ML - Day 0 - Findings

Iván Krasowski Bissio

July 19th, 2021

Kaggle

- Kaggle is a platform for Machine Learning competitions!
- *Titanic* and *House Prices* are two proposed competitions for starting.

Titanic

- *Titanic* competition aims for us to be able to predict whether a person will survive or not the crash.
- Provided data is, for each passenger: PassengerId, Pclass, Name, Sex, Age, SibSp, Parch, Ticket, Fare, Cabin, Embarked
- Expected output is passengerId $\rightarrow$ survived 0, 1.
- *train.csv* includes data for training (819 examples); *test.csv* includes data for testing (418 examples).
- *gender_submission.csv* is an example for the output expected at the competition, except it predicts that all women survive and all men do not.

Model: Random Forest

- The Random Forest classifier can be used from the scikit-learn library (*sklearn.ensemble*).
- Parameters it needs for the prediction: *n_estimators*, *max_depth*, *random_state*.
- It will fit the selected features from the input data (X: initially *['Pclass', 'Sex', 'SibSp', 'Parch']*) considering the provided labels "Survived" (y).
- Parameters and selected features can be tweaked, trying to improve performance.

House Prices

- TODO (It was too late already)