Stages of a machine learning project

Life stage of the ML project

1 Understanding the problem

2 inventory of data

3 Preparation of the data

4 Data Analyse

5 Modeling

6 Visualization and restitution

7 Deployment and maintenance

Do you have the correct data?

- No data, no machine learning
- Garbage in -> Garbage out
- Does our data look like production data?

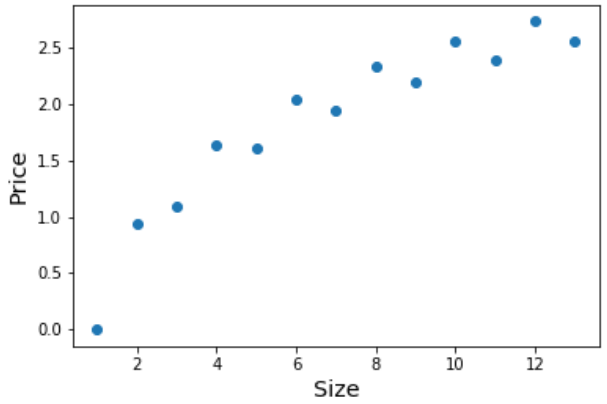
Choose the right evaluation parameters

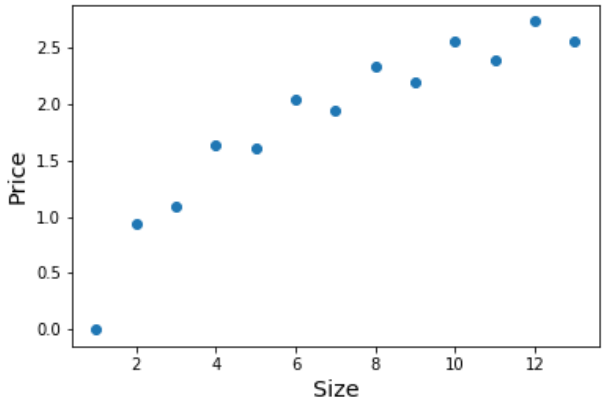
- Is it unique?
- Does it match the company's demand?
- Can it be used as a cost function in a model?

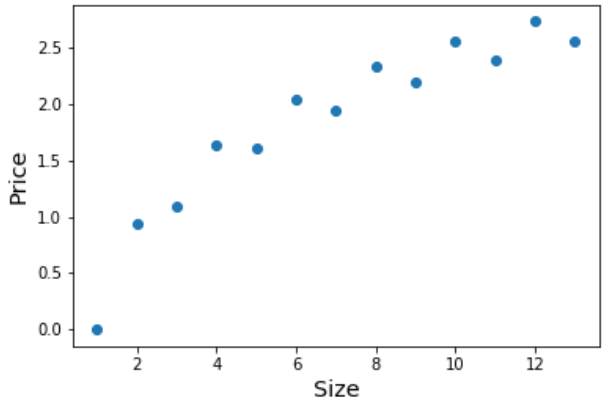
Over and under fitting

- When you use a machine learning algorithm, it is to create a model from training examples. But the goal is to apply your model to new data that your model has never seen.
- Your model is overfitted when your model performs well on your dataset, but has trouble predicting new data.
- Your model is underfitted when your model does not understand your problem well and has trouble performing on your training dataset.

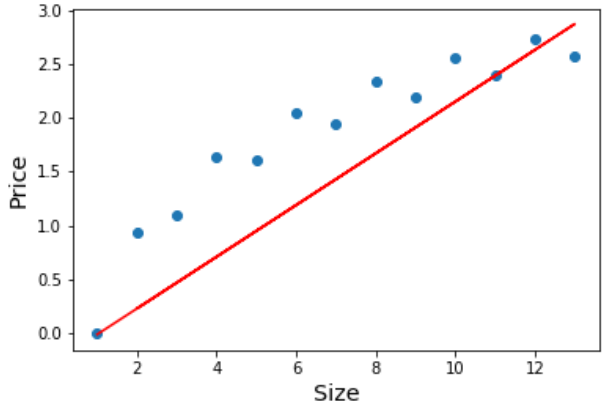
For the regression

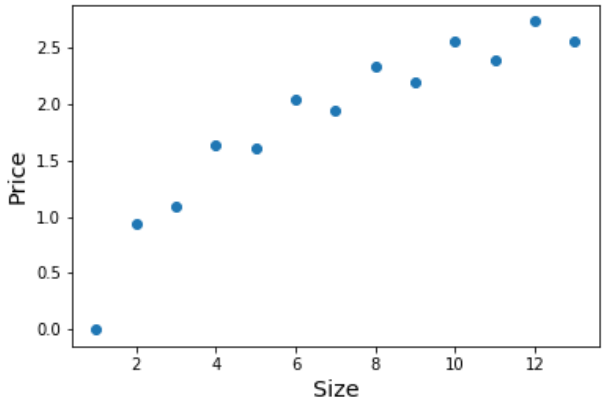


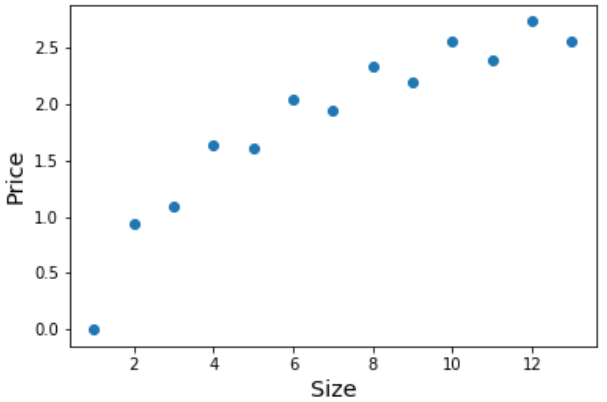




For the regression

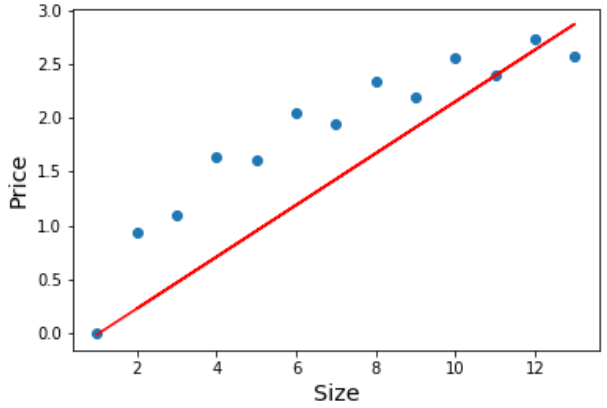




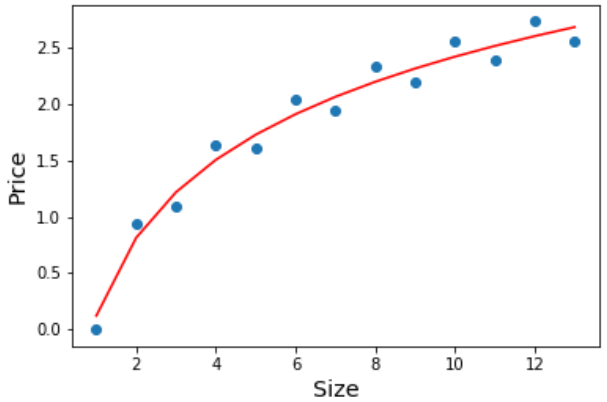


Underfitting

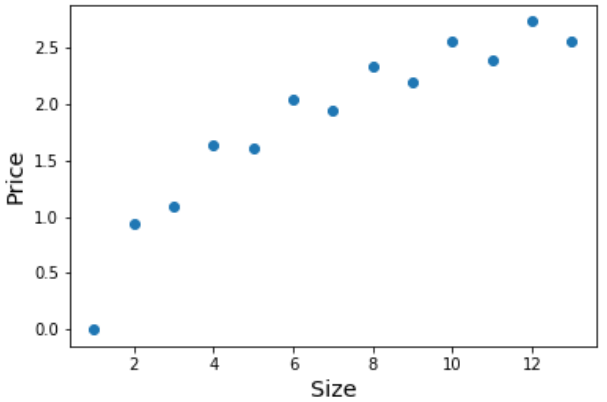
For the regression



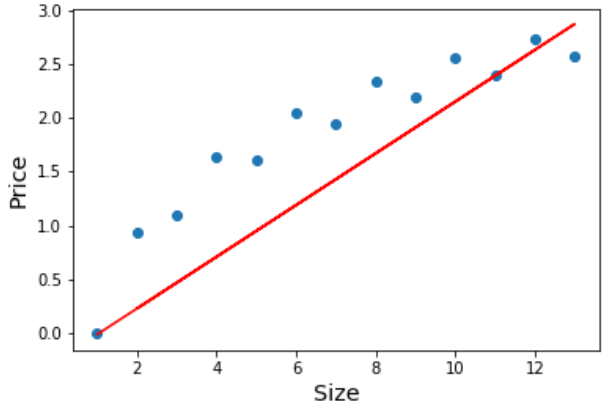
Underfitting



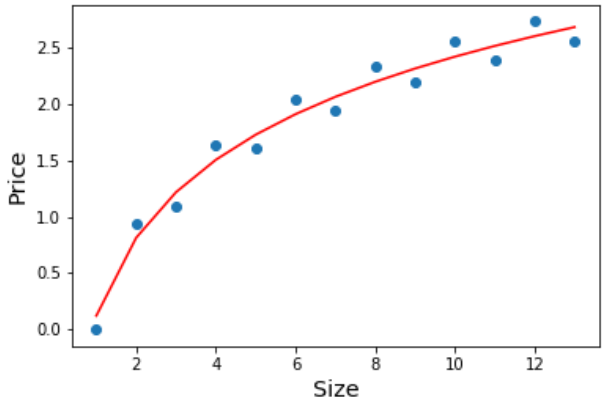
Good fitting



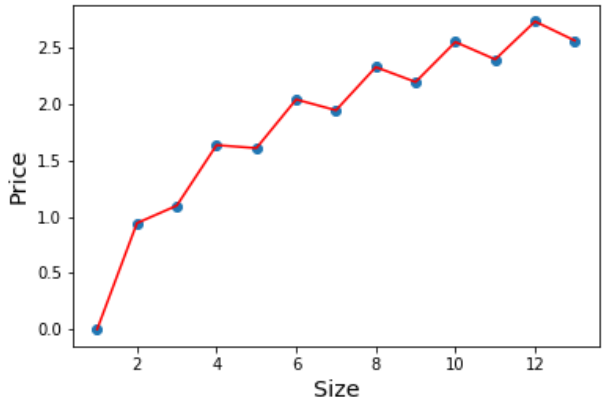
For the regression



Underfitting

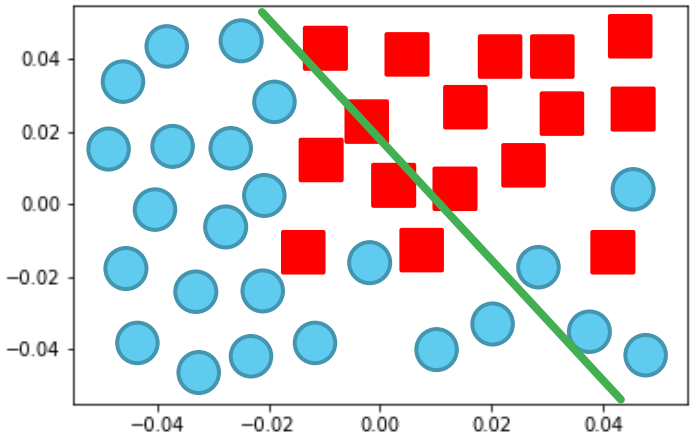


Good fitting

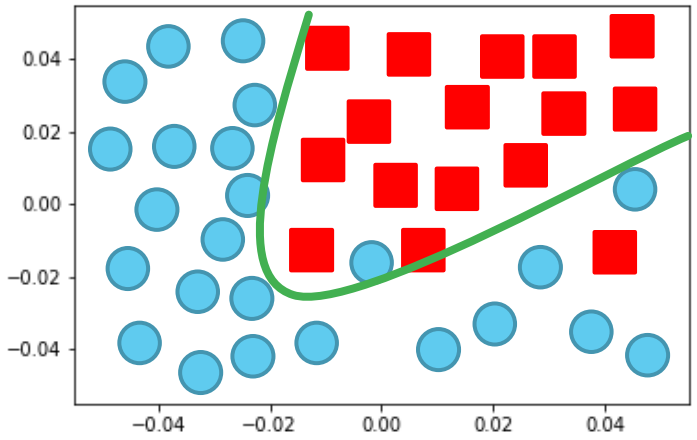


Overfitting

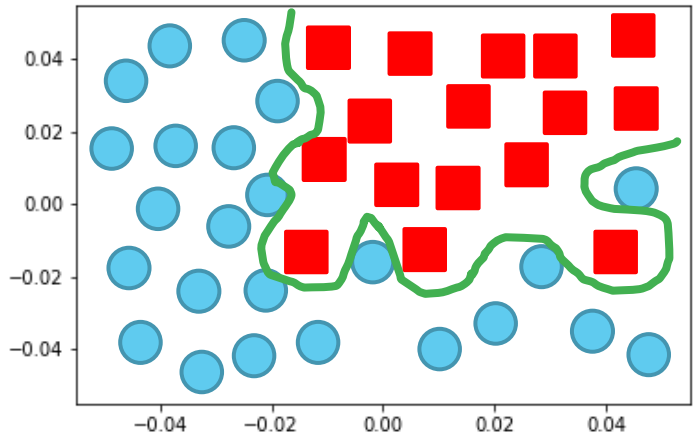
For the classification



Underfitting



Good fitting



Overfitting

Underfitting

- Change the type of model

- Create or collect more variables

Overfitting

- Add regulation
- Collect more observations
- Reduce the number of variables or the complexity of the model

Generalization

- In machine learning, the goal is to create an algorithm that performs well with new data. We call this concept the power of generalization. To measure the generalization of our model, we will predict data that our algorithm has not seen during its training and see how it performs on that set.

Train, validation and test set

	Surface (x_1)	Nb of rooms (x_2)	Years (x_3)	Price (y)
1	70	3	2010	460
2	40	3	2015	232
3	45	4	1990	315
4	12	2	2017	178
…	…	…	…	…
m-2	60	3	2010	390
m-1	35	2	1994	300
m	25	1	2005	240

Training set (70%)

Validation set (10%)

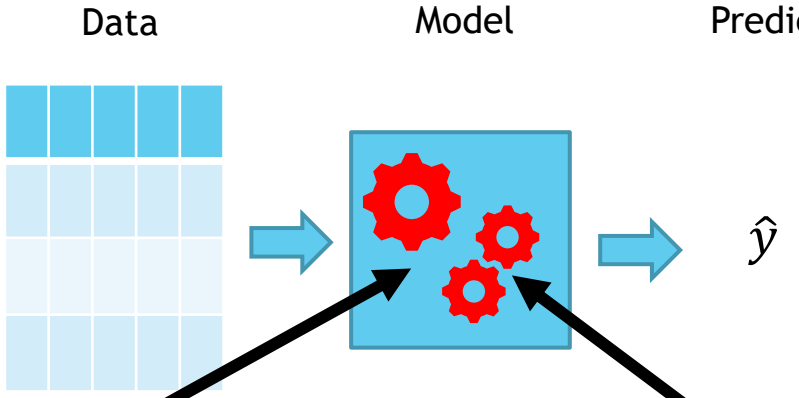
Test set (20%)

House price prediction from training data

Please note, you must create your training, validation and test set randomly!!!

Train / dev / test sets

Paramètres vs Hyperparamètres



Parameters

Changes during training

step 1: Initialize the model

Step 2: Train the model

Step 3: Deploy the model

Hyperparameters

Remains fixed during training

step 1: Initialize the model

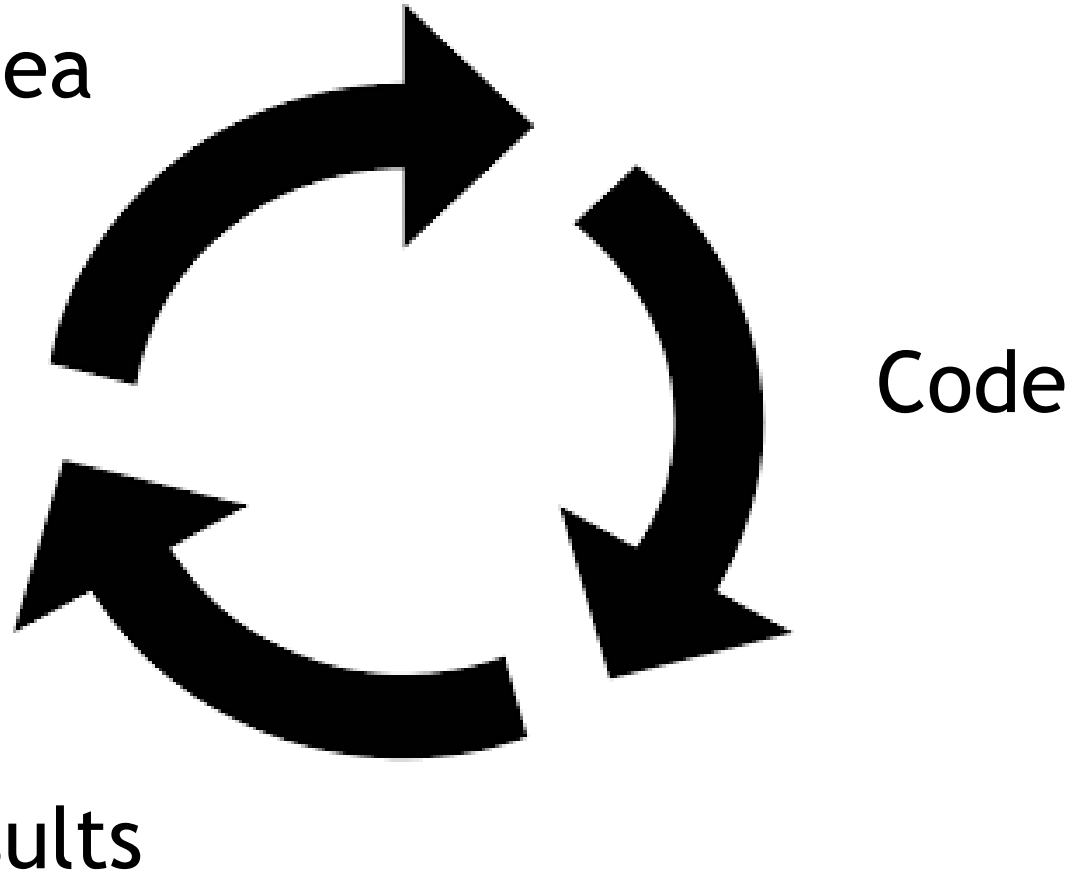
Step 2: Train the model

Step 3: Deploy the model

Reference model

- Existing process
- Expert system with simple rules
- Simplest learning machine model

Iterations



Validation with the test set

- Use the test set only at this stage to prove the performance of the model.

Model in production

- Save your model
- Put your model into production
- Set up a periodic performance evaluation of the model.
- Re-train the model when performance is too low.