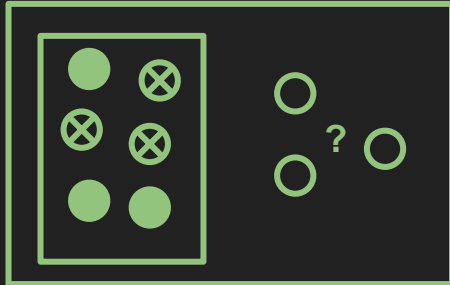# SPLINE

Machine Learning and Boosted Trees

Machine Learning

Types and Examples

Supervised

- Task driven methods.
- You have input variables X and an output variable Y and wants to learn the mapping function Y = f(X).
- Regression, Classification.
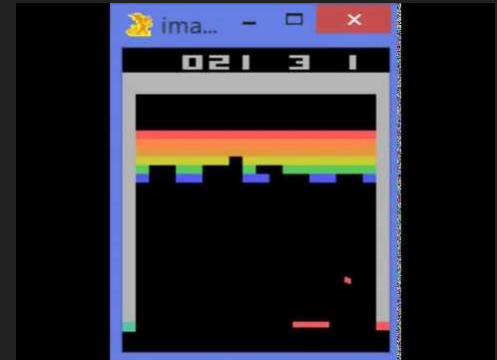- Applications: Object detection in images. Disease diagnosis.

Unsupervised

- Find hidden structure in data.
- Clustering, Dimensionality Reduction.
- Applications: Customer segmentation. Machine translation.
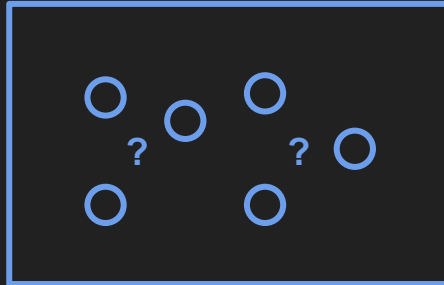
Reinforcement

- Learn an agent to act in an environment.
- Applications: Robotics, Games, Vehicle Routing
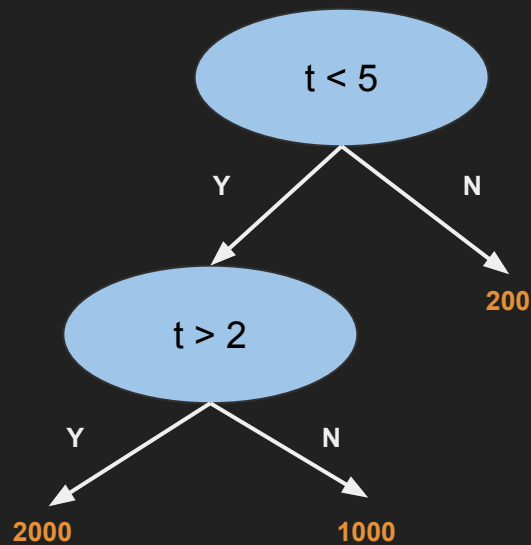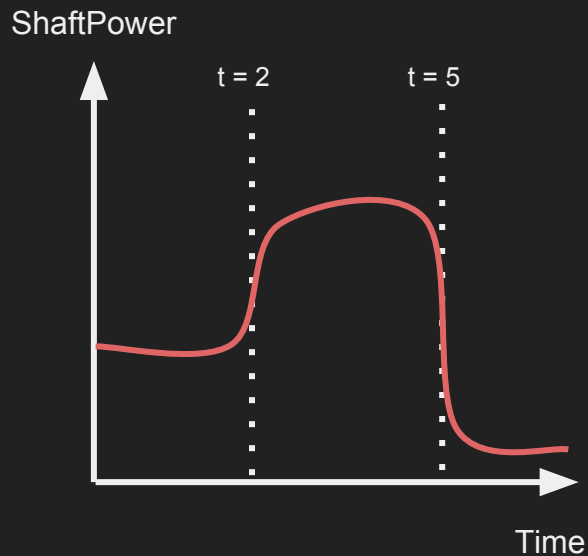
Machine Learning

Decision Trees

- Type of supervised learning algorithm for regression or classification.
- CART trees is a type of decision tree where each leaf has a score attached.
- Split variables and cut-points are learned using a greedy algorithm that minimizes a cost function.
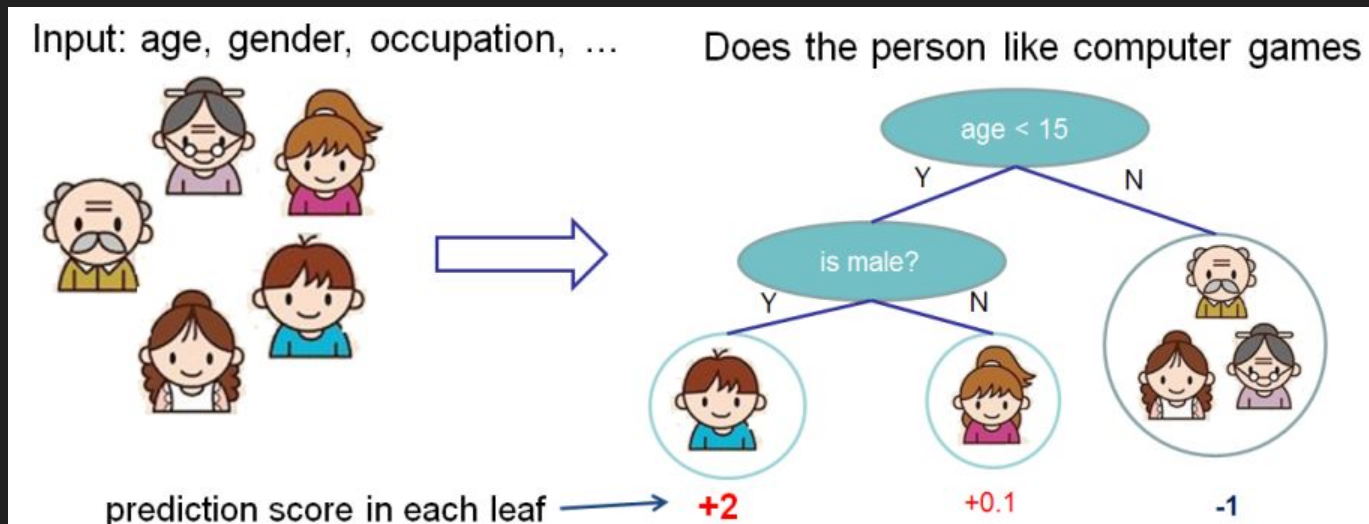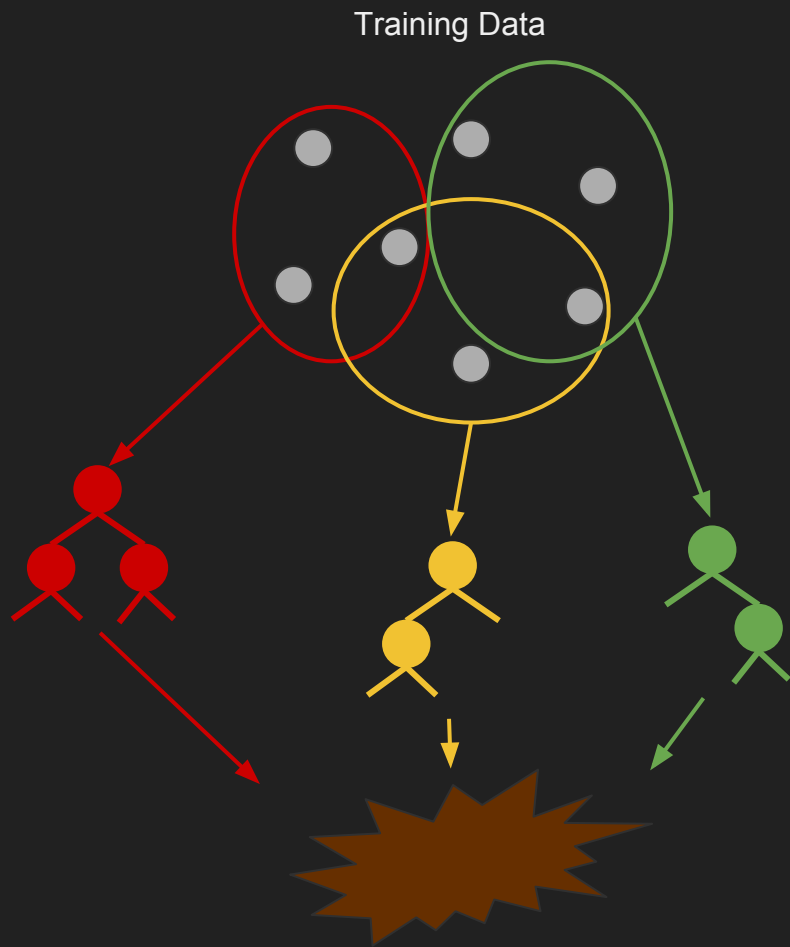
Machine Learning

Regression Tree Example

Machine Learning

## Tree Ensembles (Random Forest)

- Tree Ensembles are a set of CART trees to construct stronger and more robust models. It sums the prediction of multiple trees together.
- Benefits:
    - Widely used with good performance.
    - Feature normalization is not required.
    - Able to learn higher order feature interactions.
    - Scalable.
    - Feature ranking.

Training Data

Machine Learning

Tree Ensembles (Boosted Trees)

- Boosted Trees differs from Random Forest in that when you add a new tree, it is trained to improve the ensemble.
- For a given tree structure, you compute a score and and define an objective function to optimize for the new tree.
- XGBoost is based on on Boosted Trees and also includes other optimization strategies in learning the trees.
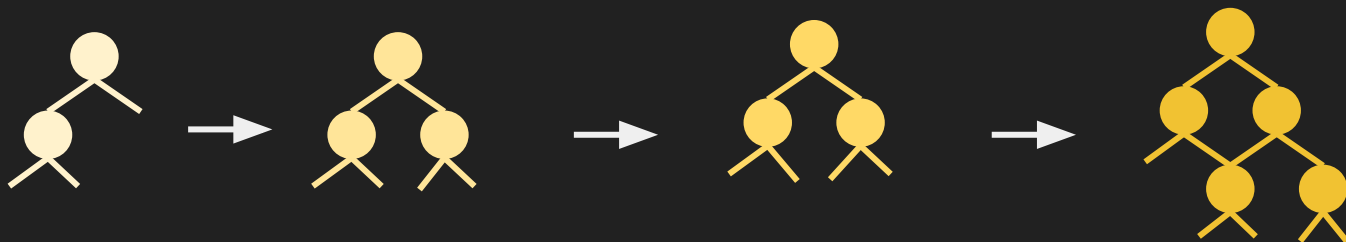
Data

Data Sources

Buoy Data

Satellite Data

Hindcast Data

Ship Data

- Only North American buoys.

- Spatial resolution 0.75/0.75 degrees.
- Temporal resolution every 6 hours.
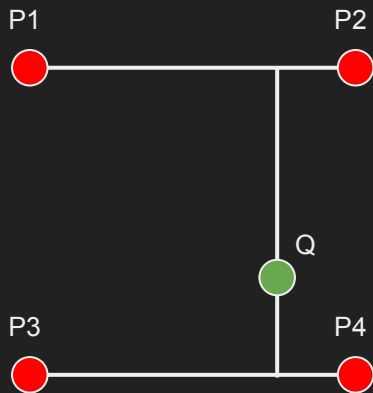
- Sensor data every 15 minutes for a voyage.
- Contains static data per voyage and static data per ship.

Preprocessing

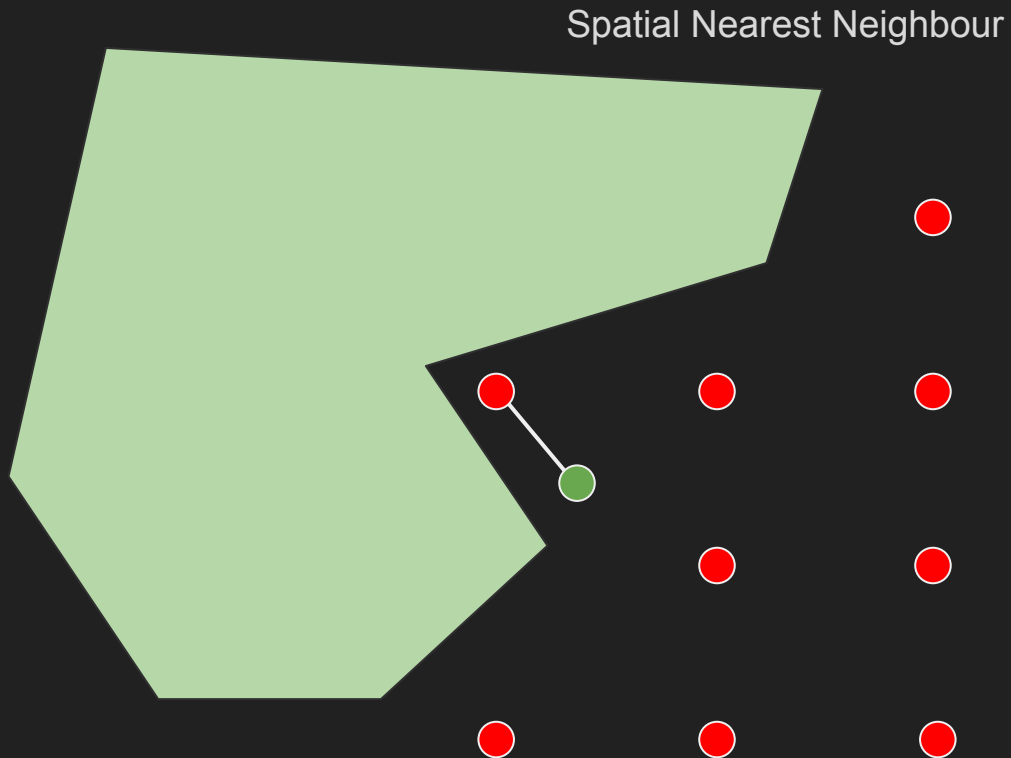Combining MetOcean data with ShipData

Spatial Interpolation

Temporal Interpolation

Preprocessing

Combining MetOcean data with ShipData

Spatial Nearest Neighbour

Preprocessing

Cleaning Data

- Removing data with NULL values.
- Filtering out data with LogSpeed < 9.7 knots.
- Filtering out data with ShaftPower < 5000 kW.
- Filtering out data with State as Manoeuvering, Anchor/Drifting, At Berth or Sea Trial.
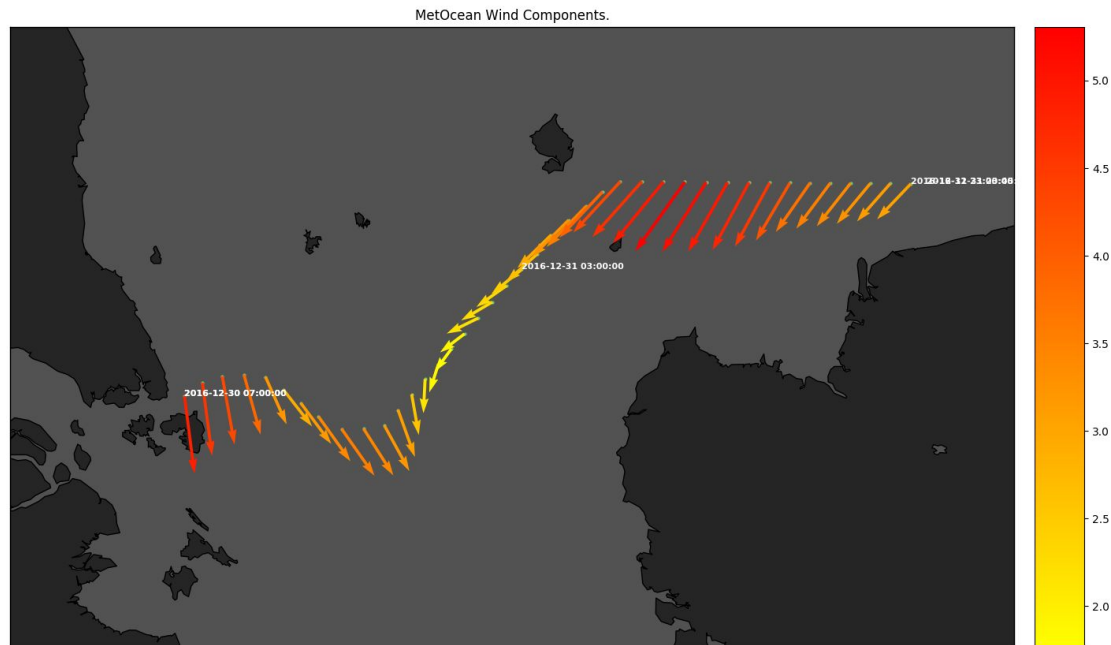- Computing component vectors for waves and ship
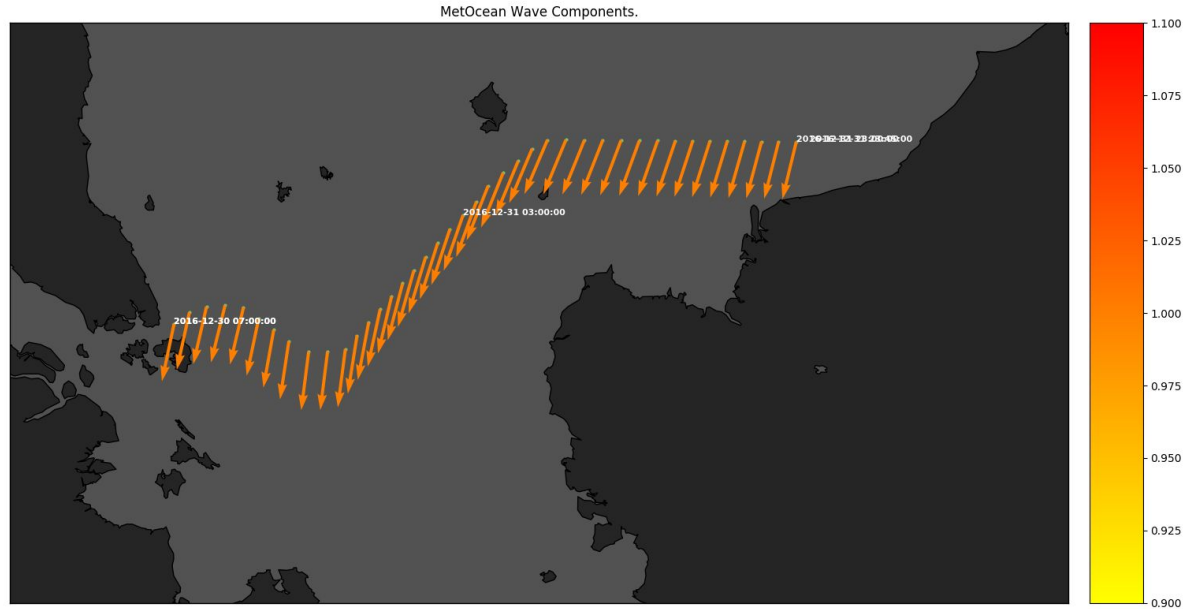
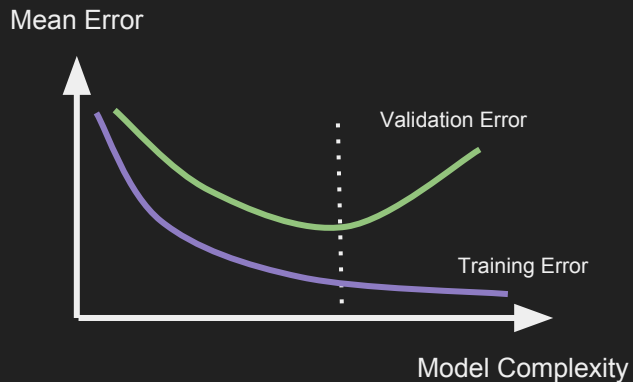# Preprocessing

Cleaning Data
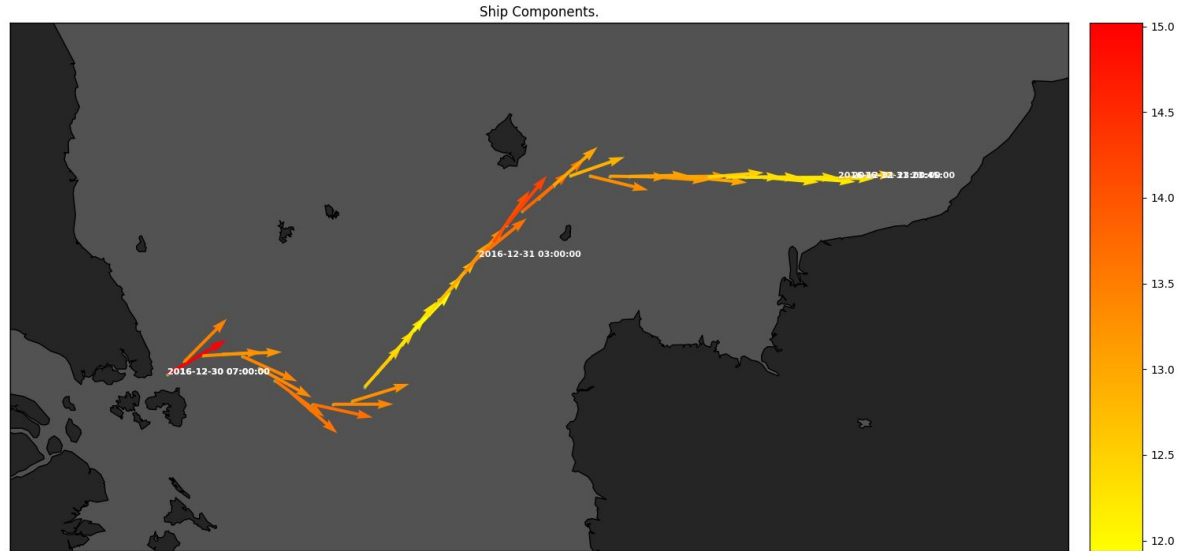
Preprocessing

Cleaning Data

MetOcean Wind Components.

Preprocessing

Cleaning Data

Preprocessing

Splitting Data

Preprocessing

Splitting Data

Random Split

Voyage Split
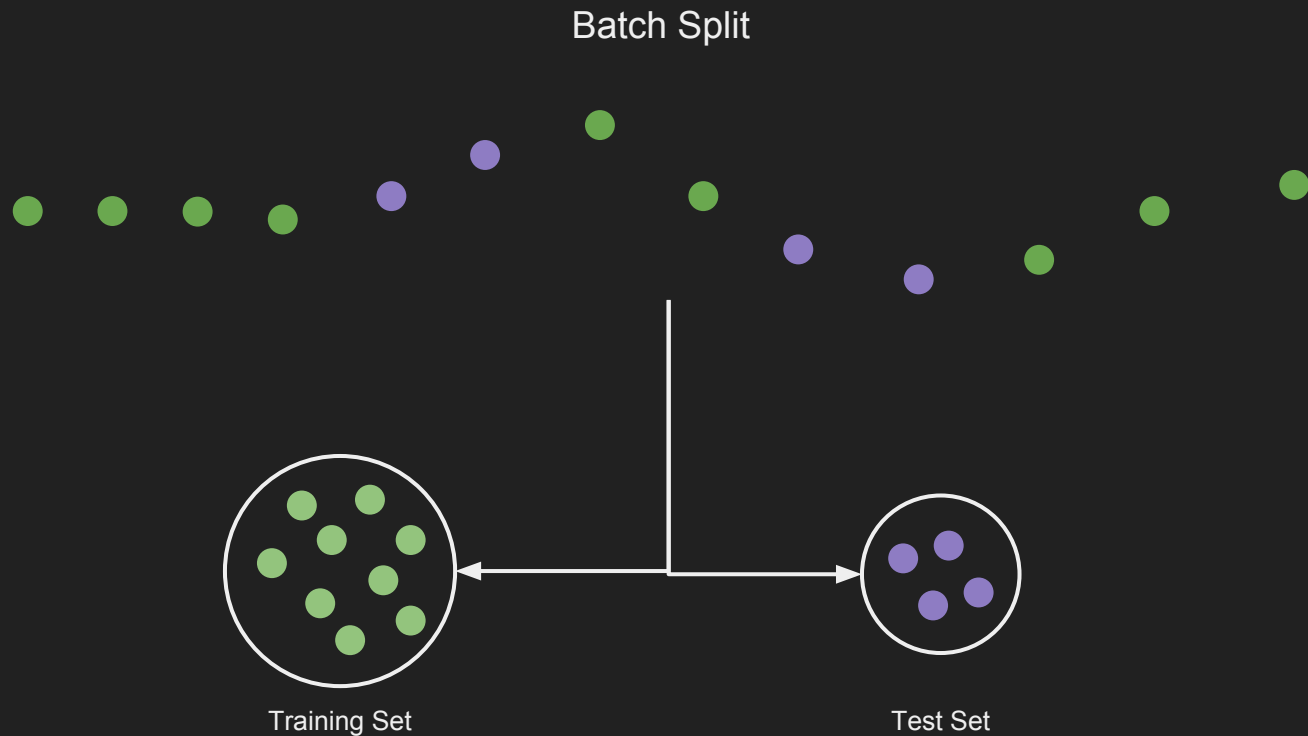
Training Set

Test Set

Training Set
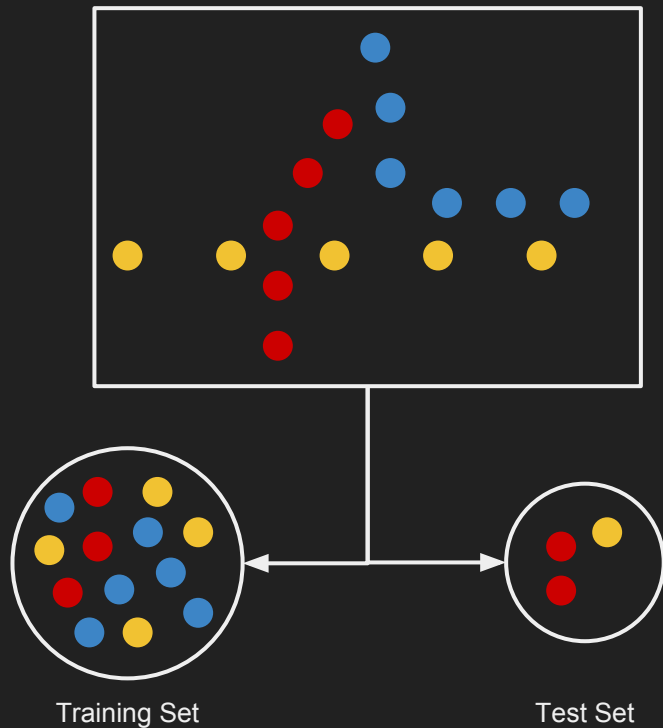
Test Set
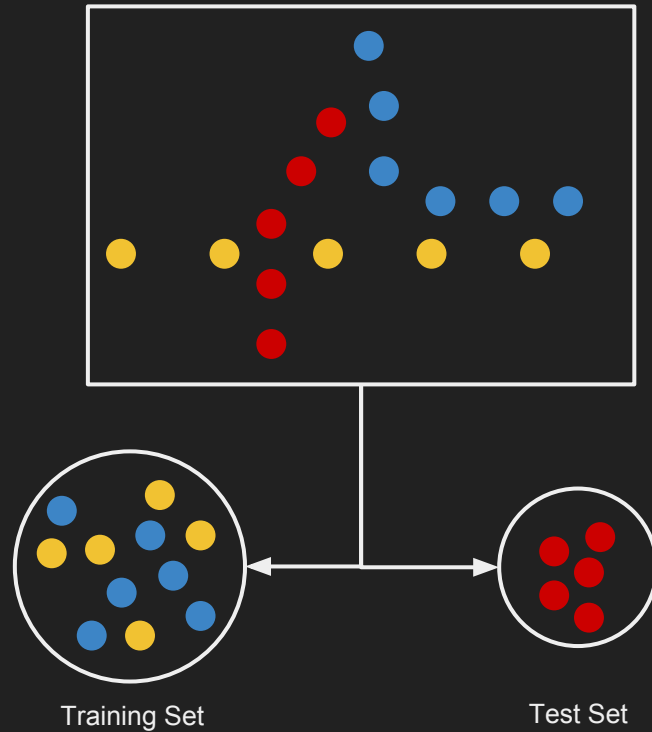
Preprocessing

Splitting Data

Results

Parameters S:

DesignDeadweight
Cargo
GPSSpeed
LogSpeed
DraftFore
DraftAft
SeaDepth
AccelerationXDirection
AccelerationYDirectio
AccelerationZDirection
Roll
Pitch
Yaw
GyroX
GyroY
GyroZ
SeaWaterTemperature (55.56%)
RudderAngle (34.78%)
WindSpeedRelative
WindDirectionRelative

Parameters H:

COGheading
mo_wind_u_comp
mo_wind_v_comp
mo_swh
mo_mwd
mo_mwp
mo_sst

Split Type	AVG ShaftPower	Test Score S	Test Score S+H
Random Split	6020.90	72.60	65.35
Voyage Split	5893.48	333.84	327.46
Batch Split 1h	6017.80	77.42	76.39
Batch Split 4h	5962.56	131.35	131.43

Results

Parameters S:

DesignDeadweight
Cargo
GPSSpeed
LogSpeed
DraftFore
DraftAft
SeaDepth
AccelerationXDirection
AccelerationYDirectio
AccelerationZDirection
Roll
Pitch
Yaw
GyroX
GyroY
GyroZ

Parameters H:

COGheading
mo_wind_u_comp
mo_wind_v_comp
mo_swh
mo_mwd
mo_mwp
mo_sst

Split Type	AVG ShaftPower	Test Score S	Test Score S+H
Random Split	6220.97	121.17	88.70
Voyage Split	6051.55	499.44	476.95
Batch Split 1h	6214.52	144.53	116.75
Batch Split 4h	6268.99	189.03	177.19

Results

Parameters S:

DesignDeadweight
Cargo
GPSSpeed

Parameters H:

COGheading
mo_wind_u_comp
mo_wind_v_comp
mo_swh
mo_mwd
mo_mwp
mo_sst

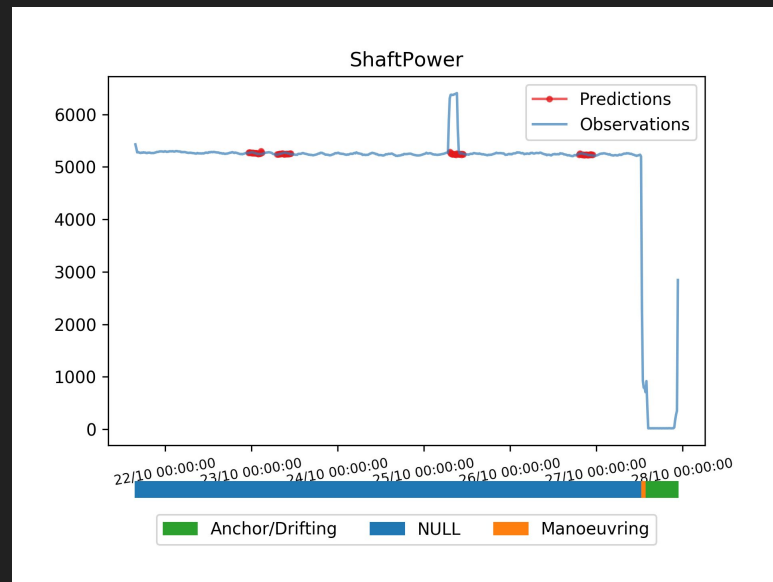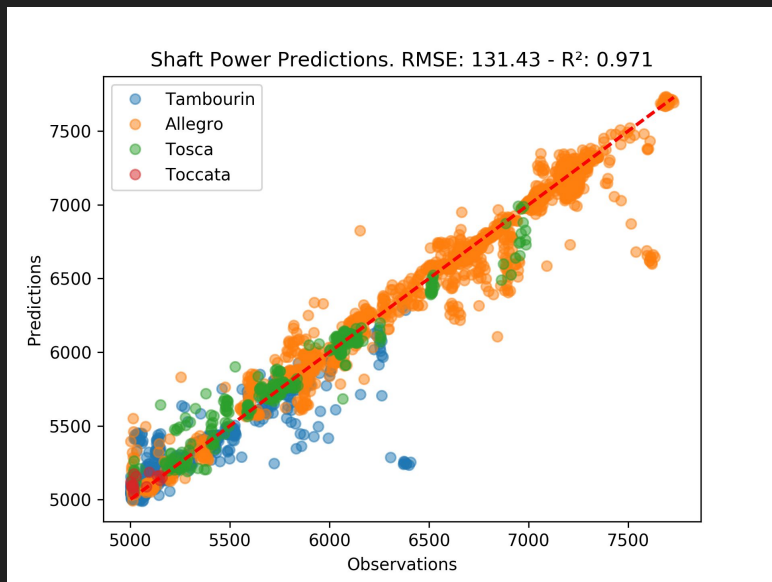| Split Type	AVG ShaftPower	Test Score S+H
Random Split	6344.00	95.51
Voyage Split	6590.51	722.12
Batch Split 1h	6317.08	139.59
Batch Split 4h	6369.05	205.88

Results

Ship Data + Hindcast Data

Voyage Split
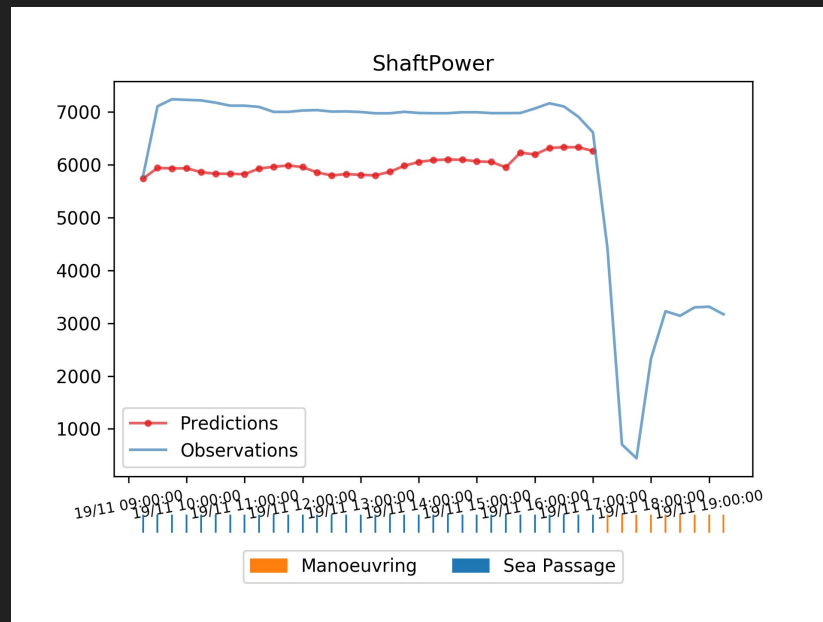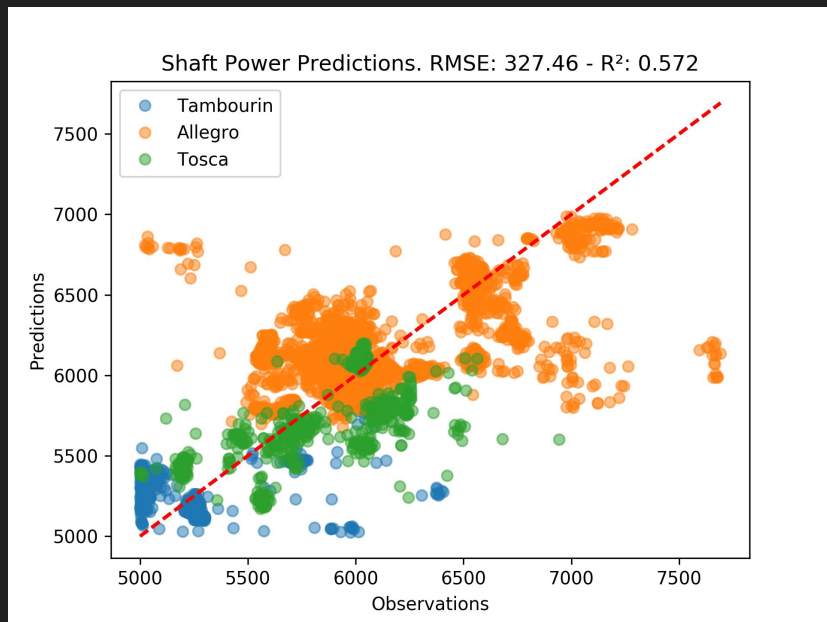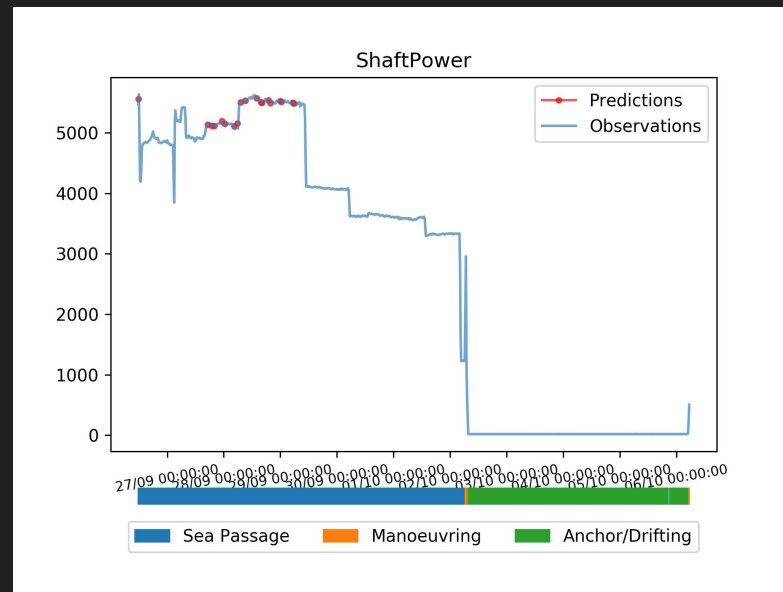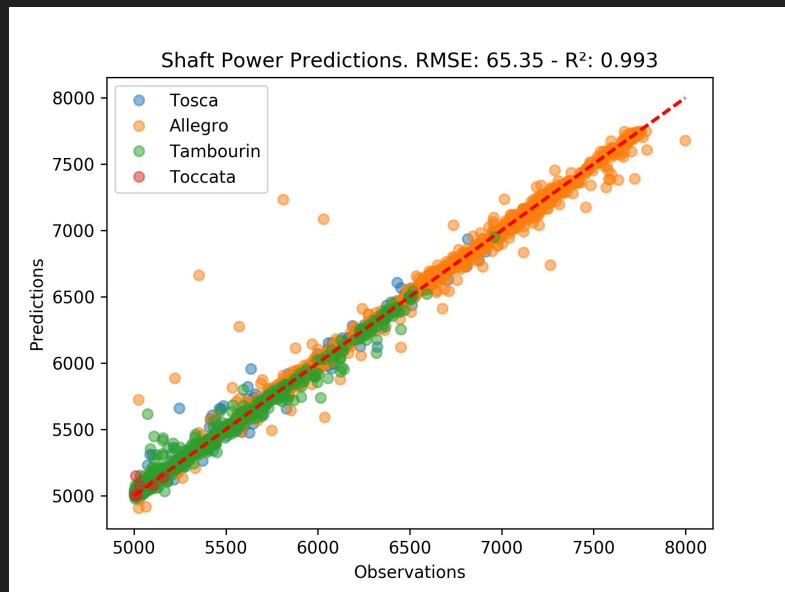
Results

Ship Data + Hindcast Data

Batch Split 4h

Results

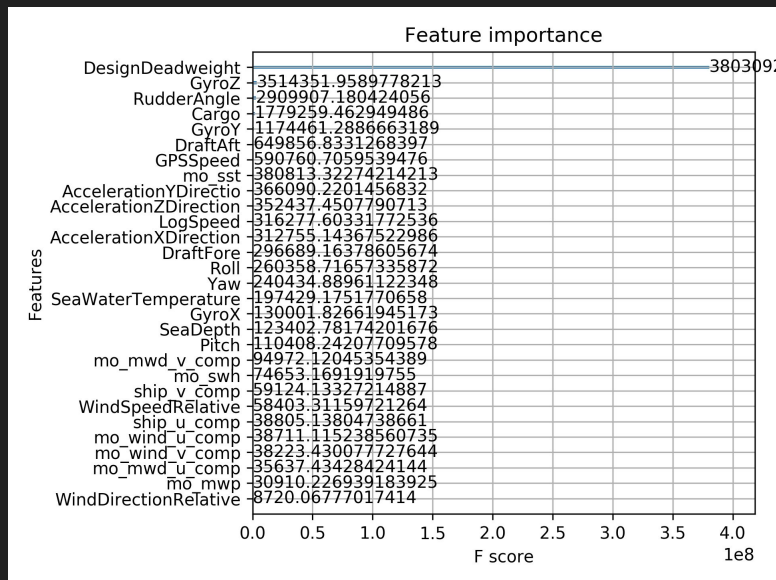## Ship Data + Hindcast Data

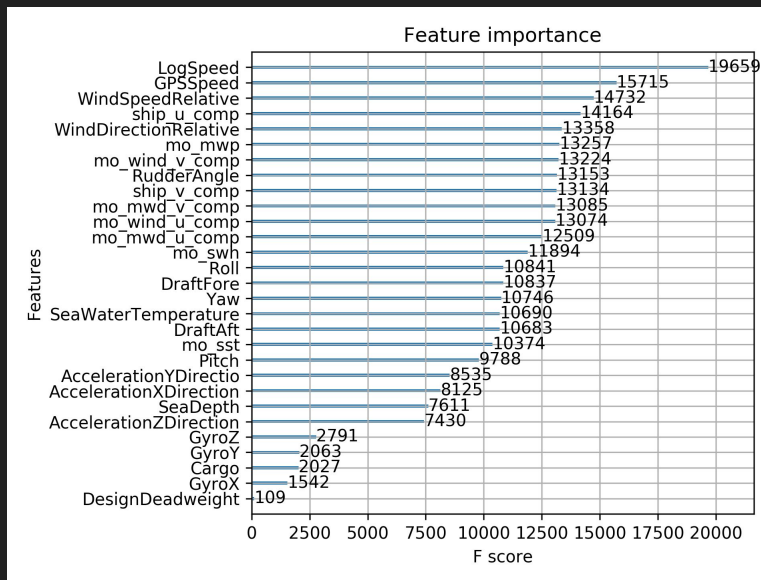Random Split

Machine Learning

Feature Importance

Weight: The number of times a feature appears in a tree.
Gain: The average gain of splits which uses the feature
(gain is a measurement of how good a split is).

Gain

Weight

Results

Possible extensions

- Try to improve Voyage/Batch split by including Lag/Window features.
- Use lon/lat to include static geographical data.
- Include satellite data as a MetOcean data source.