fueled by training examples, the algorithm finds parameters that make the linear model the best fit for your data.[1] This algorithm is used in different fields like finance,engine performance,medicine etc. It is used to predict continuous model.

In this report firstly implementation of data is expressed, secondly results are shown and finally the project is concluded.

Implementation

"car-pricing-data.csv" file includes name,year,selling price,km driven,fuel,seller type transmission,owner,mileage,engine,max power,seats variables and this dataset includes 8128 row and 12 columns.

	name	year	selling_price	km_driven	fuel	seller_type	transmission	owner	mileage	engine	max_power	seats
0	Maruti Swift Dzire VDI	2014	450000	145500	Diesel	Individual	Manual	First Owner	23.4 kmpl	1248 CC	74 bhp	5.0
1	Skoda Rapid 1.5 TDI Ambition	2014	370000	120000	Diesel	Individual	Manual	Second Owner	21.14 kmpl	1498 CC	103.52 bhp	5.0
2	Honda City 2017-2020 EXi	2006	158000	140000	Petrol	Individual	Manual	Third Owner	17.7 kmpl	1497 CC	78 bhp	5.0
3	Hyundai i20 Sportz Diesel	2010	225000	127000	Diesel	Individual	Manual	First Owner	23.0 kmpl	1396 CC	90 bhp	5.0
4	Maruti Swift VXI BSIII	2007	130000	120000	Petrol	Individual	Manual	First Owner	16.1 kmpl	1298 CC	88.2 bhp	5.0
...	...	...	...	...	...	...	...	...	...	...	...	...
8123	Hyundai i20 Magna	2013	320000	110000	Petrol	Individual	Manual	First Owner	18.5 kmpl	1197 CC	82.85 bhp	5.0
8124	Hyundai Verna CRDi SX	2007	135000	119000	Diesel	Individual	Manual	Fourth & Above Owner	16.8 kmpl	1493 CC	110 bhp	5.0
8125	Maruti Swift Dzire ZDi	2009	382000	120000	Diesel	Individual	Manual	First Owner	19.3 kmpl	1248 CC	73.9 bhp	5.0
8126	Tata Indigo CR4	2013	290000	25000	Diesel	Individual	Manual	First Owner	23.57 kmpl	1396 CC	70 bhp	5.0
8127	Tata Indigo CR4	2013	290000	25000	Diesel	Individual	Manual	First Owner	23.57 kmpl	1396 CC	70 bhp	5.0

8128 rows × 12 columns

There are two different type of data which are categorical and numerical. Name,fuel,seller type,transmission,owner are categorical and the others are numerical.In detail data types are shown in the table.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 8128 entries, 0 to 8127
Data columns (total 12 columns):
 #   Column         Non-Null Count  Dtype
---  ------         --------------  -----
 0   name           8128 non-null   object
 1   year           8128 non-null   int64
 2   selling_price  8128 non-null   int64
 3   km_driven      8128 non-null   int64
 4   fuel           8128 non-null   object
 5   seller_type    8128 non-null   object
 6   transmission   8128 non-null   object
 7   owner          8128 non-null   object
 8   mileage        7907 non-null   object
 9   engine         7907 non-null   object
 10  max_power      7913 non-null   object
 11  seats          7907 non-null   float64
dtypes: float64(1), int64(3), object(8)
memory usage: 762.1+ KB
```

In this data; mileage,engine,seats have 221 and max power has 215 missing values.

```
name              0
year              0
selling_price     0
km_driven         0
fuel              0
seller_type       0
transmission      0
owner             0
mileage         221
engine          221
max_power       215
seats           221
```

Head, tail and correlation of data are shown below.

	name	year	selling_price	km_driven	fuel	seller_type	transmission	owner	mileage	engine	max_power	seats
0	Maruti Swift Dzire VDI	2014	450000	145500	Diesel	Individual	Manual	First Owner	23.4 kmpl	1248 CC	74 bhp	5.0
1	Skoda Rapid 1.5 TDI Ambition	2014	370000	120000	Diesel	Individual	Manual	Second Owner	21.14 kmpl	1498 CC	103.52 bhp	5.0
2	Honda City 2017-2020 EXi	2006	158000	140000	Petrol	Individual	Manual	Third Owner	17.7 kmpl	1497 CC	78 bhp	5.0
3	Hyundai i20 Sportz Diesel	2010	225000	127000	Diesel	Individual	Manual	First Owner	23.0 kmpl	1396 CC	90 bhp	5.0
4	Maruti Swift VXI BSIII	2007	130000	120000	Petrol	Individual	Manual	First Owner	16.1 kmpl	1298 CC	88.2 bhp	5.0

Head of the dataset

	name	year	selling_price	km_driven	fuel	seller_type	transmission	owner	mileage	engine	max_power	seats
8123	Hyundai i20 Magna	2013	320000	110000	Petrol	Individual	Manual	First Owner	18.5 kmpl	1197 CC	82.85 bhp	5.0
8124	Hyundai Verna CRDi SX	2007	135000	119000	Diesel	Individual	Manual	Fourth & Above Owner	16.8 kmpl	1493 CC	110 bhp	5.0
8125	Maruti Swift Dzire ZDi	2009	382000	120000	Diesel	Individual	Manual	First Owner	19.3 kmpl	1248 CC	73.9 bhp	5.0
8126	Tata Indigo CR4	2013	290000	25000	Diesel	Individual	Manual	First Owner	23.57 kmpl	1396 CC	70 bhp	5.0
8127	Tata Indigo CR4	2013	290000	25000	Diesel	Individual	Manual	First Owner	23.57 kmpl	1396 CC	70 bhp	5.0

Tail of the Dataset

	year	selling_price	km_driven	seats
year	1.000000	0.414092	-0.418006	-0.009144
selling_price	0.414092	1.000000	-0.225534	0.041358
km_driven	-0.418006	-0.225534	1.000000	0.227336
seats	-0.009144	0.041358	0.227336	1.000000

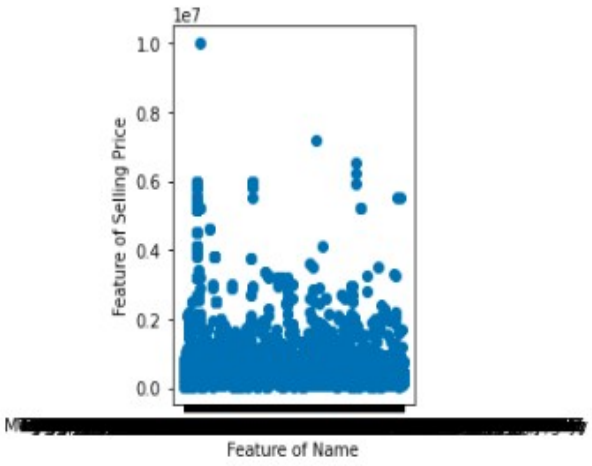
Correlation of the Dataset

In this assignment selling price is dependent and the other features are independent variables.Dependent variable will be predicted according the dependent features. Features are visualized according to the dependent variable. Some of them are shown below.

1)Name and Selling Price

	name	selling_price
0	Maruti Swift Dzire VDI	450000
1	Skoda Rapid 1.5 TDI Ambition	370000
2	Honda City 2017-2020 EXi	158000
3	Hyundai i20 Sportz Diesel	225000
4	Maruti Swift VXI BSIII	130000
...	...	...
8123	Hyundai i20 Magna	320000
8124	Hyundai Verna CRDi SX	135000
8125	Maruti Swift Dzire ZDi	382000
8126	Tata Indigo CR4	290000
8127	Tata Indigo CR4	290000

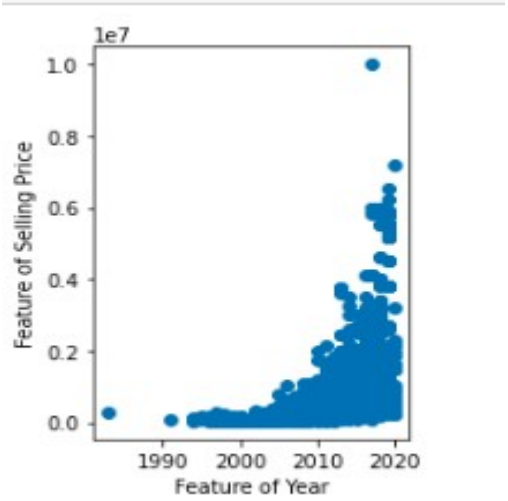
8128 rows × 2 columns



2)Year and Selling Price

	year	selling_price
0	2014	450000
1	2014	370000
2	2006	158000
3	2010	225000
4	2007	130000
...	...	...
8123	2013	320000
8124	2007	135000
8125	2009	382000
8126	2013	290000
8127	2013	290000

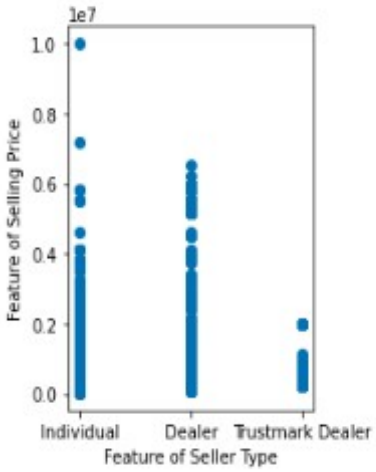
8128 rows × 2 columns



3) Seller Type and Selling Price

	seller_type	selling_price
0	Individual	450000
1	Individual	370000
2	Individual	158000
3	Individual	225000
4	Individual	130000
...	...	...
8123	Individual	320000
8124	Individual	135000
8125	Individual	382000
8126	Individual	290000
8127	Individual	290000

8128 rows × 2 columns



After observing all data according to the dependent variable, one hot encoding is done via pandas' get_dummies function. One hot encoding provides that selections of categoricals will be displayed in binary.As explained above name,fuel,seller type,transmission,owner are categorical one hot encoding is done for these features. After one hot encoding, 8128 rows × 2079 columns are formed, null values are dropped and dependent and independent variables are divided into two groups.

Independent variables:

	year	km_driven	mileage	engine	max_power	seats	name_Ambassador CLASSIC 1500 DSL AC	name_Ambassador Classic 2000 DSZ AC PS	name_Ambassador Grand 1500 DSZ BSIII	name_Ambassador Grand 2000 DSZ PW CL	...	owner_First Owner
0	2014	145500	23.4 kmpl	1248 CC	74 bhp	5.0	0	0	0	0	...	1
1	2014	120000	21.14 kmpl	1498 CC	103.52 bhp	5.0	0	0	0	0	...	0
2	2006	140000	17.7 kmpl	1497 CC	78 bhp	5.0	0	0	0	0	...	0
3	2010	127000	23.0 kmpl	1396 CC	90 bhp	5.0	0	0	0	0	...	1
4	2007	120000	16.1 kmpl	1298 CC	88.2 bhp	5.0	0	0	0	0	...	1
...	...	...	...	...	...	...	...	...	...	...	...	...
8123	2013	110000	18.5 kmpl	1197 CC	82.85 bhp	5.0	0	0	0	0	...	1
8124	2007	119000	16.8 kmpl	1493 CC	110 bhp	5.0	0	0	0	0	...	0
8125	2009	120000	19.3 kmpl	1248 CC	73.9 bhp	5.0	0	0	0	0	...	1
8126	2013	25000	23.57 kmpl	1396 CC	70 bhp	5.0	0	0	0	0	...	1
8127	2013	25000	23.57 kmpl	1396 CC	70 bhp	5.0	0	0	0	0	...	1

7907 rows × 2078 columns

Dependent variable:

```
0       450000
1       370000
2       158000
3       225000
4       130000
         ...
8123    320000
8124    135000
8125    382000
8126    290000
8127    290000
Name: selling_price, Length: 7906, dtype: int64
```

After grouping year is converted to age, all unit names are removed from columns.

Feature Dataset Creation

- All columns are used
- Most of the 5 columns are used
- Dataset is created according to correlation

Algorithms

For predicting values, used libraries:

- LinearRegression
- RandomForestRegressor

This part is used for all datasets
For using all columns in algorithms, firstly dataset is divided into train and test. Accordingly sklearn.model_selection's train_test_split is used which provides that datasets are created randomly.Train set is used to fit the machine learning model and test set is used to evaluate the fit machine learning model.[3] In this part test dataset's size is %20 of the data,train dataset's size is %80 of the data. Data are shuffled 147 times.After the division, features are transformed by scaling each feature to between 0 and 1[4]. It provides that normalizing the data.

Linear Regression and RandomForestRegressor,For these algorithms sckitlearn library is used

RandomForestRegressor

Results

1)All columns

All columns has been put into the algorithm Linear Regression and RandomForestRegressor.

	LinearRegression	**RandomForestRegressor**
Mean-squared error	6.893523388663361e+31	310093442609.11957
Mean-absolute error	993276743210525.6	276524.0135090737
Root-mean-squared error	8302724485771739.0	556860.343900621
r2 score	-1.2624132971257396e+20	0.43212481741190656

2)Most of the 5 columns are used

In this part features are selected according to the k highest scores. According to the results,

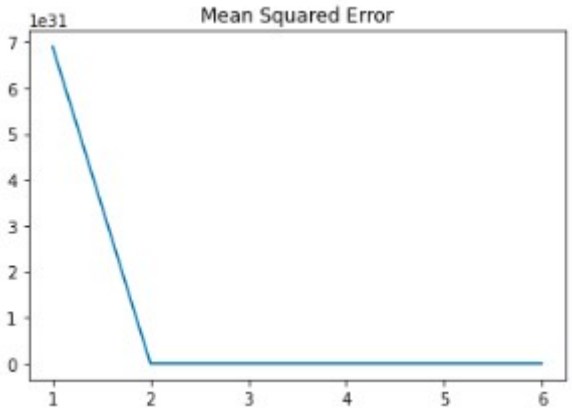
max_power,transmission_Manual,transmission_Automatic,ame_BMW X4 M Sport X xDrive20d and engine are selected as a feature set.

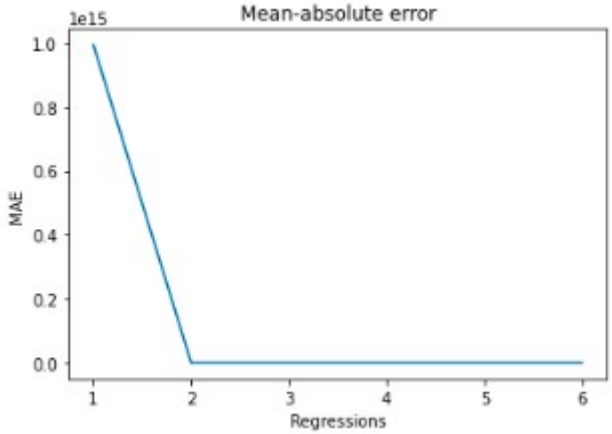
	LinearRegression	**RandomForestRegressor**
Mean-squared error	359565940366.4678	399266862319.1518
Mean-absolute error	444339.24175324745	383071.9486359539
Root-mean-squared error	599638.1745406706	631875.6699851259
r2 score	0.34152566297426623	0.26882122874598546

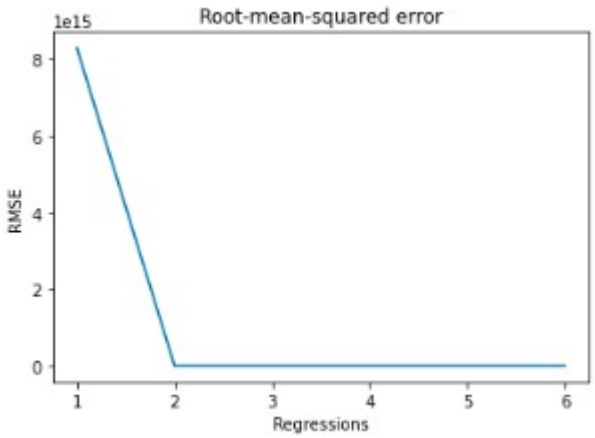
3)Dataset is created according to correlation

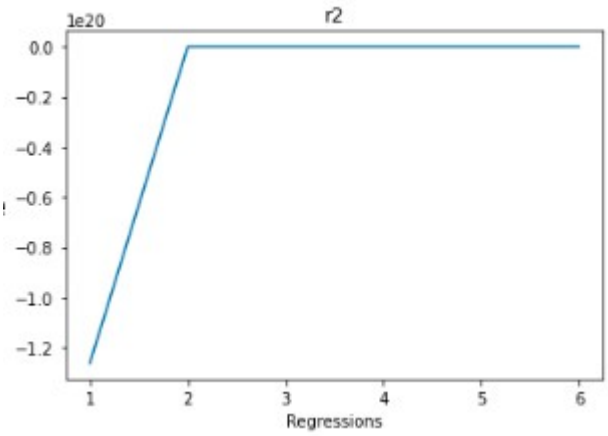
In this part features are selected according to the correlation value, correlation is selected 0,4, feaures are : year, name_BMW X4 M Sport X xDrive20d, transmission_Automatic, transmission_Manual, seller_type_Dealer

	LinearRegression	**RandomForestRegressor**
Mean-squared error	215231732871.62463	179902870305.4146
Mean-absolute error	271785.55200585164	271785.55200585164
Root-mean-squared error	463930.74145999923	424149.5848228719
r2 score	0.5692634190927124	0.7164282367885768









Conclusion

The MSE(Mean-squared error) is a risk function corresponding to the expected value of the squared error loss.[5]

MAE(Mean-absolute error) is the difference betwwen true values and predicted value.
RMSE (Root Mean Squared Error) is the standard deviation of the residuals (prediction errors).[6]

Mean Squared error,mean absolute error and root-mean-squared error are calculated for the interpret the performance of the linear regression models.

R-squared (R2) is a statistical measure that represents the proportion of variance for a dependent variable explained by variables in a regression model.[7]
According to MAE,MSE and RMSE information, 3rd Random Forest Regressor is the better and results show that **Dataset is created according to correlation** dataset is better and if we compare the algorithms mostly Random Forest Regressor algorihm's results are smaller.Based on our datasets, we can comment that this algorithm is better.

According to results,in Turkey cars are most of people's dream and every day prices are increasing. Car prices are depend on many things such as year,km_driven,owner etc. Most of sites like Sahibinden grouping cars according to these features.This model is beneficial for both car sellers and car buyers. Because, according to the formula, both of these groups are
they know the real price of the car and can limit their search accordingly.

References

[1]https://www.analyticsvidhya.com/blog/2021/05/all-you-need-to-know-about-your-first-machine-learning-model-linear-regression/#:~:text=In%20the%20most%20simple%20words,the%20dependent%20and%20independent%20variable.
[2]https://statisticsbyjim.com/regression/interpret-constant-y-intercept-regression/
3)https://machinelearningmastery.com/train-test-split-for-evaluating-machine-learning-algorithms/
4)https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.MinMaxScaler.html
5)https://en.wikipedia.org/wiki/Mean_squared_error
6)https://www.statisticshowto.com/probability-and-statistics/regression-analysis/rmse-root-mean-square-error/
7)https://www.investopedia.com/terms/r/r-squared.asp