AI for Software Engineers
Hands on Activity

SOEN 691: Engineering AI-based Software Systems

Emad Shihab, Diego Elias Costa
Concordia University

Hands on Lecture

Outline

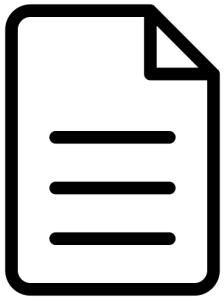
- Data
  - Exploring dataset characteristics
  - Dealing with (some) dataset problems
- Models
  - Explore different models' performance
  - Model fine-tuning
- Model Evaluation
  - Choose appropriate quality metrics
  - Establishing a baseline model
  - Understanding/Explaining the model

What is Machine Learning?

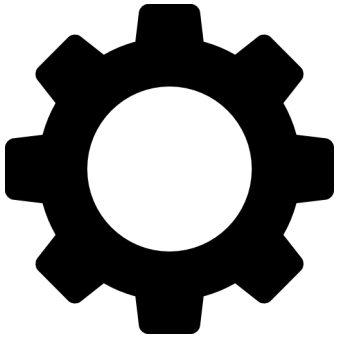
…allowing machines to **learn from the past** (data) to **produce a behaviour/decision**

Key idea: automatically learn without being programmed over and over

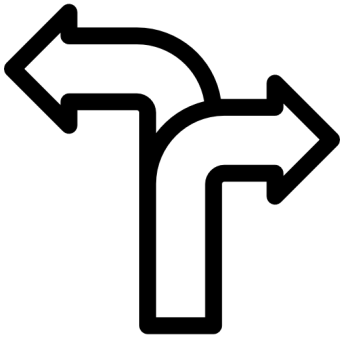
Overview of a ’Typical’ AI/ML System



Data

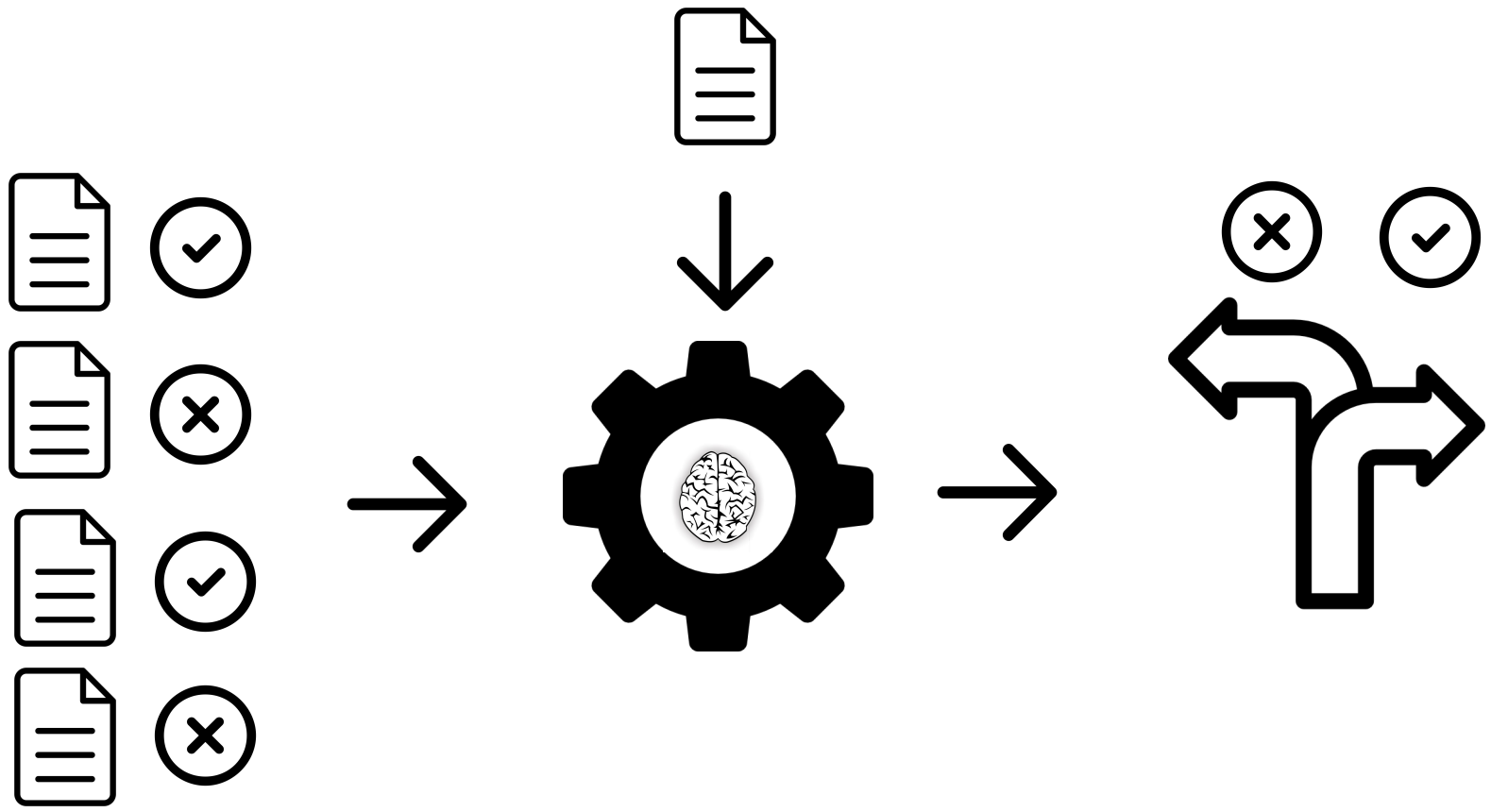


Model

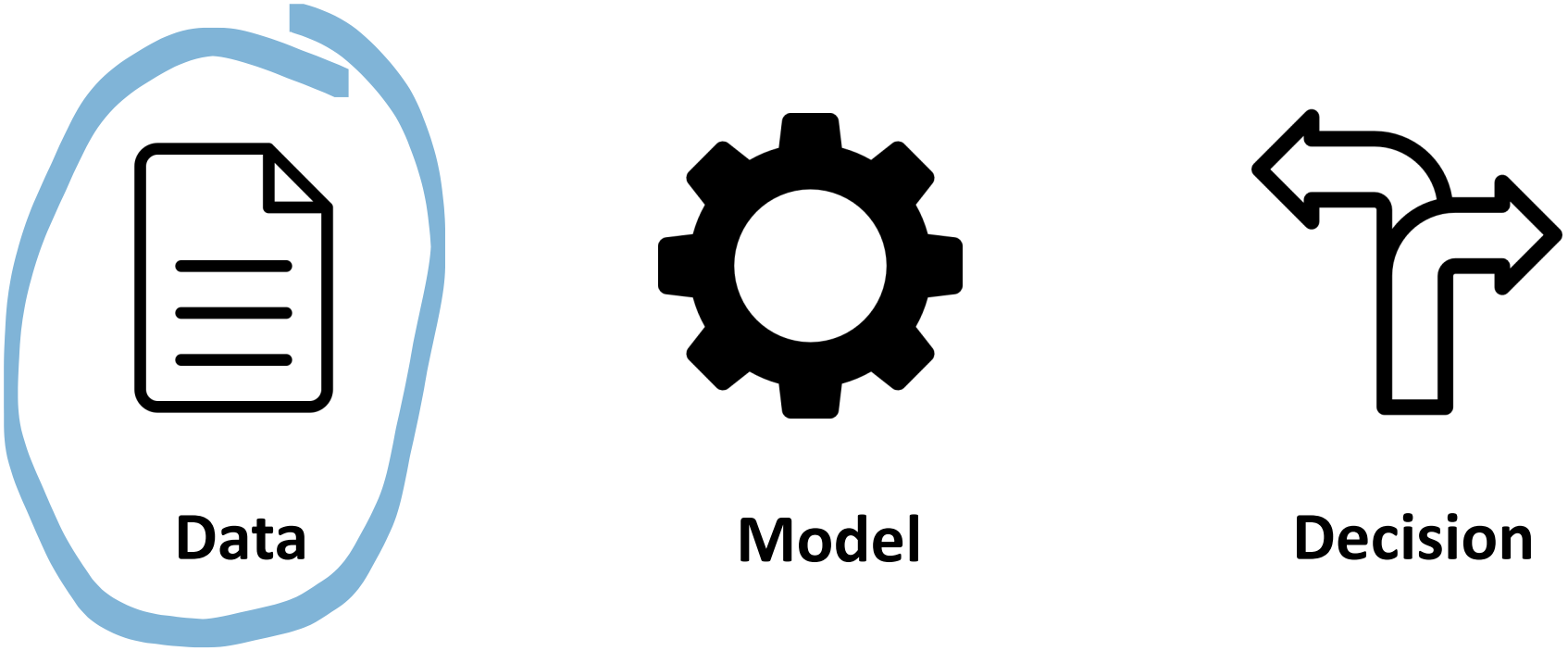


Decision

How ‘Typical’ AI/ML Systems Work



Overview of a 'Typical' AI/ML System



The Role of Data

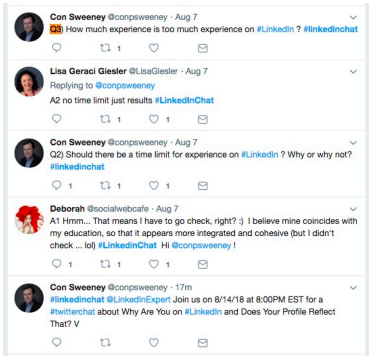
…the corner stone of any AI/ML system

CRM data
Student records
Sales logs
Usually numerical

ID	Name	Phone
1	Alice	555-000-0000
2	Bob	666-000-0000

Structured data

Social media
Audio
Articles
Usually free form text



Unstructured data

Source: https://raventools.com/blog/understanding-twitter-chats/

Structured vs Unstructured Data

Pros

- Typically quantitative
- Mostly machine generated
- Easy to analyze

Cons

- Provides limited insights

Structured data

Pros

- Typically qualitative
- Mostly human generated
- Provides very meaningful insights

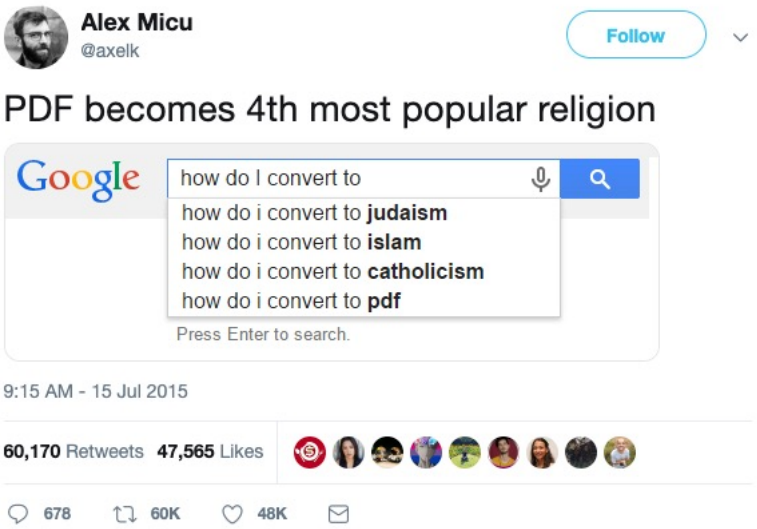
Cons

- Very, very difficult to analyze
- Unstructured -> structured

Unstructured data

A Caution About Data

…your data can significantly bias your AI system



Source: https://twitter.com/axelk/status/621352271150977025?lang=en

Important Factors to Consider About Data

Data gathering:

- Where will we get the **data from**?
- Is the collected data **reliable**?
- Does it properly **represent the observed group**?

Important Factors to Consider About Data (cont'd)

Data cleaning/pre-processing:

- Are there **outliers** in the data?
- How do we handle **missing values**?
- Do we need to **structure** some of the data better?
- Do we need to **convert or group** data?

Important Factors to Consider About Data (cont'd)

Data labeling:

- How is the data **labeled**?
- Are the labels **correct**?

- **80/20 rule:** 80% effort is spent on collecting and preparing data, 20% on machine learning

- **Data vs Analytics:** Most data in its raw form is not useful. Data becomes interesting when you use it to build analytics.

Hands-on: Credit Report

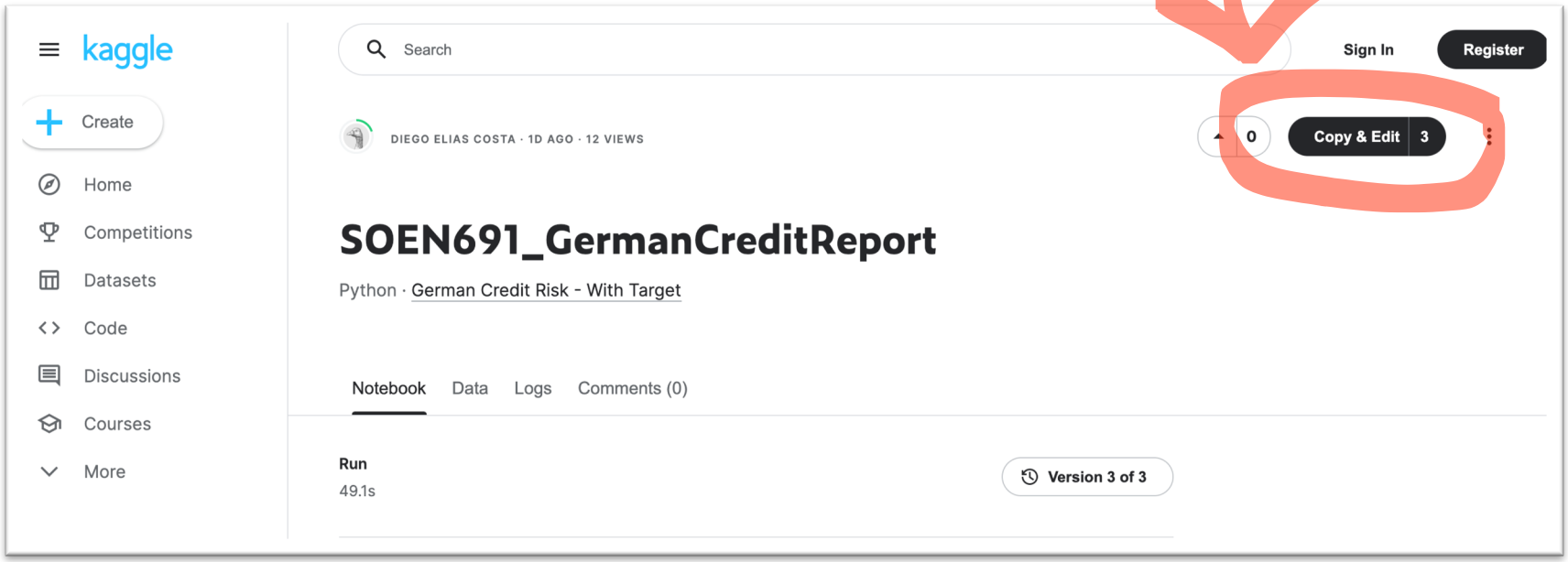
Scenario

- Bank users request a credit for a purchase
- Bank has tons of information about each client
- Analysts uses info of the client to classify the request into:
  - Good (low risk of default)
  - Bad (high risk of default)

- Can this be automated by ML?

Opening the notebook

1. Access the notebook in Kaggle (link in zoom)
   - Should have an active Kaggle user
2. Click on Copy & Edit



What is the quality of our dataset?

Explore the characteristics of the dataset to answer the following questions:

- How much data do we have?

- Do we have any missing data (Nan values)?

- What is the distribution of the target variable?

- What are the types of features in the dataset?

What is the quality of our dataset?

Explore the characteristics of the dataset to answer the following questions:

- How much data do we have?
  - 1000 records + 9 features + target variable
- Do we have any missing data (Nan values)?
  - Yes, Savings Account + Checking Account
- What is the distribution of the target variable?
  - Imbalanced ~70% good credit / 30% bad credit
- What are the types of features in the dataset?
  - 4 numerical + 5 categorical variables

Analyzing the distribution and relationship of features

Explore the distribution of features:

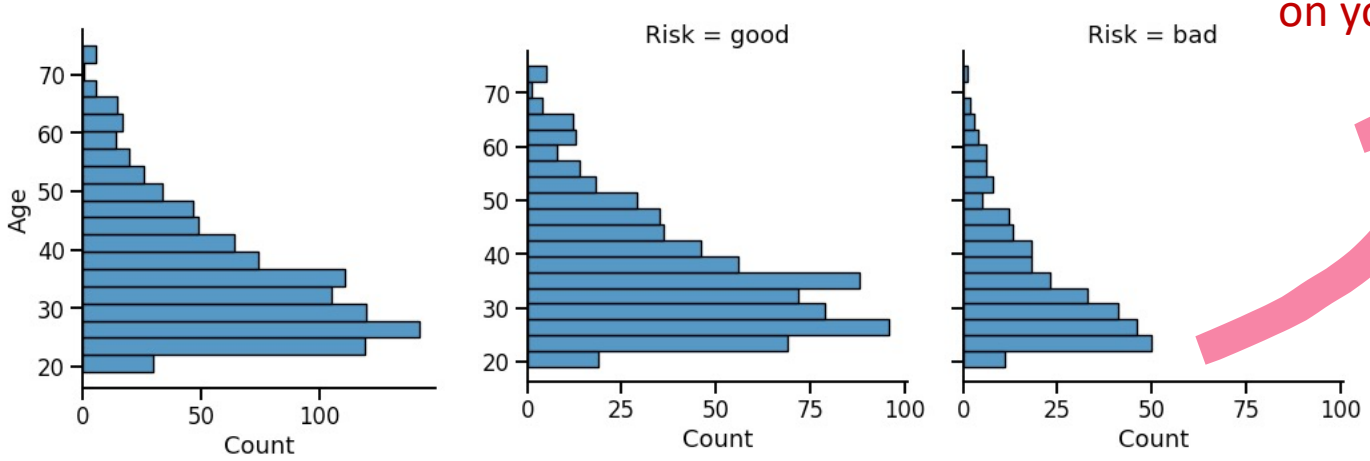
- Do we have a biased dataset?
- How some features relate to good/bad credit?

Examples of analyses:

- Age + Sex vs Risk
- Age + Checking Account vs Risk
- Age + Saving Account vs Risk
- Age + Jobs vs Risk

Analyzing the distribution and relationship of features

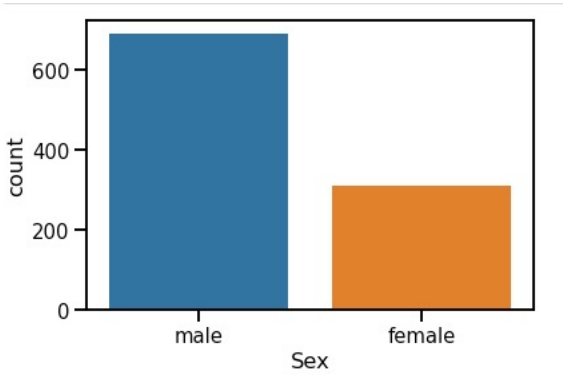
- Example of some analyses (Age)

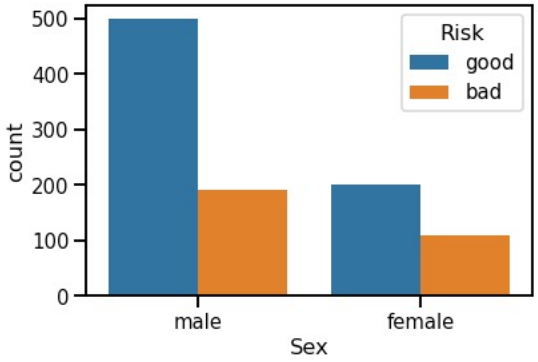


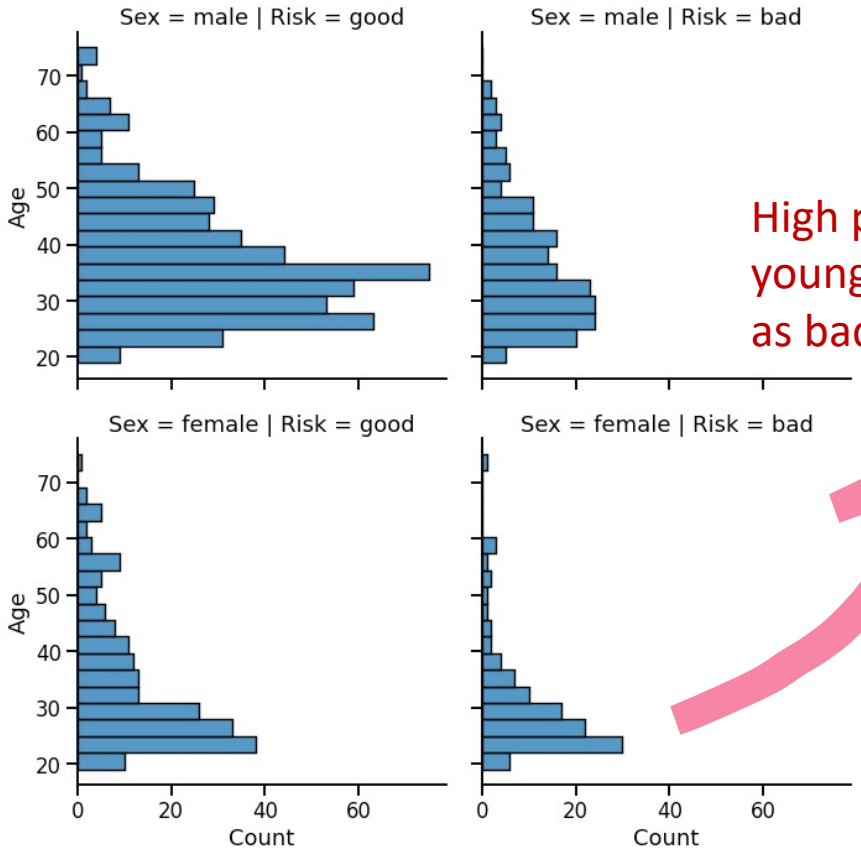
Records classified as bad are concentrated on young people (<30)

Analyzing the distribution and relationship of features

- Example of some analyses (Sex)



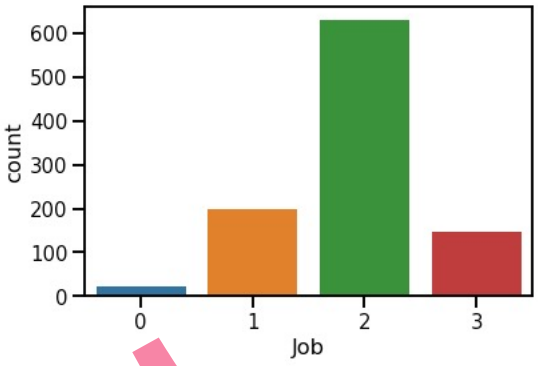




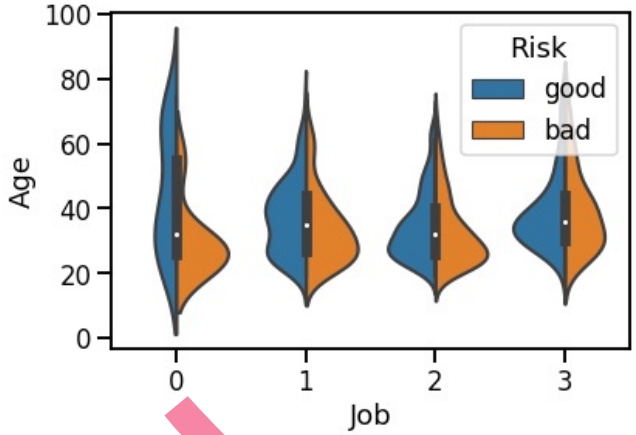
High proportion of young female classifie as bad credit

Analyzing the distribution and relationship of features

- Example of some analyses (Job)

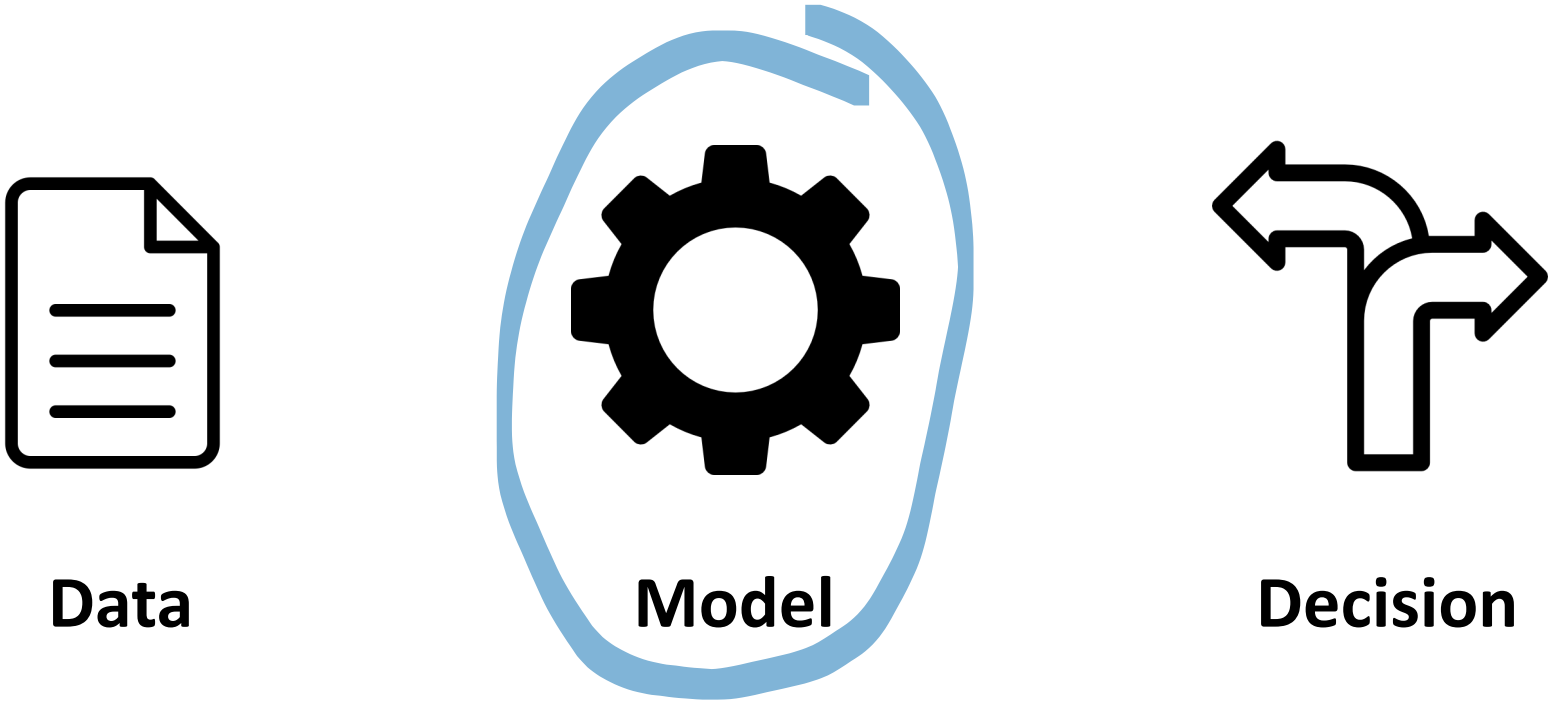


Very little data with unemployed clients



Frequently classified as bad credit

Overview of a 'Typical' AI/ML System



Main Categories of ML Models

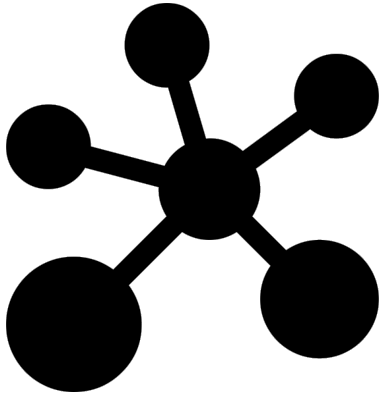
Supervised learning models: The model trains on a set of labeled training data and classifies future, unseen data based on its **training**

Unsupervised learning models: there is no training. The model **analyzes the data to find patterns** and groups similar data points

Example of ML Models

K-means Clustering (Unsupervised)

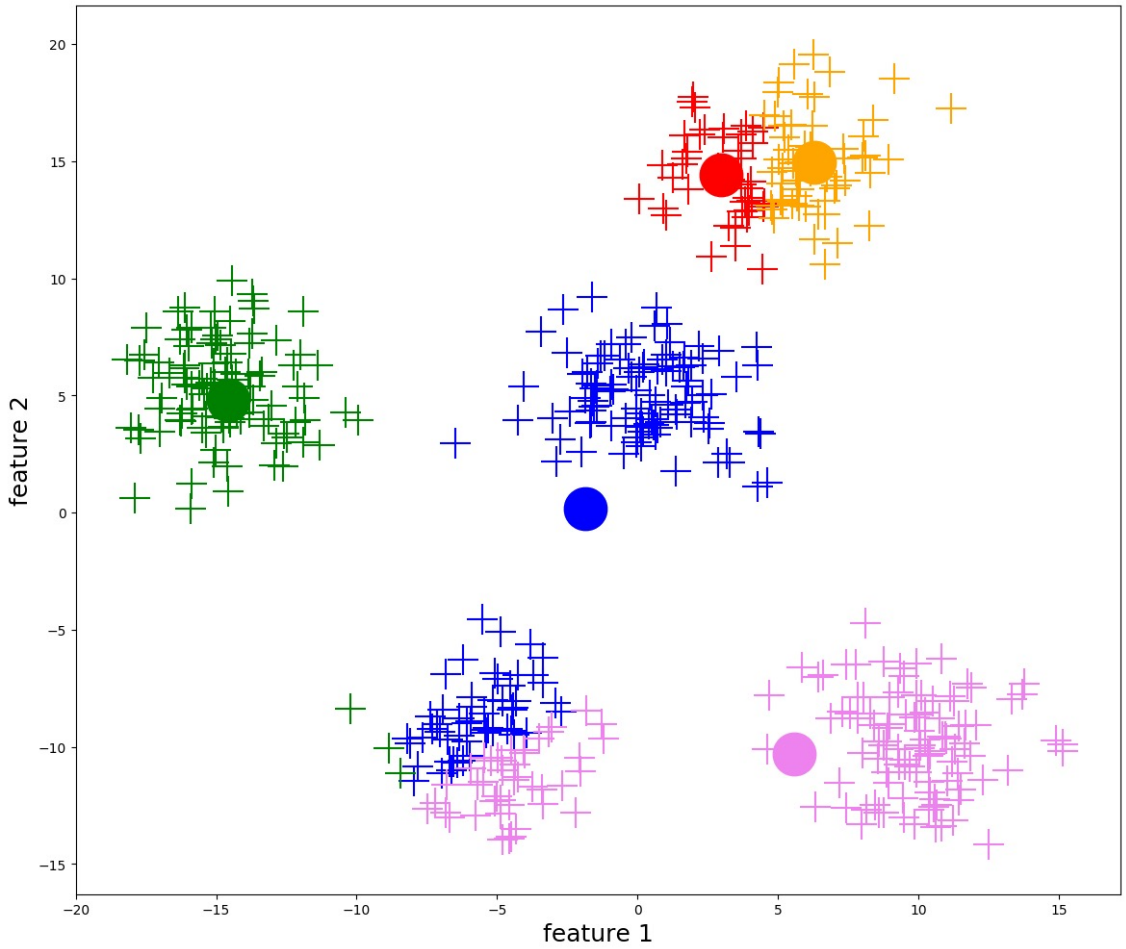
Idea: Group unlabeled data into K clusters



How?

- User provides as input K, the number of clusters
- Centroids are picked and distance is measured between each data point
- Iterate until distance is minimized and K clearly defined clusters emerge

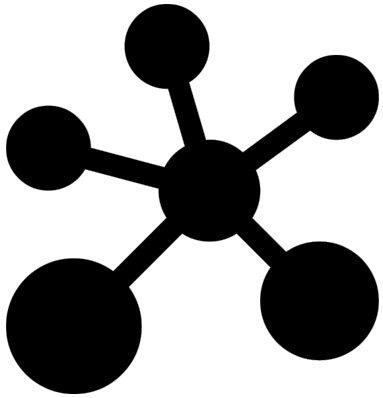
K-means Clustering



K-means clustering

Pros

- No need for labelled data
- Simple algorithm

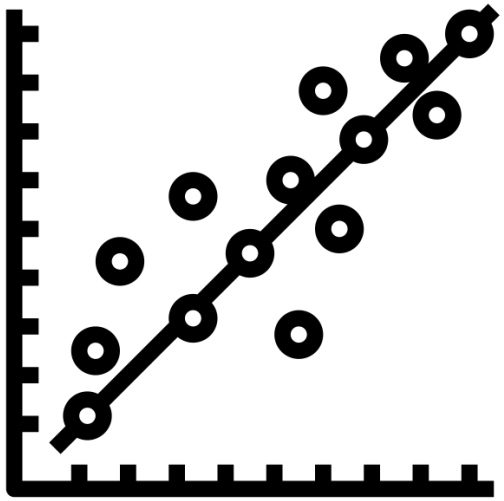


Cons

- K needs to be determined a priori
- The clusters will still need to be tagged afterwards

Linear Regression (Supervised)

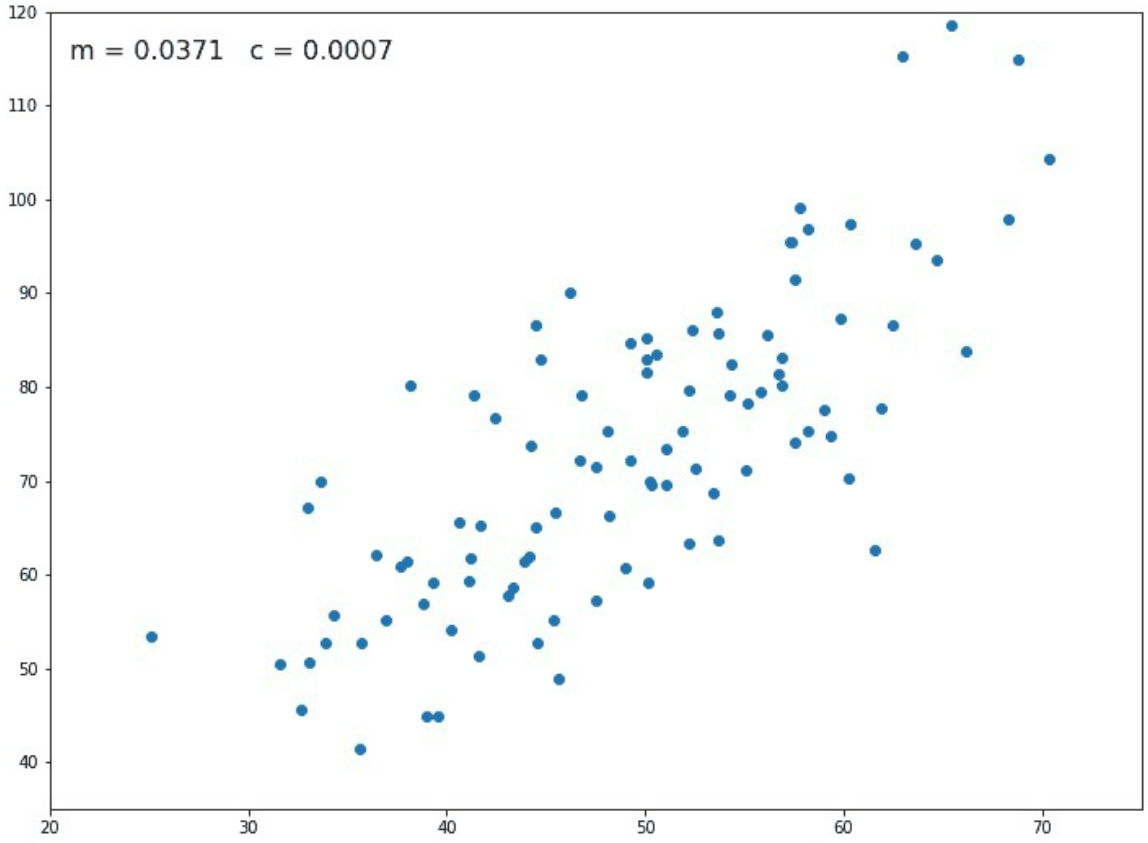
Idea: Use statistical model to represent relationship between 2 (or more) variables



How?

- Use part of the data and fit a line
- Choose line to minimize error
- Outcome is a value, e.g., height, price, etc.

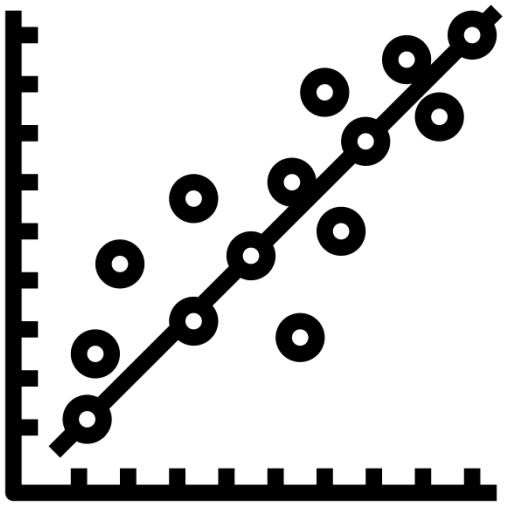
Linear Regression



Linear Regression

Pros

- Simple and explainable model
- Very popular, even today



Cons

- Assumes a **linear relationship** between the explanatory and response variables

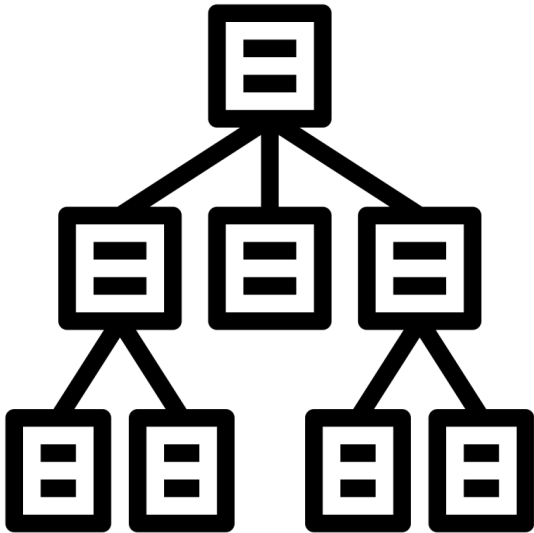
- Need to carefully consider **distribution/independence of input data**

Decision Trees (Supervised)

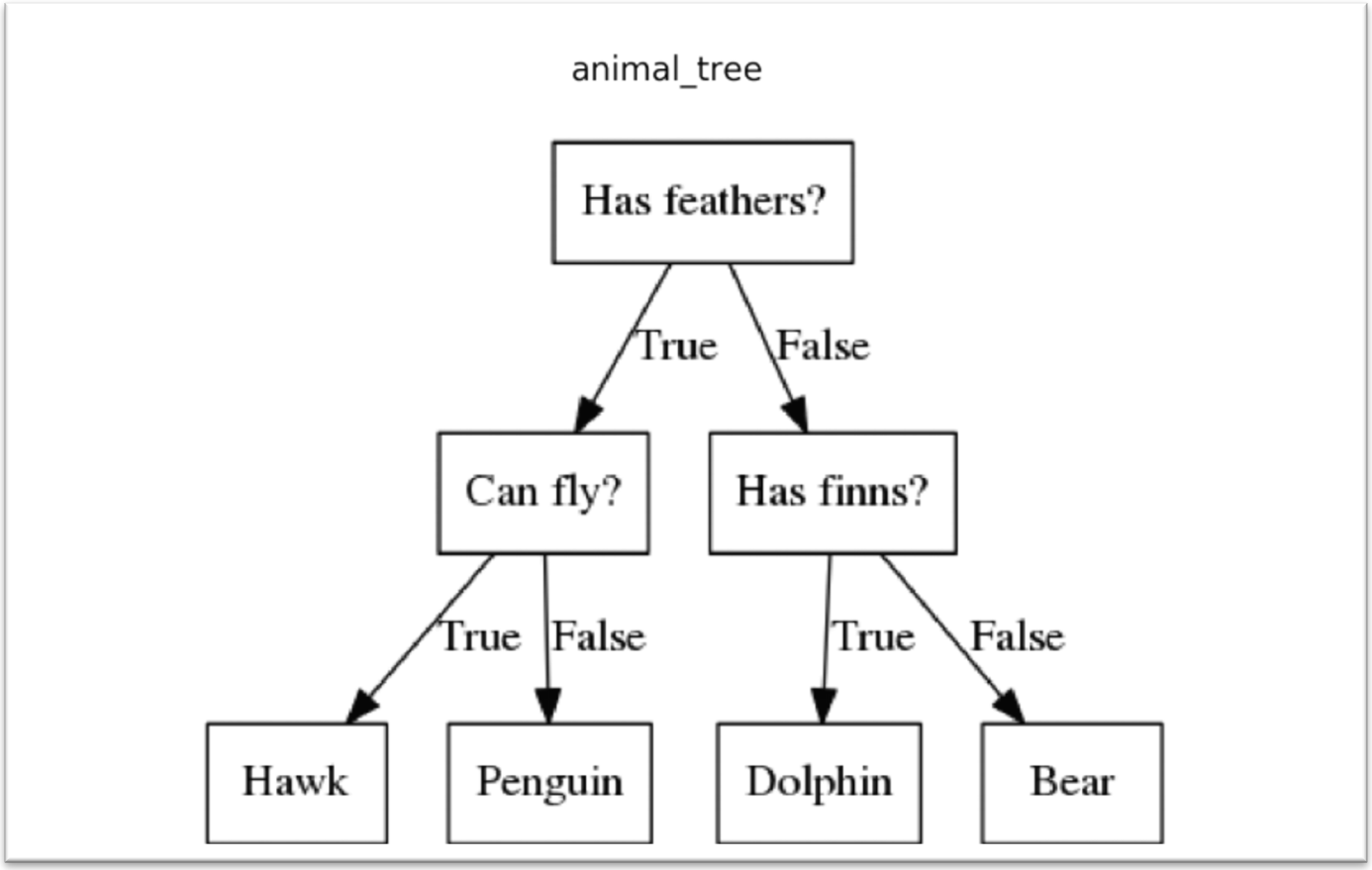
Idea: Use a flowchart tree structure to represent the relationship between features and outcomes

How?

- Select best attribute to split data into subsets
- Repeats recursively for each child
- Nodes -> features,
  Branches -> decision rules,
  Leafs -> outcomes



Decision Tree



https://towardsai.net/p/programming/decision-trees-explained-with-a-practical-example

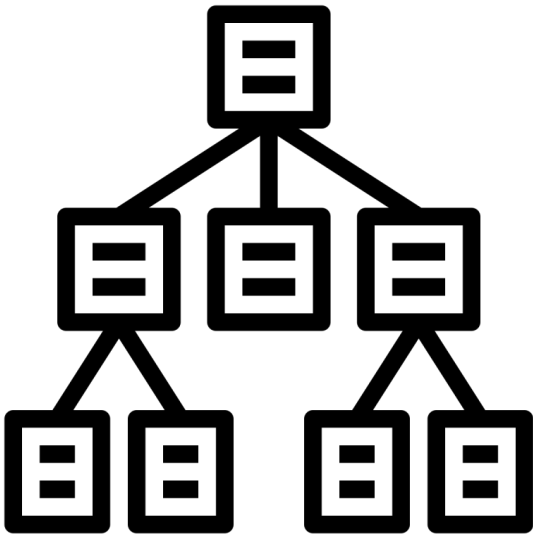
Decision Trees

Pros

- Easily explainable decisions and features
- No assumptions on data distribution
- Can capture non-linear patterns

Cons

- Biased with imbalanced datasets
- Less accurate than other ML algorithms



Different models for different problems

- Grouping unlabeled data
  - Unsupervised (K-means clustering)

- Predicting the next value (continuous)
  - Regression model (Linear Regression)

- Predicting the best class/decision
  - Classifier model (Decision Tree)

Important Factors to Consider about ML Models

- **Task to solve:**
  - Regression: output is a numerical value.
  - Classification: output is a class probability
  - Clustering: better understand unlabeled data

- **Type of input data:**
  - Structured vs unstructured
  - Numerical
  - Categorical
  - Boolean, etc.

Important Factors to Consider about ML Models (cont'd)

- **Data labelling:** do we have good quality labeled data (i.e., should we use supervised/unsupervised models)
- **Model assumptions:** are there specific assumptions on the data or the model
- **Performance:** Does the model perform well for the problem at hand?

Important Factors to Consider about ML Models (cont'd)

- **Explainability:** are the decisions being made explainable?
- **Stability:** how does the model perform over time? How can we ensure the model does not drift?
- **Overfitting:** does the model overfit the data?

Preparing the dataset for modeling

Features come with different formats

1. How to handle missing values?
2. How to encode categorical features?
3. How to extract meaningful features from raw data?

We will walk through this process together.

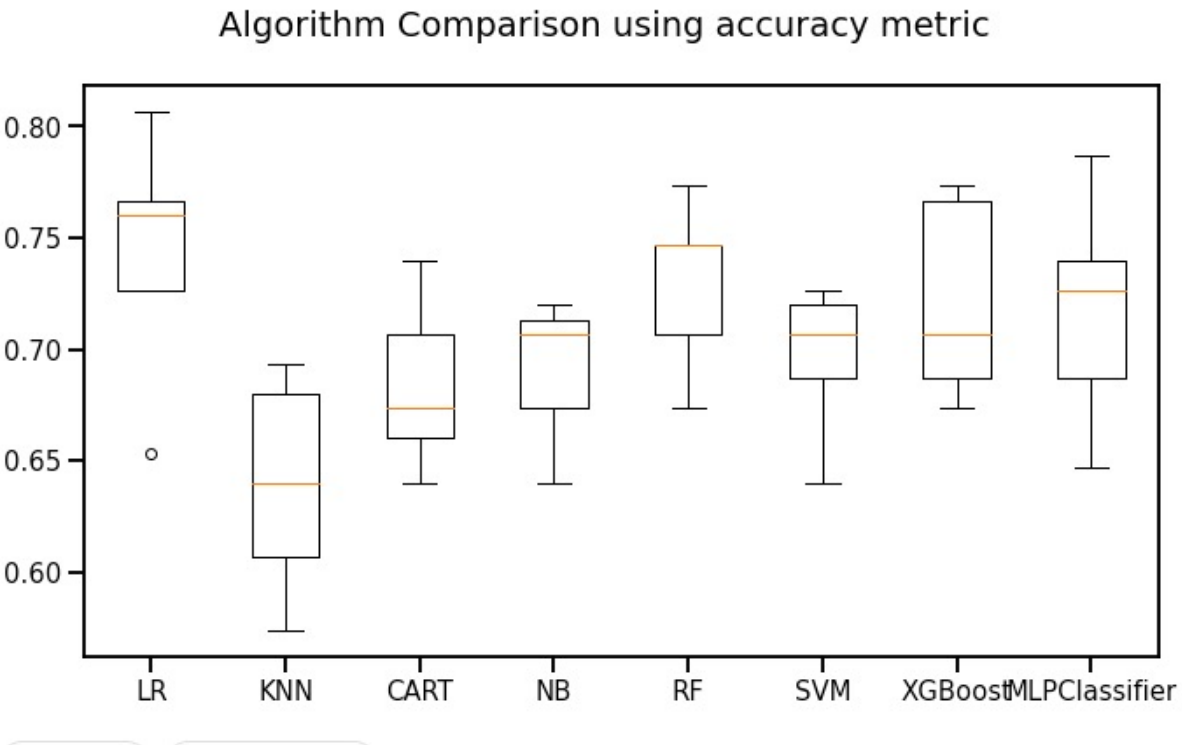
What models give the best performance?

Let us explore how some models perform in our task.

1. Choose one model from the code
2. Run the classifier and report the performance in the zoom chat!
3. Read their respective documentation and try to fine-tune some of its parameters

What models give the best performance?

Using accuracy + default parameters

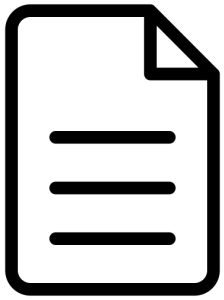


What is the real performance of our model

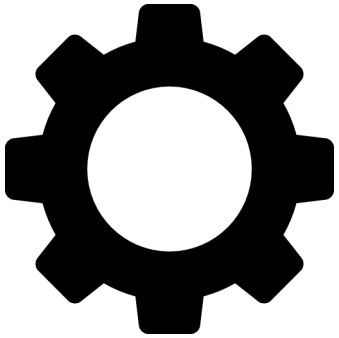
We have only explored the performance on the training data

1. Choose the best model you evaluated
2. Evaluate the performance in the test set
3. Compare the performance with some baselines

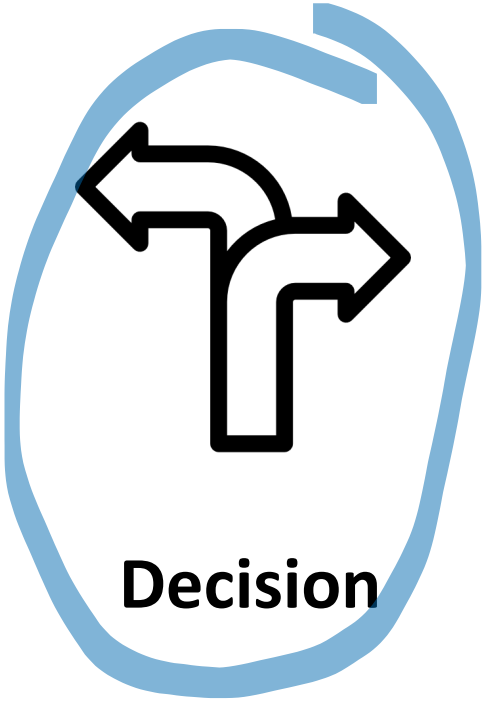
Overview of a 'Typical' AI/ML System



Data

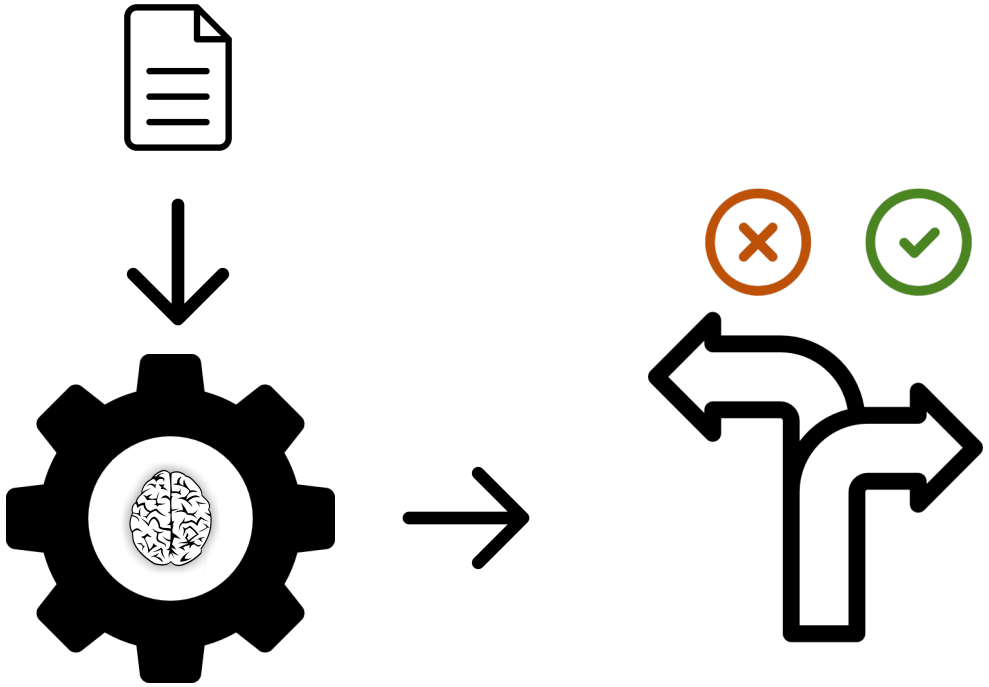


Model



Decision

Measuring Performance



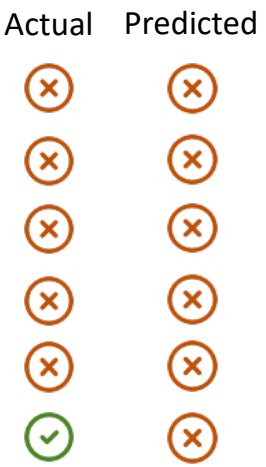
Measuring Performance Using Accuracy

Actual	Predicted	
✓	✓	TP
✗	✓	FP
✓	✓	TP
✗	✗	TN
✓	✗	FN
✗	✗	TN

Actual	Predicted
✗	✗
✗	✗
✗	✗
✗	✗
✗	✗
✓	✗

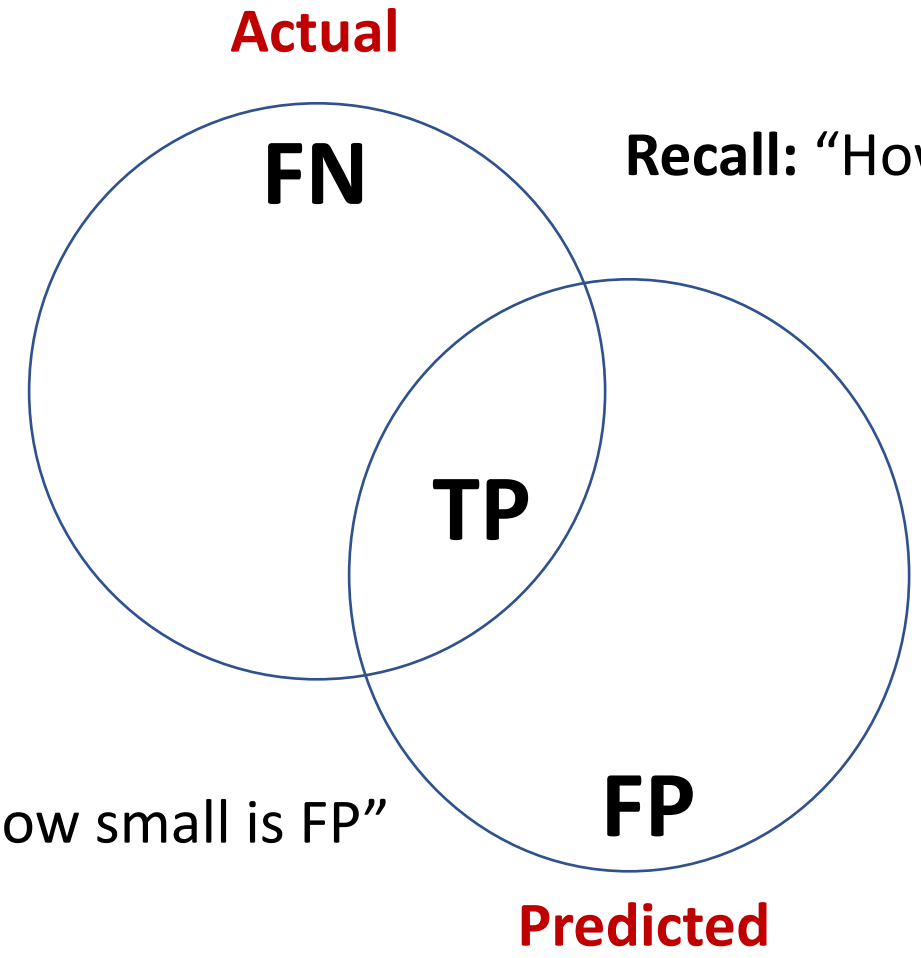
Measuring Performance Using Accuracy

Accuracy: (TP+TN)/(TP+FP+TN+FN)
= 4/6 = 66.67%



Accuracy: (TP+TN)/(TP+FP+TN+FN)
= 5/6 = 83.34%

Measuring Performance Using Precision and Recall



Recall: "How small is FN"

Precision: "How small is FP"

Measuring Performance Using Precision and Recall

Accuracy: (TP+TN)/(TP+FP+TN+FN)
= 4/6 = 66.67%

Precision: TP/(TP+FP) = 2/3 = 66.67%
Recall: TP/(TP+FN) = 2/3 = 66.67%

Accuracy: 5/6 = 83.34%

Precision: TP/(TP+FP) = 0%
Recall: TP/(TP+FN) = 0%

When to Use Different Metrics?

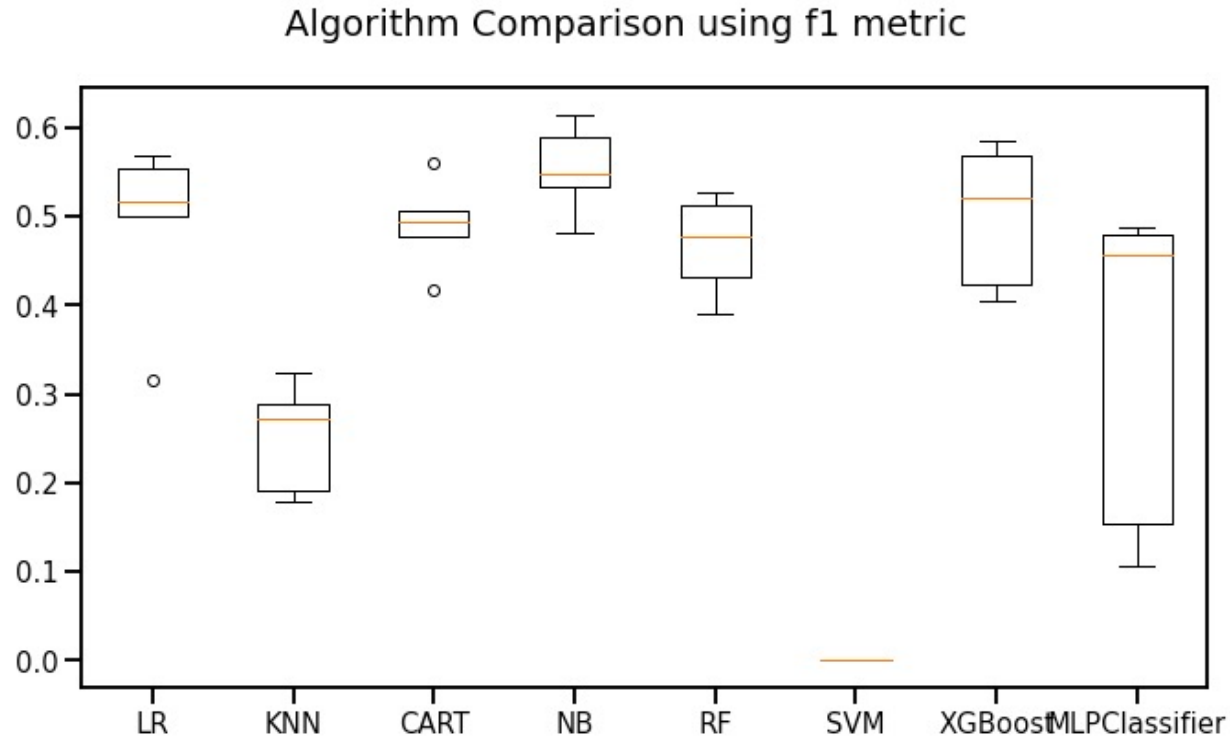
- Accuracy
  - Very informative in balanced datasets
- Precision
  - The precision of the decision is the priority
- Recall
  - Finding all the positive cases is the priority
- F1 score
  - Harmonic mean between precision and recall
  - Values equally precision and recall

What is the real performance of our model

We have only explored the performance on the training data

1. Choose the best model you evaluated
2. **Choose an appropriate performance metric**
3. Evaluate the performance in the test set
4. Compare the performance with some baselines

Revisiting the performance of models



Understanding the model

We can inspect (and learn) from the model:

- The most important features
- The probabilistic curve per feature
- Explain certain predictions

Project Homework

- Sync with your project group
- For next week (due Friday, Jan 28 at noon), I would like you to submit the following on Moodle:
  - Name and email of a project leader
  - A project title and short problem statement
  - A list of AT LEAST 3 research questions related to your problem statement
  - A list of AT LEAST 10 related papers (from 2016 and after)
    - For each paper, provide a sentence of how the paper is relevant to your problem statement

Homework

- Two papers are posted on Moodle
- For one of the papers, write a summary (aprox. 1/3 of a page)
- For the other paper, write a critique, which includes a summary, at least 3 strong points and at least 3 weaknesses (aprox. 1 page).
- Submit your summary and critique on Moodle by Friday, Jan. 28 at noon

Parting Thoughts

- Building AI systems needs careful consideration
- The data is more important than the ML algorithms*
- Choose the right algorithms, since most have many intricate assumptions
- Validate externally and look out for potential bias

*based on personal experience