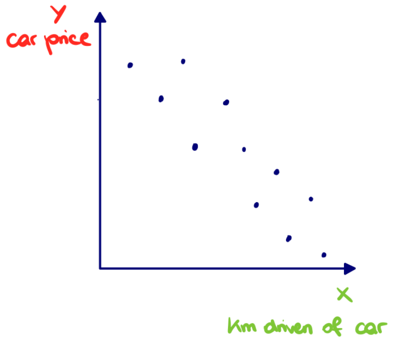
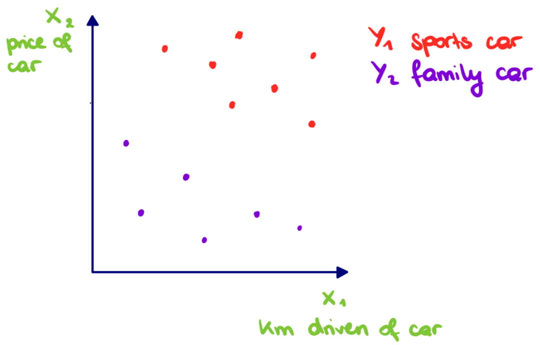
Features, Target

Data contains detail information about an object, which can be a numerical value or category. The detail information is saved as features (denoted X) and are used to explain and/or predict a dependent variable. They are also called: explanatory-, input- , independent-, predictor variables.

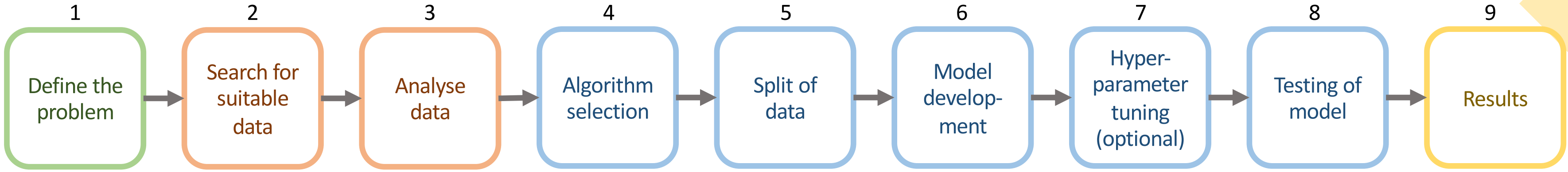


A labeled dataset contains the target (denoted Y). They are used during a supervised machine learning workflow to minimize the error between the model result and true result and to determine the model parameters. They are also called: explained-, output-, dependent-, predicted variable.



Supervised machine learning workflow

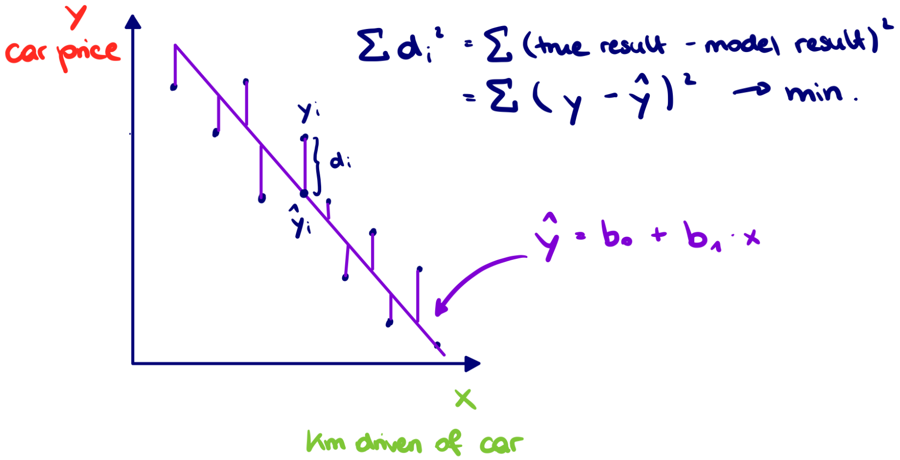
During a supervised machine learning workflow, data with features and their known target values are used. Machine learning models are developed by minimizing the error between model and true result.



1. Which research question should be investigated?
2. Is enough data with needed granularity for all wanted segments available?
3. Is data quality sufficient? How to handle missing values, outliers, invalid entries etc.?
4. Select one or more suitable algorithms
5. Data is separated into training-, test- and possible validation sample
6. Model is fitted to training sample to determine parameters
7. Models with different hyperparameter settings are developed using validation sample and compared to further improve results
8. Model is checked for overfitting on test sample
9. Derive answers for research question from results

Regression

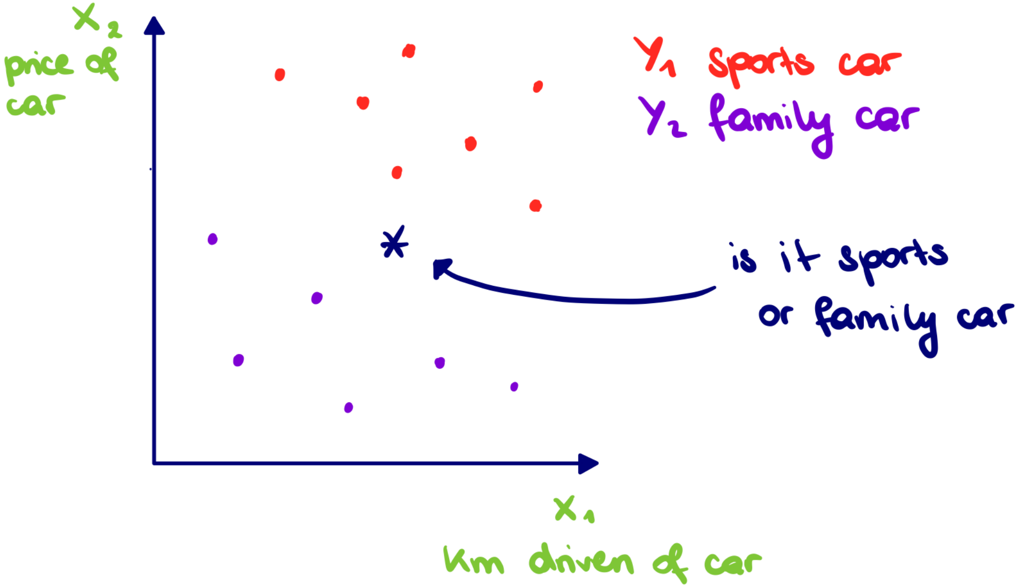
Regression is a process, where a relationship between a variable Y (target) and one or more variables X_i (features) is estimated to predict future values or to determine the strength of the dependency. One simple type is the linear regression, where the relationship between Y and X_i is linear. A common method to estimate the relationship is ordinary least squares (OLS). During this process, the sum of the squared distances between the data points and regression line is minimized and the parameters (intercept, coefficients) are determined.


Y
car price
$\Sigma d_i^2 = \Sigma(\text{true result} - \text{model result})^2$
$= \Sigma(y - \hat{y})^2 \rightarrow \text{min.}$
y_i
d_i
$\hat{y}_i$
$\hat{y} = b_0 + b_1 \cdot x$
X
Km driven of car

Classification

During a classification process, datapoints are categorized to a target class Y_i using features X_i. Examples are:

- Binary: Will it rain or not
- Multi-Class: Detection of letters, numbers
- Multi-Label: Multiple objects are detected on picture


X_2
price of car
Y_1 sports car
Y_2 family car
is it sports or family car
X_1
km driven of car