PS-EM/CIT assignment

Pedro Gomes

15/12/2023

BOSCH

01

Introduction

BOSCH

Introduction

The objective

- As an automotive supplier, PS is interested in predicting critical factors that affect its **business**.
- As such, the proposed technical challenge consisted of analysing a real-world dataset related to the **automotive industry**.
- This dataset can be found in the repository of University of California, Irvine (UCI). It consists of the specification of a vehicle in terms of various characteristics (such as size, brand, and so on), as well as its assigned insurance risk rating and insurance loss payment.
- The goal was to **predict** vehicle **prices** based on these characteristics.
- Exploratory Data Analysis (EDA) techniques were carried out, and a regression model was built for price prediction.

02

Exploratory data analysis

Exploratory data analysis

The dataset

- The provided dataset was obtained from 1985 Ward's Automotive Yearbook.
- Below is a short description of each column in the dataset, based on some research.

Column	Description
Make	The brand that produces the car (Mitsubishi, Mercedes-Benz, etc.)
Fuel type	The car’s fuel type (gas or diesel).
Aspiration	The type of engine aspiration (if “std”, the air intake depends solely on atmospheric pressure; if “turbo” it has forced induction).
Number of doors	The number of doors the car has (two or four).
Body style	The car’s body style (hatchback, sedan, etc.).
Drive wheels	How the engine’s power is delivered to the wheels (if “4WD”, the engine’s power is delivered to all wheels at the same time; if “FWD” or “RWD”, the power is delivered to the front or back wheels, respectively).
Engine location	The location of the engine in the car (front or rear).
Wheelbase	The distance between the front and back wheels (in inches).
Length	The car’s length (in inches).
Width	The car’s width (in inches).
Height	The car’s height (in inches).

Exploratory data analysis

The dataset

Column	Description
Curb weight	The car's weight without any passengers or baggage (in pounds).
Engine type	The engine type ("OHC" for "Over Head Cam", "DOHC" for "Double Over Head Cam", etc.)
Number of cylinders	The number of cylinders in the car's engine.
Engine size	The volume of the cylinders inside an engine (in cubic inches).
Fuel system	The car's fuel system ("MPFI" for "Multi Point Fuel Injection", "IDI" for "Indirect Injection", etc.).
Bore	The diameter of each cylinder in the engine (in inches).
Stroke	The distance travelled by the piston in the cylinder during one engine cycle (in inches).
Compression ratio	The ratio between the volume of the cylinder when the piston is at its minimum position *vs* when it is at its maximum position.
Horsepower	A measure of the engine's output. It is related to its power.
Peak RPM	The number of the engine's revolutions per minute (RPM) at peak engine power.
City MPG	The average distance that the car travels per gallon of fuel, measured on average city conditions (in miles).
Highway MPG	The average distance that the car travels per gallon of fuel, measured on average highway conditions (in miles).
Symbolling	The risk factor associated to the car based on its price (a value of +3 indicates that the car is risky, a value of -3 means that it is safe).
Normalized losses	The relative average loss payment per insured vehicle year (in dollars). In other words, the average loss per car per year.
Price	The vehicle's price (in dollars). This is the target column.

Exploratory data analysis

The dataset

#	Column	Has missing values	Data type
1	Symbolling	No	Integer
2	Normalized losses	Yes	Float
3	Make	No	Text
4	Fuel type	No	Text
5	Aspiration	No	Text
6	Number of doors	Yes	Text
7	Body style	No	Text
8	Drive wheels	No	Text
9	Engine location	No	Text
10	Wheelbase	No	Float
11	Length	No	Float
12	Width	No	Float
13	Height	No	Float
14	Curb weight	No	Integer
15	Engine type	No	Text
16	Number of cylinders	No	Text
17	Engine size	No	Integer
18	Fuel system	No	Text
19	Bore	Yes	Float
20	Stroke	Yes	Float
21	Compression ratio	No	Float
22	Horsepower	Yes	Float
23	Peak RPM	Yes	Float
24	City MPG	No	Integer
25	Highway MPG	No	Integer
26	Price	Yes	Float

- The following columns, marked in yellow on the left, have **missing values**:
  - Normalized losses
  - Number of doors
  - Bore
  - Stroke
  - Horsepower
  - Peak RPM
  - Price
- As the goal was to predict the Price, the records for which it was missing were not considered.
- The features “Number of doors” and “Number of cylinders” were converted to numbers (instead of using their text description).
- **Duplicate** records were found in the data. For these records, the vehicles’ characteristics were the same, despite having a different Price value.

Exploratory data analysis

The dataset

- An example of duplicate observations is shown below, for two records concerning Alfa-Romero vehicles:

symboling	normalized-losses	make	fuel-type	aspiration	num-of-doors	body-style	drive-wheels	engine-location	wheel-base	...	engine-size	fuel-system	bore	stroke	compression-ratio	horsepower	peak-rpm	city-mpg	highway-mpg	price
3	NaN	alfa-romero	gas	std	two	convertible	rwd	front	88.6	...	130	mpfi	3.47	2.68	9.0	111.0	5000.0	21	27	13495.0
3	NaN	alfa-romero	gas	std	two	convertible	rwd	front	88.6	...	130	mpfi	3.47	2.68	9.0	111.0	5000.0	21	27	16500.0

- These vehicles have the exact same properties yet have a different price. This happens for 16 records.
- The following approaches were considered to overcome this:
  - Keep the data as it is. This can be deceiving to the model, as seeing two (or more) equal records with different target values during training can lead to learning wrong patterns and making wrong predictions.
  - Keep only one of the records. This can lead to information loss.
  - **Aggregate** the duplicate records into one by computing an average. For the example given above, this means having a single record, where the price and normalized losses are the mean of the values of these two records. Ideally, this should be done in Feature Engineering, as to avoid data leakage.

- The chosen approach was the last one. After this, the dataset was left with a total of **192 observations**.

Exploratory data analysis
Feature analysis

- The EDA process was divided into 2 parts:
  - A **univariate analysis**, in which each feature was analysed individually using descriptive statistics and graphical visualizations.
  - A **multivariate analysis**, in which the relationships between the different features and the features and the target were analysed.

- Histograms, box plots and QQ plots were the preferred plot types for numerical variables, as they allow a clear visualization of their distribution.

- Categorical features were analysed using bar charts, pie charts and count plots.

- The analysis of each feature can be found in more detail in the Jupyter Notebook.

Exploratory data analysis

Feature analysis

- To assess the relationship between the numerical features and the target, three approaches were carried out:
  - Plotting such variables against the target variable (for example, using scatter plots).
  - Computing the **correlation** between the variables and the target.
  - Computing **Feature Importance (FI)** using a simple Random Forest model.

- The correlation between two variables is a statistical measure that expresses how much two variables are linearly related.
  - A correlation of +1 means there is a perfect positive correlation.
  - A correlation of -1 means that there is a perfect negative correlation.

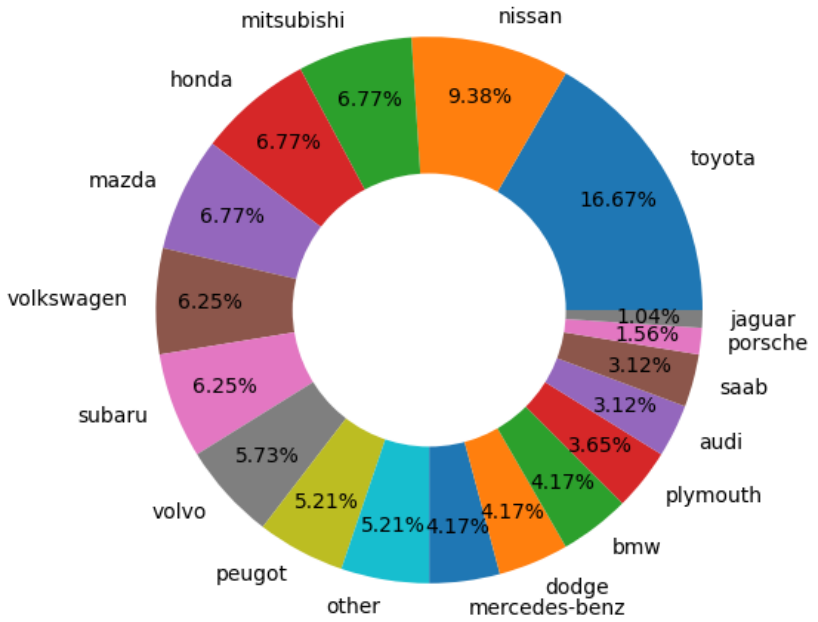
- Feature importance allows us to know which features are more useful in predicting the target variable; in other words, the higher the value of FI, the more useful the feature is in the prediction.

- In the next slides a summary of the main finds are presented.

Exploratory data analysis

Summary of vehicle characteristics

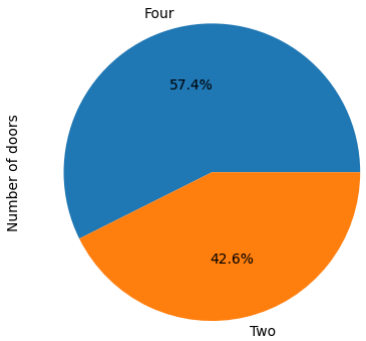
- An analysis of the vehicle's make revealed that the most frequent brands are Japanese, namely Toyota, Nissan, Mazda, Mitsubishi, Honda and Subaru.
- Toyota is the most common brand, with a frequency of 16.7 % in the data.
- Other brands (marked as "other" in the pie) include brands that are present in less than 2% of the observations, namely Chevrolet, Alfa-Romero, Isuzu, Renault and Mercury.

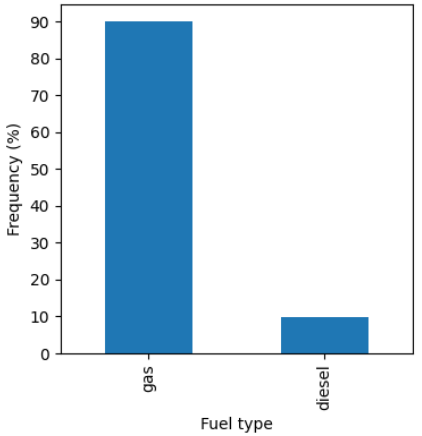


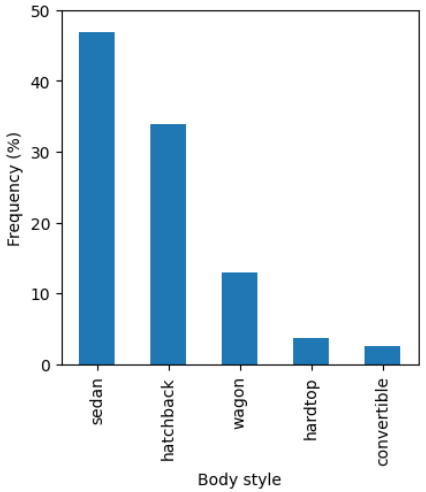
BOSCH

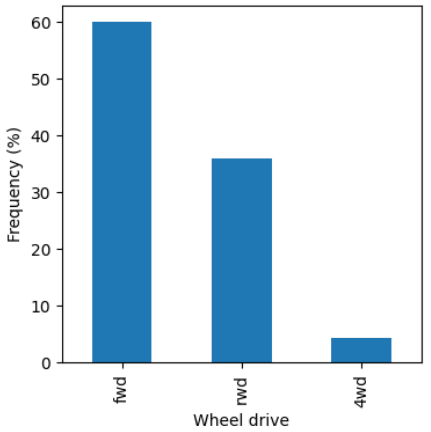
Exploratory data analysis

Summary of vehicle characteristics









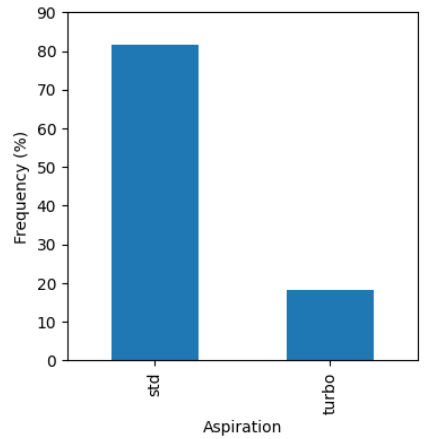
- Most vehicles have four doors, as can be seen on the pie chart on the left, with a percentage of 42.6% having only two doors.
- Diesel vehicles are a minority in the data, having a presence of 11% in the observations. The vast majority are gas cars.
- Sedans are the most common type, followed by hatchbacks. Convertibles are the less frequent ones.
- The majority of the vehicles have engine power delivered to the front wheels, and less than 5% are 4x4.

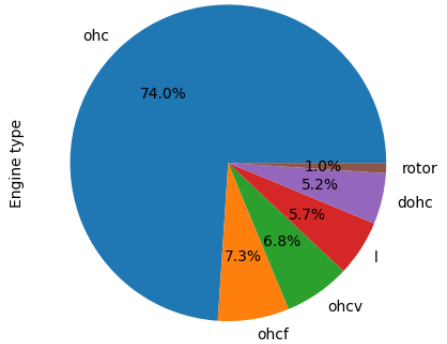
BOSCH

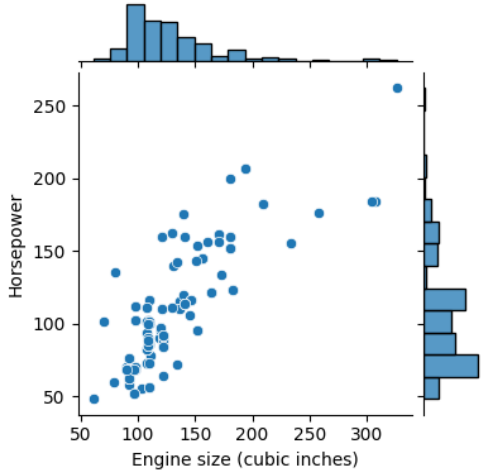
Exploratory data analysis

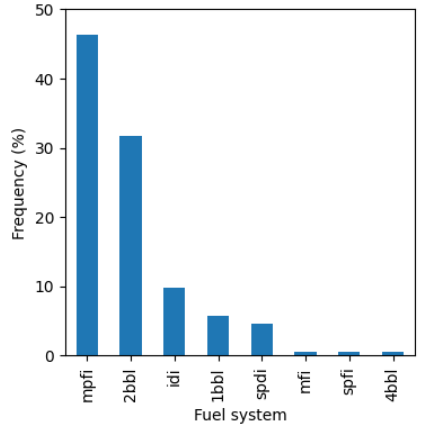
Summary of vehicle characteristics

- In what concerns engine properties, most engines have standard aspiration, whereas around 20% have turbo aspiration.
- The most common engine type is OHC (Over Head Cam). The other types are relatively rare, with a frequency less than 10% in the observations.
- The engine size, typically below 150 cubic inches, scales linearly with horsepower, which is usually in the range 50 – 125.
- MPFI (Multi Point Fuel Injection) is the most common fuel system. MFI, SPFI and 4BBL have little representation.





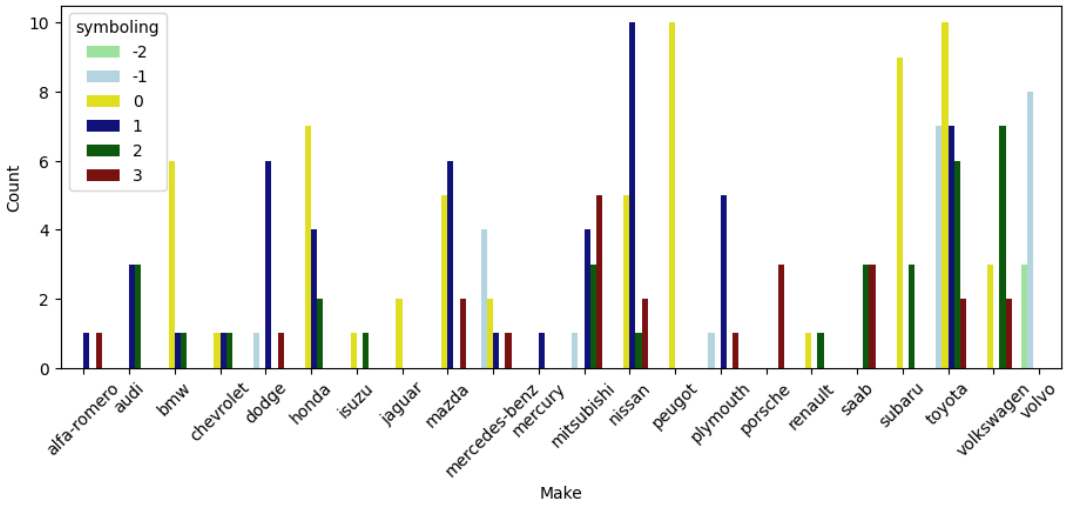




Exploratory data analysis

Summary of insurance-related properties

- In the bar chart below is shown the count of vehicles per make and symbolling.
- In what concerns risk assessment, the riskiest vehicles – those with a risk factor greater than 1 (on average) – are Alfa-Romero, Audi, Mitsubishi, Saab, and Volkswagen. All Porsche cars are extremely risky, with a symbolling score of +3.
- All Volvo cars are considered safe from the insurance point of view, having a negative risk score.
- BMW, Jaguar, Peugeot and Subaru vehicles have, on average, a risk factor close to or equal to 0, meaning that for these vehicles the risk factor is in line with its price.

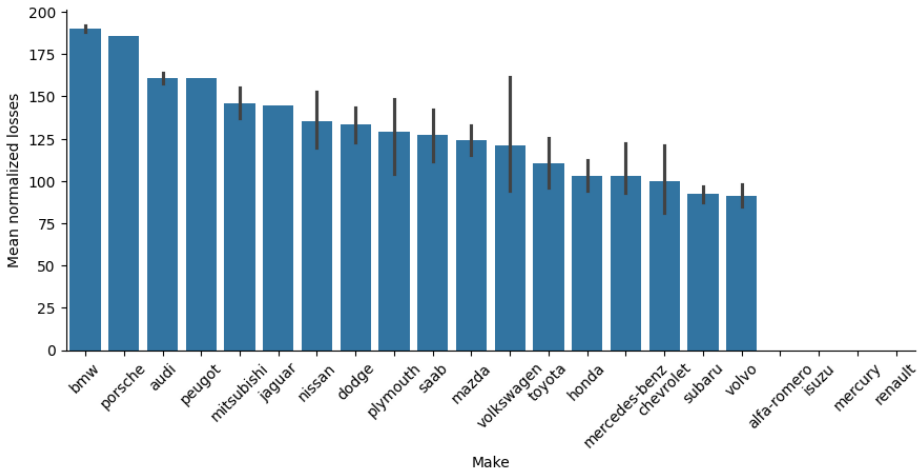


Exploratory data analysis

Summary of insurance-related properties

- Below is shown the mean normalized loss per make (the average loss per car per year).

- The vehicles which represent a greater loss are BMWs, followed by Porsche and Audi. For Alfa-Romero, Isuzu, Mercury and Renault data regarding the loss is not available.



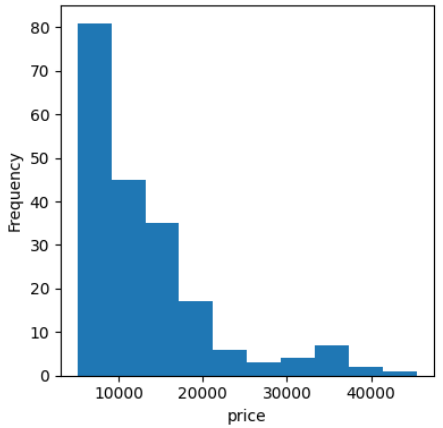
- Volvo and Subaru vehicles have, on average, the smaller loss.

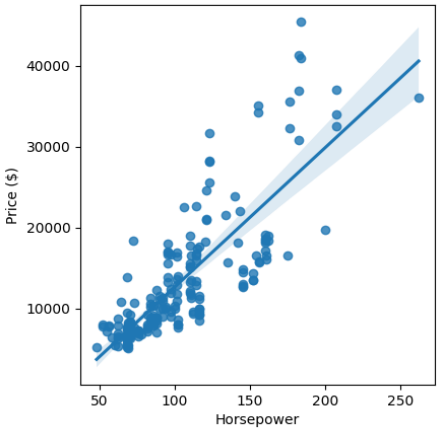
BOSCH

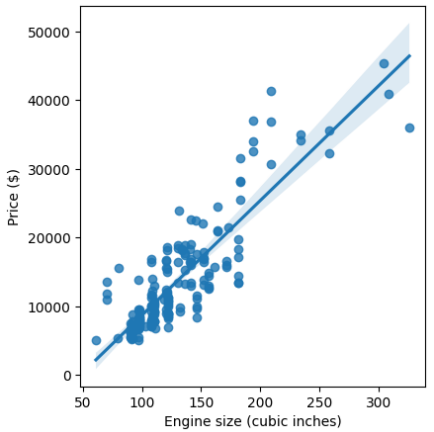
Exploratory data analysis

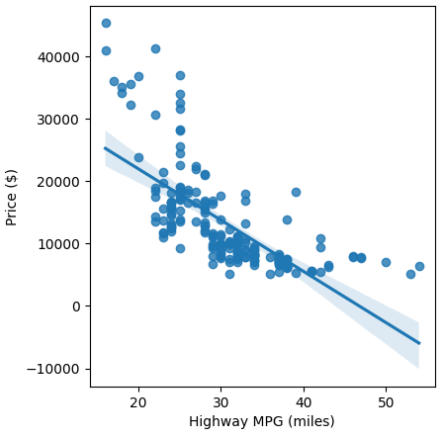
Relationships between numerical variables and the target

- The target variable – Price – is in the range 5000 – 45 400 $. It has a right-skewed distribution, as can be seen from the histogram shown on the right, with a median of 9992 $.
- It is positively correlated with the vehicle's horsepower. This is expected, as vehicles with higher horsepower are more expensive. It is also correlated with the engine size.
- It has a positive correlation with wheelbase, length and width (these plots are not shown here). This also makes sense, as larger cars (e.g., sedans) are often more expensive.
- In turn, it is negatively correlated with highway MPG (although the shape of this relationship is not linear; it resembles an exponential curve).









BOSCH

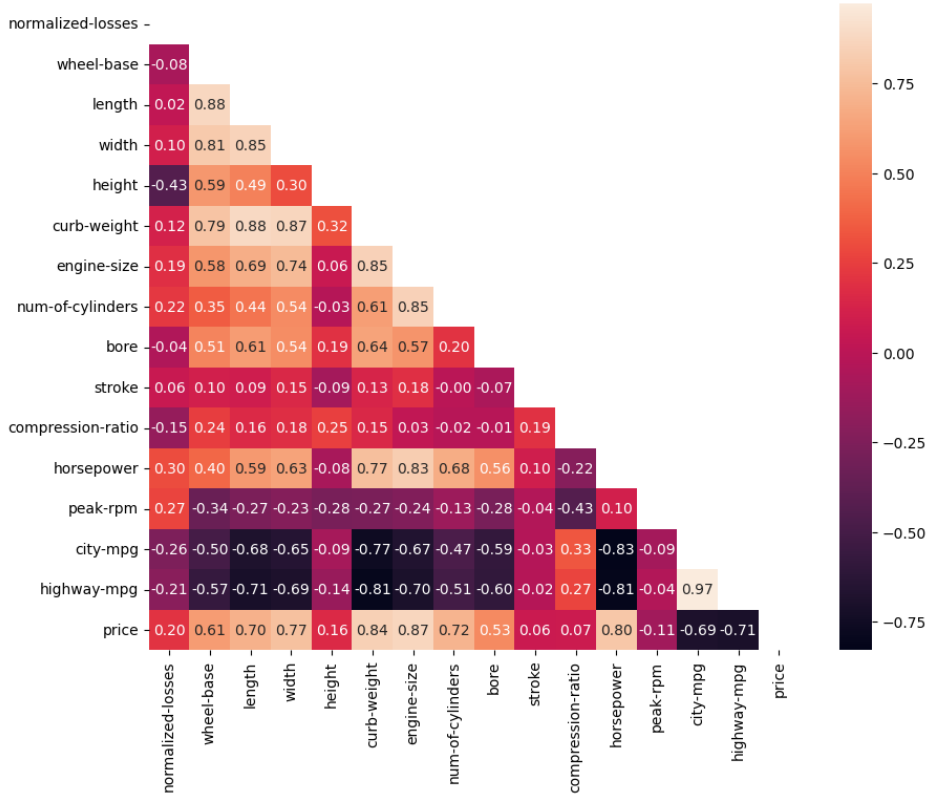
Exploratory data analysis

Relationships between numerical variables and the target

- On the right is shown the correlation matrix for the different variables and the target.

- The following variables are strongly correlated (considering that a strong correlation is ≥ 0.70):
  - Curb weight, wheelbase, length and width. This makes sense, as bigger vehicles will have bigger length, width, wheelbase and will be heavier.
  - Engine size with horsepower, curb weight and number of cylinders.
  - Highway MPG and City MPG.

- In turn, the following features are negatively correlated:
  - City MPG/Highway MPG with curb weight and horsepower.
  - Highway MPG with length.



BOSCH

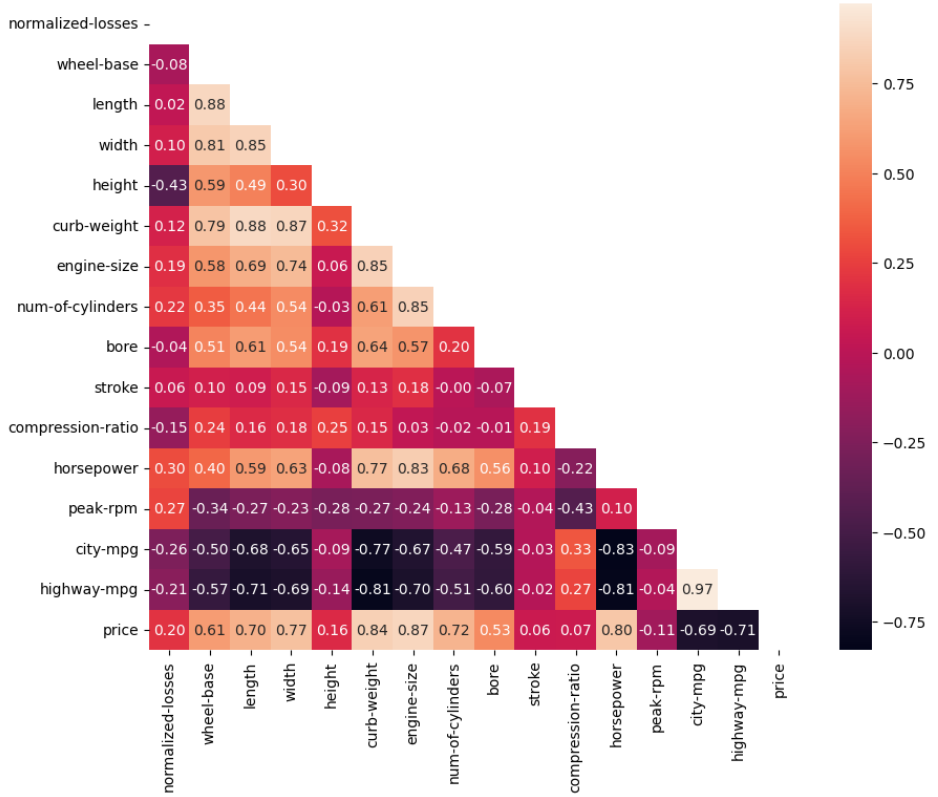
Exploratory data analysis

Relationships between numerical variables and the target

- Price is strongly correlated with:
  - Length
  - Width
  - Curb weight
  - Number of cylinders
  - Engine size
  - Horsepower

- In turn, it is negatively correlated with Highway MPG.

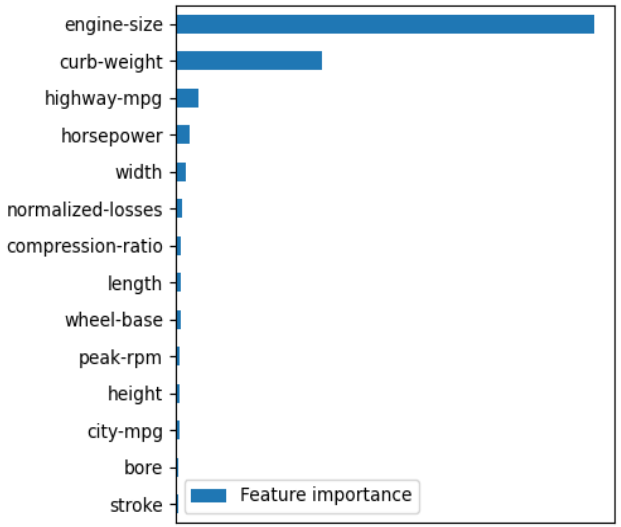
- Based on these results and the plots shown in the previous slides, these features were considered good predictors of the target variable.



BOSCH

Exploratory data analysis

Relationships between numerical variables and the target

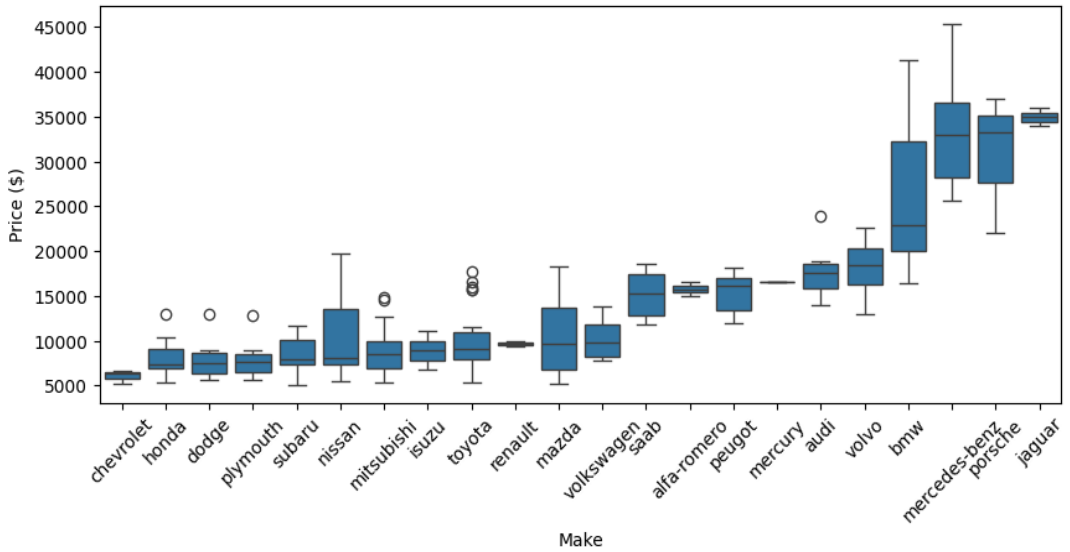


- A quick assessment of the **importance** of the numerical features in predicting the target was done with a simple (default) Random Forest model.
- This model was fit to the data, and the respective **feature importances** extracted. The result is shown on the bar chart on the left.
- Engine size is the most helpful feature in predicting Price, followed by curb weight (these are, however, strongly correlated) and highway MPG.
- It is important to note that the purpose of this approach was a simple evaluation of which features are most helpful in predicting the target; the model was not tuned, and it was not built for predictions.

Exploratory data analysis

Relationships between categorical variables and the target

- The Price distribution for the different brands is shown in the box plot below.
- Jaguar, Porsche, Mercedes-Benz and BMW are the most expensive vehicles, with median prices above 20k $.
- The lower price models are Chevrolet, Honda, Dodge, Plymouth and Subaru.
- This feature was considered a good predictor of Price.

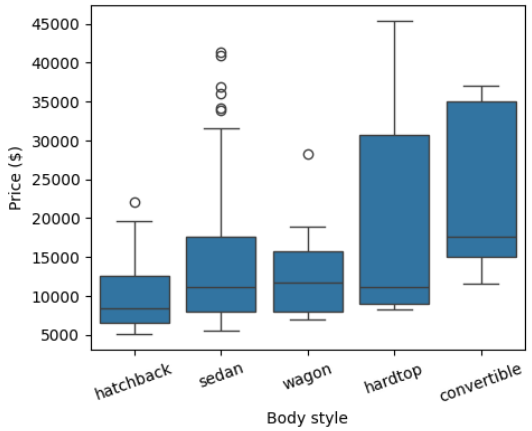


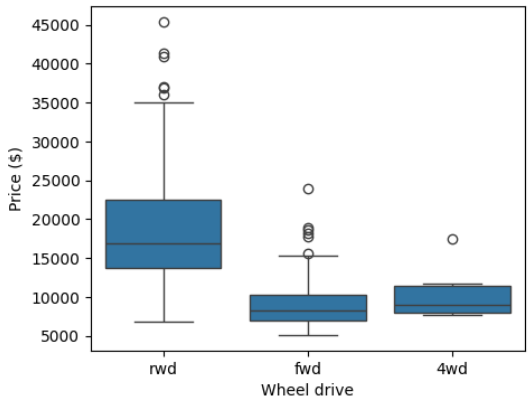
BOSCH

Exploratory data analysis

Relationships between categorical variables and the target

- In what concerns the different body styles, the most expensive vehicles are convertibles, followed by hardtops, although there is some overlap in the distributions of the different styles.
- Hatchbacks are the cheapest models.
- Between the different wheel drives, RWD (vehicles for which the engine power is transmitted to the back wheels) automobiles are more expensive.
- Given these results, these features were considered in the modelling.
- The mean price for 2-door vehicles and 4-door vehicles is quite similar, without a big difference in price between the groups. This feature was not considered as a useful predictor.

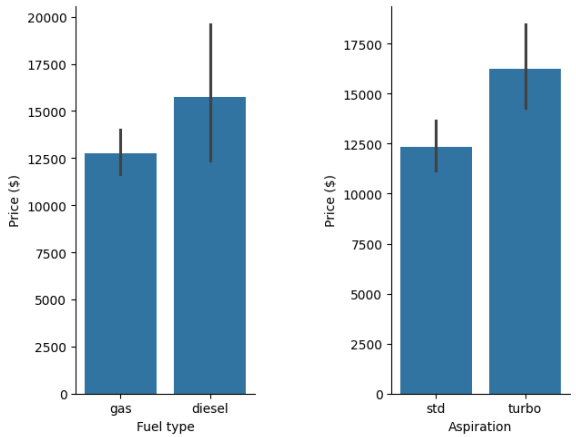




BOSCH

Exploratory data analysis

Relationships between categorical variables and the target



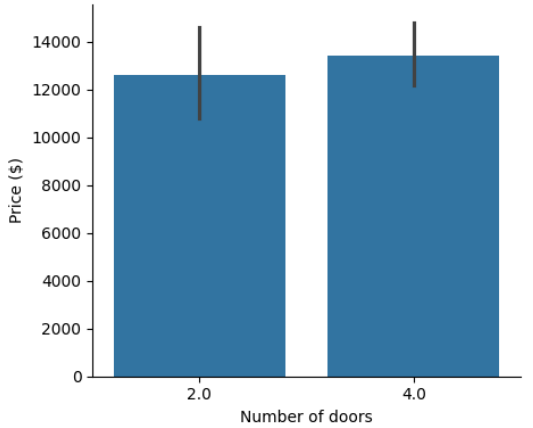
- Diesel vehicles are, on average, more expensive than gas vehicles.
- The same is verified for cars with turbo aspiration compared with standard aspiration. Therefore, these two features were considered as useful predictors.
- Generally, price increases with the number of cylinders in the engine, as can be seen on the bottom plot. This makes sense, as more cylinders translate in larger engines and more power.
- The type of engine seems to be related with the number of cylinders, as, for example, rotor engines have only 2 cylinders, whereas OHCF engines have 4 or 6 cylinders. OHCV engines are more expensive on average. Engine type was also considered as a feature in the model.

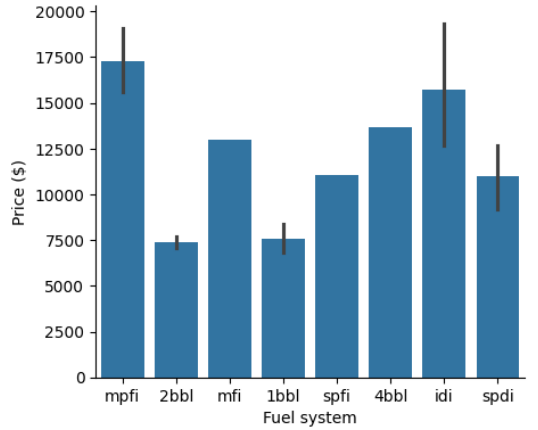
Exploratory data analysis

Relationships between categorical variables and the target

- The mean price for 2-door vehicles and 4-door vehicles is quite similar, and there isn't big difference in price between the groups. This feature was not considered in iterations over the baseline model, as discussed in the following section.

- In turn, the different Fuel systems have different price means between them, with MPFI being the most expensive one. Fuel system was considered as a feature in the model.





BOSCH

03

Feature Engineering & modelling

BOSCH

Feature Engineering & modelling

Baseline model

- The baseline approach to this problem considered **all** the features in the dataset, and simple **Linear Regression.**
- The data was split in an **80-20 fashion**, i.e., 80% of the dataset was used for training purposes and 20% for testing.
- The preferred metric to evaluate model performance was the **Root Mean Squared Error (RMSE)**. This is the average squared difference between the predictions and the true values. It can be interpreted as the “amount of error” (in dollars) the model makes when making predictions.
- The **Mean Absolute Percentage Error (MAPE)** and the **R-squared** were also considered. MAPE consists of the average percentage error of the model, whereas R-squared is used to measure the “goodness of fit” and indicates how much of the variance in the independent variable can be explained by the independent variables.

Feature Engineering & modelling

Baseline model

- The feature engineering and modelling pipeline for the baseline approach consisted of the following steps:
  1. **Median imputation**. Numerical variables with missing values are imputed with their median value. The median was chosen over the mean as most of the variables have skewed distributions.
  2. **Mode imputation**. Categorical variables are imputed with the most frequent value (the mode).
  3. **Dummy encoding** of categorical text variables. Text variables are one-hot encoded, except for the last category, as to avoid redundancy.
  4. **Linear Regression** model fit.

Below are shown the scores for the baseline:

	RMSE	MAPE	R-squared
Train	1475,80	9,67	0,96
Test	2432,37	17,28	0,90

- From these results it is clear the baseline is **overfitting** the training data – the difference in scores between the training set and the test set is large. The fitted model is too complex and fails to generalize to new data.

Feature Engineering & modelling

Iteration 1

- A first iteration in building a better model consisted of dropping the following features, for the mentioned reasons:
  - Low correlation with the target/small differences between groups:
    - Stroke
    - Peak RPM
    - Compression ratio
    - Number of doors
    - Symbolling
  - Multicollinearity:
    - Width and Wheelbase (strong correlation with Length)
    - City MPG (strong correlation with Highway MPG)
    - Curb weight, Horsepower and Number of cylinders (high correlation with Engine size)
  - Missing values:
    - Normalized losses (almost 20% of the values in this column are missing).

- One of the goals was the removal of features that don't correlate with the target and collinear features, as the absence of **multicollinearity** is one of the main assumptions of linear regression.

Feature Engineering & modelling

Iteration 1

- Another goal consisted of **removing rare categories** from the data. Since “Make”, “Body style” and “Fuel system” have several rare categories (appearing in less than 5% of the records), these were grouped into an “Other” category.

- Rare categories can be problematic as they can cause models to overfit; also, when encoding the feature, if a category is rare, its corresponding encoded version will be very sparse (although this is most worrying for tree models).

- For “Make”, the threshold below which a category is considered rare was set to 2.5%; for “Body style” and “Fuel system”, it was 5%.

- The remaining steps of the pipeline were maintained (i.e., imputation of missing values with the median, etc.), adjusted to the changes previously mentioned.

Feature Engineering & modelling

Iteration 1

- The results for this approach are shown below.

- In comparison with the baseline, the **RMSE and MAPE decreased substantially**; in turn, the gap between the train and test sets is much smaller, indicating that the model was simplified and is not overfitting.

- The MAPE value indicates that, on average, the predictions differ from the real values by 13%, which is reasonable.

	Set	RMSE	MAPE	R-squared
Baseline	Train	1475,80	9,67	0,96
	Test	2432,37	17,28	0,90
Iter. #1	Train	2059,22	11,28	0,93
	Test	2029,79	12,96	0,92

- The R-squared score is acceptable, and the model can explain 92% of the variance in the data.

- However, the RMSE value is still considerably high; the lowest vehicle prices are in the order of the thousands of dollars, which is the order of magnitude of the RMSE.

Feature Engineering & modelling

Iteration 2

- A second iteration consisted of using **Ridge regression**. This is similar to Linear Regression, with the difference that a penalty term is introduced in the algorithm's loss function. This penalty term is meant to shrink the magnitude of some of the coefficients, leading to a simpler model and reducing the influence of some features.

- **Hyperparameter tuning** was carried out to find the best regularization strength (also called **alpha**). This was done with 3-fold cross-validation (**3CV**).

- An additional scaling step was included in the pipeline, to scale the features and facilitate model convergence.

- The results are shown on the right. As can be seen, the approach that performed best was this one (marked in green), although the scores differ little from iteration 1.

	Set	RMSE	MAPE	R-squared
Baseline	Train	1475,80	9,67	0,96
	Test	2432,37	17,28	0,90
Iter. #1	Train	2059,22	11,28	0,93
	Test	2029,79	12,96	0,92
Iter. #2	Train	2059,41	11,28	0,93
	Test	2027,11	12,97	0,92

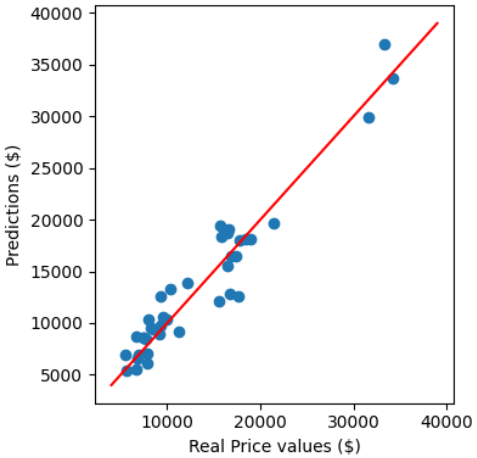
Feature Engineering & modelling

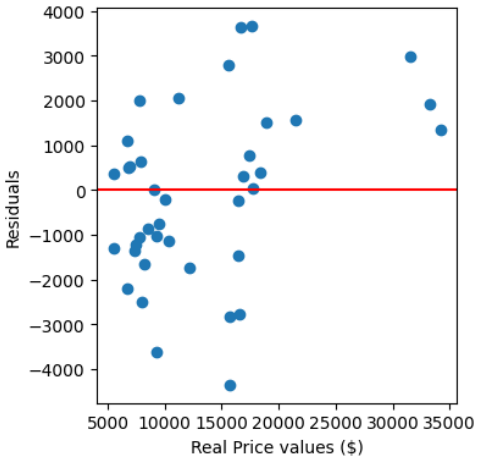
Iteration 2

- For the best performing model, a comparison of the real values with the predictions shows that the model performs fairly well, with the points being distributed around the identity line (shown in red).

- A scatter plot of the residuals (shown below) shows, as the price increases, the residuals increase (although slightly). This means that, to some extent, the assumption of **homoscedasticity** in linear regression – that the variance of the residuals is constant – is not met.

- Nevertheless, for prediction purposes this is not worrying; the cause for this difference is related to the distribution of the price (which is right-skewed) and only when assessing the statistical significance of the regression parameters does the assumption of homoscedasticity have to be met.





BOSCH

04

Conclusions and next steps

Conclusions & future steps

- The purpose of the current assignment was to build a regression model capable of predicting car prices based on their characteristics, such as horsepower, brand, engine type, and so on.

- The strongest numerical predictors of price were found to be length, width, curb weight, engine size and horsepower. Categorical features, such as Make and Body style, were also deemed relevant in the prediction.

- The best performing model consisted of a Linear Regression model with feature selection and feature grouping. This model achieved a RMSE of 2027.11 $, and a MAPE of 12.97%, meaning that on average the predictions differ from the real values by 12.97%.

- In the next slide are presented some limitations of the current approach and the steps that could be taken to improve the model and/or the interpretation of its results.

Conclusions & future steps

Limitations of the current approach

- The feature selection process was mostly based on a visual or mathematical analysis of the relationship between the features and the target (e.g., via Pearson's correlation).
  - From the business point of view, some features might be more important than the ones selected but these ended up being discarded because of their lower correlation with the target and high correlation with other features.
  - For example, the Horsepower feature might be more relevant from the business point of view than Engine size; still, it was not considered because the latter had a strong correlation with Price.
  - This makes the task of explaining the model to a stakeholder more complicated.

- The importance of categorical features in predicting the target was mainly assessed by visualization.
  - More robust feature selection could be carried out using hypothesis testing. For example, *t*-test and ANOVA statistics are useful in understanding if categorical features are helpful in predicting the target.
  - The relationship between the different categorical variables could be explored further, as the focus was mostly put on understanding the relationship of these features with the target.

- A thorough examination of the model's coefficients was not carried out.
  - The coefficients of a regression model provide valuable information on the features the model considers important.

Conclusions & future steps

Future steps

- **Engineer features**.
  - For the sake of interpretation, features concerning length and mass can be converted to SI units.
  - Create new features. For example, using the bore, stroke and number of cylinders one can compute the engine displacement (the utility of such a feature would still have to be assessed, as the price is already highly correlated with several engine properties); or using the symbolling column, create a flag indicating if the vehicle is risky or not.
  - Apply variance-stabilizing transformations (e.g., Box-Cox) on numerical features. This can help in making the relationship with the target more linear (and mitigate the effect of outliers).

- Test other **feature selection** techniques, such as backward or forward selection.

- Test **other models** (e.g., the Decision Tree regressor or the Support Vector regressor).
  - For other algorithms, the feature engineering process would have to be adjusted accordingly (for example, tree-based models do not work well with sparse data from one-hot encoding).

- Gather **more data**. The current dataset is quite small, with only 192 records; this leaves little data to train the model, making it harder to generalize. Building a more complex model (using a Decision Tree regressor, for example), would be a challenge for this reason.