Introduction to Machine Learning

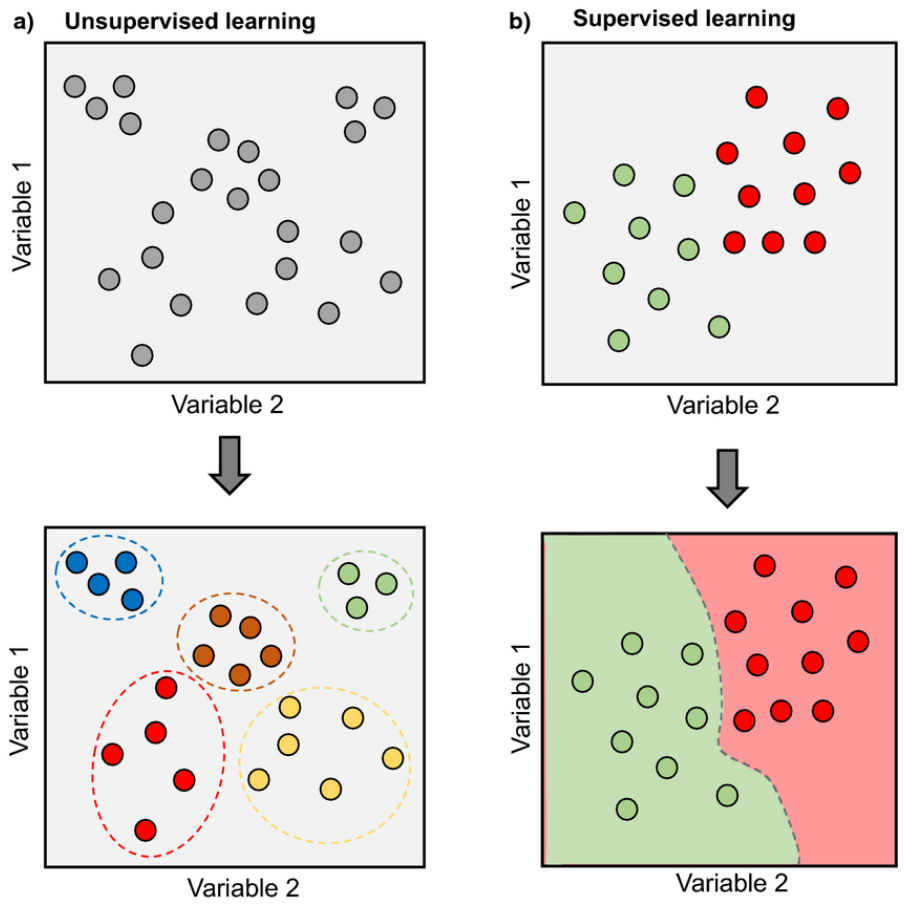
Daniel Hopp

UN Datathon 2023

November 2023

What is machine learning?

- Machine learning creates new outputs from data without having to be specifically programmed to do so
- Unsupervised learning creates outputs without knowing the answer, so to say (e.g., clustering, anomaly detection, etc.)
- Supervised learning creates outputs where a ground truth is known, this ground truth is what the model is trained on



Classification vs. regression

- There are two main types of problems in supervised learning, classification and regression
- Classification: the target variables can be grouped into discrete categories (cat, dog, etc.)
- Regression: the target variables occur in continuous values
- In the official statistics context, regression problems are more widespread (e.g., even OLS can be considered machine learning)

What's the difference between ML and statistics or econometrics?

- There is a large degree of overlap between ML and statistics/econometrics, with a fuzzy boundary between the two
- Some people say it is down to the types of models used (e.g., a decision tree is ML but OLS is econometrics)
- But that is subjective, and the larger difference is down to how the approaches assess performance and the claims they make about causality
- Econometrics usually makes use of hypothesis testing and statistical tests to assess model performance, and as a result frequently makes claims about causality
- ML usually uses out-of-sample testing to assess model performance, and is not usually concerned with causality

When is ML appropriate/applicable?

- ML is useful for a few main use cases in official statistics:
  - When you have a target variable that has a long publication lag (i.e., nowcasting)
  - When you want to produce forecasts (e.g., model the relationship between GDP and investment, get a forecast of investment based on published GDP forecasts)
  - When you want to increase coverage (e.g., have health data and child mortality data for country 1, but only health data for its neighbor, country 2, train a model on country 1's data, run inference on country 2's health data)
  - When you want to run scenario analysis (e.g., model the relationship between ODA and deforestation in the past to see what impact different levels of ODA may have on future deforestation) or risk assessment

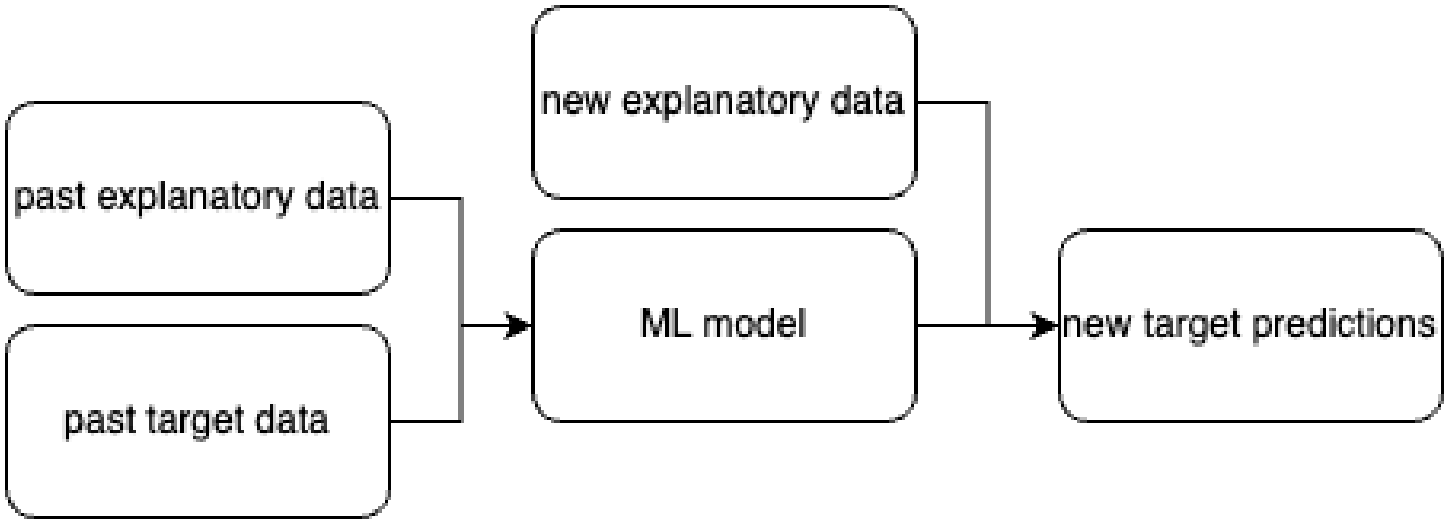
How can ML be useful for AIS data?

- AIS data is special because of its coverage, granularity, and timeliness
- These characteristics give the data a lot of use on their own
- But extra value can be unlocked by modelling their relationship with other indicators, that's where machine learning is useful
- It can expand its applications from the maritime to other domains, like trade, emissions, and the economy

When is ML not appropriate/applicable?

- Supervised ML is not a data generation process, it simply models the relationship between existing data
- Therefore, it cannot for instance populate a new SDG indicator
- As it learns from data, it needs a certain amount to work properly
- Just a handful of observations is usually not enough to produce an ML model

The machine learning pipeline

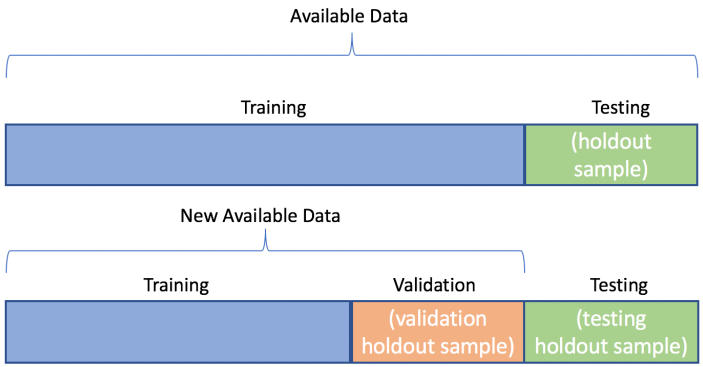


Getting your data ready

- The first step in an ML project is gathering the data, both explanatory and target
- Once your data is gathered, it needs to be cleaned and pre-processed to work with the type of algorithm you will use (e.g., normalized, scaled, made to be the same frequency, etc.)
- Additional features (i.e., explanatory variables) can also be extracted from your data (also called 'feature engineering')
- For instance, whether a day is a bank holiday or not, using GDP per capita instead of GDP and population alone, average country elevation from regional elevation data, etc.

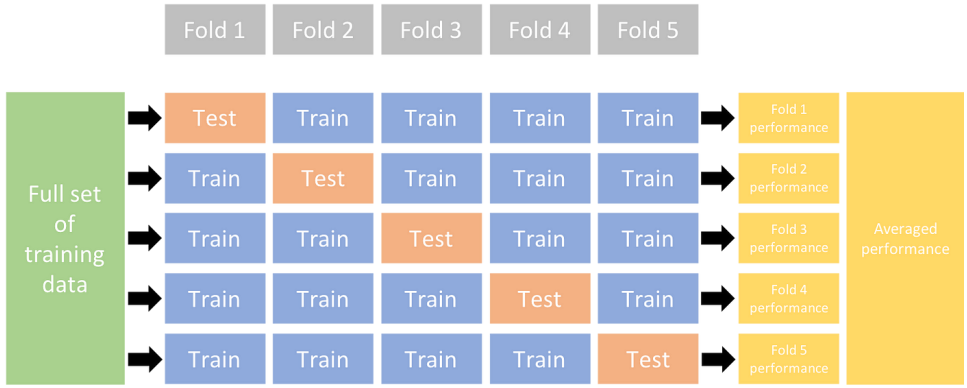
Train and test split

- The train set is used to train the model, while the test set is used to assess its performance on a new, unseen dataset (out-of-sample)
- this is to avoid overfitting
- A usual split is 80% of your data for training, 20% for testing
- The validation set is similar to the test set, but used to select hyperparameters
- If you use the test set to choose hyperparameters, you still may be overfitting the data



Cross validation

- Cross validation is a method to further decrease overfitting
- The training data is shuffled and trained and tested on different subsets to increase robustness of performance statistics
- The observations need to be independent to use the method, for e.g. time series, rolling-basis cross validation can be used

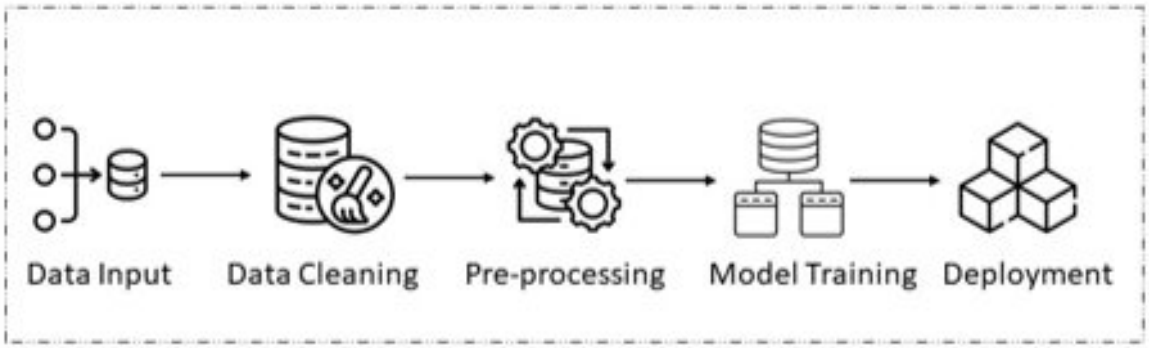


Variable selection and hyperparameter tuning

- Variable selection refers to choosing which variables go into the model
- In econometrics, input variables are usually taken as given, while in ML, input data can often have many features
- Not all of these are necessarily useful and can lead to overfitting
- While the parameters of an ML model refer to the weights, coefficients, etc. that change based on the training data, hyperparameters refer to macro-structures of the algorithms which do not change based on the data
- Examples include the maximum depth of a decision tree, the number of layers in a neural network, etc.
- These hyperparameters need to be tuned/chosen using the validation set or cross validation

Final model, inference, and production

- Once a model has been selected (i.e., its input variables and hyperparameters chosen), it can be put into production
- This entails setting up the data collection and transformation pipeline, setting up the model to be retrained on the latest data on a regular interval, and generating predictions on new data



Some specific ML algorithms

- Up to this point, we have been discussing ML models as a monolithic abstract
- In reality, there are dozens of algorithms to choose from, each with their own unique characteristics, requirements, use cases, and hyperparameters
- Some of the most common include:
  - Decision Trees
  - Gradient Boost
  - Logistic Regression
  - Long Short-term Memory Artificial Neural Networks (LSTM)
  - Multi-layer Perceptrons (MLP)/Feedforward Artificial Neural Networks (ANN)
  - OLS
  - Random Forest
  - Support Vector Machines (SVM)

Conclusion

- In conclusion, ML is a powerful tool which can increase the power and scope of your data
- Since ML pipelines, especially the data collection phase, can be complex, solid programming fundamentals are necessary to properly leverage it
- Combining the content of yesterday's exercise with today's, you should be familiar with the basics of implementing ML pipelines
- In the exercise session we will attempt to create our own ML model for predicting Russian exports