Leveraging Data to Gain Competitive Edge in the Automotive Industry: A case study

Identifying critical features that influence car prices

Samuel Aduroja, August 2023

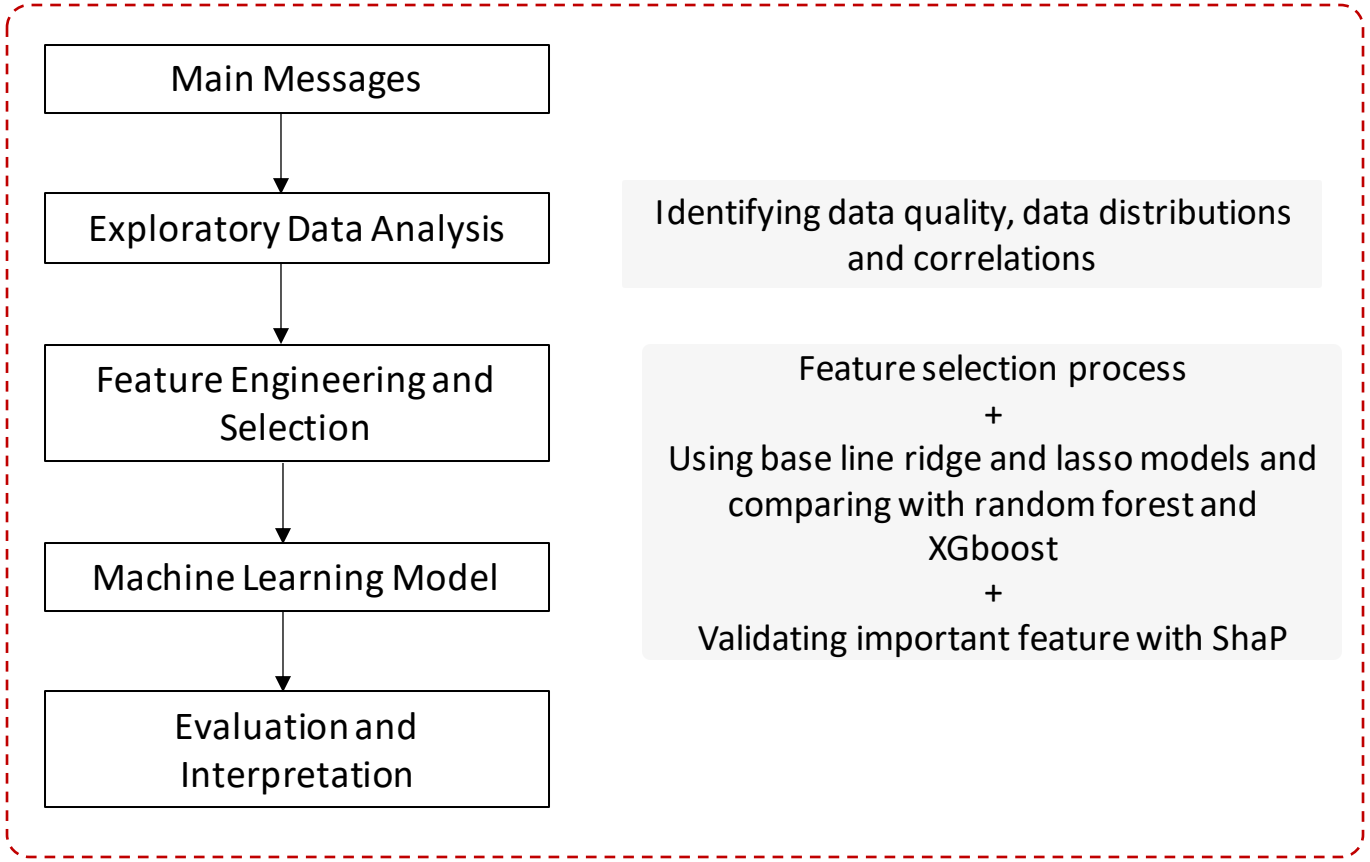
BOSCH

Main Messages

- By understanding important features in a model, organizations and practitioners can make more informed decisions and gain competitive edge.

Agenda

1. Main Messages
2. Exploratory Data Analysis (EDA)
3. Feature Engineering and Selection
4. Machine Learning Model
5. Evaluation and Interpretation



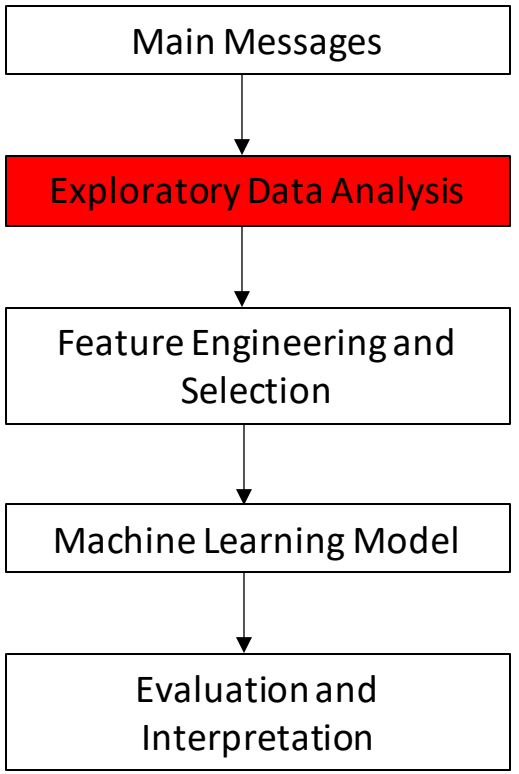
BOSCH

Exploratory Data Analysis – Dataset Description

- Given two files named 'imports-85.names' and 'imports-85.data:
  - imports-85.names:
    - Describes:
      - The dataset and its features(205 observations, 26 features – 15 continuous, 1 integer, 10 nominal)
      - Characteristics and range of possible values the features can take
      - Missing values information(norminal-losses has the highest missing values instances, 41)

Feature	Number of Missing Values
normalized-losses	41
num-of-doors	2
bore	4
stroke	4
horsepower	2
peak-rpm	2
price	4

  - Imports-85.data:
    - Provides information about:
      - Cars physical and mechanical properties, make, fuel type and body style, and market price
      - Insurance – related data such as risk rating and normalized losses



BOSCH

Exploratory Data Analysis – Dataset Description

Numerical Features

15 Continuous value features

Engine related features
- engine-size
- bore
- stroke
- peak-rpm
- compression-ratio
- city-mpg
- highway-mpg

Physical/mechanical features
- wheel-base
- width
- length
- height
- weight

Insurance related feature
- normalized-losses

Target
- price

1 integer value feature

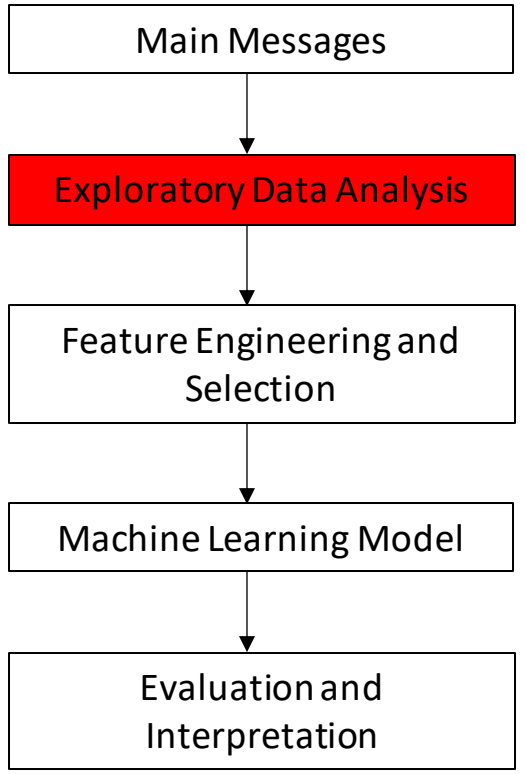
Insurance related feature
- symboling

Categorical Features

10 Nominal value features

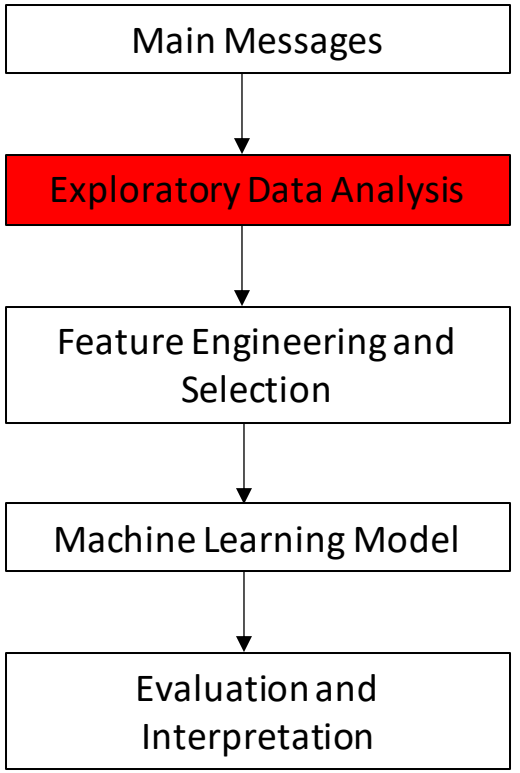
Engine related features
- fuel-type
- engine-type
- num-of-cylinders
- fuel-system
- aspiration

Physical/mechanical features
- engine-location
- body-style
- num-of-doors
- drive-wheels
- make

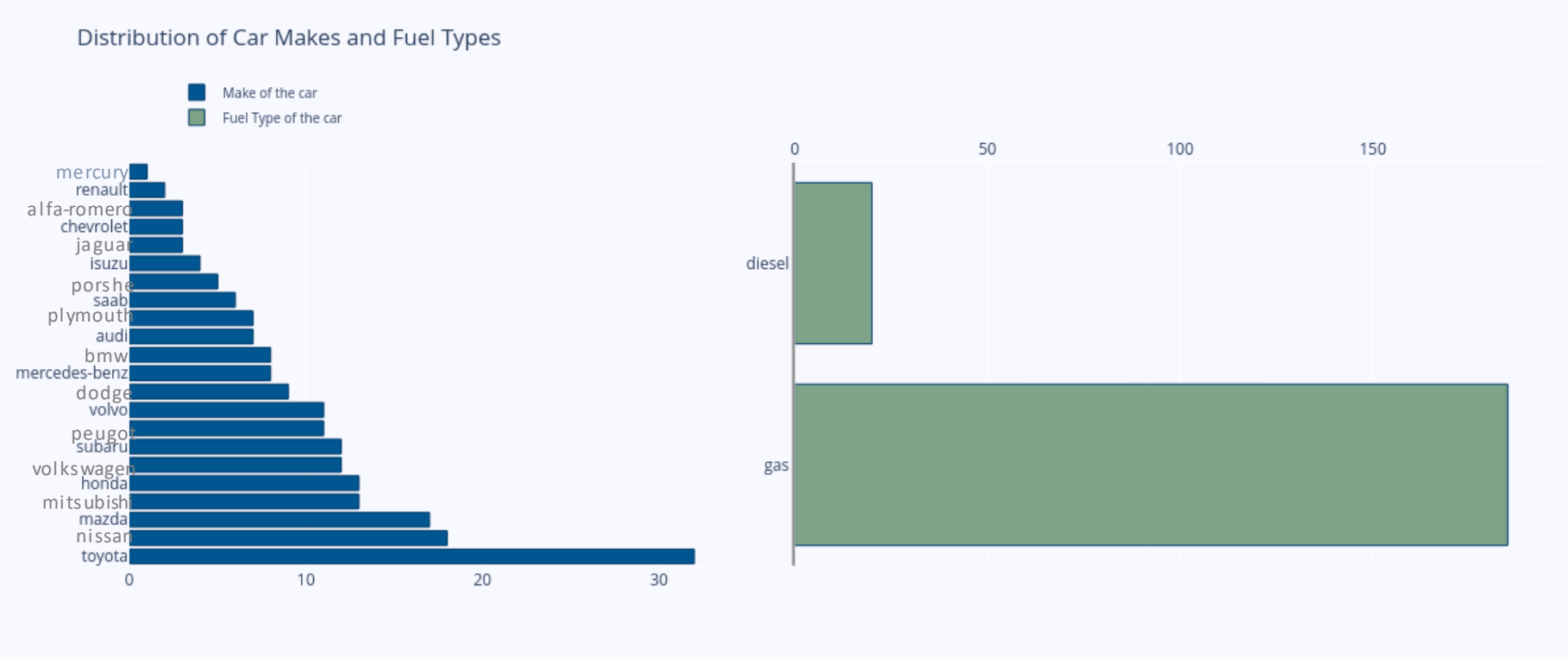


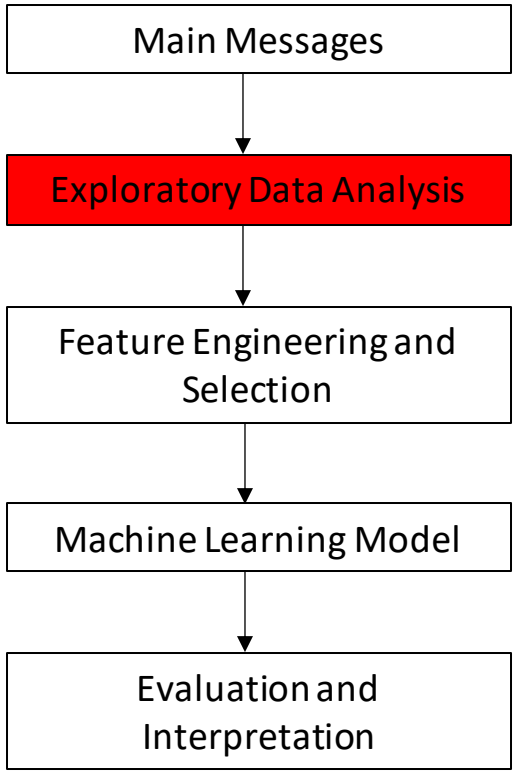
Exploratory Data Analysis – Categorical Data Description

Categorical Features	Unique	Top	Frequency
make	22	toyota	32
fuel-type	2	gas	185
aspiration	2	std	168
num-of-doors	2	four	114
body-style	5	sedan	96
drive-wheels	3	fwd	120
engine-location	2	front	202
engine-type	7	ohc	148
num-of-cylinders	7	four	159
fuel-system	8	mpfi	94



Exploratory Data Analysis – Categorical Data Description

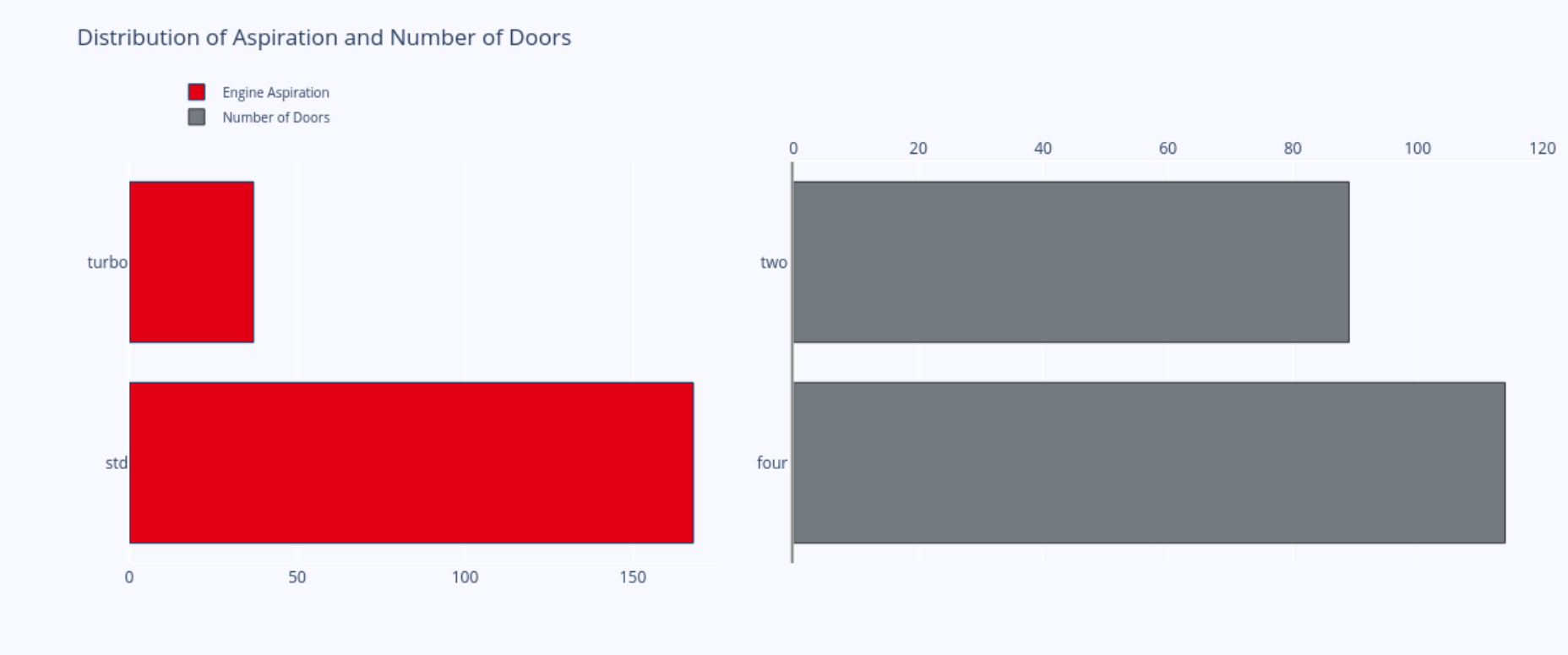




- Key take aways
  - Toyota is the most common make, followed by Nissan, Mazda
  - Gas is the most common fuel type

BOSCH

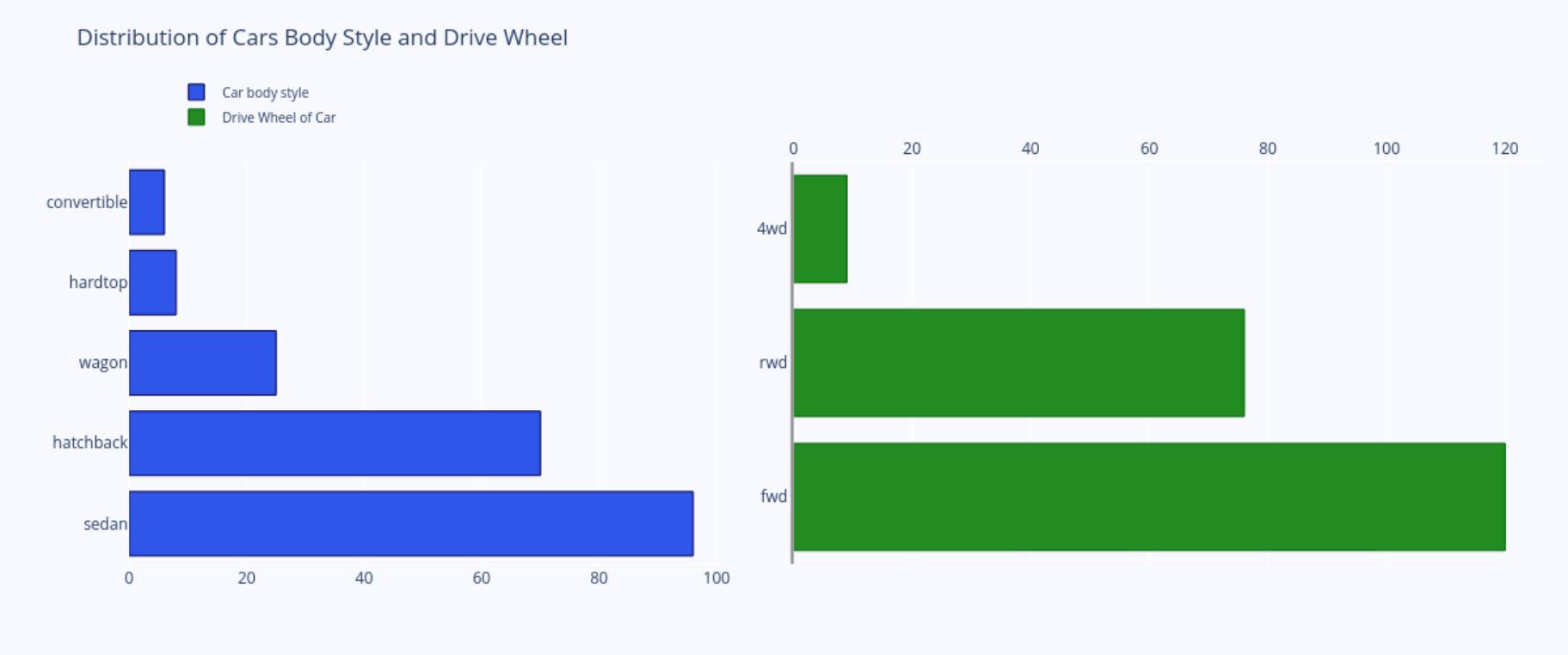
Exploratory Data Analysis – Categorical Data Description

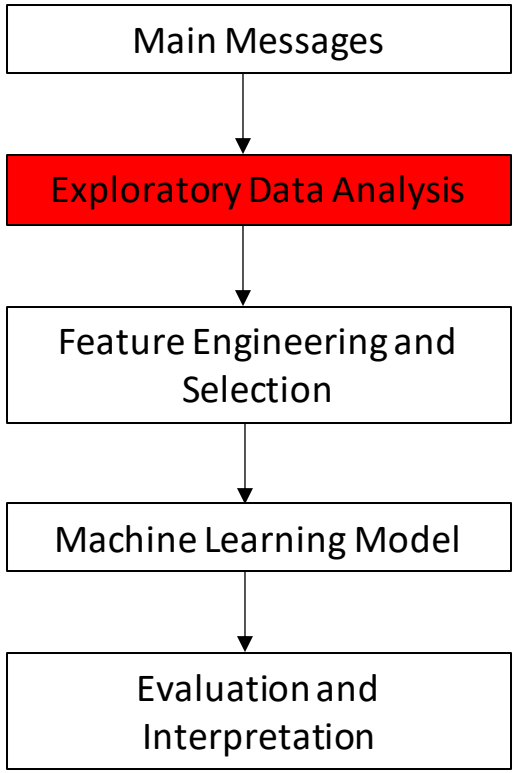


- Main Messages
- **Exploratory Data Analysis**
- Feature Engineering and Selection
- Machine Learning Model
- Evaluation and Interpretation

- Key take aways
  - Standard aspiration is more than turbo
  - Four-door cars are more common than two-door cars

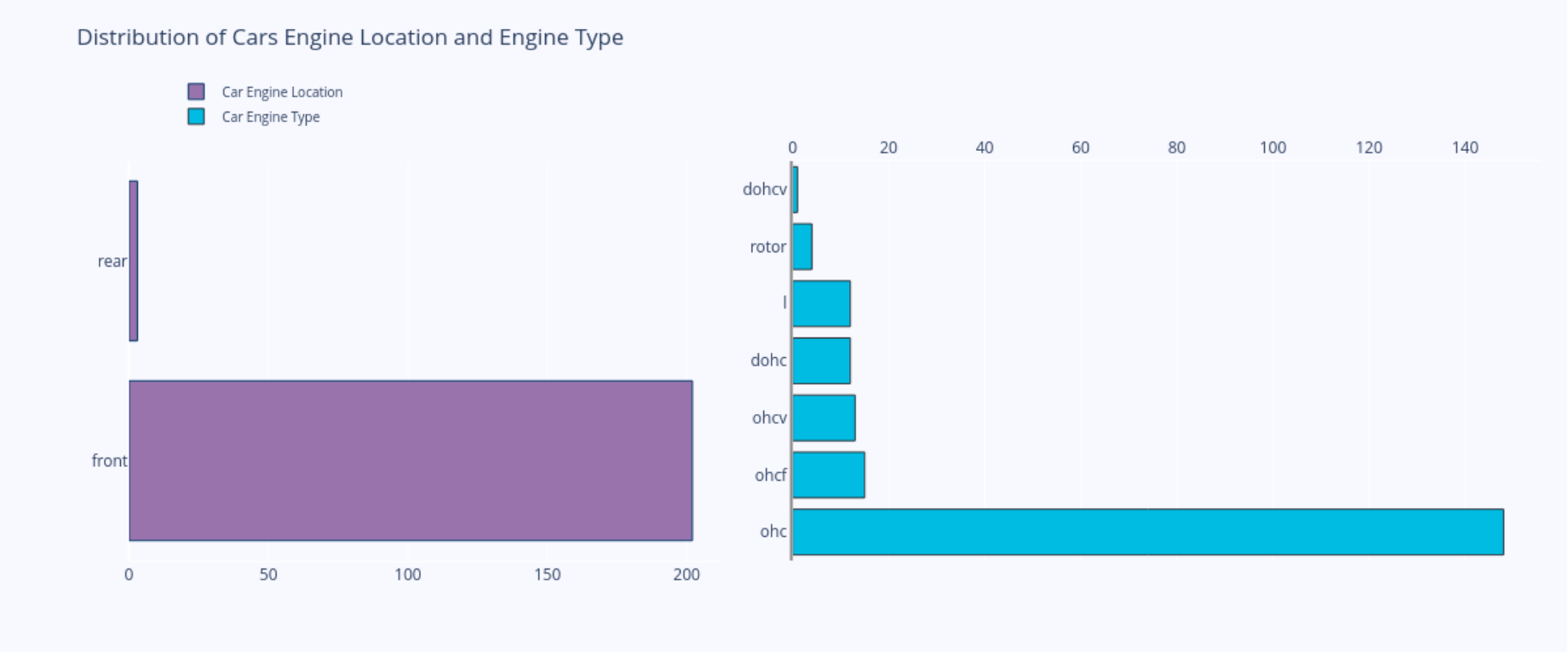
Exploratory Data Analysis – Categorical Data Description





- Key take aways
  - Sedan is the most common body style, followed by hatchback
  - Front-wheel is the most common, followed by rear-wheel drive

Exploratory Data Analysis – Categorical Data Description



Main Messages

Exploratory Data Analysis

Feature Engineering and Selection

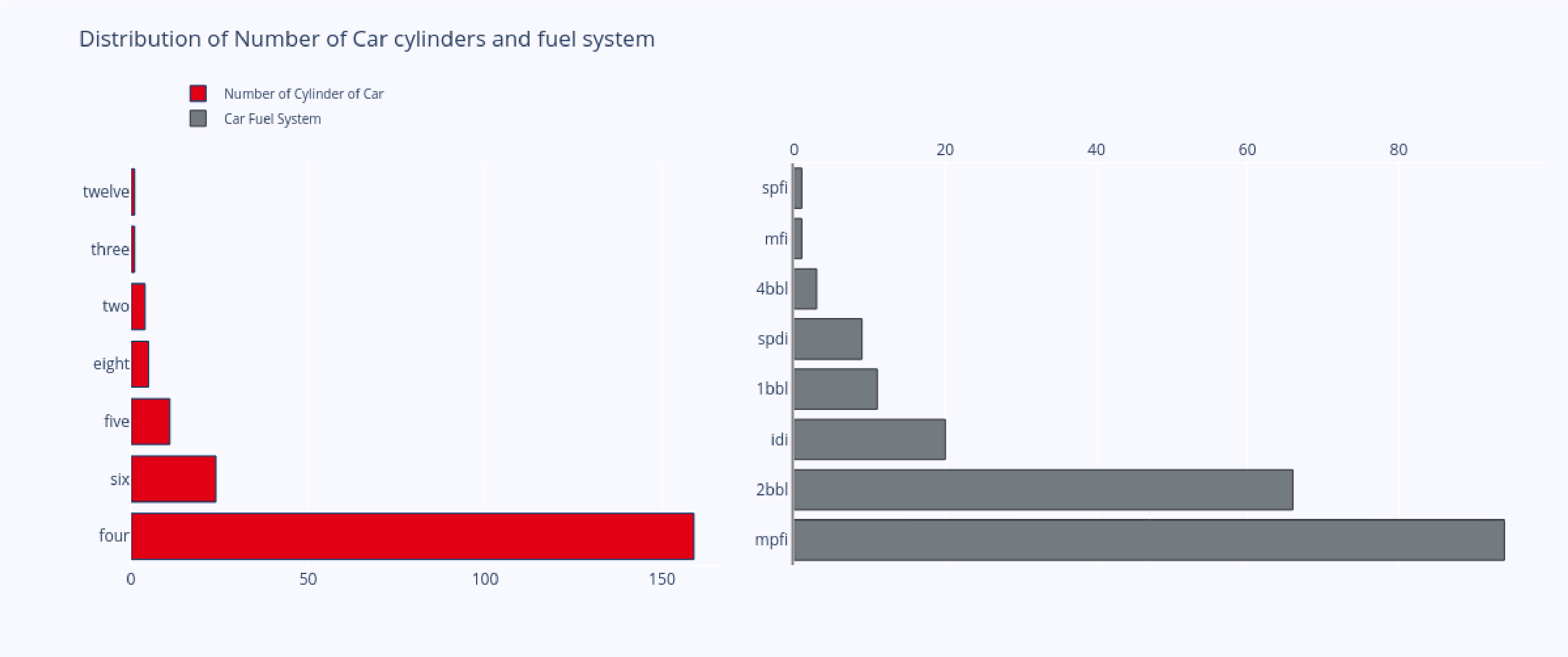
Machine Learning Model

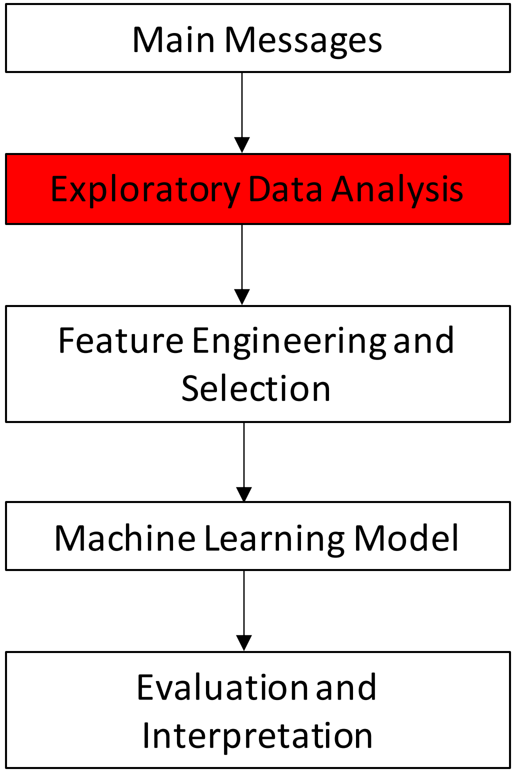
Evaluation and Interpretation

- Key take aways
  - Almost all cars have the engine in the front
  - Overhead cam(ohc) is the most common engine type

BOSCH

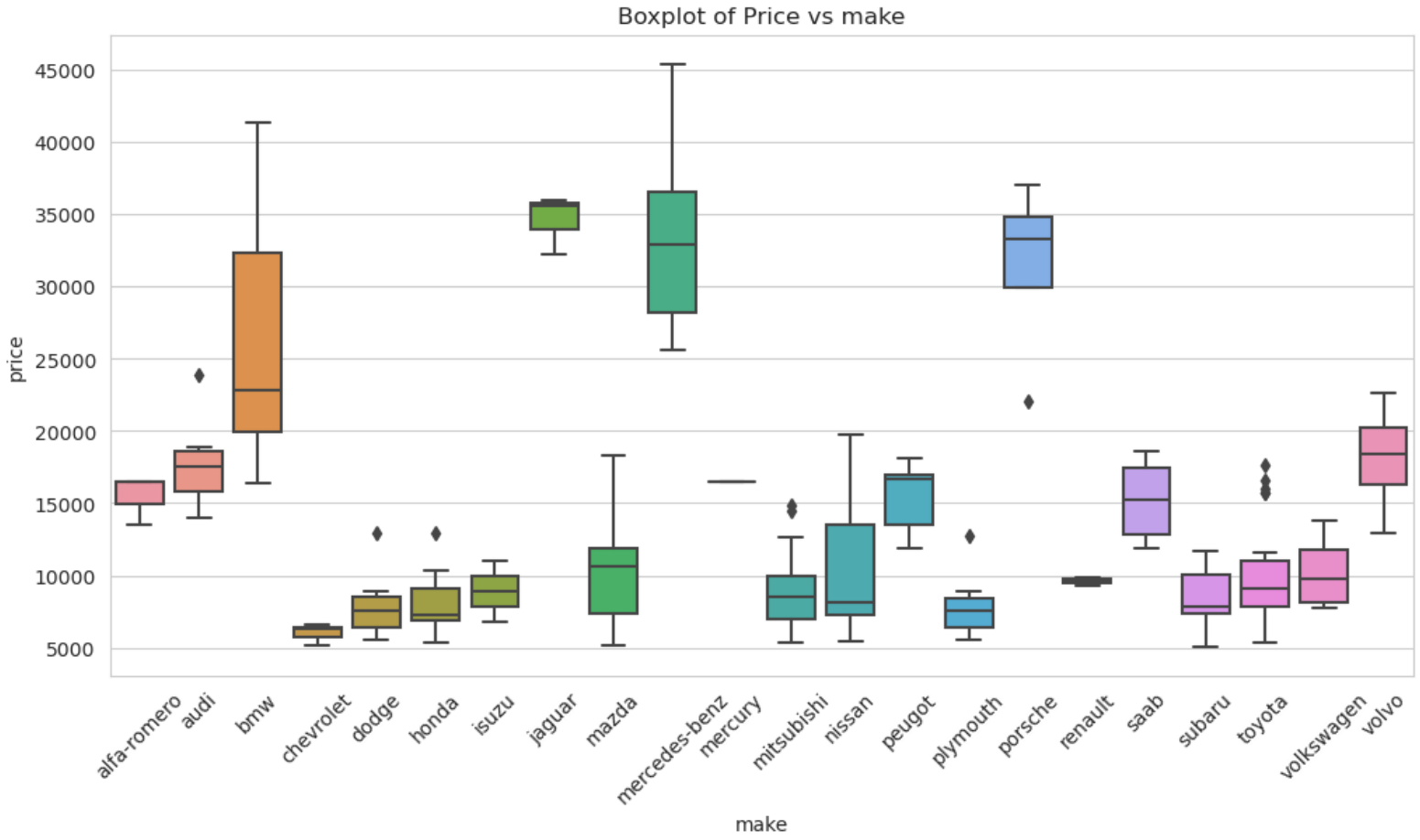
Exploratory Data Analysis – Categorical Data Description





- Key take aways
  - Most cars have four cylinders, followed by six and five
  - Multi-point fuel injection(mpfi) and two-barrel carburetor(2bbl) are the most common fuel systems.

Exploratory Data Analysis – Categorical Data Description

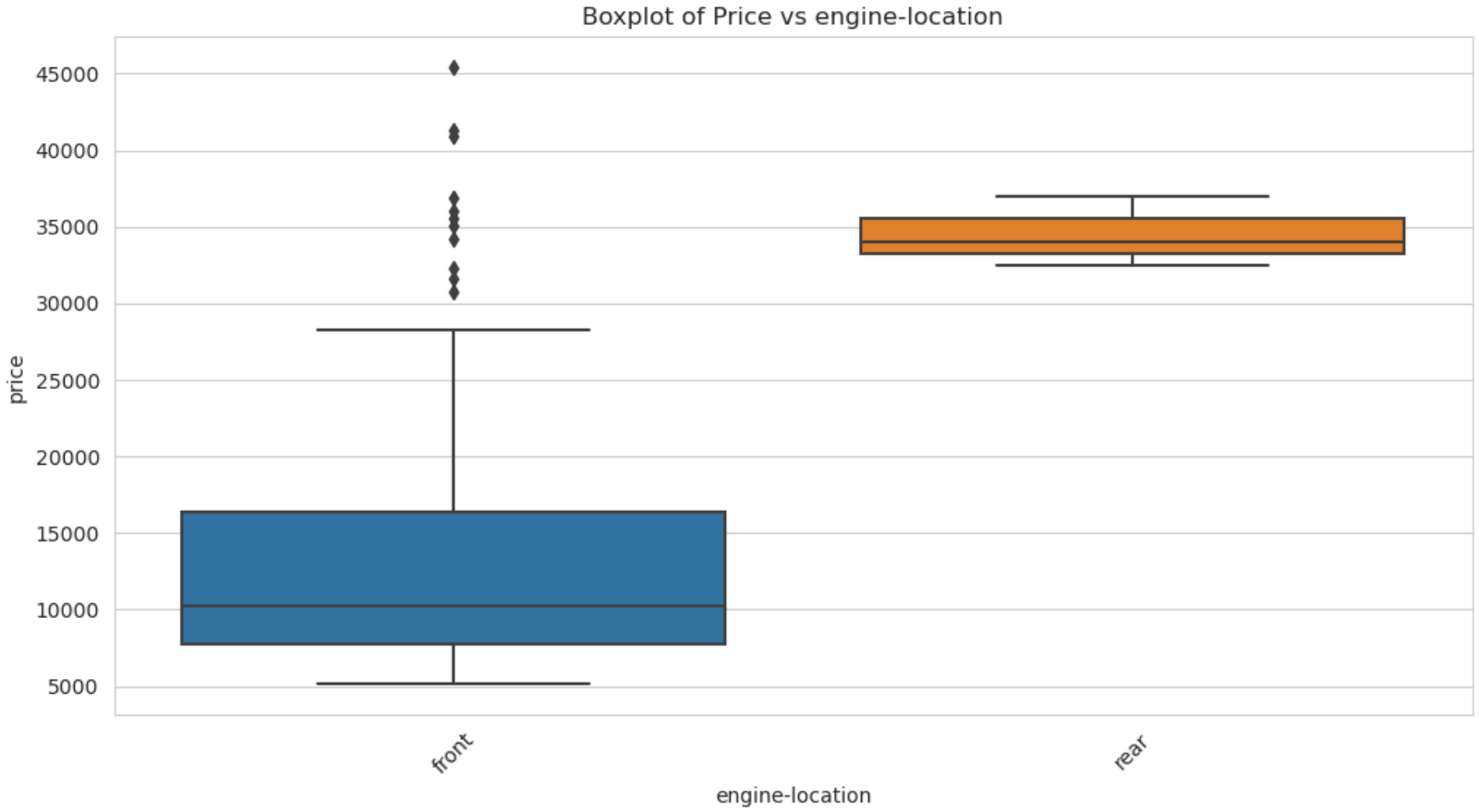


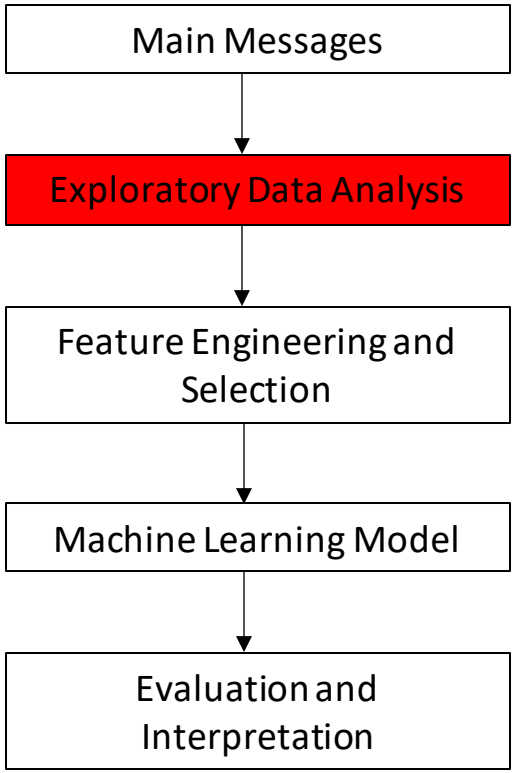
- Main Messages
- Exploratory Data Analysis
- Feature Engineering and Selection
- Machine Learning Model
- Evaluation and Interpretation

- Key take aways
  - Different car manufacturers have various price ranges, with some brands being more luxurious and expensive.

BOSCH

Exploratory Data Analysis – Categorical Data Description

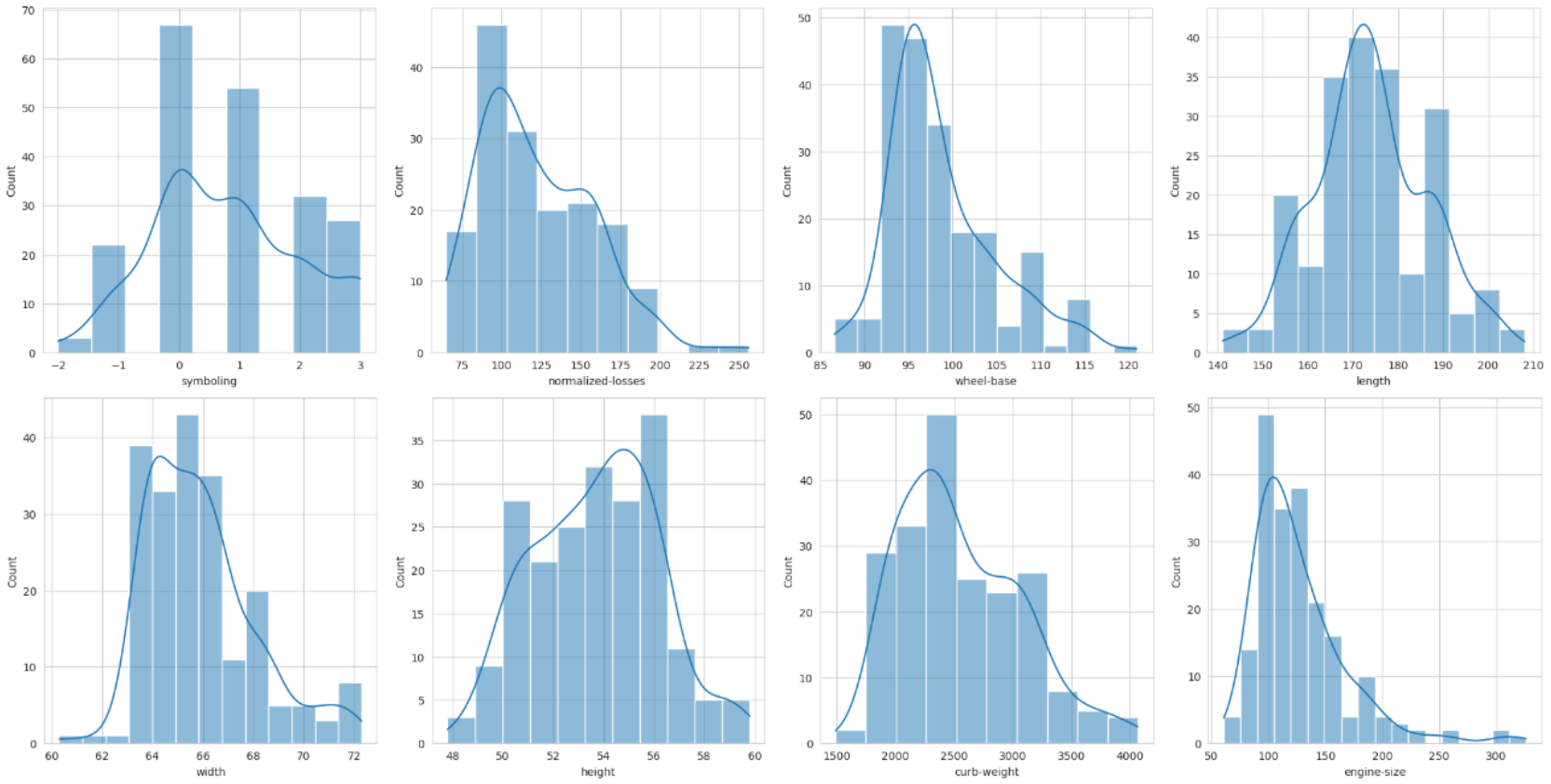


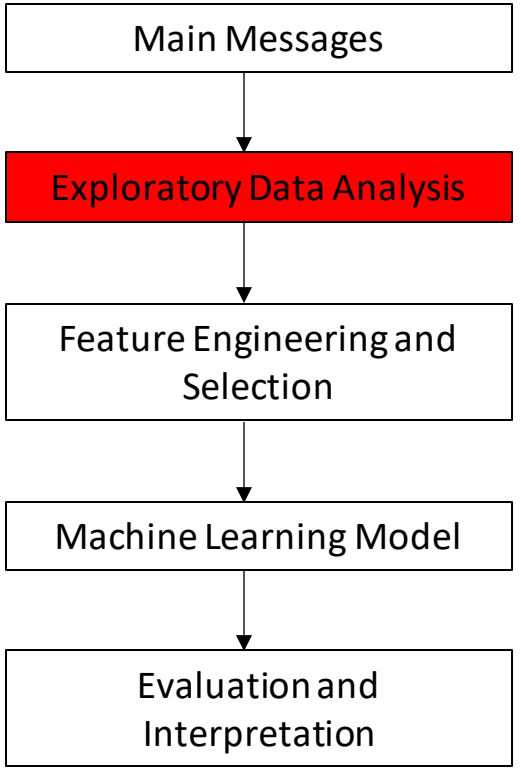


- Key take aways
  - The location of the engine (front or rear) affects the price.

BOSCH

Exploratory Data Analysis – Numerical Data Description

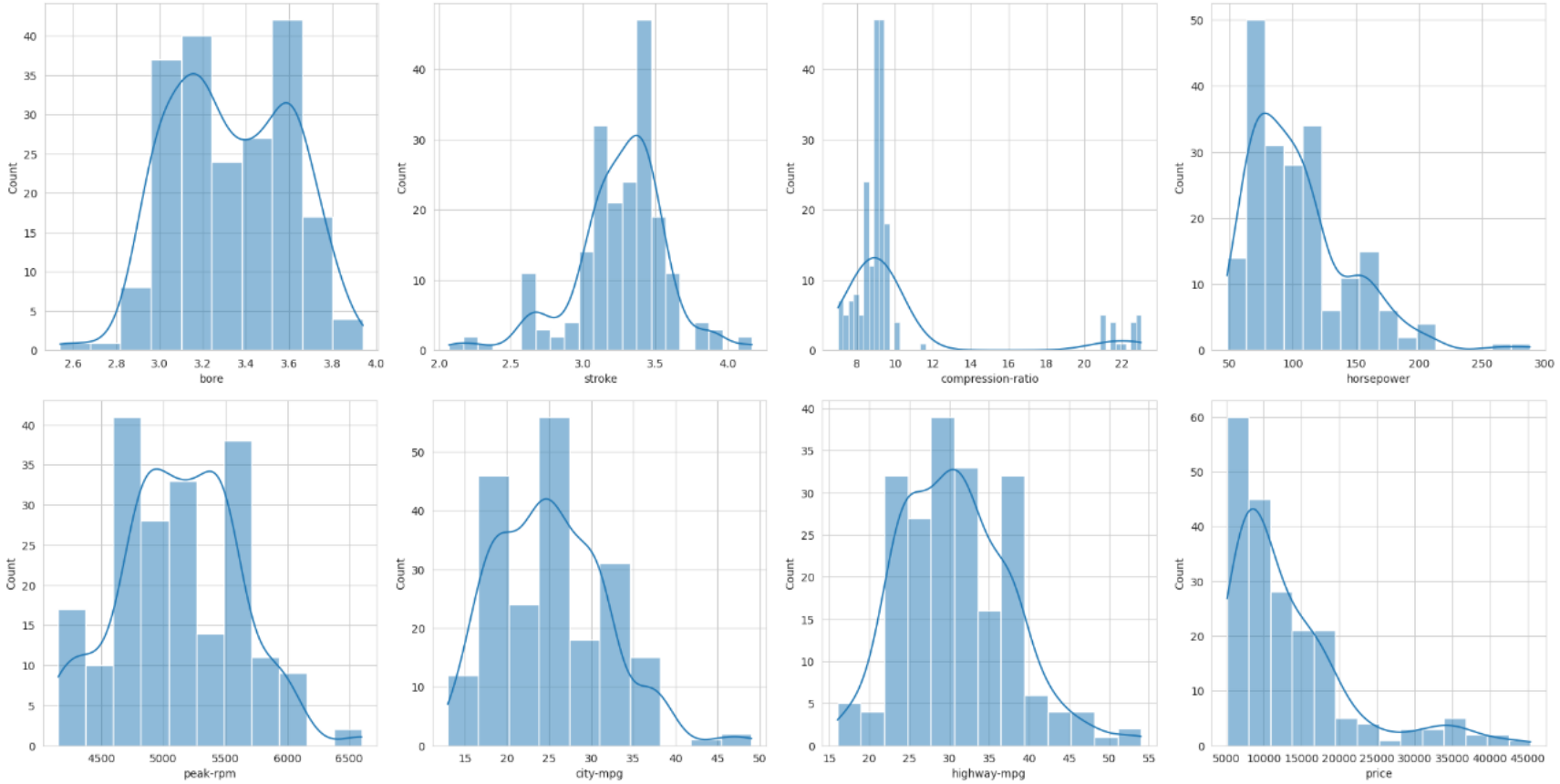


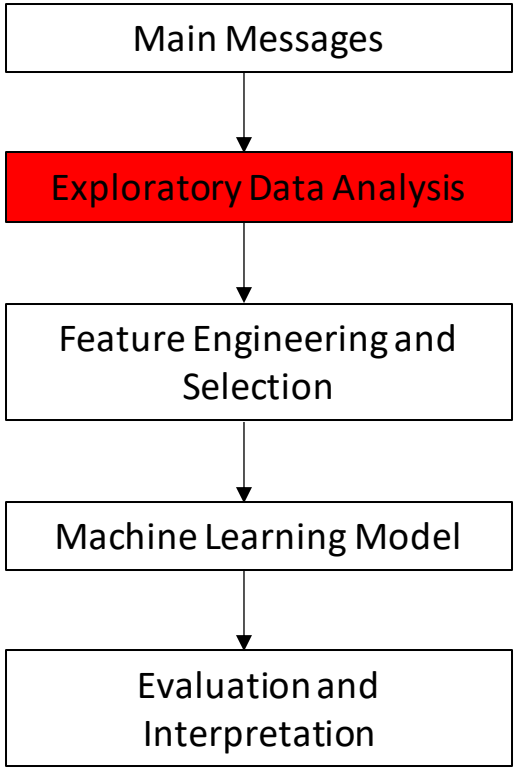


- Key take aways
  - The average insurance risk, 'symboling', rating is around 0.83 showing that the average auto is safe.
  - normalized-losses: The average normalized loss is around 122, but there are missing values in this column.

BOSCH

Exploratory Data Analysis – Numerical Data Description

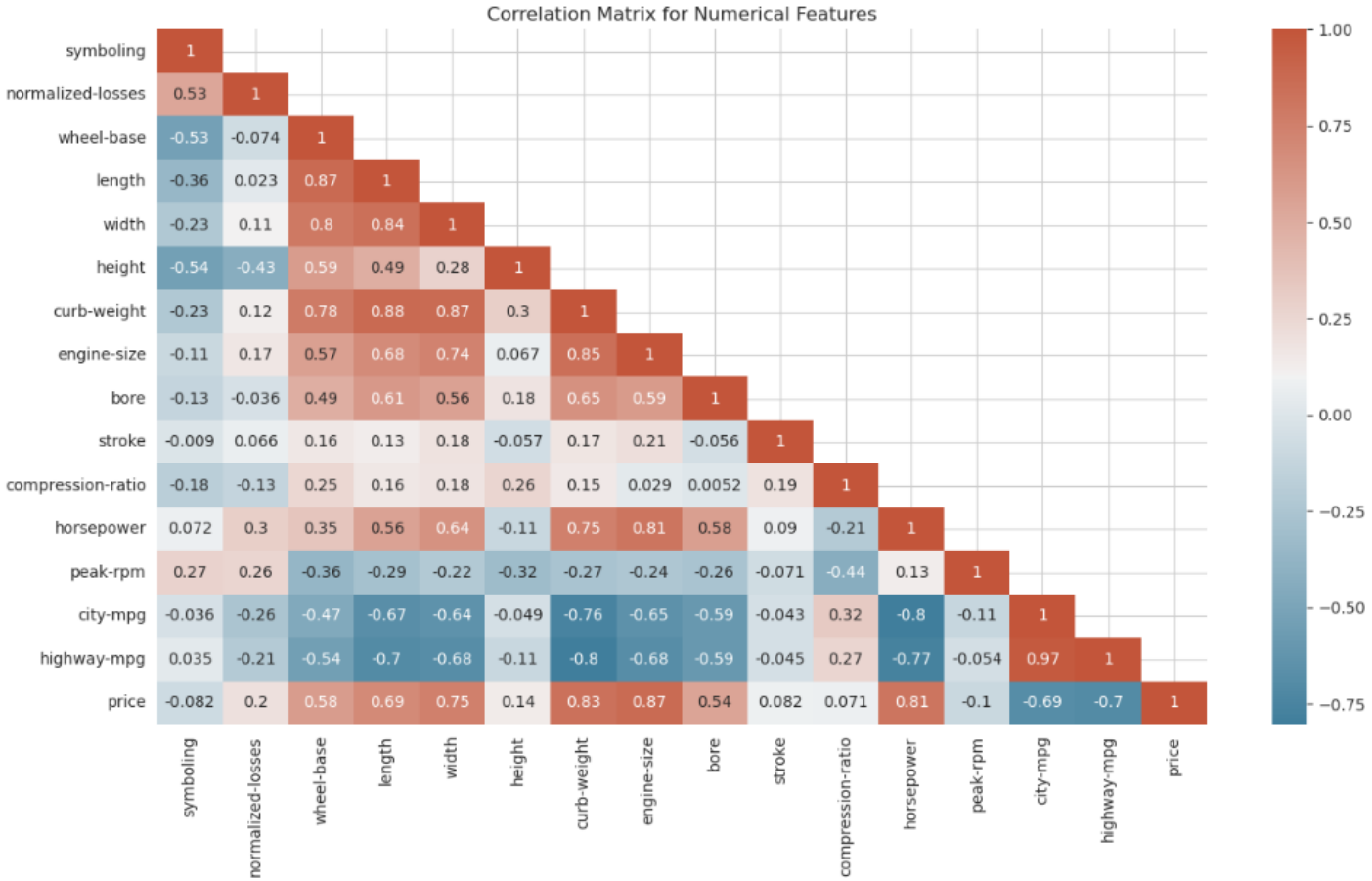


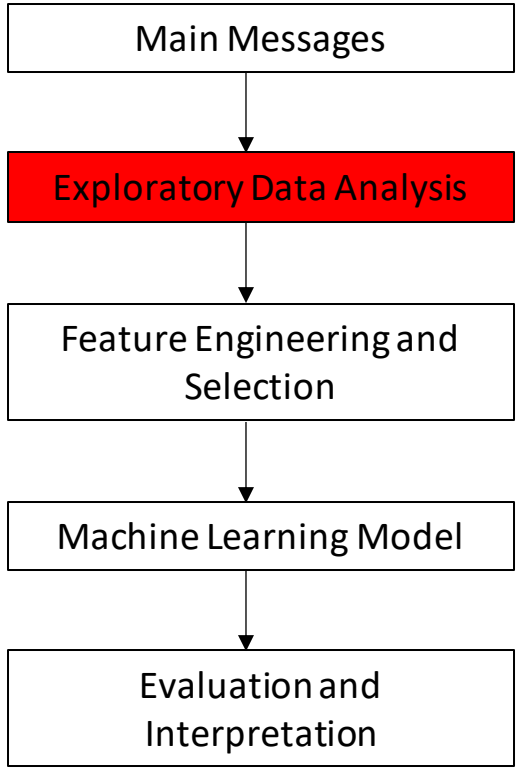


- Key take aways
  - wheel-base, length, width, height, curb-weight, engine-size, bore, stroke, compression-ratio, horsepower, peak-rpm, city-mpg, highway-mpg, and price: These numerical attributes vary significantly in their means, standard deviations, and ranges, reflecting the variety of car types in the dataset.
  - The price is highly skewed to the right, indicating that most cars are in the lower to mid-price range, with a few high-priced outliers.

BOSCH

Exploratory Data Analysis – Numerical Data Description

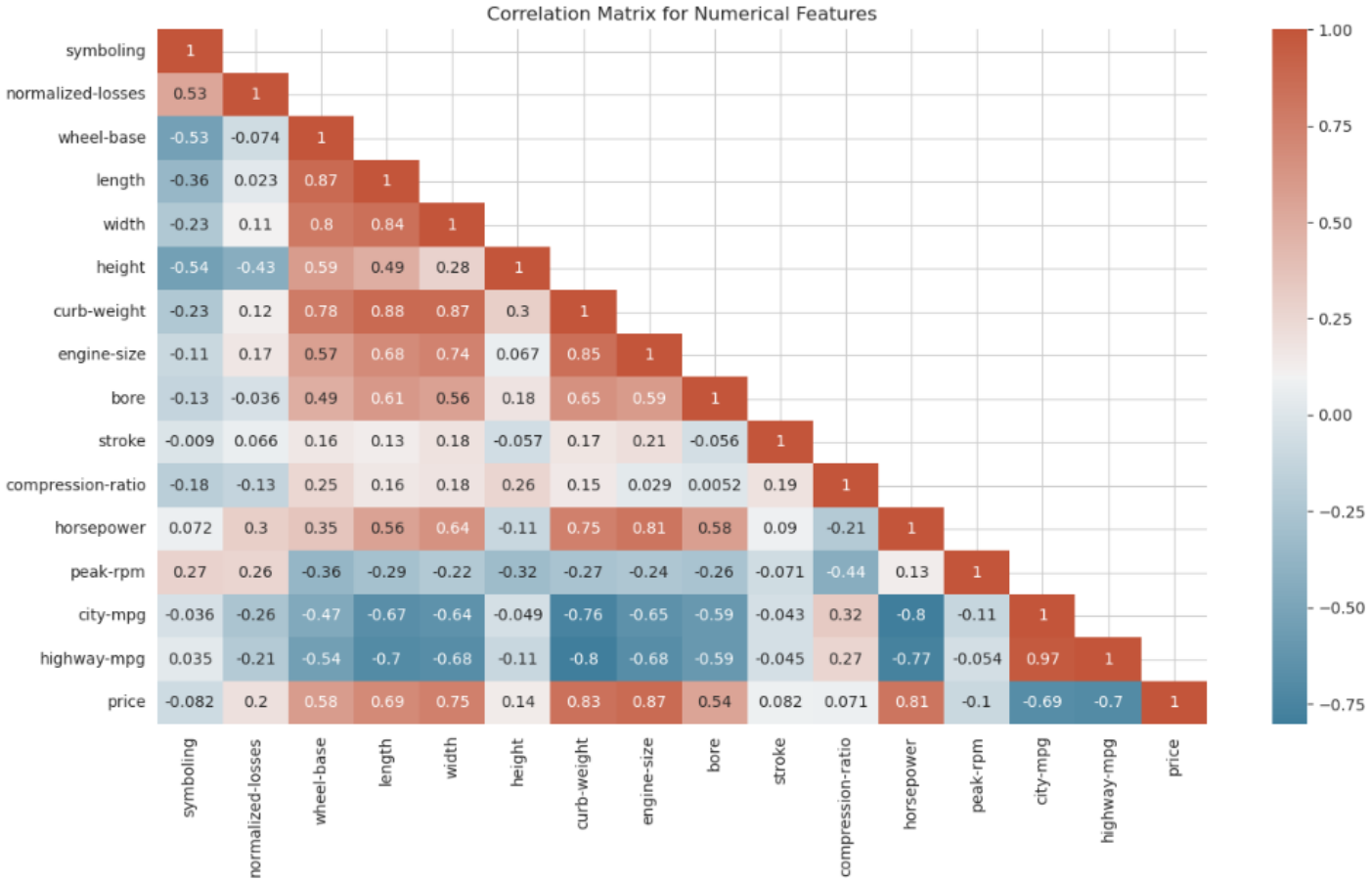


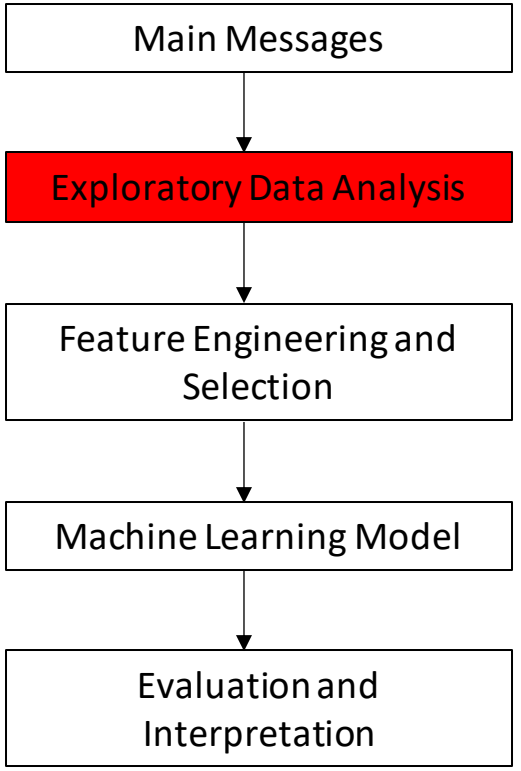


- Key take aways
  - wheel-base, length, width, curb-weight, engine-size, bore and horsepower are positively correlated to price.
  - city-mpg and highway-mpg are strongly correlated, which is expected as both represent fuel efficiency.
  - city-mpg and highway-mpg are negatively correlated to price. Does not necessarily mean higher MPG causes reduction in price. This is a clear case proving correlation is not causation. However, city-mpg and highway-mpg negative correlation to horsepower make sense.

BOSCH

Exploratory Data Analysis – Numerical Data Description

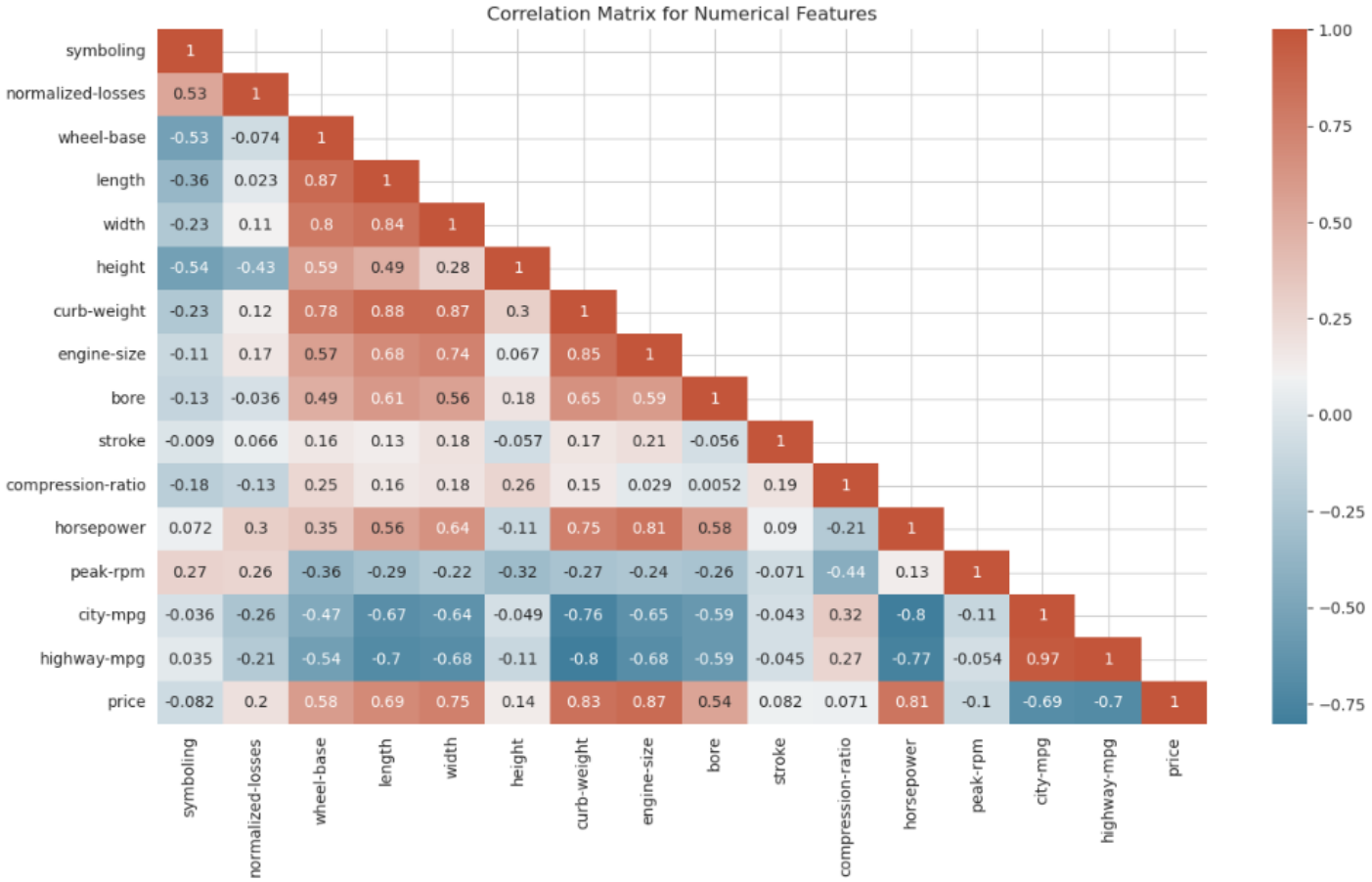


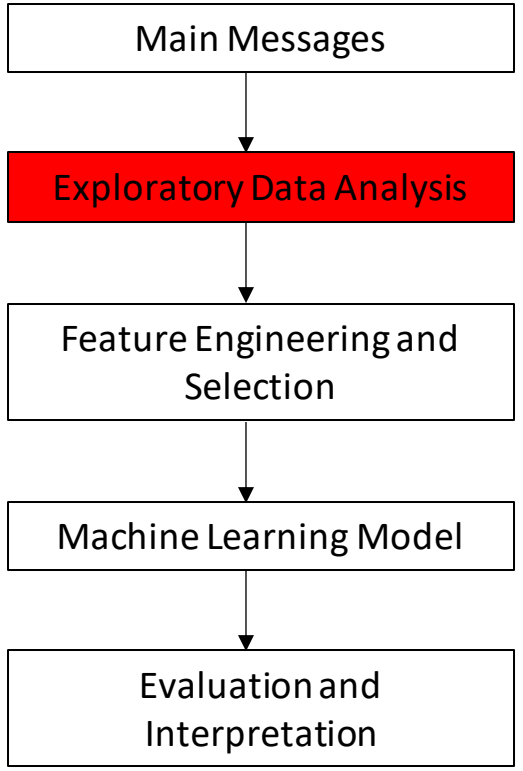


- Key take aways
  - wheel-base, length, width, curb-weight, engine-size, bore and horsepower are positively correlated to price.
  - horsepower is positively correlated with engine-size and curb-weight, and negatively correlated with city-mpg and highway-mpg. This makes sense, as more powerful cars tend to have larger engines, be heavier, and have lower fuel efficiency.
  - curb-weight is positively correlated with length, width, and engine-size, suggesting that larger cars with larger engines tend to be heavier.

BOSCH

Exploratory Data Analysis – Numerical Data Description

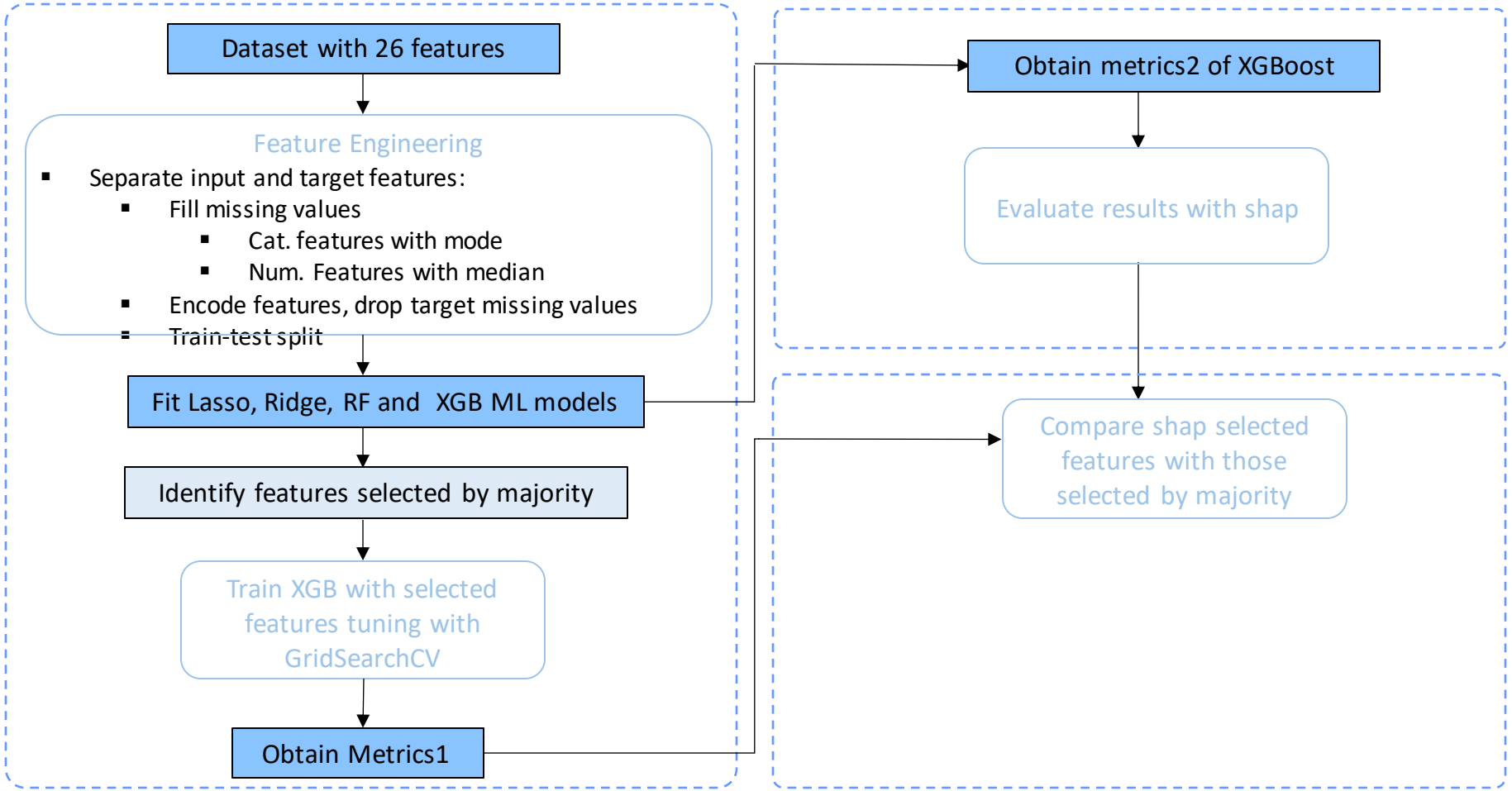


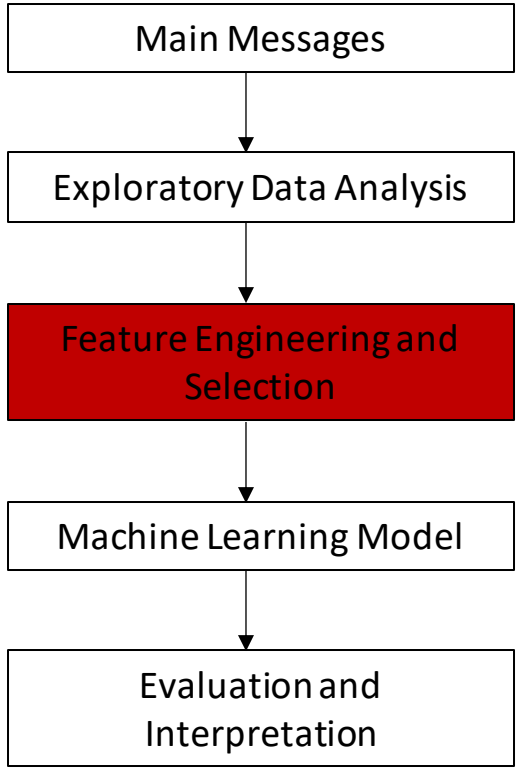


- Key take aways
  - price shows a positive correlation with features such as engine-size, curb-weight, width, and length. This suggests that larger, heavier cars with bigger engines tend to be more expensive.

BOSCH

Feature Engineering and Selection: Framework





BOSCH

Feature Engineering and Selection: Baseline Models

RIDGE top 10 features

Features	Importance
engine-size	3833.141414
compression-ratio	2435.306144
fuel-type	2308.452375
width	1024.153134
wheel-base	1013.378337
engine-location	872.769000
drive-wheels	682.770731
aspiration	678.418059
height	542.184809
symboling	418.783165

LASSO top 10 features

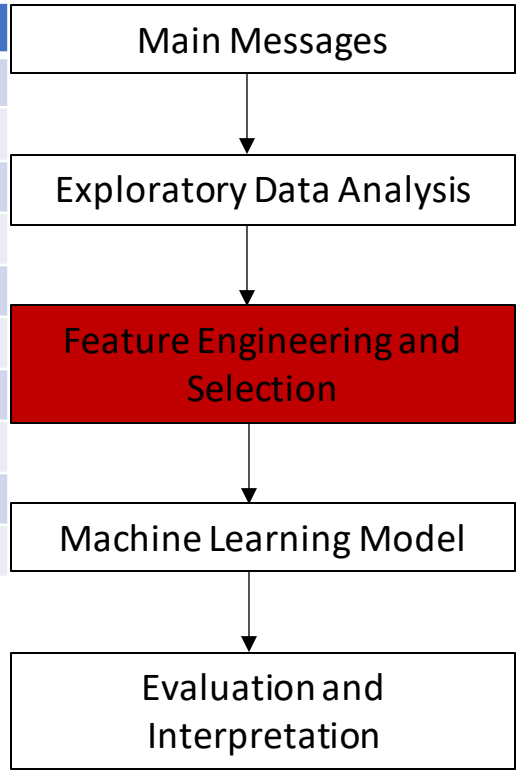
Features	Importance
engine-size	3859.380886
compression-ratio	2664.912434
fuel-type	2550.567706
wheel-base	1020.867719
width	1016.054429
engine-location	874.438810
aspiration	711.063428
drive-wheels	679.828506
height	549.329592
symboling	418.948469

RF top 10 features

Features	Importance
curb-weight - 1	0.429652
engine-size - 2	0.333241
highway-mpg -3	0.052631
horsepower - 4	0.045938
width - 5	0.029043
city-mpg	0.026042
make	0.018965
wheel-base	0.011357
peak-rpm	0.008360
length	0.006878

XGB top 10 features

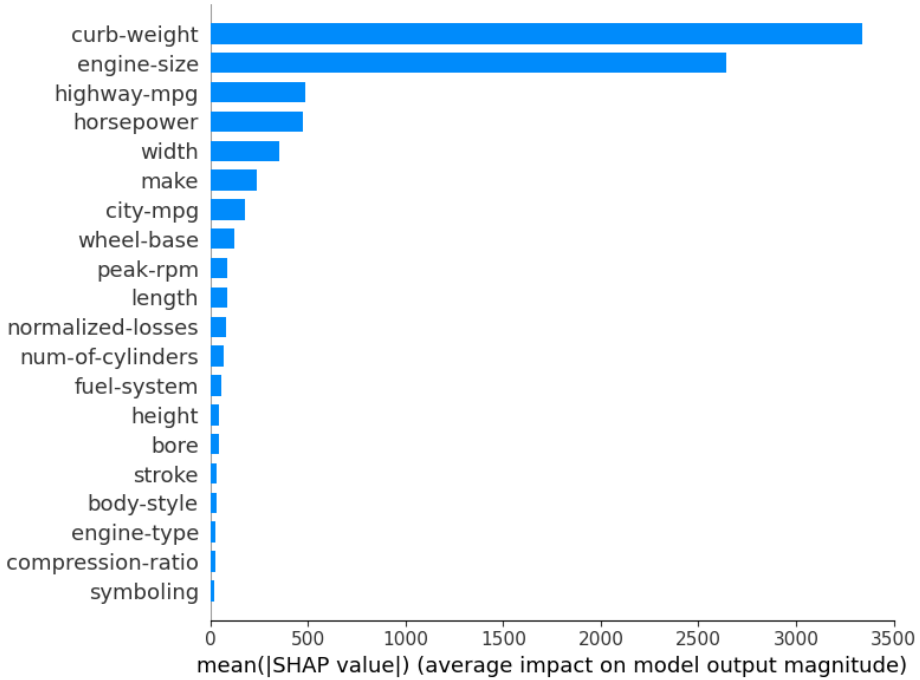
Features	Importance
engine-size -2	0.516384
curb-weight -1	0.107390
width - 5	0.091722
fuel-system	0.058653
highway-mpg - 3	0.056983
bore	0.034710
horsepower - 4	0.029415
num-of-cylinders	0.019615
drive-wheels	0.019013
stroke	0.013233



LASSO RMSE	RIDGE RMSE	RF RMSE	XGB RMSE
4212.3095	4200.8664	2816.8152	2344.4011

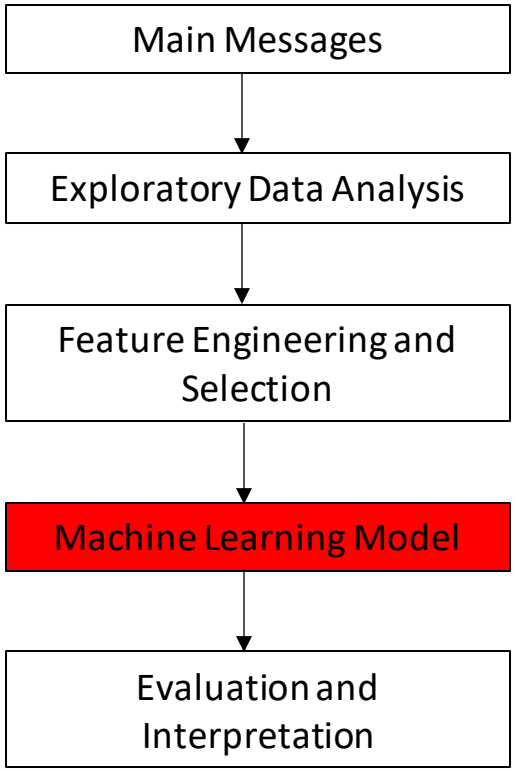
- Key take aways
  - The two tree-based models only have 5 similar top 10 features while non-tree-based models have same top 10 features
  - Tree – based models have better RMSE than non-tree-based models
  - Since tree-based models performed better, we train models with these

Machine Learning Model: Random Forest Model



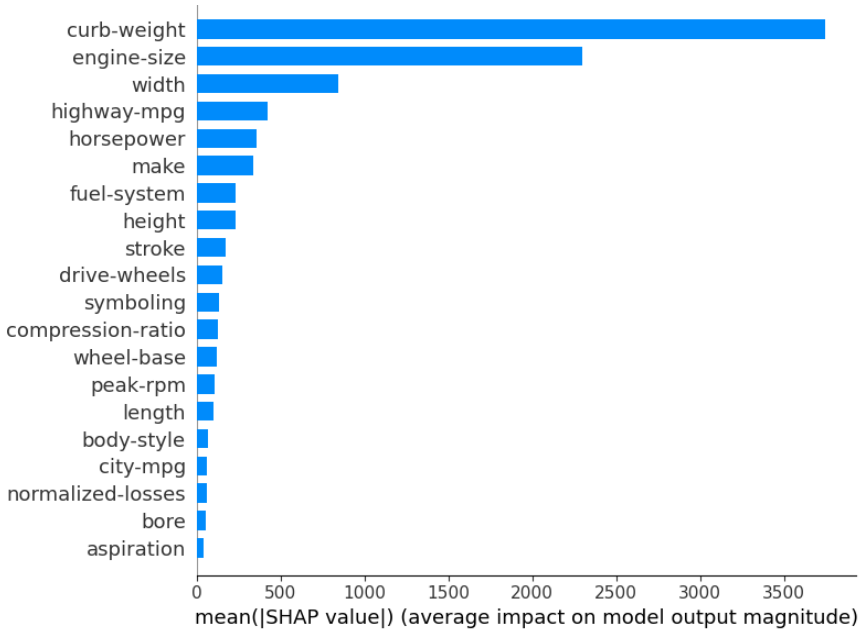
RF top 10 features

Features	Importance
curb-weight	0.429652
engine-size	0.333241
highway-mpg	0.052631
horsepower	0.045938
width	0.029043
city-mpg	0.026042
make	0.018965
wheel-base	0.011357
peak-rpm	0.008360
length	0.006878



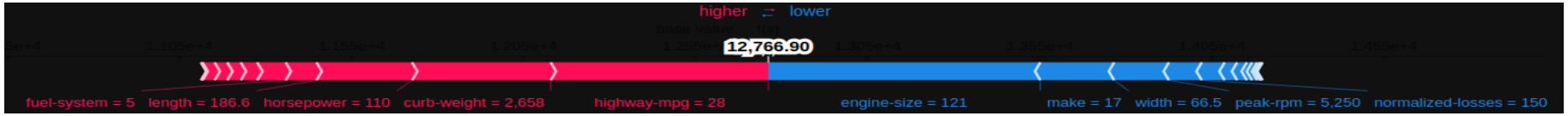
- Key take aways
  - Hyperparameters were tune with GridSearchCV and RandomizedSearchCV
  - ShAP validated previously identified top 10 features with the first four feature the same in both cases

Machine Learning Model: XGB Model



XGB top 10 features

Features	Importance
engine-size	0.516384
curb-weight	0.107390
width	0.091722
fuel-system	0.058653
highway-mpg	0.056983
bore	0.034710
horsepower	0.029415
num-of-cylinders	0.019615
drive-wheels	0.019013
stroke	0.013233



Main Messages

↓

Exploratory Data Analysis

↓

Feature Engineering and Selection

↓

Machine Learning Model

↓

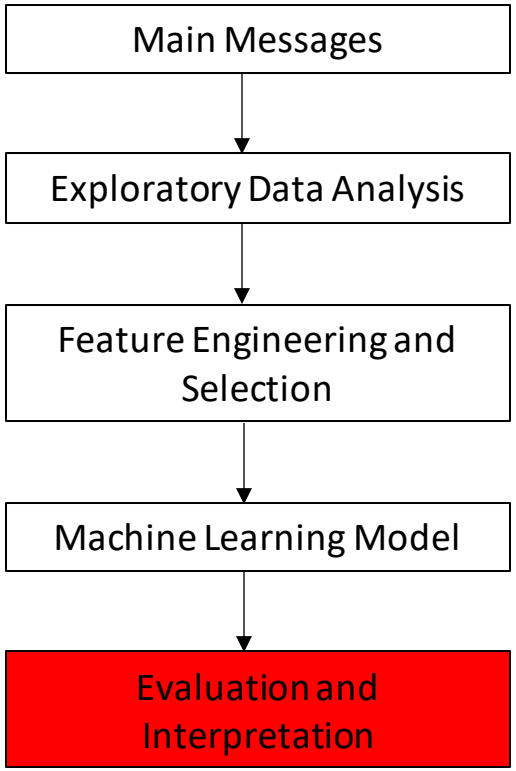
Evaluation and Interpretation

- Key take aways
  - The top 10 features in the previous XGB did not consider features 'make' and 'height'
  - In terms of the rmse, the previous XBG gave a better result without the need for hyperparameter. Tuning makes the results worse.

BOSCH

Evaluation and Interpretation: XGB Model

- Model Strength:
  - The selected tree-based models have features that depend on the physical/mechanical characteristics as well as engine characteristics of cars.
  - Models dependent on the features identified can be help an automotive supplier like Bosch to give more attention to these parts through research and development for possible improvement. This can give theorganization an edge since this is also their core – they have more control over it.
- Model Weakness:
  - The features of the model does not include insurance related features. This could limit an automotive supplier autonomy in determine vehicle price.
  - The model does not consider the effect of seasonality, environmental sustainability and regions as these are in turn not provided in the data.



Thank You