Engine Size as a predictor in used cars in the UK

Report based on CRISP-DM

Libraries used:

Pandas, Numpy, Seaborn, Matplotlib, Sklearn,

Pyplot, Scikit learn.

DAB-FT-D921-004

Mariana Yokoyama

2021299

Gabriela Fernandez Perez

2021305

Nuno Alfredo Ribeiro Teixeira de Almeida

2021310

SUMMARY

- Business Understanding
- Data Understanding
- Data Preparation
- Data Visualization
- Modelling
- Run Models
- Evaluation

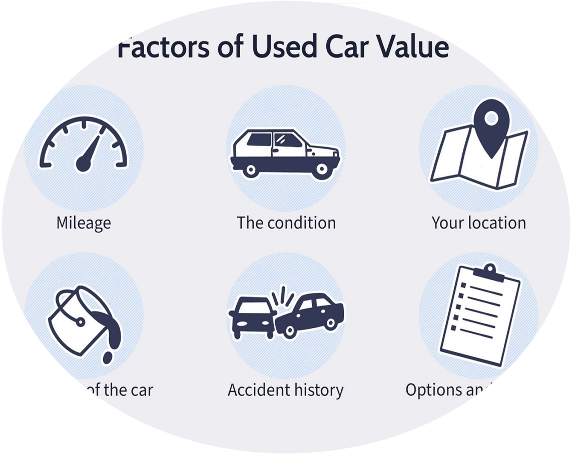
1

Business Understanding

What is the company's requirement?

What is the impact of the engine size on the price of used cars in the UK?

Business Understanding



2

Data Understanding

Data created by Aditya, collected from Kaggle as csv. Format and downloaded in Jupyter notebook.

17, 965
Observations

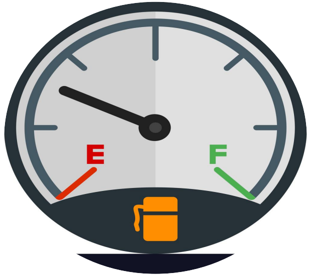
9
Variables

Variables

VS
1 3 5
2 4 6 R
P
R
N
D

60
80
100
120
140
160
180
200
km
160648
shutterstock.com · 369991682

GB
UK PL8TE
UK PL8TE


E
F

£50
£10
£5

Petrol
Petrol
Diesel

02 9013984 5
28 02 15
HN05 PPP
Y
£ 220.00
FIAT
PC
12
DVLA/EVL
18/02/14

FINDINGS

51

“Engine Size” values are Zero values

86.37%

Manual Cars

154

Duplicated entries

43.22%

Engine Size 1

0

Missing Values

67.79%

Petrol

4,039

Outliers

Most Popular Values (Mode)

Price: 10,000£

Model: Fiesta

Registration Year: 2017

Cars Deliver: 65.7mpg

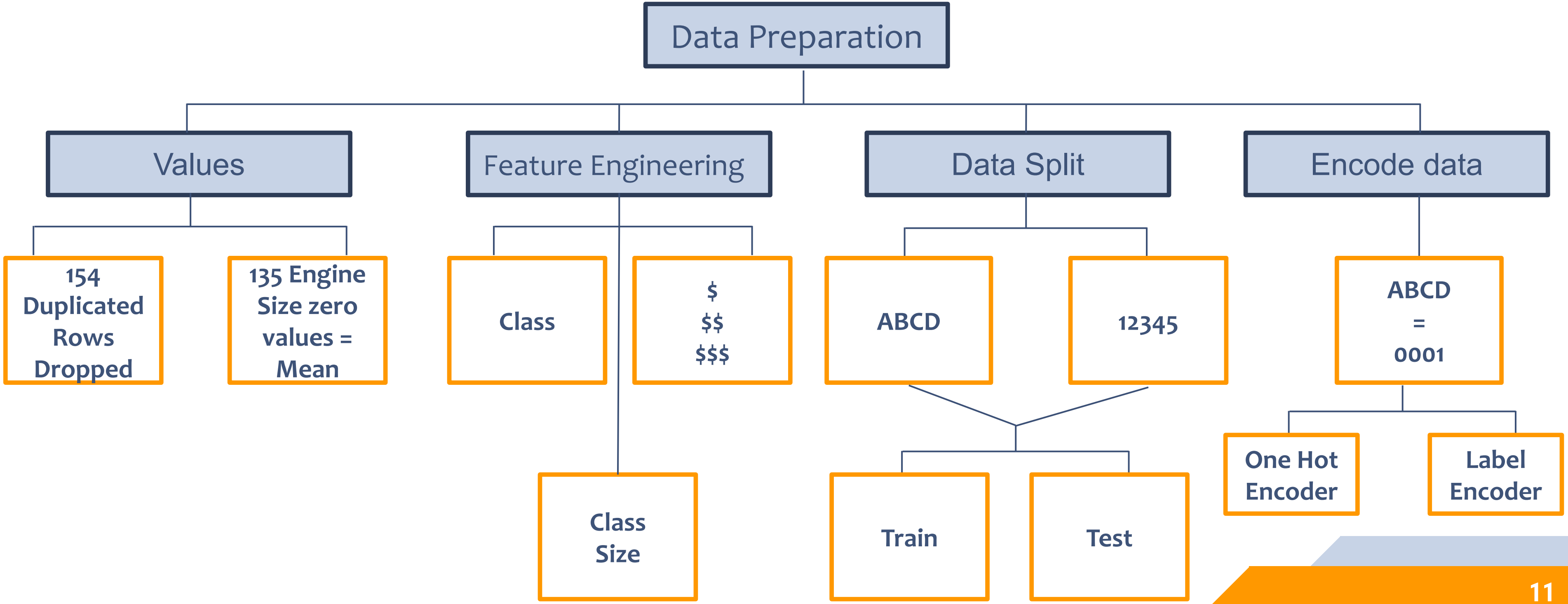
Mileage 10

3

Data Preparation

Preparing the data set for modelling

Data Preparation


Data Preparation

Values

154 Duplicated Rows Dropped

135 Engine Size zero values = Mean

Feature Engineering

Class

\$
\$\$
\$\$\$

Class Size

Data Split

ABCD

12345

Train

Test

Encode data

ABCD
=
0001

One Hot Encoder

Label Encoder

4

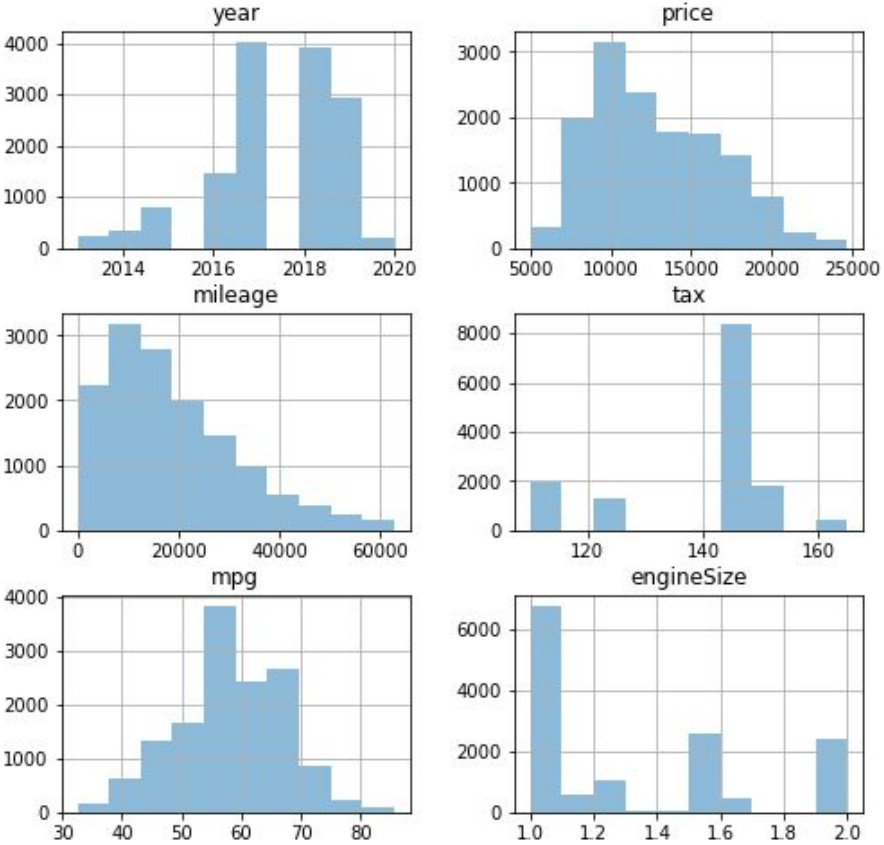
Data Visualization

Findings by visualizations

Descriptive Statistics

Quantitative Variables:

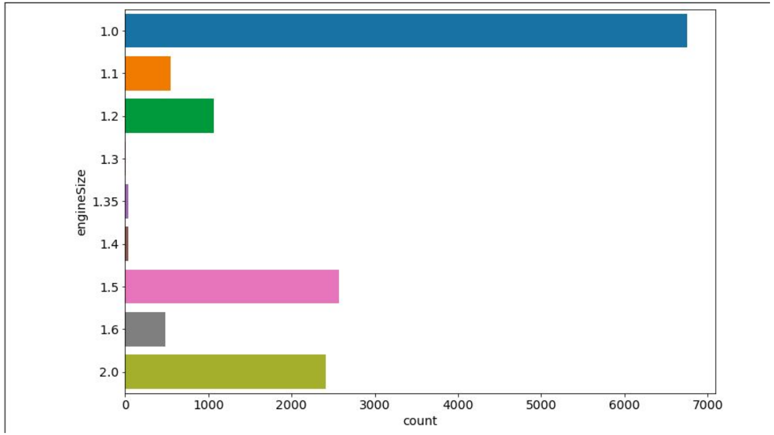
- Price: closest value of £10.000.
- Mileage: right skewed distribution with most of the values clustered from 0 to £25.000.
- Year left tail distribution and clustered between 2016 and 2020
- Engine size, 1.0 is the most frequent followed by 1.5 and 2.0.
- Tax values are clustered between the values 140 and 160

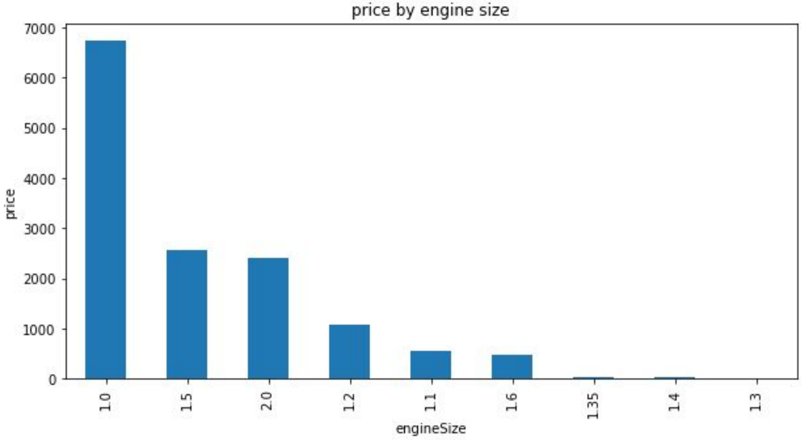


Qualitative Variables:

- Average highest price is for cars with Engine Size 1.

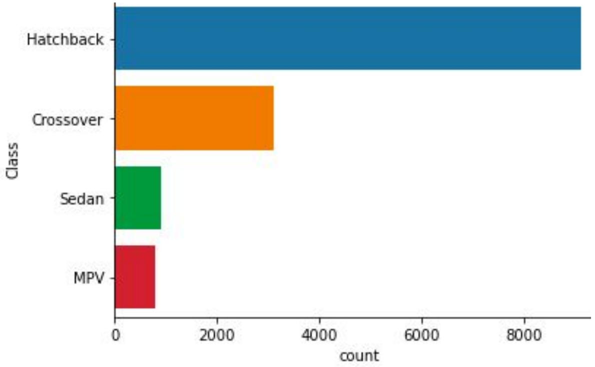
- Cars with Engine Size between 1 and 3 are the most popular.



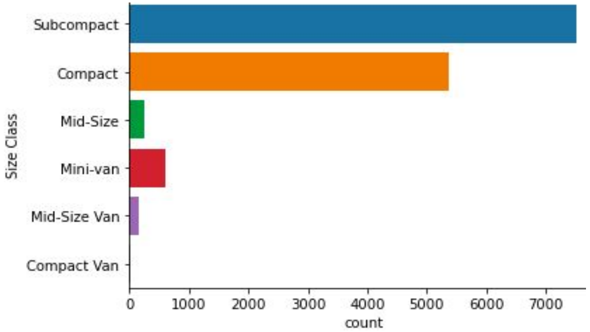


Class sizes:

- Ford has four classes and Hatchback is the most popular one

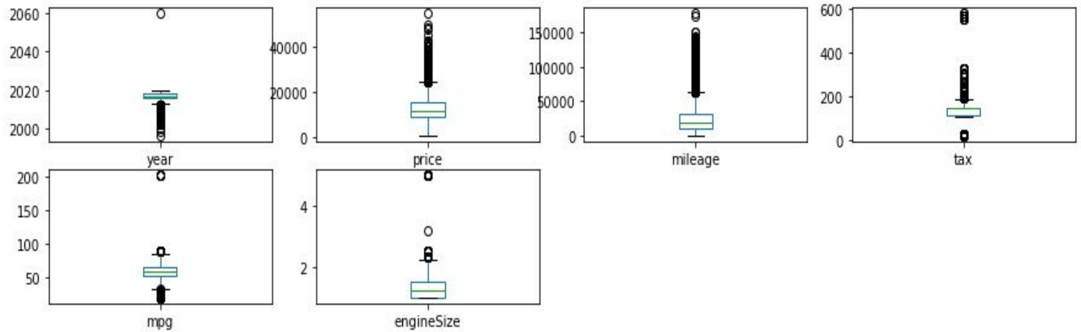


- Each class has sub classes like subcompact, Compact and Mini-van.



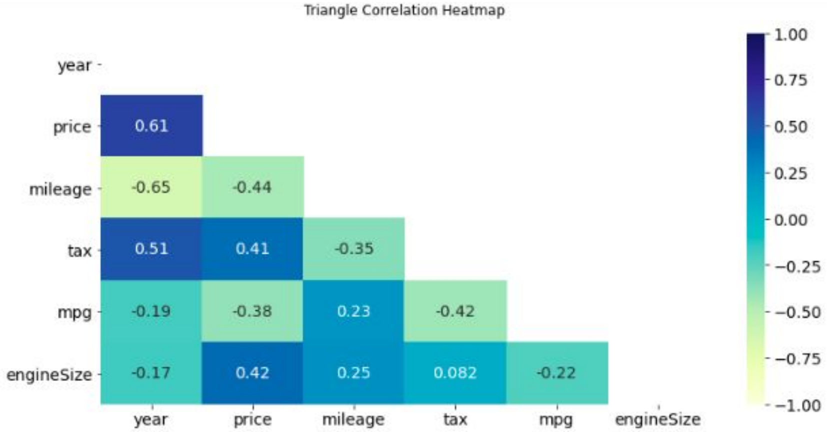
Outliers:

- 4,039 outliers removed



Heatmap:

- Price and Year are the most correlated variables
- Lowest relationship between Tax and Engine Size
- Engine size and Price have a moderate correlation.

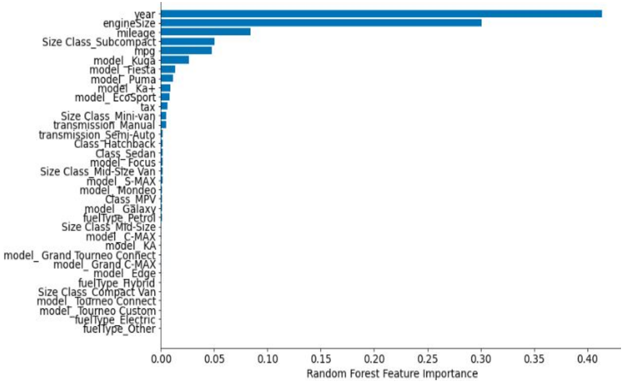


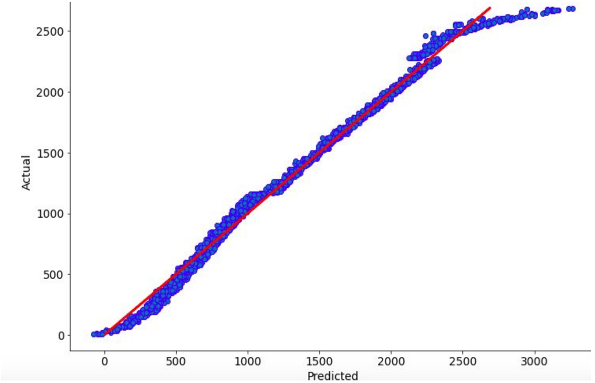
5

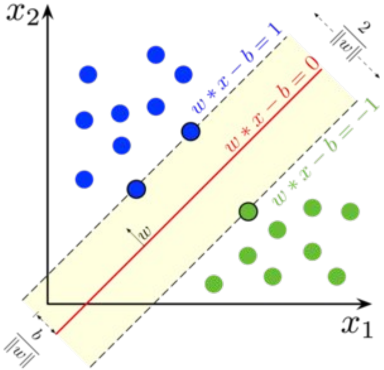
Modelling

Machine Learning Models explanation

MODELS







Random Forest Feature Importance

Multilinear Regression

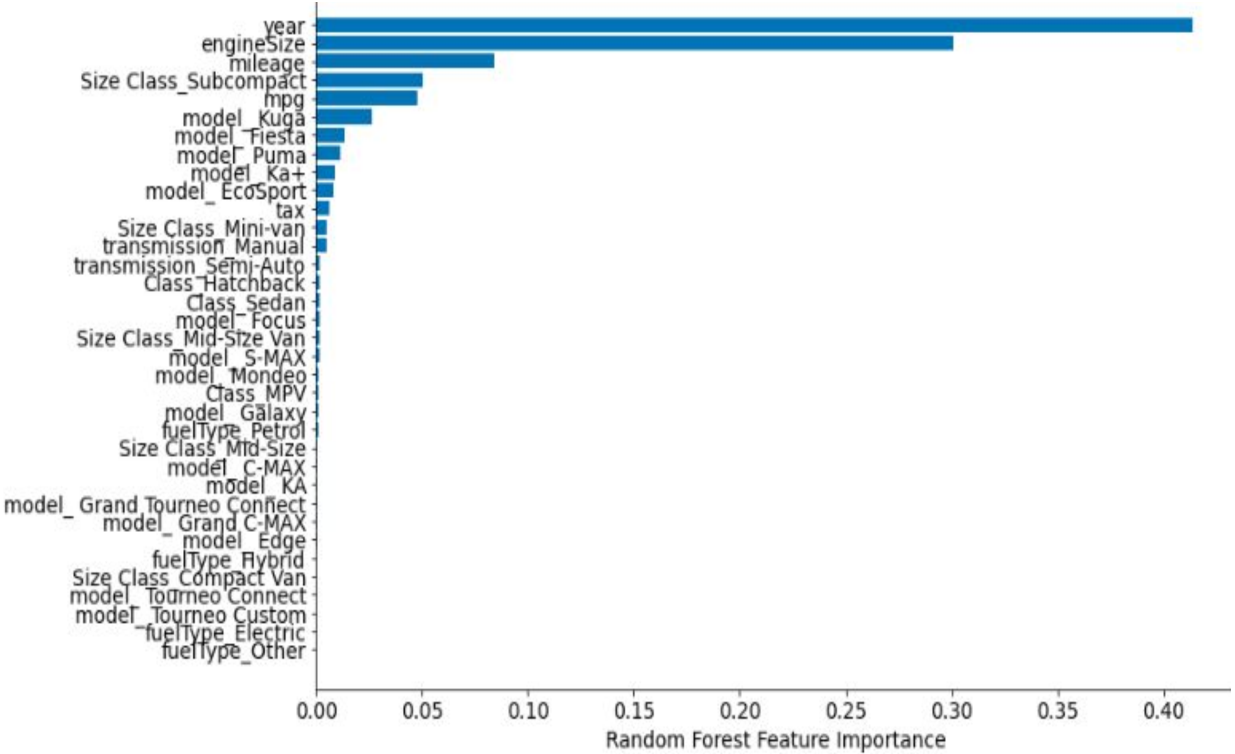
Support Vector Regression

6

Models Evaluation

Evaluation Metrics and Accuracy

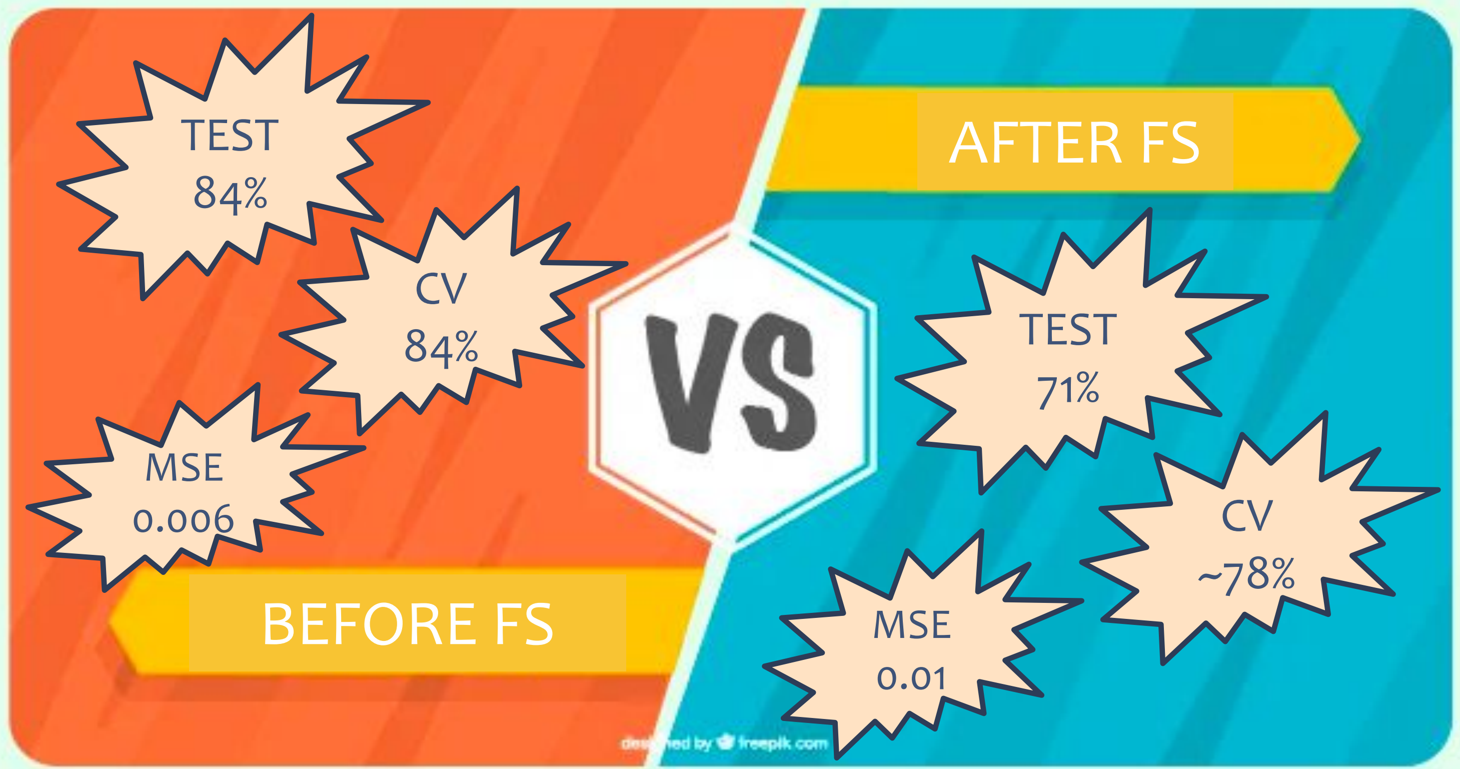
Random Forest Feature Selection



Selected Features:

- Year
- Engine Size
- Mileage
- Subcompact
- MPG
- Other car models

Linear Regression



BEFORE FS

TEST
84%

CV
84%

MSE
0.006

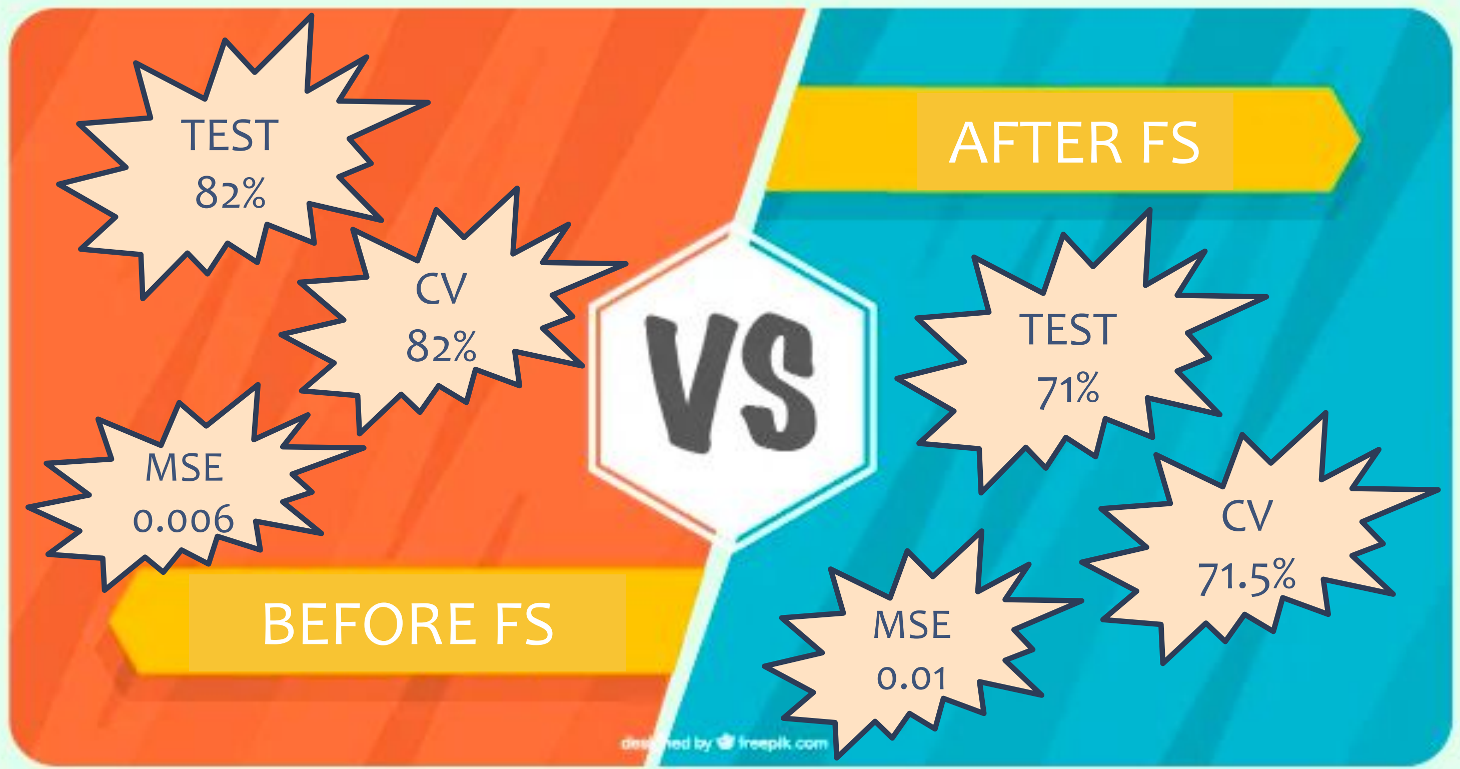
AFTER FS

TEST
71%

CV
~78%

MSE
0.01

Supported Vector Regressor



BEFORE FS

TEST
82%

CV
82%

MSE
0.006

AFTER FS

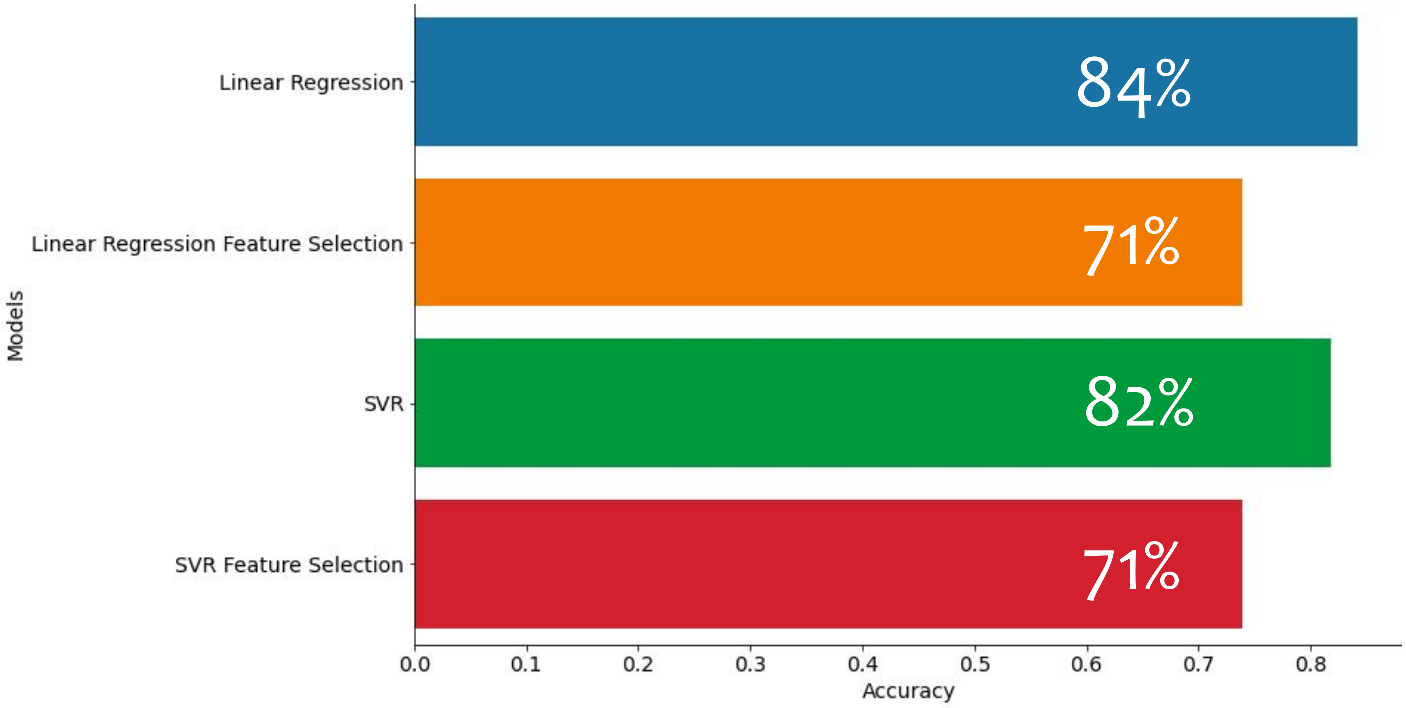
TEST
71%

CV
71.5%

MSE
0.01

des…ned by freepik.com

Models Accuracy



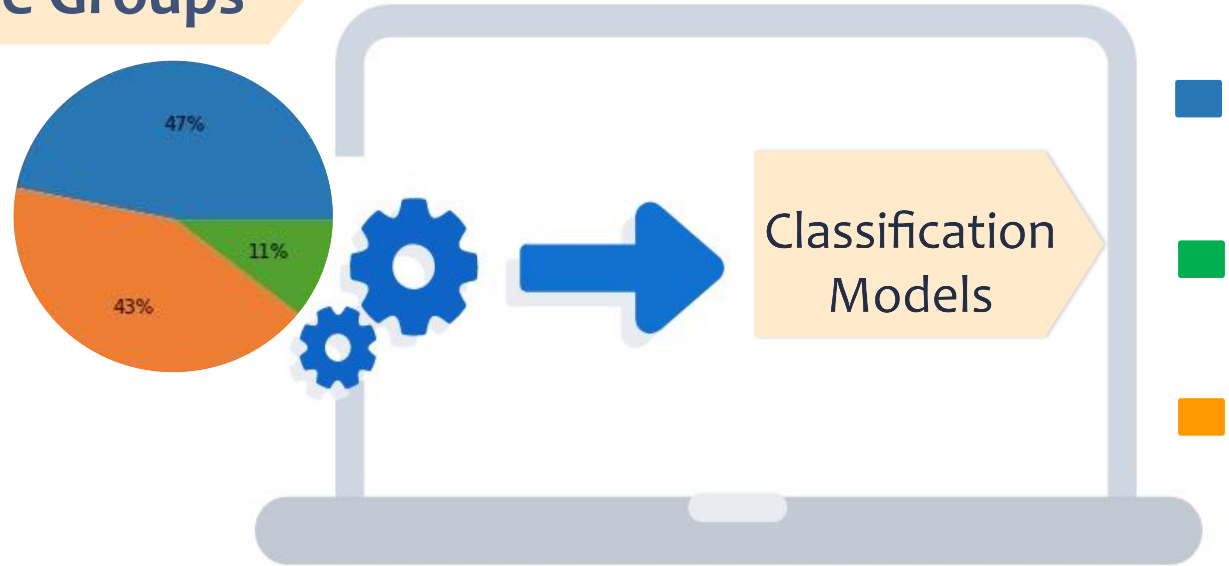
7

Deployment or Next Steps

Machine Learning Models explanation

Next Steps

Price Groups



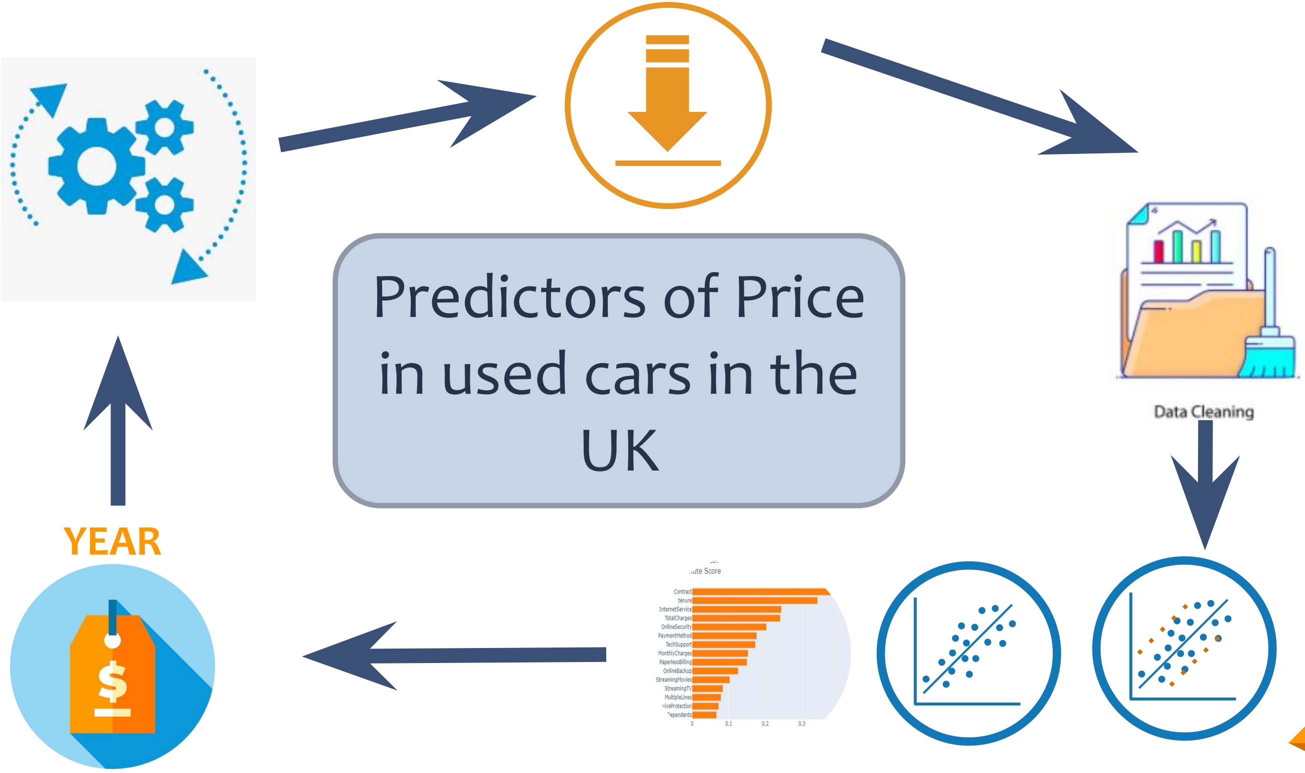
Classification Models

🟦 Low Price

🟩 Medium Price

🟧 High Price

Conclusions



Predictors of Price in used cars in the UK

- Aditya (2020) *100,000 UK Used Car Data set*. Available at: https://kaggle.com/adityadesai13/used-car-dataset-ford-and-mercedes (Accessed: 19 November 2021).

- Amat, R. (2017) *Máquinas de Vector Soporte (Support Vector Machines, SVMs)*. Available at: https://www.cienciadedatos.net/documentos/34_maquinas_de_vector_soporte_support_vector_machines (Accessed: 19 November 2021).

- APA, D. of P. (2020) *garbage in, garbage out – APA Dictionary of Psychology*. Available at: https://dictionary.apa.org/garbage-in-garbage-out (Accessed: 6 December 2021).

- APD, R. (2019) '¿Cuáles son los tipos de algoritmos del machine learning?', *APD España*, 4 April. Available at: https://www.apd.es/algoritmos-del-machine-learning/ (Accessed: 18 November 2021).

- Brownlee, J. (2017) 'Difference Between Classification and Regression in Machine Learning', *Machine Learning Mastery*, 10 December. Available at: https://machinelearningmastery.com/classification-versus-regression-in-machine-learning/ (Accessed: 19 November 2021).

- Iqbal, M. (2019) 'Gaussian Naïve Bayes and Support Vector Machine.'

- Jones, T. (2018) ‘Supervised learning models’, *IBM Developer*, 26 February. Available at: https://developer.ibm.com/articles/cc-supervised-learning-models/ (Accessed: 18 November 2021).

- Lopez, F. (2018) *Teorema de Bayes - Definición, qué es y concepto*, *Economipedia*. Available at: https://economipedia.com/definiciones/teorema-de-bayes.html (Accessed: 18 November 2021).

- Majumder, P. (2020) *Gaussian Naive Bayes*, *OpenGenus IQ: Computing Expertise & Legacy*. Available at: https://iq.opengenus.org/gaussian-naive-bayes/ (Accessed: 19 November 2021).

- Oxford (no date) *garbage in garbage out*, *Oxford Reference*. doi:10.1093/oi/authority.20110803095842747.

- Rombauts, W. (2021) ‘List of Ford vehicles’, *Wikipedia*. Available at: https://en.wikipedia.org/w/index.php?title=List_of_Ford_vehicles&oldid=1057606756 (Accessed: 30 November 2021).

- Sethi, A. (2020) ‘Support Vector Regression In Machine Learning’, *Analytics Vidhya*, 27 March. Available at: https://www.analyticsvidhya.com/blog/2020/03/support-vector-regression-tutorial-for-machine-learning/ (Accessed: 6 December 2021).

THANKS !

¡GRACIAS!

OBRIGADO!

Any questions?