Machine Leaning using Python

Capstone Project 4

Audi

Used Car Price Prediction

by : Kenny Lim/Cohort 3

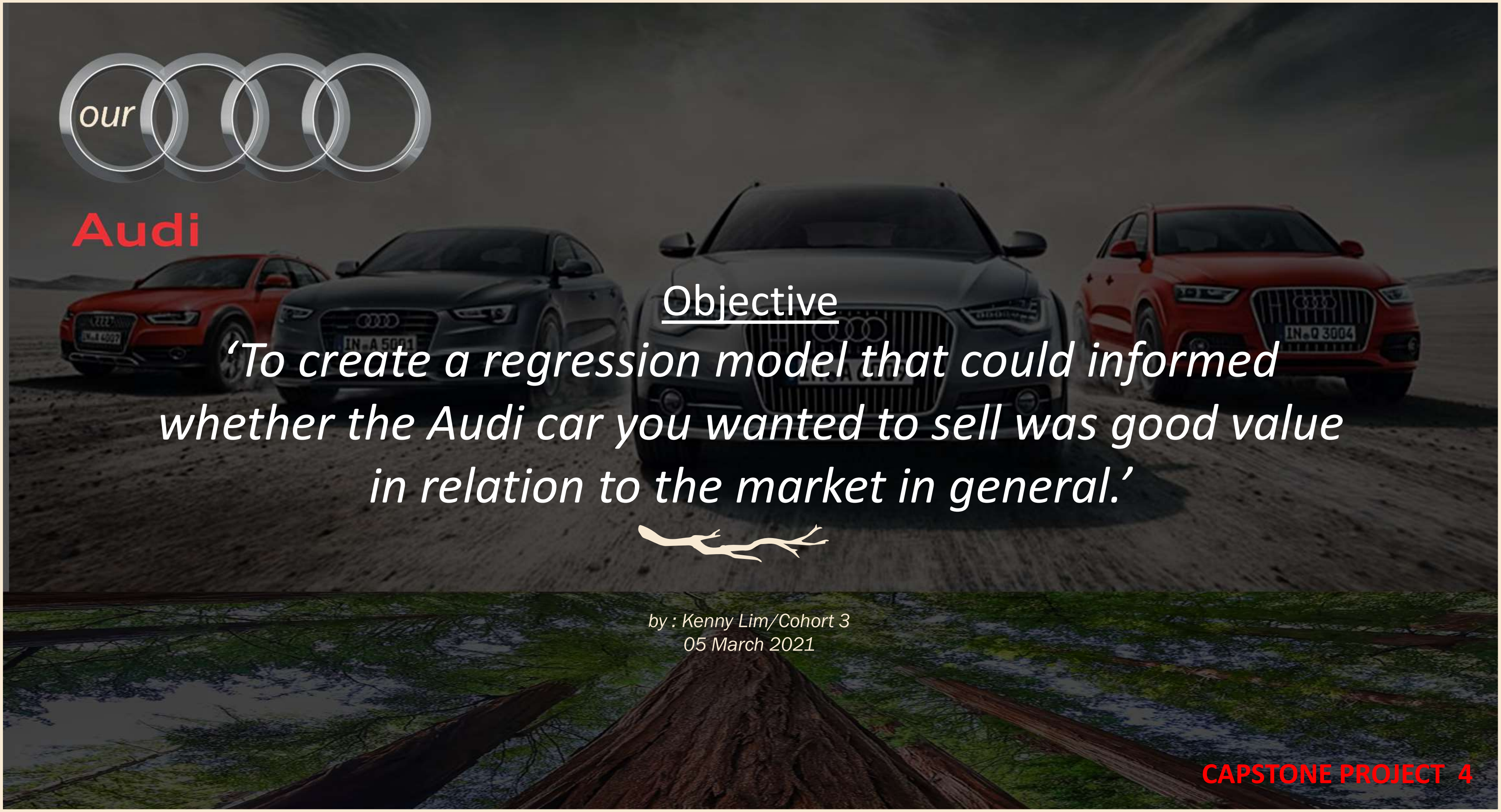
05 March 2021

our Audi

CONTENT

1. Introduction & Objective
2. Business Context
3. Methodology
4. Process Workflow
5. Results
6. Conclusions

by : Kenny Lim/Cohort 3
05 March 2021



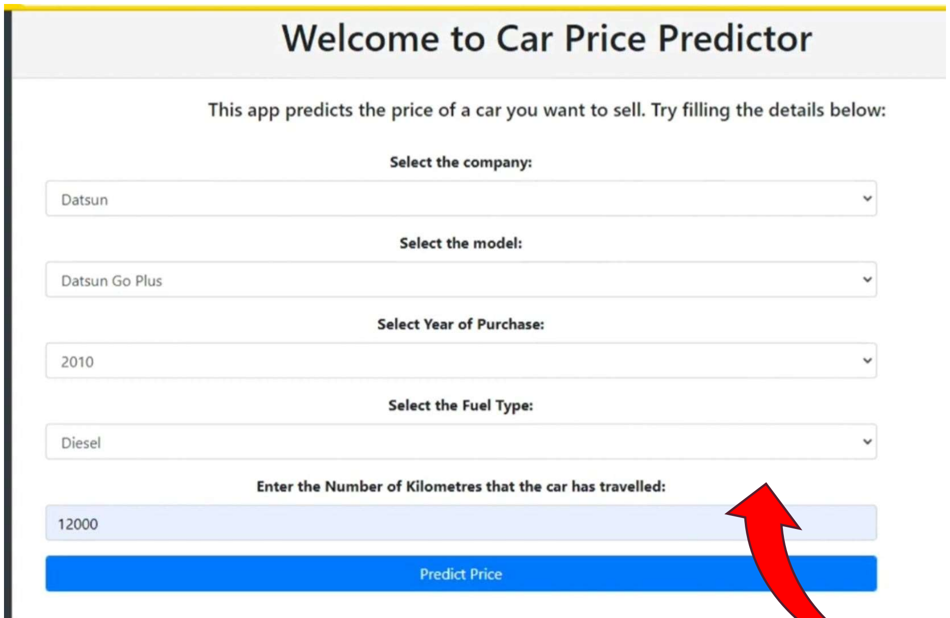
our
Audi

Objective

'To create a regression model that could informed whether the Audi car you wanted to sell was good value in relation to the market in general.'

by : Kenny Lim/Cohort 3
05 March 2021

CAPSTONE PROJECT 4



Business Context :

To build a used car price predictor using Linear Regression model.

To find info such as when is the ideal time to sell certain cars (i.e. at what age & mileage are there significant drops in resale value).

Then convert it into a full-fledged website using the flask framework.

CAPSTONE PROJECT 4

by : Kenny Lim/Cohort 3
05 March 2021

Used Car Database

The scraped data of used cars listing.

Listing has more than 100,000 used car info separated into files corresponding to each car manufacturer

Sources : 100,000 UK Used Car Data set | Kaggle

by : Kenny Lim/Cohort 3

05 March 2021

CAPSTONE PROJECT 4

audi.csv (521.55 KB)

Detail Compact Column

9 of 9 columns

About this file

Audi used car listings.

model	year	price	transmission	mileage	fuelType	tax	mpg	engineSize
audi model	registration year	price in £	type of gearbox	distance used	engine fuel	road tax	miles per gallon	size in litres
A3 18% Q3 13% Other (7322) 69%	1997 – 2020	1490 – 145k	Manual 41% Semi-Auto 34% Other (2708) 25%	1 – 323k	Diesel 52% Petrol 47% Other (28) 0%	0 – 580	18.9 – 188	0 1.26 - 1.57 Count: 2,338
A1	2017	12500	Manual	15735	Petrol	150	55.4	1.4
A6	2016	16500	Automatic	36203	Diesel	20	64.2	2.0
A1	2016	11000	Manual	29946	Petrol	30	55.4	1.4
A4	2017	16800	Automatic	25952	Diesel	145	67.3	2.0
A3	2019	17300	Manual	1998	Petrol	145	49.6	1.0
A1	2016	13900	Automatic	32260	Petrol	30	58.9	1.4
A6	2016	13250	Automatic	76788	Diesel	30	61.4	2.0
A4	2016	11750	Manual	75185	Diesel	20	70.6	2.0
A3	2015	10200	Manual	46112	Petrol	20	60.1	1.4

Audi Used Car Listing

Records : 10,668

Features : 9 (columns)

The cleaned data set contains information of price, transmission, mileage, fuel type, road tax, miles per gallon (mpg), and engine size.

Used Car Database

Listing with more than 100,000 used car that also includes other car manufacturer

Sources : [100,000 UK Used Car Data set | Kaggle]

by : Kenny Lim/Cohort 3
05 March 2021

CAPSTONE PROJECT 4

by : Kenny Lim/Cohort 3
05 March 2021

Audi

Methodology

Audi

Model

- Linear Regression Model (Baseline)
- Support Vector Machine (SVM) Model (Alternative)

Metrics

- Linear Regression R2 (Coefficient of determination),
- Mean Squared Error (MSE)
- Mean Absolute Error (MAE)

Tools

scikit learn pandas NumPy matplotlib seaborn jupyter

```
import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt
import sklearn as sk
```

Capstone Project 4

by : Kenny Lim/Cohort 3
05 March 2021

Process Workflow :

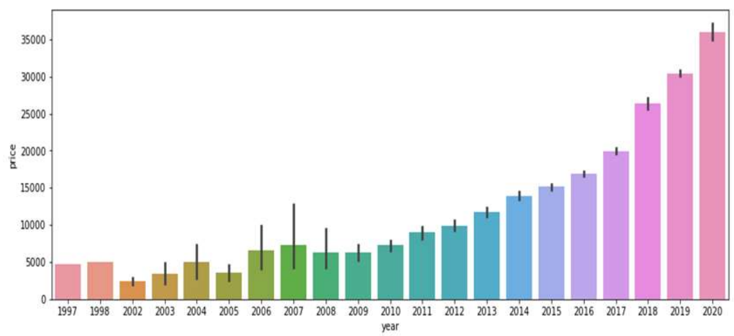
* Exploratory Data Analysis (EDA)
  - Descriptive statistic/correlation/visualization

* Data Preparation & Preprocessing
  - Cleaning/ Data Transformation /Feature Engineering

* Create Machine Learning Models
  - Regression model

* Training Machine Learning Model

* Evaluating Performance
  - Regression problem : MSE, MAE & R2



Visualize manufactured cars (eg year = 2018, 2019) are sold for more average price when compared to the cars that are manufactured earlier.

by : Kenny Lim/Cohort 3
05 March 2021

EDA & Data Preparation

- Upload .csv file and store in Dataframe ➔ car_audi
- Extract information enfolded in the dataset and summarize the main characteristics of the data

```
car_audi.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 10668 entries, 0 to 10667
Data columns (total 9 columns):
 #   Column        Non-Null Count  Dtype
---  ------        --------------  -----
 0   model         10668 non-null  object
 1   year          10668 non-null  int64
 2   price         10668 non-null  int64
 3   transmission  10668 non-null  object
 4   mileage       10668 non-null  int64
 5   fuelType      10668 non-null  object
 6   tax           10668 non-null  int64
 7   mpg           10668 non-null  float64
 8   engineSize    10668 non-null  float64
dtypes: float64(2), int64(4), object(3)
memory usage: 750.2+ KB
```

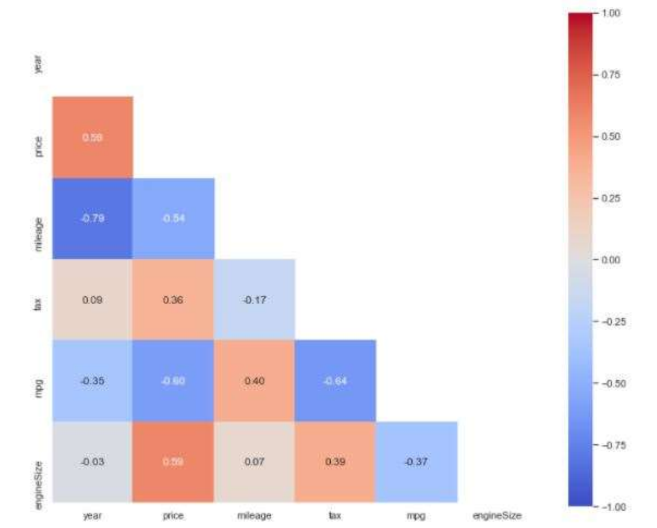
- Get an overall sense of the data shape with the mean/median, min, max, q1, q3 values

	year	price	mileage	tax	mpg	engineSize
count	10668.000000	10668.000000	10668.000000	10668.000000	10668.000000	10668.000000
mean	2017.100675	22896.685039	24827.244001	126.011436	50.770022	1.930709
std	2.167494	11714.841888	23505.257205	67.170294	12.949782	0.602957
min	1997.000000	1490.000000	1.000000	0.000000	18.900000	0.000000
25%	2016.000000	15130.750000	5968.750000	125.000000	40.900000	1.500000
50%	2017.000000	20200.000000	19000.000000	145.000000	49.600000	2.000000
75%	2019.000000	27990.000000	36464.500000	145.000000	58.900000	2.000000
max	2020.000000	145000.000000	323000.000000	580.000000	188.300000	6.300000

- Ensure each feature have non-zero & missing values. (Drop/fill N.A when necessary).

```
model           0
year            0
price           0
transmission    0
mileage         0
fuelType        0
tax             0
mpg             0
engineSize      0
dtype: int64
```

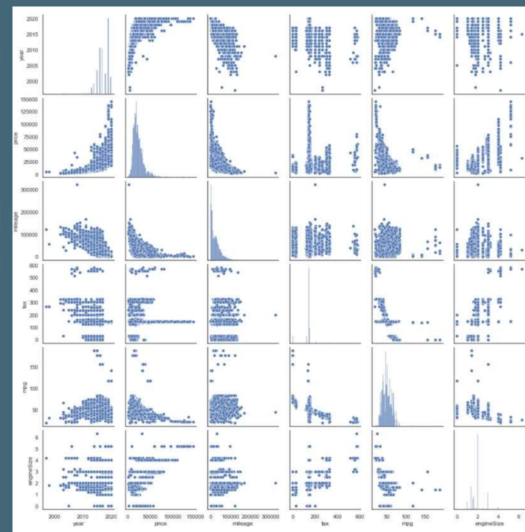
Capstone Project 4

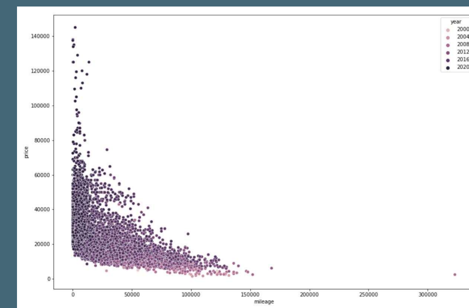


Correlation Matrix heatmap Visualization

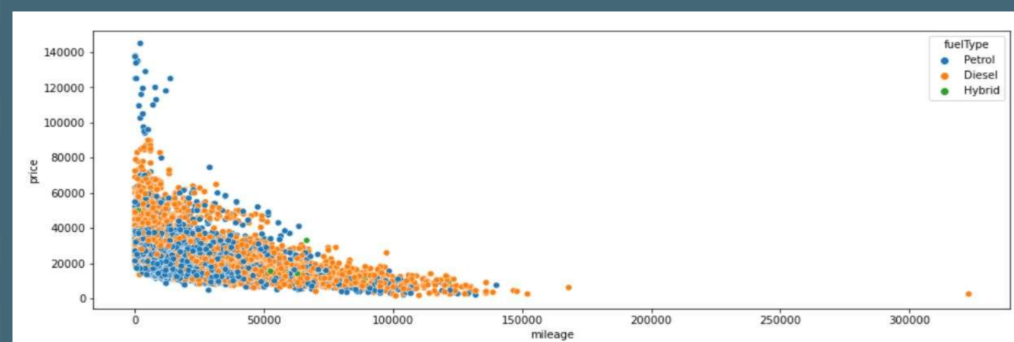
by : Kenny Lim/Cohort 3
05 March 2021

Data Processing / Visualizing Data

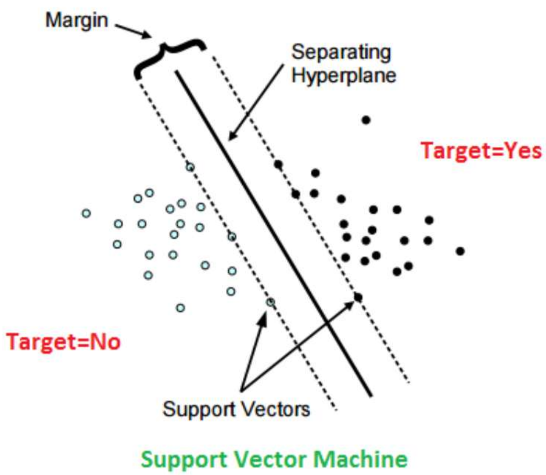




To determining important features that have strong relationship with the target by identifying high correlation values (both positives and negatives)



Capstone Project 4



by : Kenny Lim/Cohort 3
05 March 2021

Create & Training Machine Learning Models

Prepare & split data into training and testing datasets

Apply Method 1 : Support Vector Machine (SVM) Regression problem

Evaluating Performance (Result)

Result for MSE, MAE & R2 (Coefficient of determination)

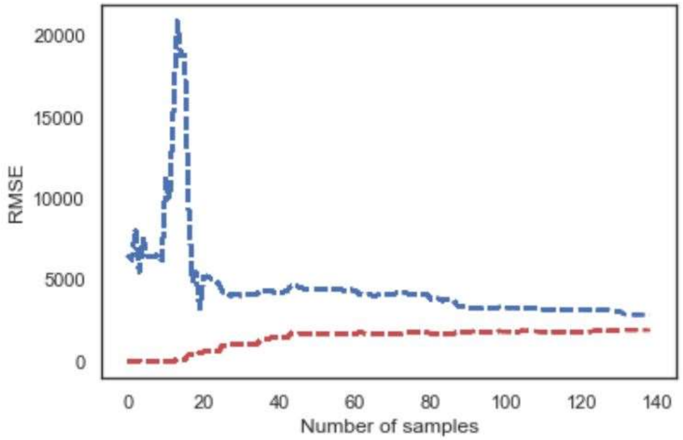
```
# from sklearn.metrics import mean_squared_error, r2_score, mean_absolute_error

# The mean squared error & mean absolute error
print('Mean squared error: {:.2f}'.format(mean_squared_error(y_test, pred)))    #y_pred = pred
print('Mean absolute error: {:.2f}'.format(mean_absolute_error(y_test, pred)))  #y_pred = pred


# The coefficient of determination: 1 is perfect prediction
print('Coefficient of determination: {:.2f}'.format(r2_score(y_test, pred)))   #y_pred = pred
```

```
Mean squared error: 141866344.17
Mean absolute error: 7704.22
Coefficient of determination: 0.04
```

R-Squared (R^2 or the coefficient of determination) is a statistical measure in a regression model that determines the proportion of variance in the dependent variable that can be explained by the independent variable. In other words, r-squared shows how well the data fit the regression model (the goodness of fit).

R2 can take values from 0 to 1. A value of 1 indicates that the regression predictions perfectly fit the data. Results look like the model is not a good fit for car price prediction.

Capstone Project 4



Learning curves above tell us that the RMSE stabilizes as long as the number of samples grow in volume.

by : Kenny Lim/Cohort 3
05 March 2021

Create Machine Learning Models

Apply Method 2 : Linear Regression Model

Data Transformation for categorical features

- To avoid mis-interpretation of feature correlation by ML algorithm
- Apply One-Hot Encoding for Nominal Features
- Split the data into training & testing dataset at test size = 0.2
- Apply Data Normalization using Standard Scaler

	year	price	mileage	tax	mpg	engineSize	model_ A1	model_ A2	model_ A3	model_ A4	...
0	2017	12500	15735	150	55.4	1.4	1	0	0	0	...
1	2016	16500	36203	20	64.2	2.0	0	0	0	0	...
2	2016	11000	29946	30	55.4	1.4	1	0	0	0	...
3	2017	16800	25952	145	67.3	2.0	0	0	0	1	...
4	2019	17300	1998	145	49.6	1.0	0	0	1	0	...

5 rows × 38 columns

Root mean squared error or RMSE is a measure of the difference between actual values and predicted values of a machine learning model like Linear Regression. Root mean squared error is a measure of how well the machine learning model can perform. The lower the RMSE, the better the model.

	Desired Output (Price)	Predicted Output (Price)
3099	10595	10350.260011
4354	35995	32455.219773
516	50414	48475.273956
1634	12798	10637.212491
10372	20000	19024.977247
3603	18495	17551.492430
7001	37888	41436.290935
3969	21990	22854.075767
8948	20350	18530.428444
5277	15490	16176.089667

Compare the trained output for price prediction

by : Kenny Lim/Cohort 3
05 March 2021

Evaluating Performance (Result)

Linear Regression Model

- Result for MSE, MAE & R2 (Coefficient of determination)

```
r2_score(y_test,y_pred)

0.9093927416646835
```

```
from sklearn.metrics import mean_squared_error, r2_score, mean_absolute_error

# The coefficients
print('Coefficients: \n', lr.coef_)

# The mean squared error
print('Mean squared error: {:.2f}'.format(mean_squared_error(y_test, y_pred)))
print('Mean absolute error: {:.2f}'.format(mean_absolute_error(y_test, y_pred)))


# The coefficient of determination: 1 is perfect prediction
print('Coefficient of determination: {:.2f}'.format(r2_score(y_test, y_pred)))

Coefficients:
 [-7930.74401387 -8709.75622336 -6213.85481867 ... -6422.7770615
 30219.00686415 10753.10195535]
Mean squared error: 14232561.58
Mean absolute error: 2605.12
Coefficient of determination: 0.91
```

```
print("Regression model's training score = {:.2f}".format(pipe.score(X_train, y_train)))
print("Regression model's test score     = {:.2f}".format(pipe.score(X_test, y_test)))

Regression model's training score = 0.98
Regression model's test score     = 0.91
```

the great

CONCLUSIONS

Both Support Vector Machine model & Linear Regression model were trained & tested successfully !!!

✓ Using **Linear Regression model**,
- ➢ $R2$ (Coefficient of determination) = 0.9 (are much higher)
- ➢ MAE = 2605.12 (are much lower),

COMPARE TO

✓ Using **Support Vector Machine (SVM) model**
- ➢ $R2$ - Coefficient of determination = 0.04 (are much lower)
- ➢ MAE = 7704.22 (are much higher)

by : Kenny Lim/Cohort 3
05 March 2021

Therefore, analysis concludes that Linear Regression model is more accurate/better model for predicting the used car prices/value

THANK YOU

by : Kenny Lim/Cohort 3
05 March 2021

Capstone Project 4

Q & A

by : Kenny Lim/Cohort 3
05 March 2021

Capstone Project 4