Machine Learning Model Pipeline
Feature Selection

Why Do We Select Features?

- Simple models are easier to interpret
- Shorter training times
- Enhanced generalisation by reducing overfitting
- Easier to implement by software developers ➔ Model production
- Reduced risk of data errors during model use
- Data redundancy

Reducing features for model deployment

- Smaller json messages sent over to the model
  - Json messages contain only the necessary variables / inputs
- Less lines of code for error handling
  - Error handlers need to be written for each variable / input
- Less information to log
- Less feature engineering code

Variable Redundancy

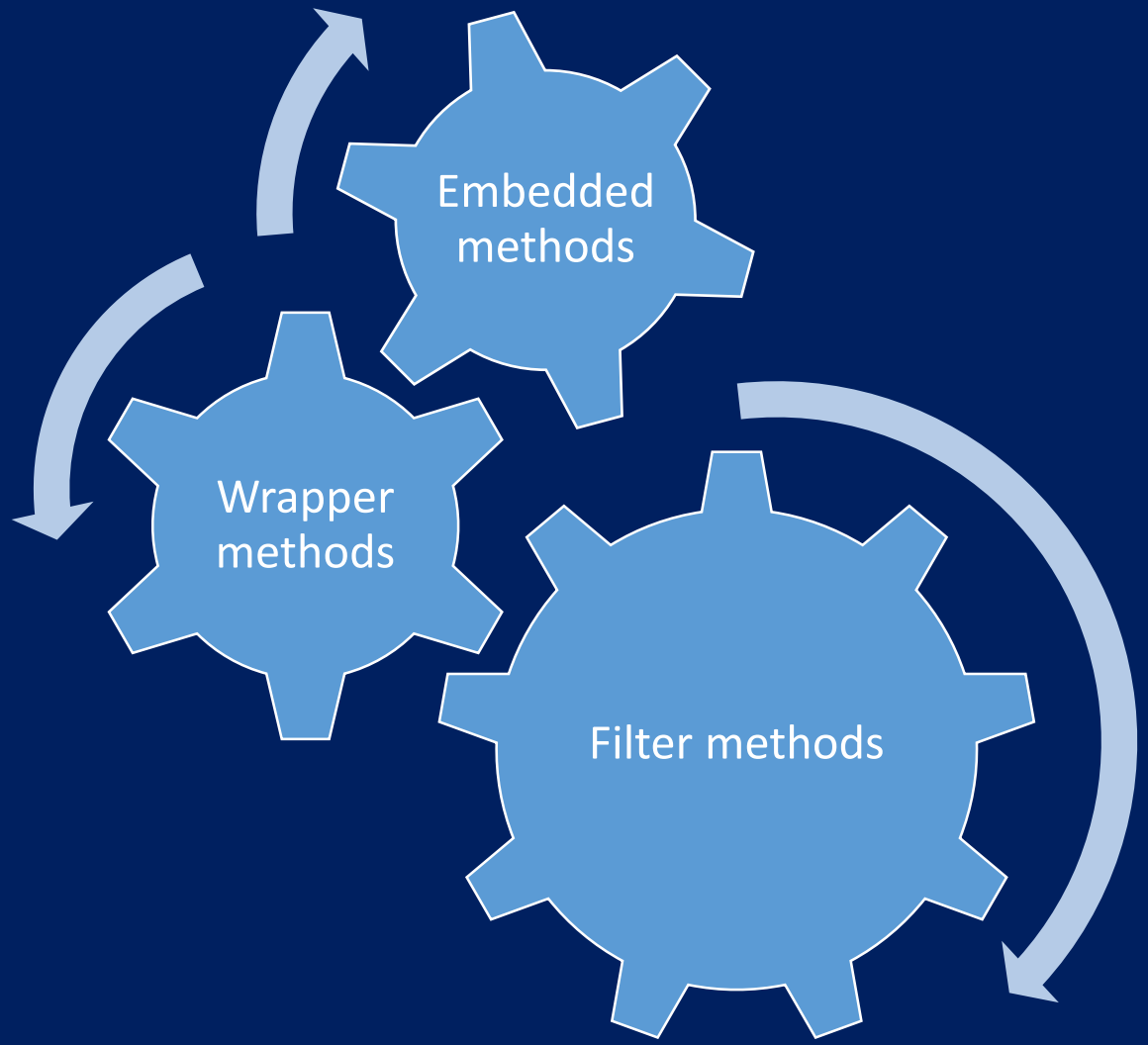
Constant variables
Only 1 value per variable

Quasi – constant Variables
> 99% of observations show same value

Duplication
Same variable multiple times in the dataset

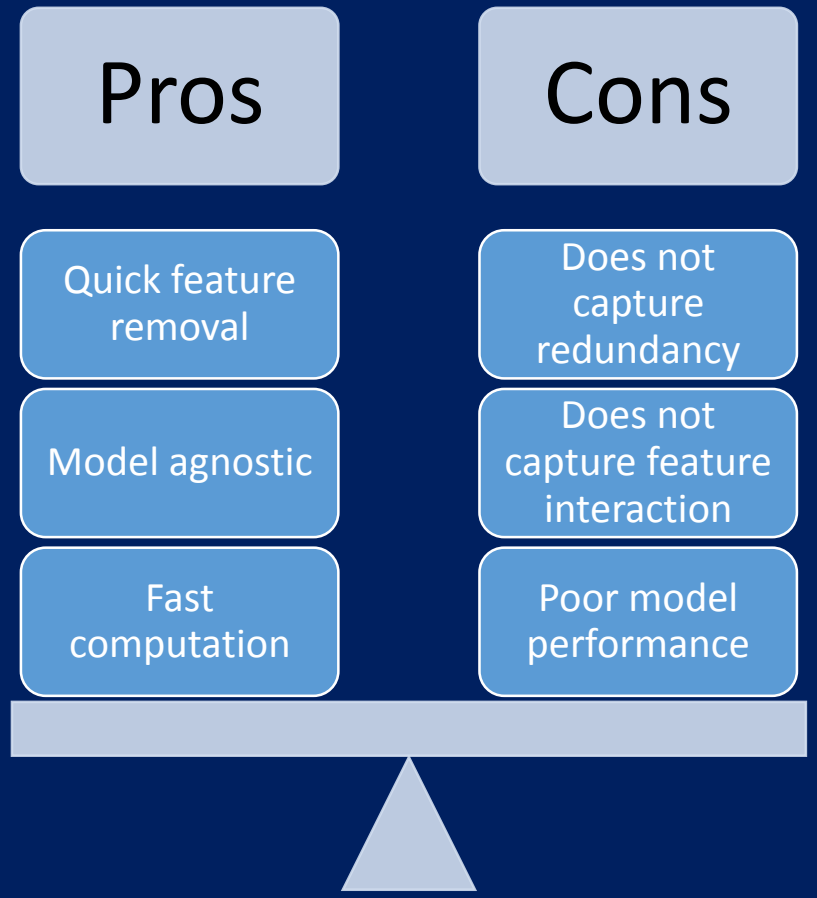
Correlation
Correlated variables provide the same information

Feature Selection Methods

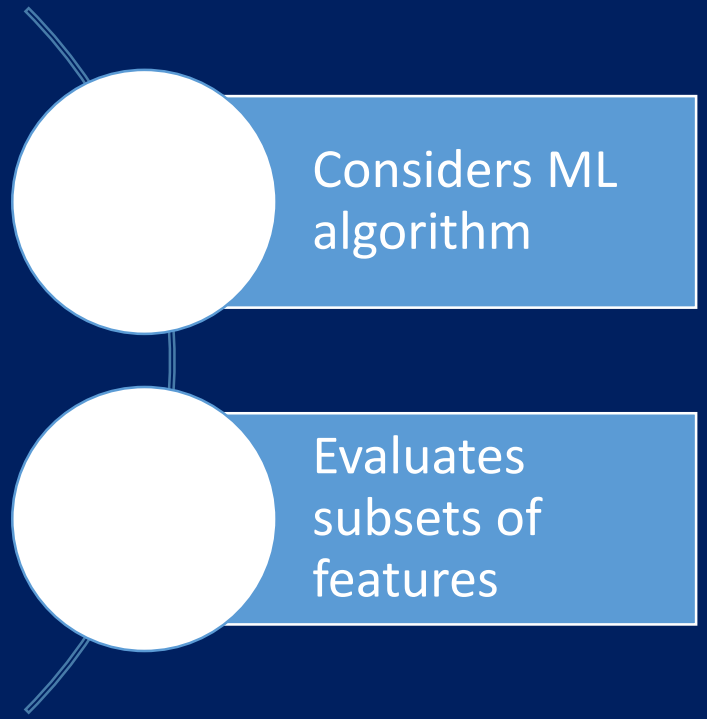


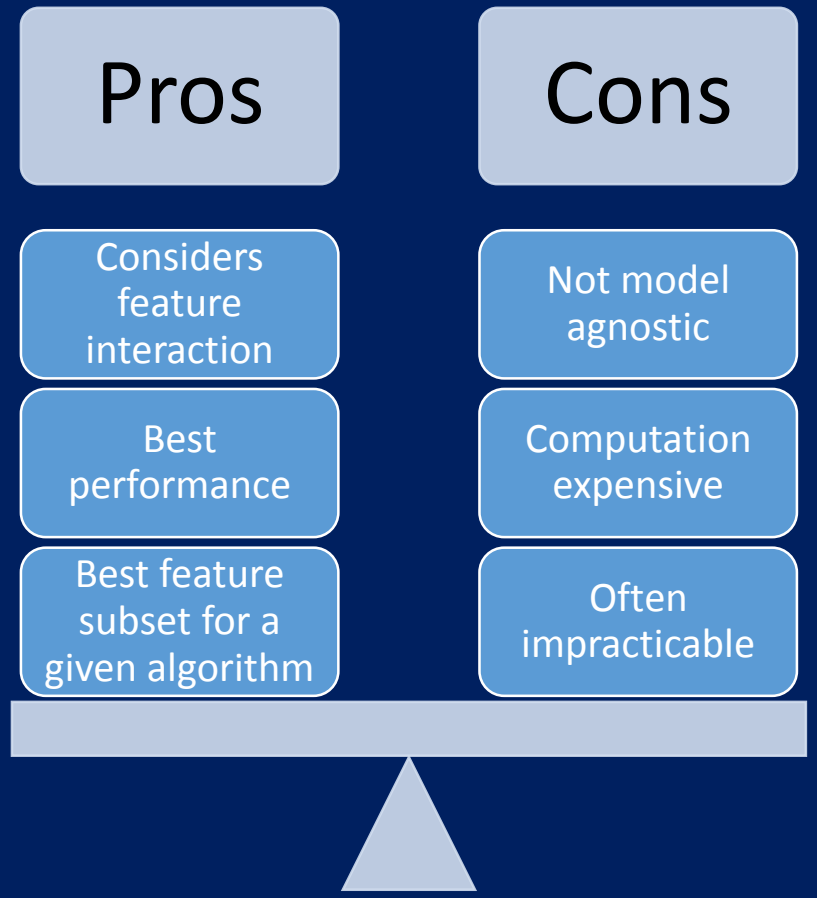
Filter methods

- Independent of ML algorithm
- Based only on variable characteristics



Wrapper methods





Embedded methods