Machine Learning Workflow

Day	Monday
Tasks	Familiarize yourself with Kaggle; explore the dataset; run a dummy model

What is Machine Learning?

Q: What examples of machine learning use cases did you encounter this weekend or this morning so far?

A:

- Shopping on amazon: recommendations
- Alexa / Siri / virtual assistants
- Voice recognition
- Cancer detection / computer vision / deep learning

> “Field of study that gives computers the ability to learn **without being explicitly programmed**.”
>
> Arthur Samuel (1959)

Q: What does it mean for a machine to learn?

A: Use *past data* to detect *patterns*, to predict *future outcomes*, improve from *experience*.

Types of Machine Learning

Supervised Learning

- We have prior knowledge on the output values for our dataset — *labels,* `y` (cats / dogs, survived / did not survive, price of a house, bike sharing demand)
- Two types of problems:
    - Classification — predicting a *class* for data point: cat / dog / mouse / horse, survived / did not survive; **week 2**
    - Regression — predicting a (continuous) *value* for a data point: price of a house, bike sharing demand; **week 3**
- Data is divided *training* / *validation* / *test* datasets. We do this to make sure the model is generalizable.

Unsupervised Learning

- Data is *without labels*.
- Models look for inherent structure (clusters, dimensionality reduction).
- More difficult to evaluate than supervised learning models.

→ *whiteboard: draw the supervised ML model* ←

Your Project for the Week

Goals:

- Based on passenger information data, build a model that predicts whether a person had survived the Titanic disaster or not.
- Submit your predictions to Kaggle.

Data:

`train.csv` : Your dataset for this weeks project. This is the dataset you are training and evaluating your model on.

`test.csv` : The dataset Kaggle uses to evaluate your model. It *does not* have labels.

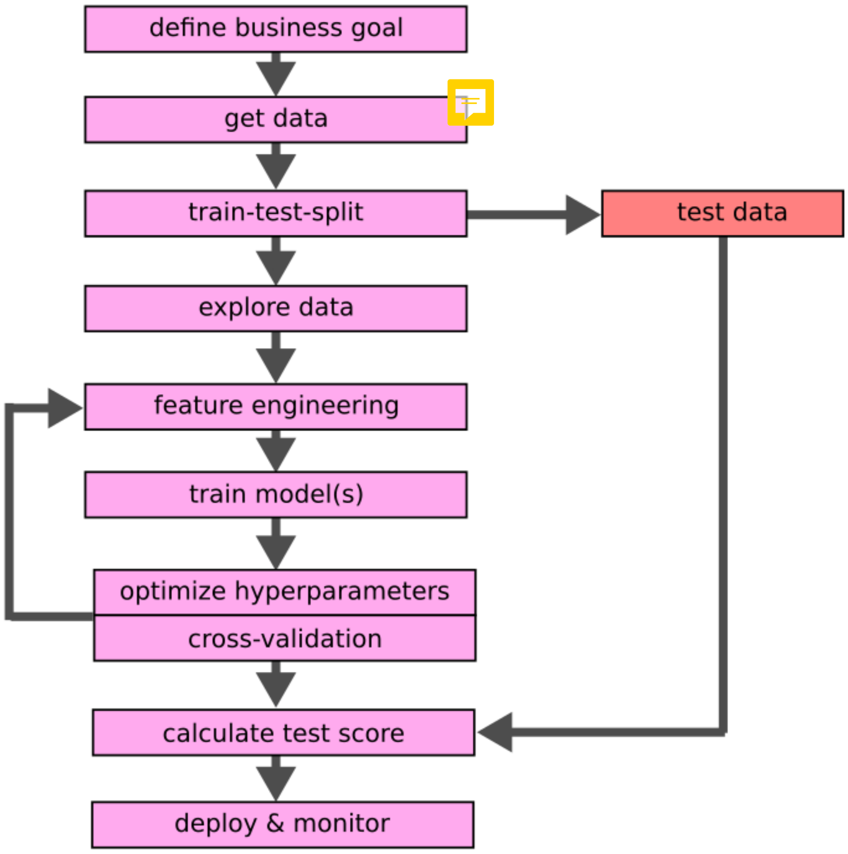
Note: The naming of datasets this week is **confusing**.

- The `train.csv` file is your **full** dataset for this week. *You will need to split this dataset* into training and test sets — training set to build your model on, and test set to evaluate how your model is doing. In real life (non-Kaggle competition life) this would just be called something like `data.csv`.
- The `test.csv` is not the "real" test set in that it doesn't give you labels. This is just a dataset Kaggle uses to evaluate your model. In real-life/job you won't be submitting your ML models to competitions and so you won't have an equivalent of this (but you will still do train/test split on your whole dataset as explained in the previous bullet point). You can think of this as `kaggle_submission.csv`.

Evaluation:

- Don't pay attention to all the perfect scores on Kaggle — they are all cheaters!
- These are the scores you can expect from the model you build:
    - EASY: `70%`
    - MEDIUM: `75%`
    - MEDIUM-HARD: `77.5%`
    - HARD: `79%`
    - VERY HARD: `over 80%`

Machine Learning Workflow



1. Define business goal ✔️
    - Create a model that predicts if a person survived the Titanic disaster or not; optimize for accuracy.
2. Get data ✔️
    - From Kaggle, or from `week_02/data` folder.
3. Train-test-split: ✔️
    - Use `scikit-learn` library in Python
    - Use `train.csv` for this step
    - If you want to keep your labels together with your data while doing EDA, you can use the code example on the bottom, then separate your X and y later.

```
from sklearn.model_selection import train_test_split
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.20, random_state
=42)
```

```
from sklearn.model_selection import train_test_split
df_train, df_test = train_test_split(df, test_size=0.20, random_state=42)
```

4. Explore data ✔️
   - Plot all the things!
5. Feature engineering ✔️
   - Tomorrow
6. Train model ✔️
   - Figure out how to run the dummy model from scikit-learn on your data
   - This afternoon you'll learn how to run logistic regression

7-...

- Later this week

→ *notebook: ML workflow* ←

Your Tasks

- Download and explore the data (do *Exploratory Data Analysis challenge* in 2.1. What is Machine Learning)
- Familiarize yourself with Kaggle.
- Run the scikit-learn dummy model on your data (or logistic regression after the afternoon encounter).