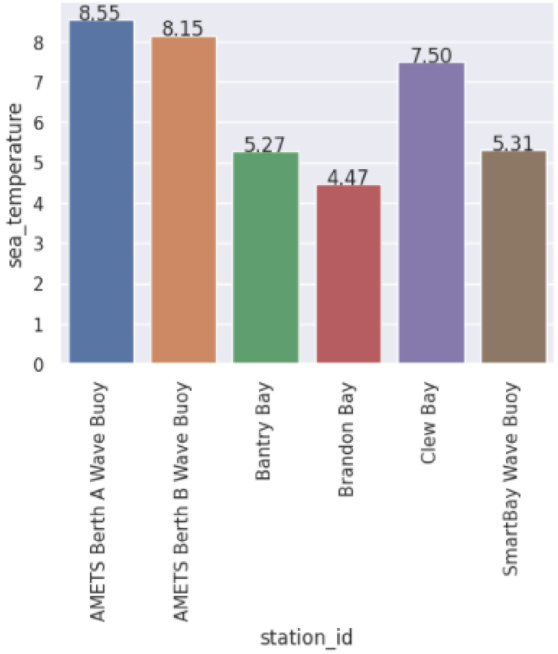
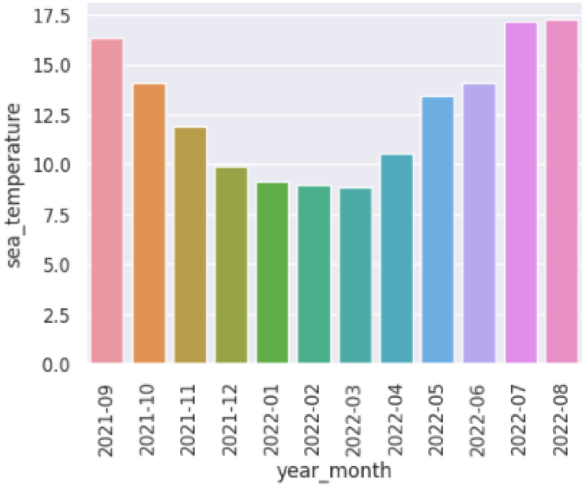
Dadosfera Data Science Case: Time Series Regression
Vitor Hugo Martins Ferreira

Minimum Requirements: Questions about the dataset

Question 1: What is the lowest temperature of each one of the Bouys? Which usually month it occurs?

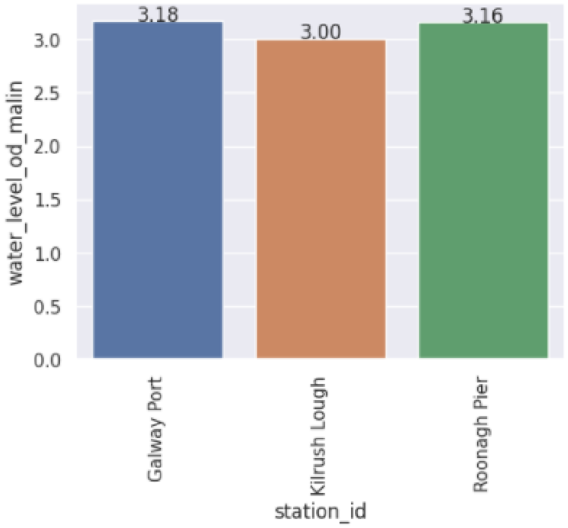
The graph below shows the lowest temperatures (ºC) recorded in each buoy. Throughout the year, the lowest temperatures are normally recorded between the months of January and March.

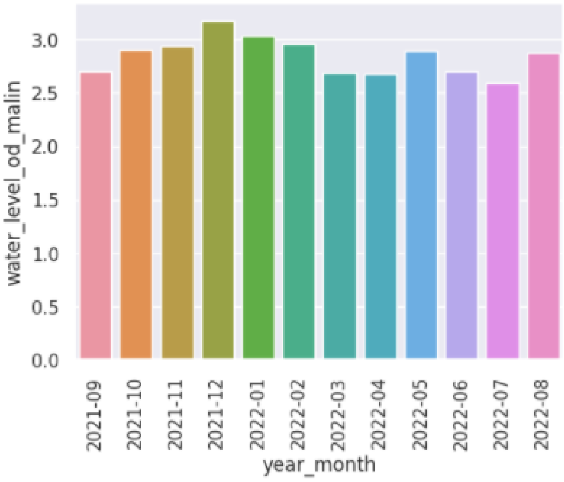


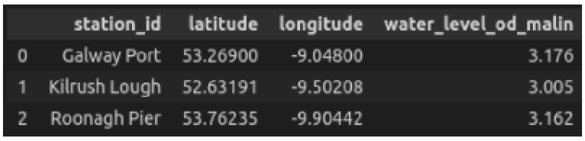


Question 2: Where (lat/long) do we have the biggest water level? Which usually month it occurs?

The graph below shows the highest sea levels (meters) recorded in the buoys that had the highest records. Throughout the year, the Irish Sea reaches its highest levels between the months of December and February.





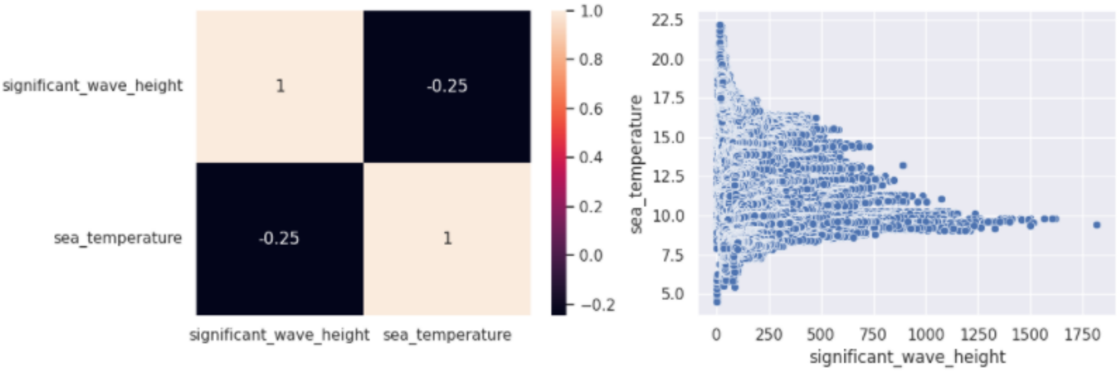


	station_id	latitude	longitude	water_level_od_malin
0	Galway Port	53.26900	-9.04800	3.176
1	Kilrush Lough	52.63191	-9.50208	3.005
2	Roonagh Pier	53.76235	-9.90442	3.162

Question 3: How the Wave Lenghts correlates with Sea Temperature? It is possible to predict with accuracy the Wave Lenght, based on the Sea Temperature and the Bouy location?

The correlation between Wave Length and Sea Temperature is negative, which means that with a lower temperature, there is a tendency to form larger waves. However, by Pearson's method, this correlation presented a low index, indicating that the sea temperature does not have a very relevant impact on the size of its waves (it can be observed too in the scatter plot).

This indicates that using only this feature and the buoy position, it would not be possible to build a predictive model with high accuracy.



Bonus Item: Time Series Regression

5. Data Preparation

5.1 Feature Transformation:

- **Standarditazion:** not used because none of the variables showed a normal curve;
- **Rescaling:**
  - MinMax Scaler: peak_period, upcross_period, year;
  - Robust Scaler: peak_direction, significant_wave_height;
- **Encoding:** applied to the categorical variable station_id
- **Nature Transformation:** for cyclic variables (month, day of month, week of year, day of week and hour) the sine and cosine transformation was applied

5.2 Feature Selection:

In the first version of the model, the Boruta algorithm was used to indicate the relevant variables. By this method all dataset variables were selected.

Note: thinking about a practical application for the predictive model, where users would only have information about the date, time and place where they would like to see the temperature forecast, a new model was built using only these features

6. Machine Learning Modeling

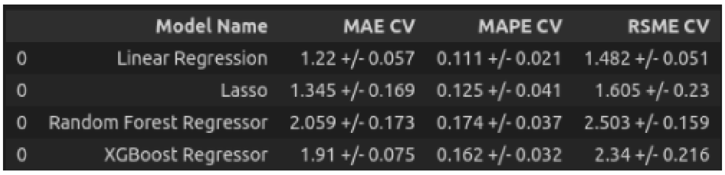
For the construction of the regression model, linear and non-linear regression algorithms were tested:

- Linegar Regression: Linear Regression Model and Linear Regression Regularized Model (Lasso).
- Nonlinear Regressions: Random Forest Regressor and XGBoost Regressor.

And to evaluate the performance of each model, the following error metrics were used:

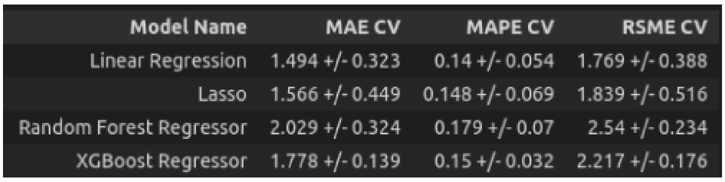
- MAE: Mean Absolute Error
- MAPE: Mean Absolute Percentage Error
- RSME: Root Mean Square Error

6.1 All Features



	Model Name	MAE CV	MAPE CV	RSME CV
0	Linear Regression	1.22 +/- 0.057	0.111 +/- 0.021	1.482 +/- 0.051
0	Lasso	1.345 +/- 0.169	0.125 +/- 0.041	1.605 +/- 0.23
0	Random Forest Regressor	2.059 +/- 0.173	0.174 +/- 0.037	2.503 +/- 0.159
0	XGBoost Regressor	1.91 +/- 0.075	0.162 +/- 0.032	2.34 +/- 0.216

6.1 Time Features

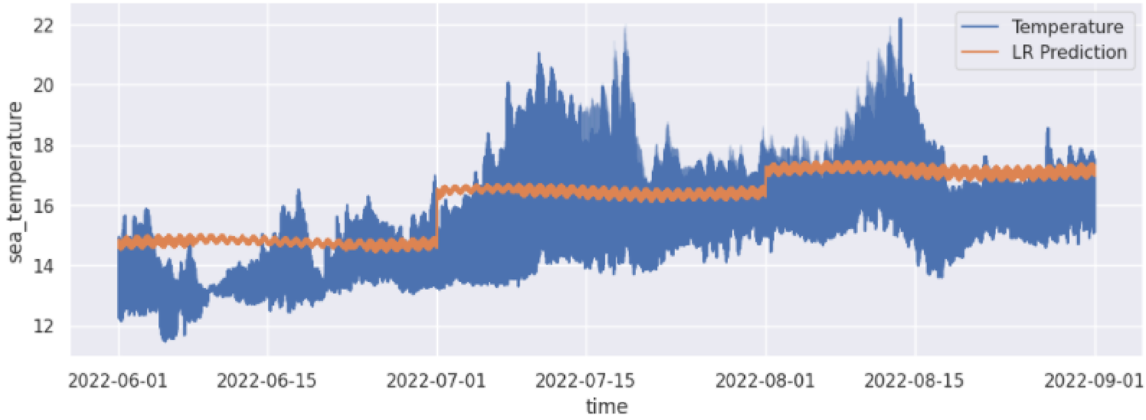


Model Name	MAE CV	MAPE CV	RSME CV
Linear Regression	1.494 +/- 0.323	0.14 +/- 0.054	1.769 +/- 0.388
Lasso	1.566 +/- 0.449	0.148 +/- 0.069	1.839 +/- 0.516
Random Forest Regressor	2.029 +/- 0.324	0.179 +/- 0.07	2.54 +/- 0.234
XGBoost Regressor	1.778 +/- 0.139	0.15 +/- 0.032	2.217 +/- 0.176

Note: the above results were obtained with Cross-Validation (CV). It was done by defining 3 different validation periods (each one with 3 months) and defining the test period as the interval before the validation period. Because it is a time series model, the Cross Validation cannot be done by randomly defining intervals.

7. Evaluation

For the final evaluation of the model, we will proceed with the version in which only the date, time and location variables are used and we will use the linear regression model algorithm.



Comparing the temperature records present in the validation dataset with the values predicted by the linear regression model, it is clear that this is not a very accurate result and that, depending on the practical application of this model, it will be necessary to improve this modeling.