Brandeis | INTERNATIONAL BUSINESS SCHOOL

4/17/2020

Analysis of Tropical Atmosphere Ocean Data

Big Data II Final Project

Team Invincibles
Brandeis University

Table of Contents

Introduction

We retrieved the dataset from the UCI Machine Learning Repository. This data was collected with the Tropical Atmosphere Ocean (TAO) array which was developed by the international Tropical Ocean Global Atmosphere (TOGA) program. The array consists of nearly 70 buoys positioned throughout the equatorial Pacific from the 1980s to the 1990s measuring oceanographic and surface meteorological information to understand and predict season-to-interannual climate variations originating in the tropical areas, most notably those related to El-Nino/Southern Oscillation (ENSO) Cycles.

According to Wikipedia, “El Niño–Southern Oscillation (*ENSO*) is an irregularly periodic variation in winds and sea surface temperatures over the tropical eastern Pacific Ocean, affecting the climate of much of the tropics and subtropics. The warming phase of the sea temperature is known as *El Niño* and the cooling phase as *La Niña*. The *Southern Oscillation* is the accompanying atmospheric component, coupled with the sea temperature change: *El Niño* is accompanied by high air surface pressure in the tropical western Pacific and *La Niña* with low air surface pressure there. The two periods last several months each and typically occur every few years with varying intensity per period.” [1]

The program was initiated partly owing to an urgent need to understand the severe 1982-1983 El Nino events, one of the strongest recorded in history.[2] According to an article from Washington Post: “It [1982-1983 El Nino event] led to droughts in Indonesia and Australia, widespread flooding across the southern United States, lack of snow in the northern United States, and an anomalously warm winter across much of the mid-latitude regions of North America and Eurasia. The estimated global economic impact was over US$8 billion. This El Niño event also led to an abnormal number of hurricanes in the Pacific Ocean during this time span; the strongest hurricane up to 1983 hit Hawaii during the event.” [3] The ENSO cycle during

[1] https://en.wikipedia.org/wiki/El_Ni%C3%B1o%E2%80%93Southern_Oscillation

[2] https://en.wikipedia.org/wiki/1982%E2%80%9383_El_Ni%C3%B1o_event

[3] https://www.washingtonpost.com/news/capital-weather-gang/wp/2015/06/12/how-the-super-el-nino-of-1982-83-kept-itself-a-secret/

this period was not detected until it was near its peak. This highlighted the need for an ocean observing system (i.e. the TAO array) to support studies of large scale ocean-atmosphere interactions on seasonal-to-interannual time scales by collecting data on date, location, zonal wind (west<0), meridional winds (south<0), humidity, sea-surface temperature and air temperature.

The dataset was initially purported for forecasts of tropical Pacific Ocean temperatures and predictions of El Nino events in the next 1 to 2 years period. However, this process would require ad hoc knowledge in environmental science, which goes beyond our domain. So instead of working on a project beyond our capabilities, which requires a lot of other data as well, we are set to investigate the relationship among variables and buoys.

Data Preparation

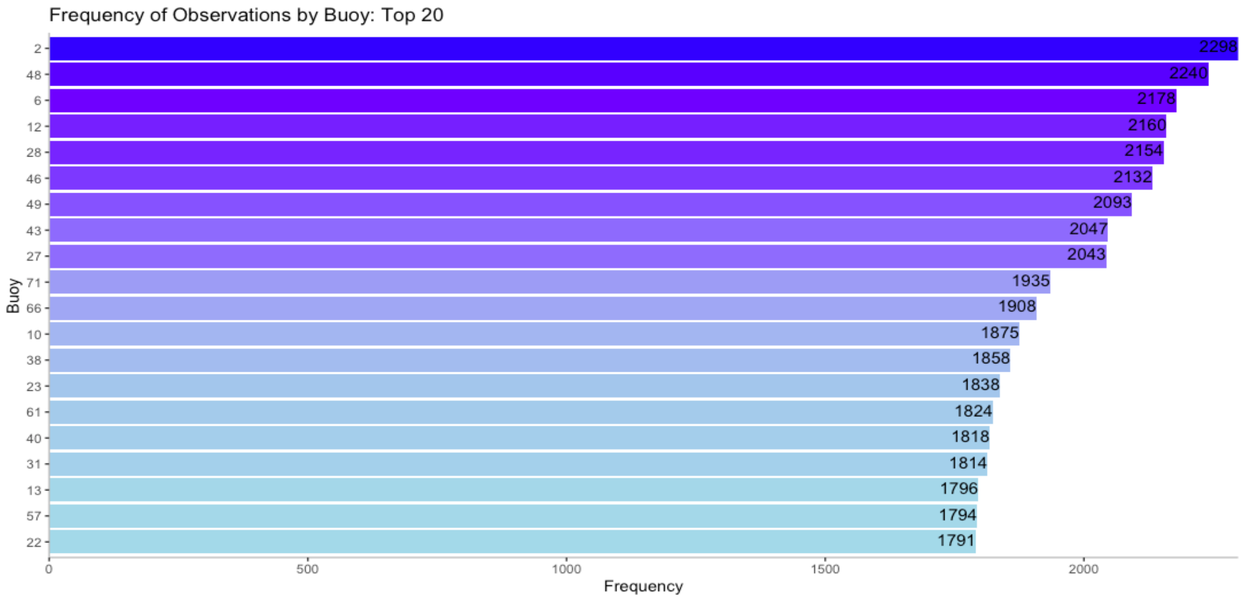
Upon receiving the data, we encountered a lot of missing data in this dataset mostly with a **Missing not at Random (MNAR)** pattern. Firstly, the buoys were commissioned at varying dates; hence the amount of data collected depends on the year. The range of the year at which the buoys were commissioned is 18 years with 1980 being the earliest and 1998 being the latest. In addition, the amount of data available is also dependent on the buoys' reliability. Weather conditions, such as currents, rainfall, and solar radiation, as well as the machine's mechanical issues hindered the buoys' ability to collect high-quality data consistently. Many buoys were unable to record variables because of these external reasons, especially for measuring humidity.

Trifacta detected that 14% of observations in "Zonal Winds" and "Meridional Winds ", 37% of observations in "Humidity", as well as 10% of observations in "Air Temp" and "Sea Surface Temp" was missing. The dataset uses "." to represent these missing values. We applied the "Replace" function to replace them with "null" and got rid of most rows with null values. The resulting dataset had a size roughly 60% of the original dataset.

Our study consists of two parts: Regression and Classification.

We were specifically curious about the five relevant and important variables that the buoys measured, including Zonal. Winds, Meridional Winds, Humidity, Air Temp and Sea.Surface.Temp. For the regression part of our study, we were initially interested in exploring what could contribute to the El Nino effect in the tropical Pacific. We had a hypothesis that these variables could showcase significant signs of El Nino. As a key indicator of the El Nino effects, Sea.Surface.Temp was therefore the appropriate target variable to kick off our exploration. We hoped to get a sense of the association between Sea.Surface.Temp and the other four variables; therefore, we predicted sea surface temperature by other variables. If those variables were not statistically significantly associated with Sea.Surface.Temp, in the future, we could consider alternative measures to improve the Sea.Surface.Temp predictability.

The dataset was measured by 72 buoys distributed in significantly different time periods. The different properties of each buoy would introduce omitted variable bias if we would regress all of them. As a result of this, we chose only one buoy to conduct the regression analysis. We use buoy 2 for our analysis since it contains the most observations (2298 entries) as shown in the bar plot below:



By the regression analysis, we would eventually have a better understanding of what factors could be associated with the El Nino effects. However, the data complexity held us from developing meaningful research. The raw data were not labeled. For the classification part of our study, we needed to classify the data into different buoys as the data source claimed.

Because each buoy contained continuous geographical and time information, the ambiguity in buoy labeling raises critical questions: how reliable the data are if a buoy could travel five longitude degrees away from its original location? what if data of different buoys are collected and recorded at the same time? then how can we possibly differentiate a buoy from another? It would be challenging to perform further predictive analysis without clear buoy labeling. Therefore, the classification part of our report serves as a preliminary investigation of buoy labeling. The location data alone were not a reliable indicator to predict the buoy index. We needed to use more complicated tools to clearly label buoys.

After we wrangled the data in the first place, we concluded total 72 buoys in the dataset, whereas missing data prevailed over almost all buoys. Only buoy 2 and buoy 48 did relatively well to retain data, recording 2298 and 2240 entries, respectively. Thus, we conducted our classification analysis on these two buoys. Our report will further explain how to group data to buoy 2 and buoy 48 by applying KNN classification and classification trees. Models we use in this report set an example for classification of more than 2 buoys in the future.

After subseting the data, we would use for regression and classification analysis, we applied the `normalize` function in R to standardize all continuous variables in the dataset and converted these variables in the scale from 0 to 1. Meanwhile we developed interaction terms based on variable pairs, as well as squared terms. Moreover, we constructed a monthly dummy variable to describe the month in which the data point is collected. Since the data collected by a buoy is time series data and it presents strong seasonal characteristics, a variable to capture the seasonal effect would prove to be necessary in both regression and classification analysis. It enhances the accuracy of prediction especially in the case of regression as shown later.

Descriptive Statistics

Since we would be using data collected by Buoy 2 for regression, data collected by Buoy 2 and 48 for classification, we would be showing descriptive statistics to offer a broad perspective into these two buoys. Buoy 2 and Buoy 48 are similar in zonal winds and humidity. However, Buoy 48 seems to moor at a relatively warmer region, with both higher air temperature and sea surface temperature. In addition, buoy 48 recorded on average stronger meridional winds.

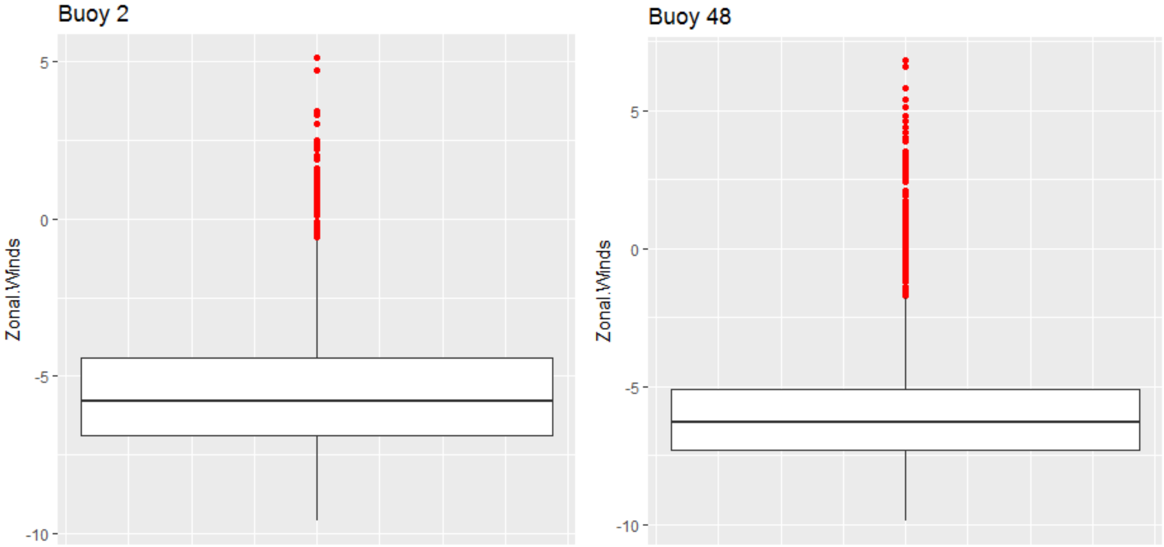
Statistics Table

Buoy 2							
Variables	**Min**	**Max**	**Mean**	**Median**	**Mode**	**Sd**	**Var**
ZonalWinds	-9.60	5.10	-5.44	-5.80	-6.80	2.08	4.33
MeridionalWinds	-7.40	5.20	-0.17	-0.10	0.40	1.81	3.28
Humidity	67.8	95.80	81.61	81.50	79.80	3.81	14.48
AirTemp	22.32	29.56	26.06	26.03	25.96	1.26	1.58
SeaSurfaceTemp	22.24	30.07	26.53	26.38	26.05	1.47	2.16
Buoy 48							
Variables	**Min**	**Max**	**Mean**	**Median**	**Mode**	**Sd**	**Var**
ZonalWinds	-9.90	6.80	-5.81	-6.30	-6.60	2.33	5.43
MeridionalWinds	-6.20	4.30	-1.00	-1.00	-1.20	1.66	2.75
Humidity	65.10	91.80	79.19	79.30	80.5	4.04	16.28
AirTemp	25.26	29.59	27.14	27.09	26.76	0.80	0.65
SeaSurfaceTemp	25.20	30.70	27.64	27.54	26.85	1.03	1.06

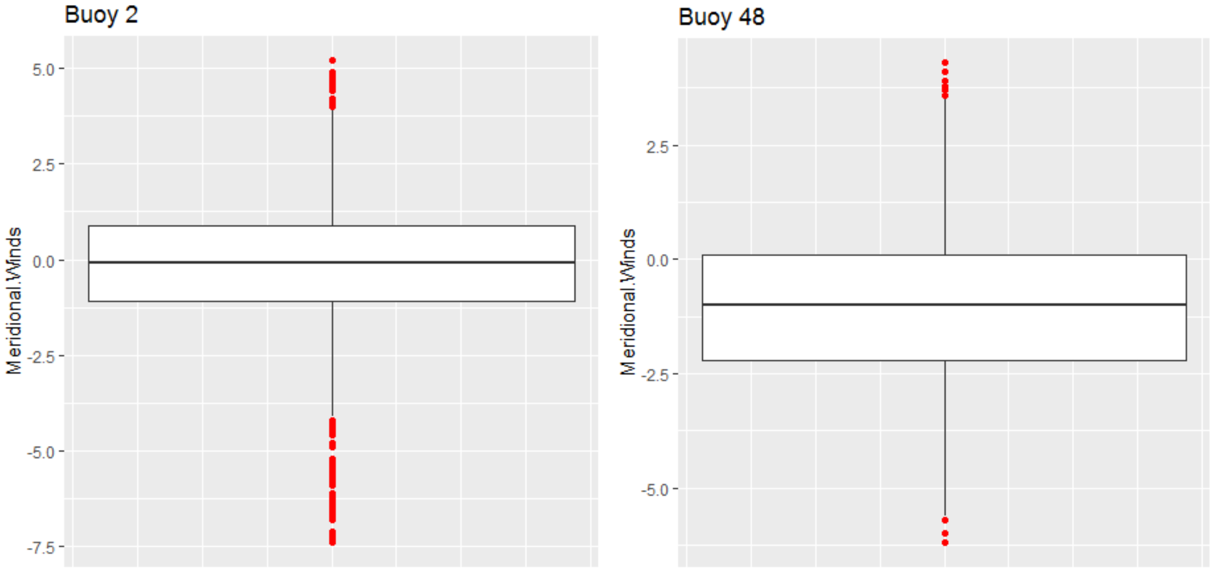
Visualization

In this part, we visualized all five variables for each buoy and put them side by side for better comparison.

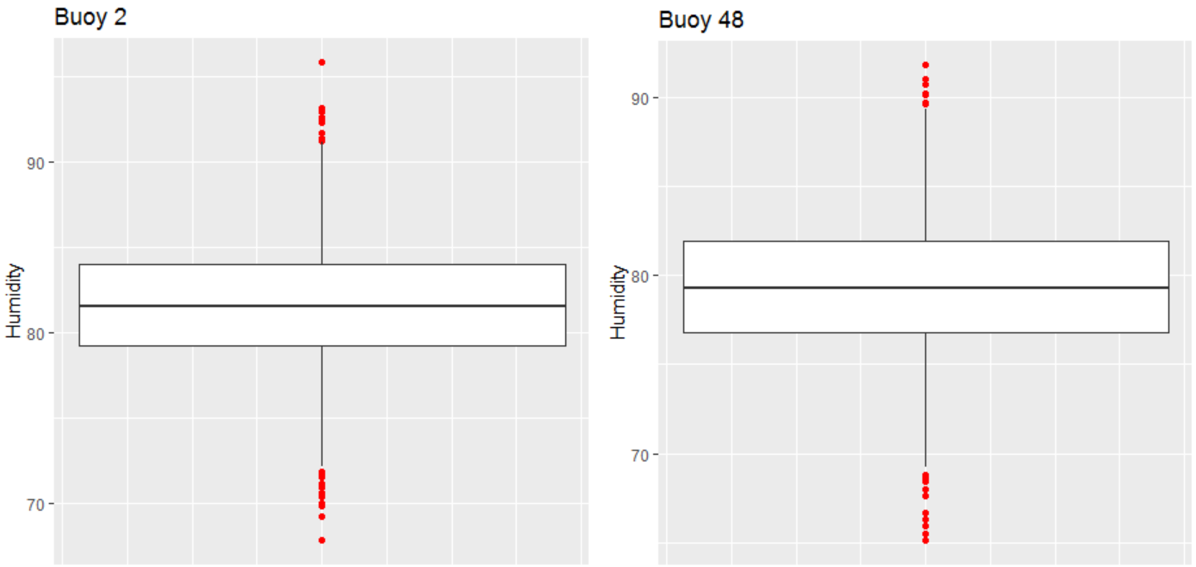
Zonal Winds



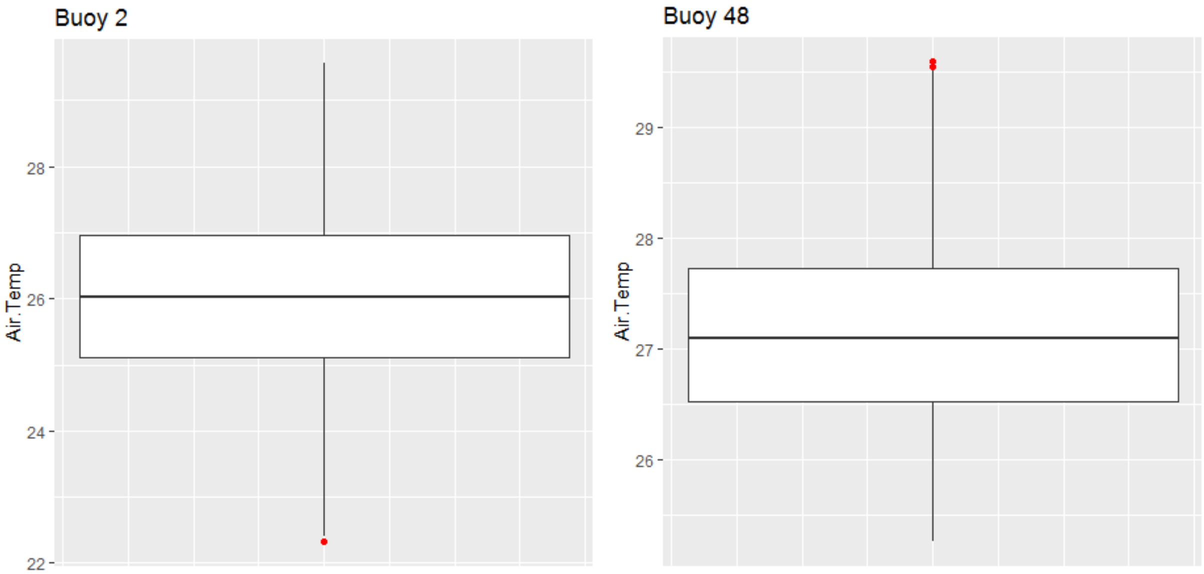
Meridional Winds



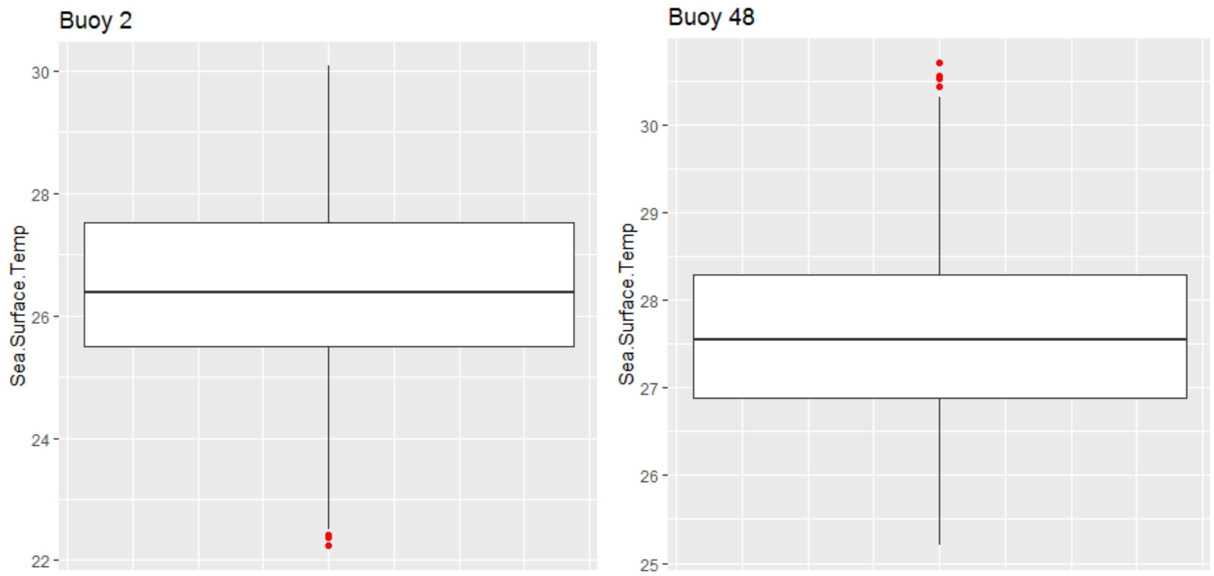
Humidity



Air Temp



Sea Surface Temp

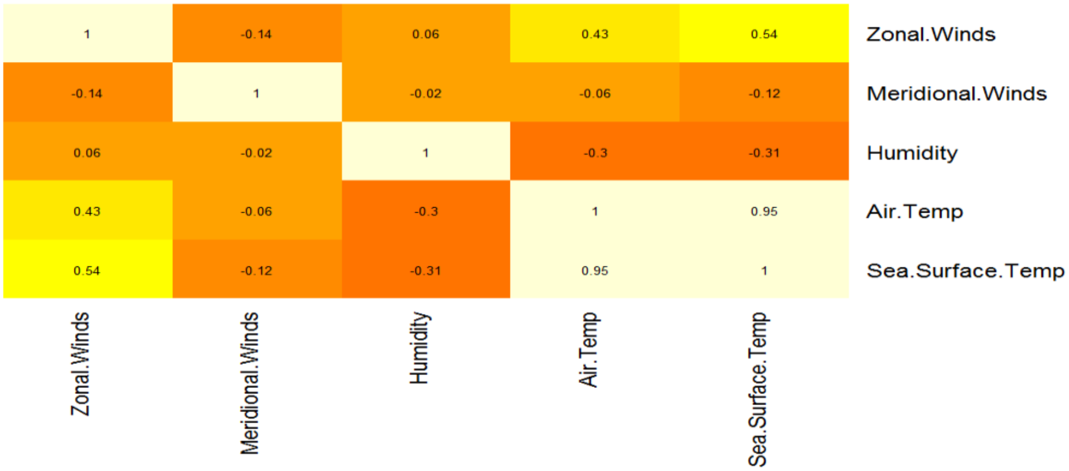


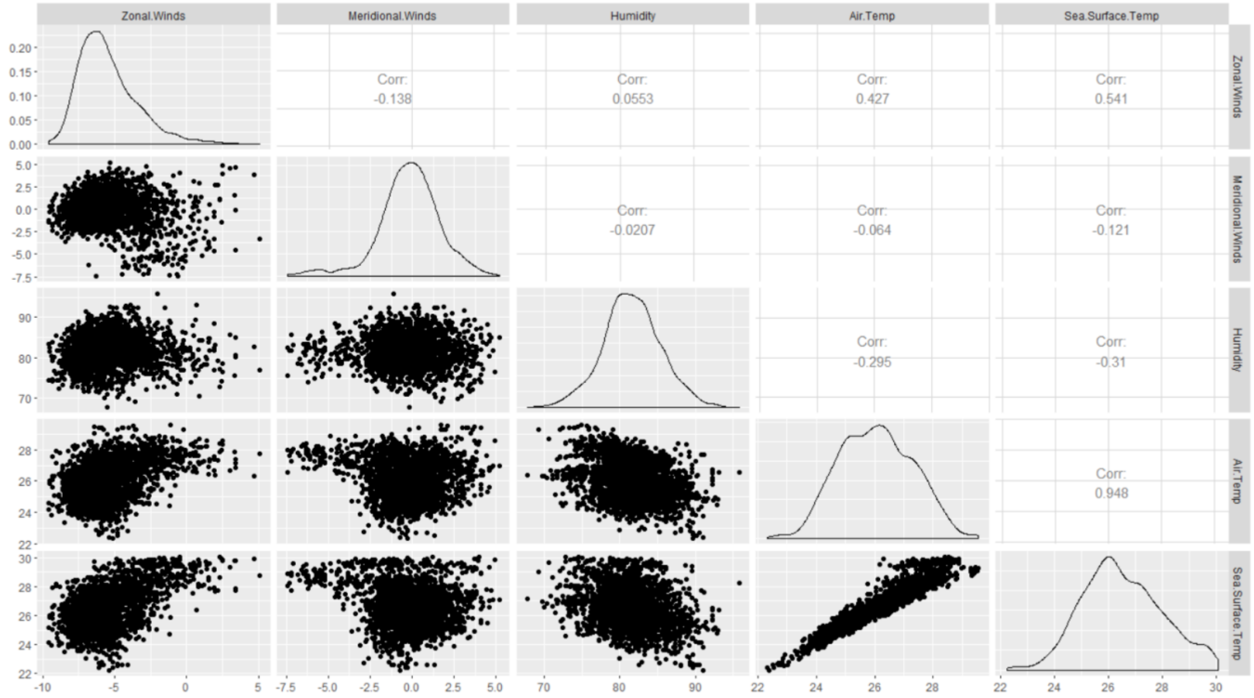
Correlation Matrix

In this part, we examined the correlation among independent variables for each buoy. The results would be crucial to help us determine if we would run into multicollinearity when conducting regression analysis.

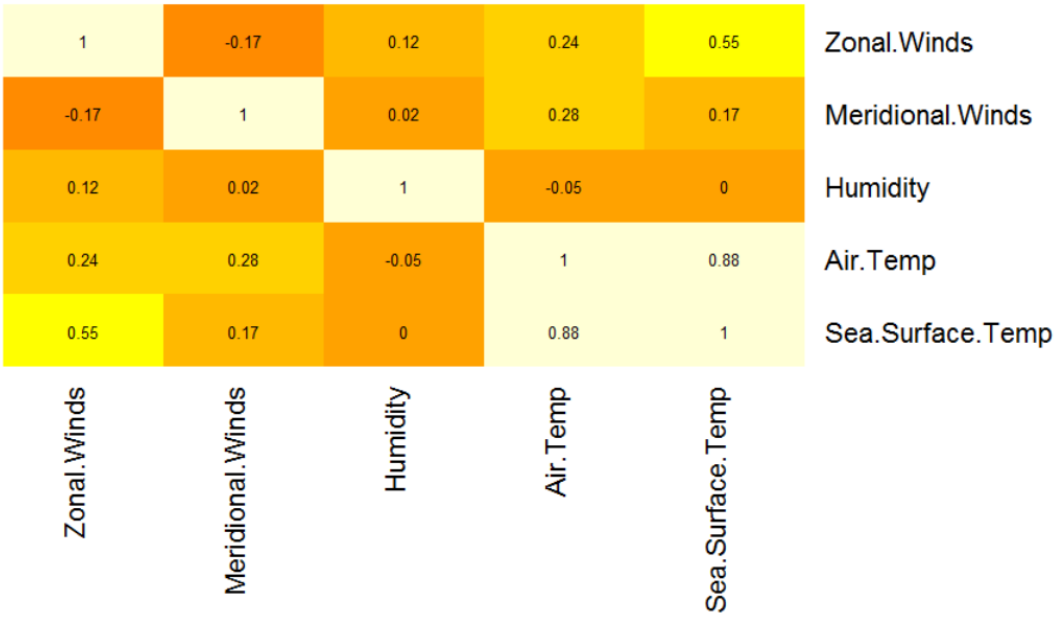
In general, sea surface temperature is highly correlated with air temperature as expected. However, the former is the dependent variable, so the high correlation between the two would not put us in trouble. Zonal wind is correlated with sea surface temperature as well. For both buoys, the correlation between zonal wind and sea surface temperature is above 0.5.

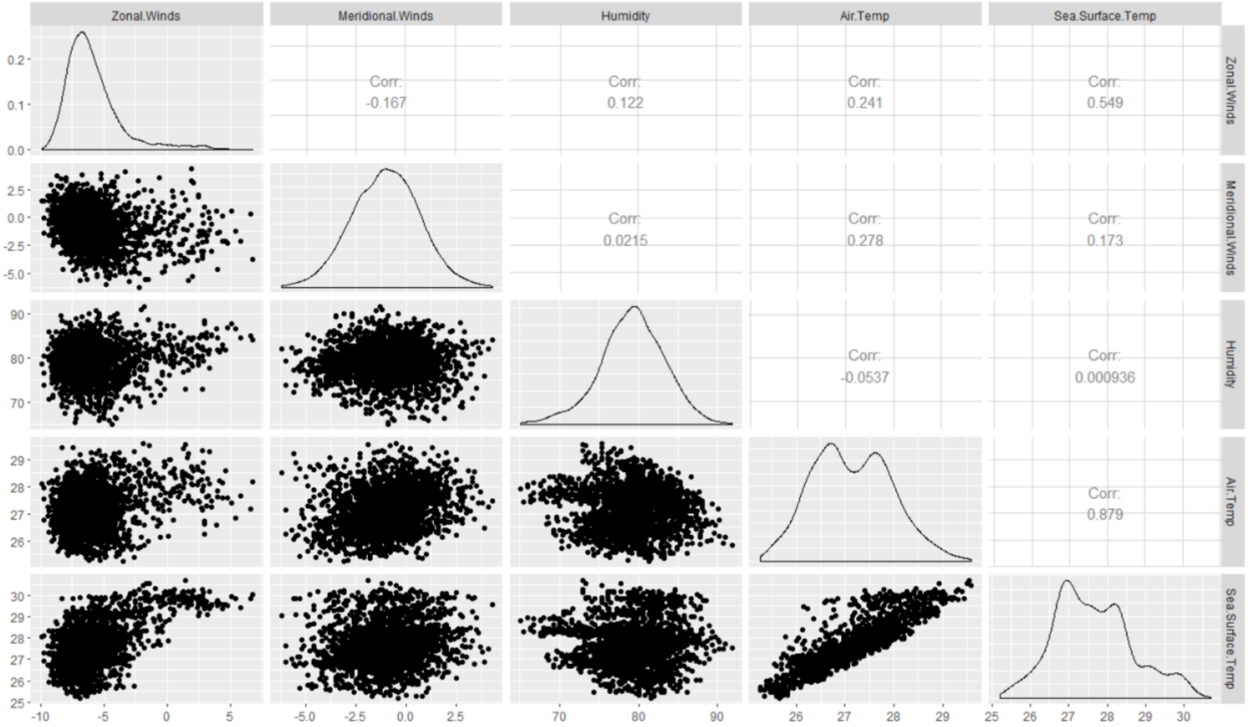
Buoy 2


1 -0.14 0.06 0.43 0.54 Zonal.Winds
-0.14 1 -0.02 -0.06 -0.12 Meridional.Winds
0.06 -0.02 1 -0.3 -0.31 Humidity
0.43 -0.06 -0.3 1 0.95 Air.Temp
0.54 -0.12 -0.31 0.95 1 Sea.Surface.Temp
Zonal.Winds
Meridional.Winds
Humidity
Air.Temp
Sea.Surface.Temp


Zonal.Winds
Meridional.Winds
Humidity
Air.Temp
Sea.Surface.Temp
Corr:
-0.138
Corr:
0.0553
Corr:
0.427
Corr:
0.541
Corr:
-0.0207
Corr:
-0.064
Corr:
-0.121
Corr:
-0.295
Corr:
-0.31
Corr:
0.948
Zonal.Winds
Meridional.Winds
Humidity
Air.Temp
Sea.Surface.Temp
0.20
0.15
0.10
0.05
0.00
5.0
2.5
0.0
-2.5
-5.0
-7.5
90
80
70
28
26
24
22
30
28
26
24
22
-10
-5
0
5
-7.5
-5.0
-2.5
0.0
2.5
5.0
70
80
90
22
24
26
28
22
24
26
28
30

Buoy 48


1
-0.17
0.12
0.24
0.55
Zonal.Winds
-0.17
1
0.02
0.28
0.17
Meridional.Winds
0.12
0.02
1
-0.05
0
Humidity
0.24
0.28
-0.05
1
0.88
Air.Temp
0.55
0.17
0
0.88
1
Sea.Surface.Temp
Zonal.Winds
Meridional.Winds
Humidity
Air.Temp
Sea.Surface.Temp


Zonal.Winds
Meridional.Winds
Humidity
Air.Temp
Sea.Surface.Temp
Corr:
-0.167
Corr:
0.122
Corr:
0.241
Corr:
0.549
Corr:
0.0215
Corr:
0.278
Corr:
0.173
Corr:
-0.0537
Corr:
0.000936
Corr:
0.879
Zonal.Winds
Meridional.Winds
Humidity
Air.Temp
Sea.Surface.Temp
0.2
0.1
0.0
2.5
0.0
-2.5
-5.0
90
80
70
29
28
27
26
30
29
28
27
26
25
-10
-5
0
5
-5.0
-2.5
0.0
2.5
70
80
90
26
27
28
29
25
26
27
28
29
30

Model Comparison

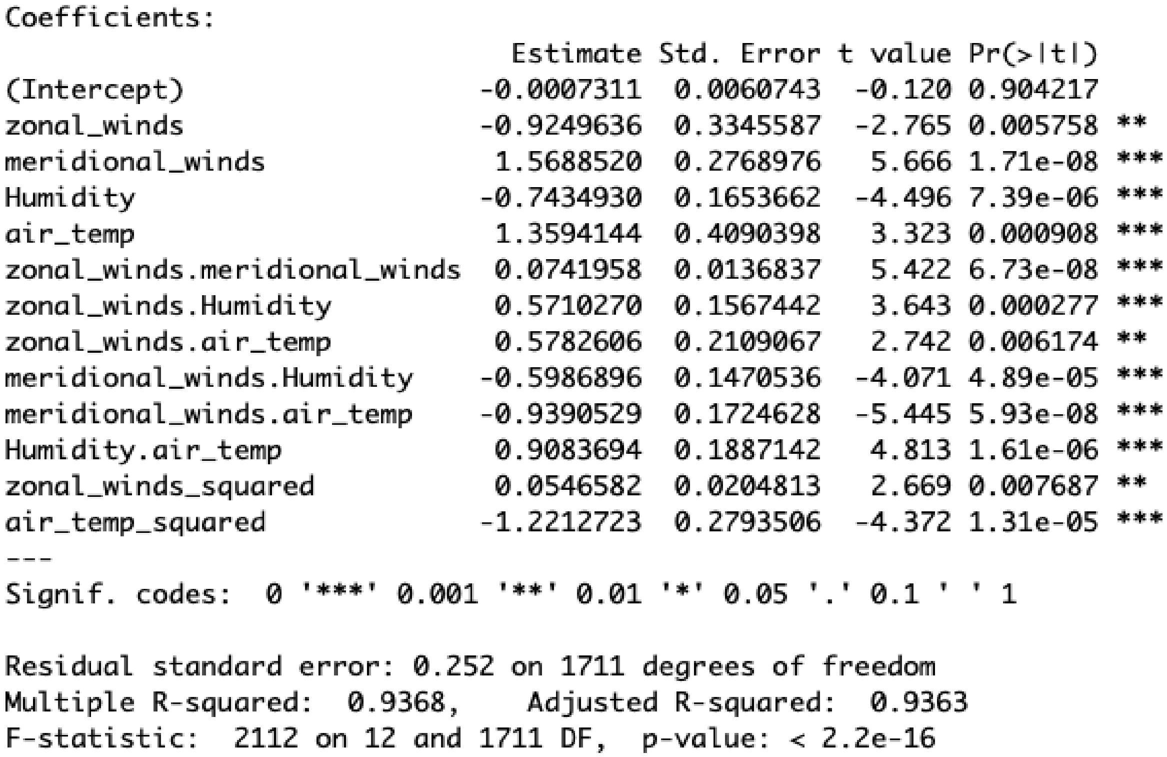
We will compare 3 models in total: Linear Regression, K-Nearest Neighbors and Decision trees. The discussion below shows the comparison of the results between our models and DataRobot's models.

Regression

Because buoy 2 had the most complete data with 2,298 entries out of 72 buoys, we only used the data from buoy 2 to apply the linear regression. Sea surface temperature is regressed against monthly dummies, zonal winds, meridional winds, humidity, air temperature, as well as their interaction and squared terms. We have four models running on different variable combinations: a model only on four key variables, a model with monthly dummies and four key variables, a model with interaction and squared terms, and a model including all the variables. As measured by r-squared metric, all four models yield similar results, but the model with interaction and squared terms provides the lowest RMSE. We therefore would focus on this model. The results of the regression are presented below:

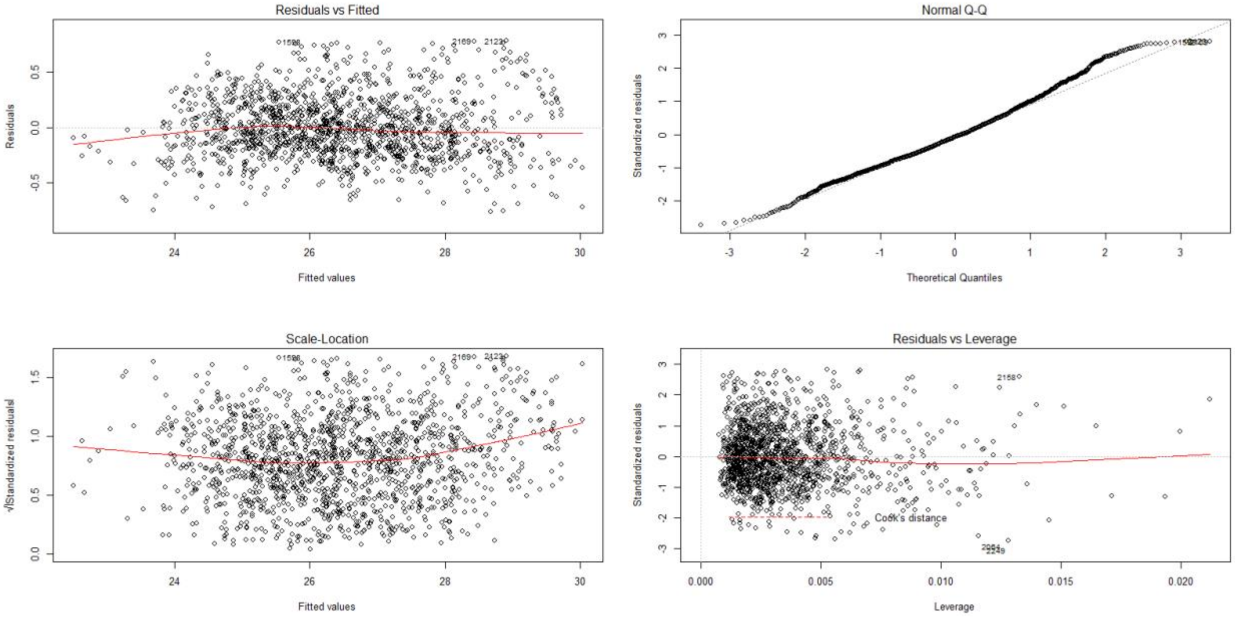
Comparison between the results of linear regression

Our Linear Regression:



```
Coefficients:
                              Estimate Std. Error t value Pr(>|t|)    
(Intercept)                 -0.0007311  0.0060743  -0.120 0.904217    
zonal_winds                 -0.9249636  0.3345587  -2.765 0.005758 ** 
meridional_winds             1.5688520  0.2768976   5.666 1.71e-08 ***
Humidity                    -0.7434930  0.1653662  -4.496 7.39e-06 ***
air_temp                     1.3594144  0.4090398   3.323 0.000908 ***
zonal_winds.meridional_winds  0.0741958  0.0136837   5.422 6.73e-08 ***
zonal_winds.Humidity         0.5710270  0.1567442   3.643 0.000277 ***
zonal_winds.air_temp         0.5782606  0.2109067   2.742 0.006174 ** 
meridional_winds.Humidity   -0.5986896  0.1470536  -4.071 4.89e-05 ***
meridional_winds.air_temp   -0.9390529  0.1724628  -5.445 5.93e-08 ***
Humidity.air_temp            0.9083694  0.1887142   4.813 1.61e-06 ***
zonal_winds_squared          0.0546582  0.0204813   2.669 0.007687 ** 
air_temp_squared            -1.2212723  0.2793506  -4.372 1.31e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.252 on 1711 degrees of freedom
Multiple R-squared:  0.9368,	Adjusted R-squared:  0.9363 
F-statistic:  2112 on 12 and 1711 DF,  p-value: < 2.2e-16
```

Residual Plotting (After Removing Outliers)



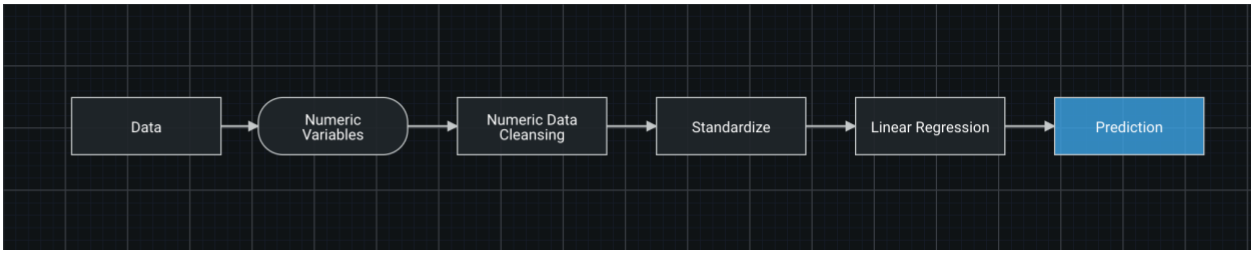
Adjusted R-squared: 0.9363

RMSE: 0.2673

In the linear regression study, our model concludes that not only four key variables but also higher power and interaction terms are statistically significant by large t values (> 1.96). They are critical for predicting the sea surface temperature. Therefore, these four measurements should remain to be collected by buoys for the purpose of analyzing the change of sea surface temperature. Our high Adjusted R-squared and low RMSE suggest that the residuals are distributed closely and evenly around the line. These values show that the variables are closely correlated to each other, which means that our model is a good fit.

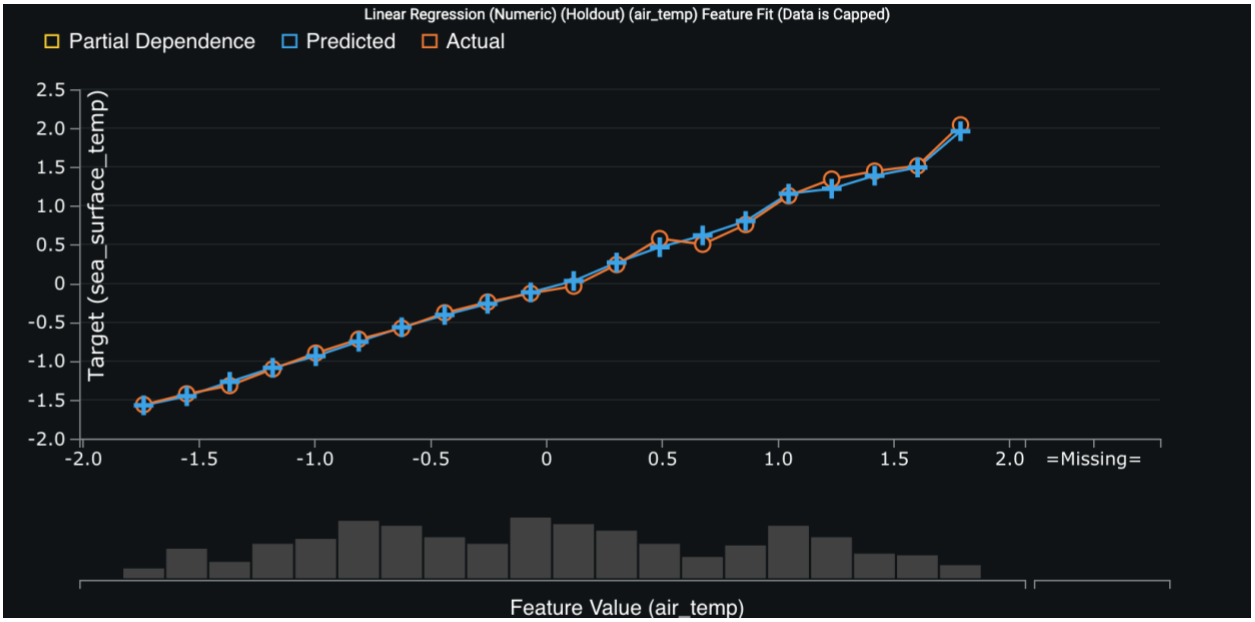
DataRobot’s Linear Regression:

Here is the blueprint of the model:

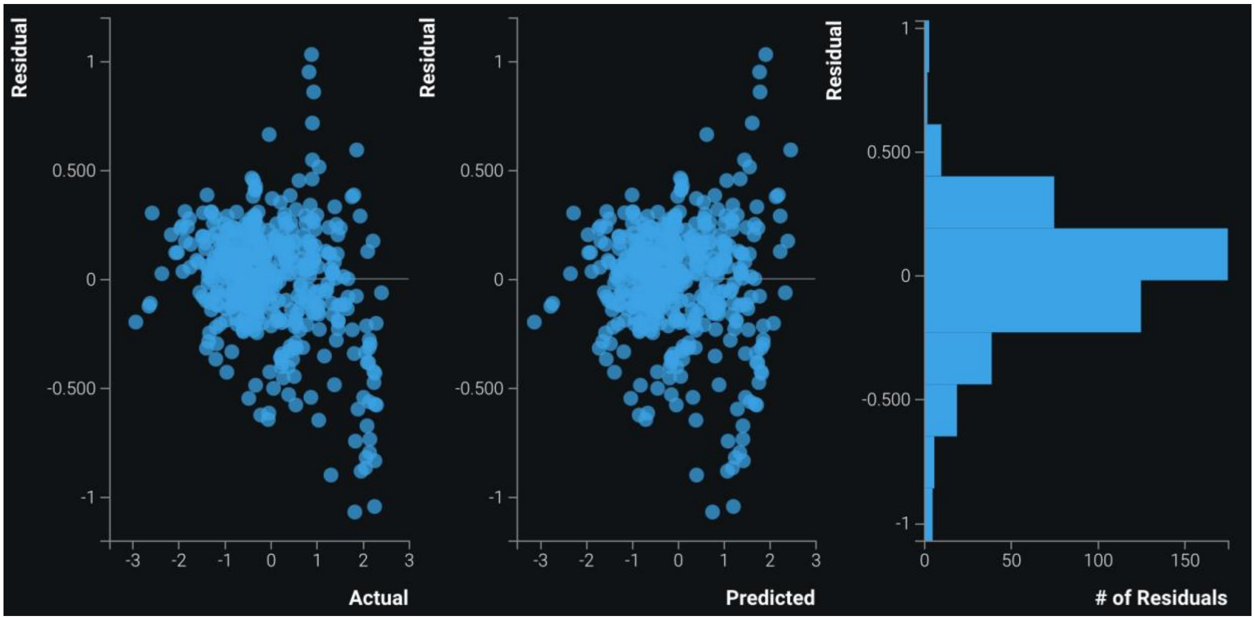


Here are how features contribute to the model:

Linear Regression Feature Fit



Residuals Distribution for regression



Adjusted R-squared: 0.9323

RMSE: 0.2701

DataRobot’s high Adjusted R-squared and low RMSE suggest that the residuals are distributed closely and evenly around the line. These values show that the variables are closely correlated to each other, which means that DataRobot’s model is a good fit. Because these values are nearly equivalent to our model’s, the two models are nearly equally efficient.

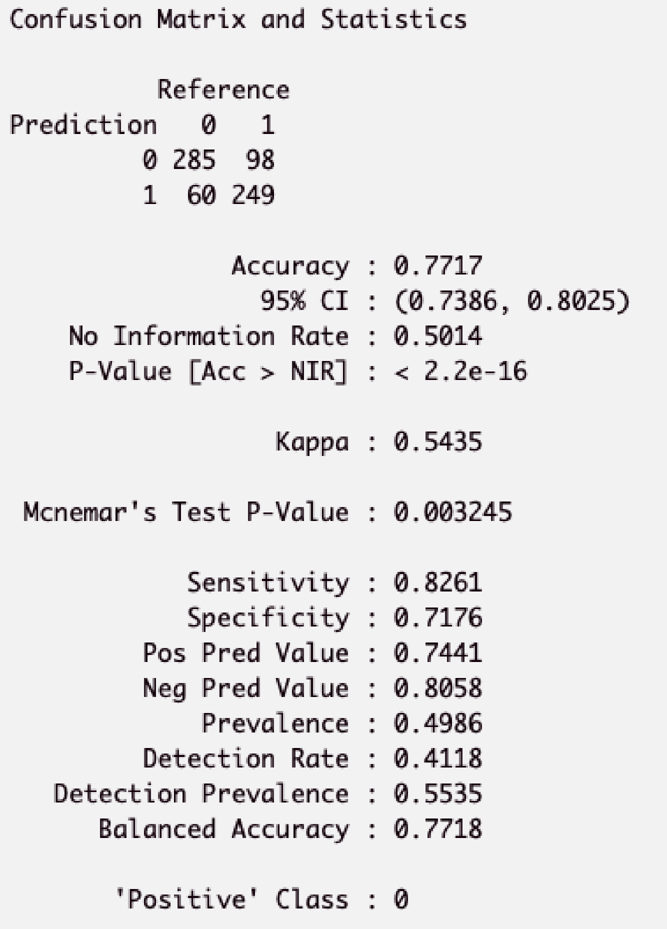
Classification

Our classification models run zonal winds, meridional winds, humidity, air temperature, and sea surface temperature against buoy, trying to classify the data according to the buoy which measured them. According to accuracy, our best model is Decision Trees. The second best is K-Nearest Neighbors. DataRobot’s best is K-Nearest Neighbors. Its second best is Decision Trees.

Comparison between the results of K-Nearest Neighbors

Our K-Nearest Neighbors:

In this analysis, we increased the complexity of the model by increasing the number of predictors to include the initial six predictors, their squared and interaction terms leading to 21 predictors. We then calculated the distance between two data points using the Euclidean distance method. Attached below is the result for our KNN model. The accuracy is 0.77, sensitivity 0.82 and specificity 0.71. In general, the model does well to classify the two buoys. As we discussed in the descriptive section, the two buoys present different patterns in air temperature and meridional winds, and we infer that these two variables actually help the model to differentiate one buoy from the other.

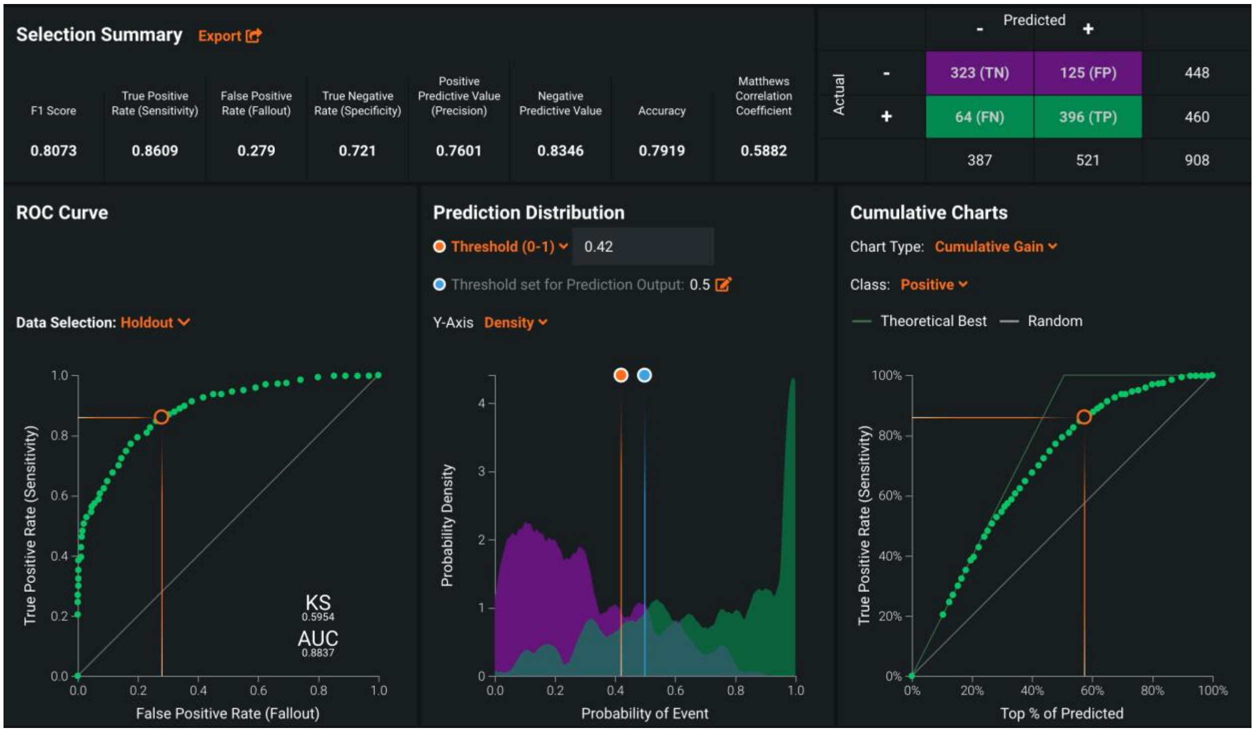


```
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 285  98
         1  60 249

               Accuracy : 0.7717
                 95% CI : (0.7386, 0.8025)
    No Information Rate : 0.5014
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.5435

 Mcnemar's Test P-Value : 0.003245

            Sensitivity : 0.8261
            Specificity : 0.7176
         Pos Pred Value : 0.7441
         Neg Pred Value : 0.8058
             Prevalence : 0.4986
         Detection Rate : 0.4118
   Detection Prevalence : 0.5535
      Balanced Accuracy : 0.7718

       'Positive' Class : 0
```

DataRobot’s K-Nearest Neighbors:

We used the Auto-tuned K-Nearest Neighbors Classifier (Euclidean Distance) model to analyze the initial 6 predictor variables.

Here is the result:



Selection Summary Export

F1 Score	True Positive Rate (Sensitivity)	False Positive Rate (Fallout)	True Negative Rate (Specificity)	Positive Predictive Value (Precision)	Negative Predictive Value	Accuracy	Matthews Correlation Coefficient
0.8073	0.8609	0.279	0.721	0.7601	0.8346	0.7919	0.5882

		Predicted -	Predicted +	
Actual	-	323 (TN)	125 (FP)	448
Actual	+	64 (FN)	396 (TP)	460
		387	521	908

ROC Curve

Data Selection: Holdout

KS 0.5954

AUC 0.8837

Prediction Distribution

Threshold (0-1) 0.42

Threshold set for Prediction Output: 0.5

Y-Axis Density

Cumulative Charts

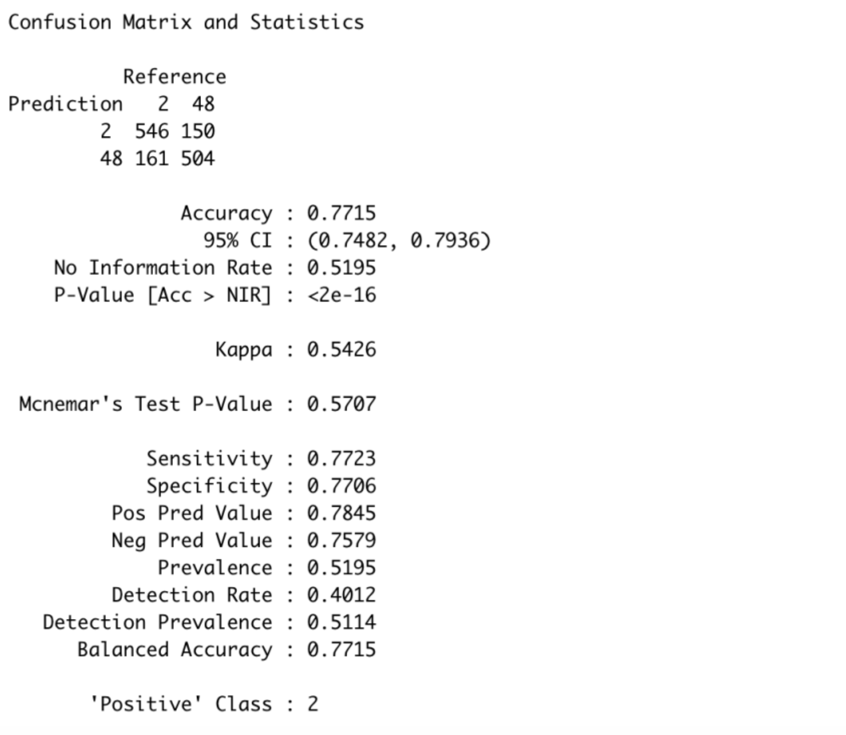
Chart Type: Cumulative Gain

Class: Positive

Accuracy is calculated as the total number of correct predictions. The Accuracy of our K-Nearest Neighbors is 0.7717, which means that our KNN can predict 77.17% of the whole dataset correctly. The Accuracy of DataRobot’s K-Nearest Neighbors is 0.7999, which means that DataRobot’s KNN can predict 79.99% of the whole dataset correctly. DataRobot’s model is slightly more accurate than ours. It is not surprising that the difference in accuracy is minimal. In KNN, the only parameter that can be tuned is the *number of neighbors*. This can easily be done in R as demonstrated in our hand-crafted models where we found that the optimal K = 3.

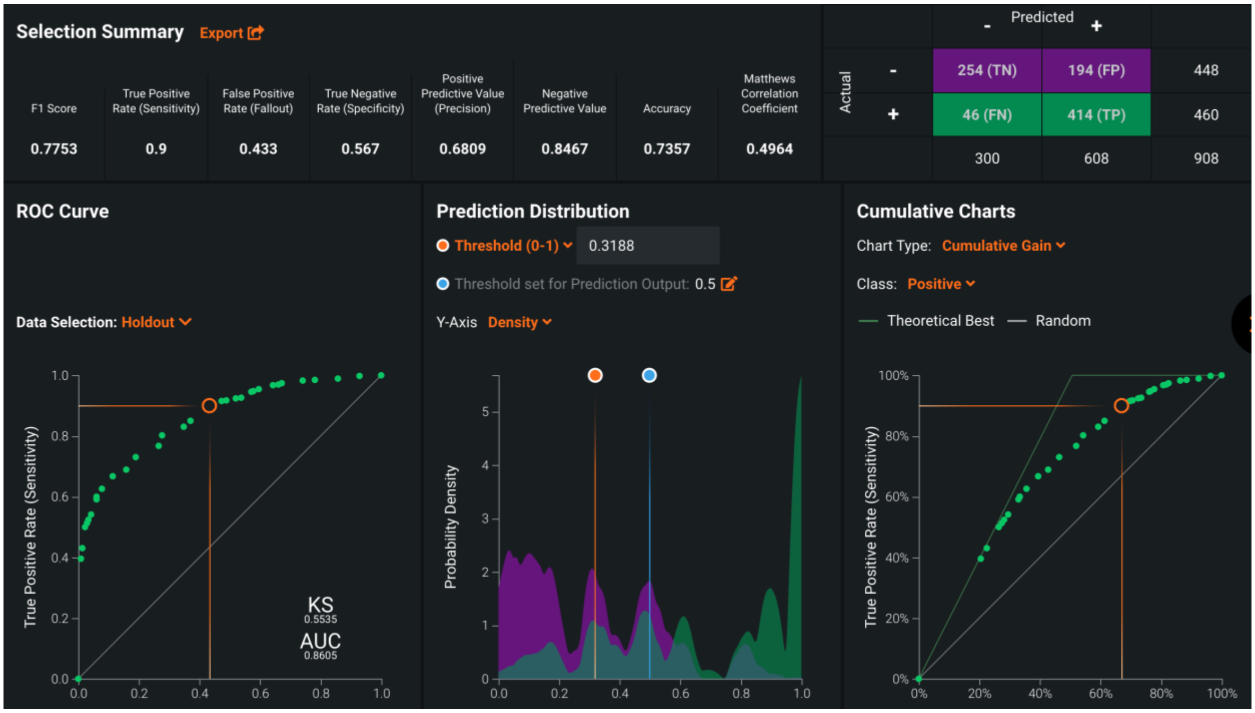
Comparison between the results of Decision Trees

Our Decision Trees:



```
Confusion Matrix and Statistics

          Reference
Prediction   2  48
        2  546 150
        48 161 504

               Accuracy : 0.7715
                 95% CI : (0.7482, 0.7936)
    No Information Rate : 0.5195
    P-Value [Acc > NIR] : <2e-16

                  Kappa : 0.5426

 Mcnemar's Test P-Value : 0.5707

            Sensitivity : 0.7723
            Specificity : 0.7706
         Pos Pred Value : 0.7845
         Neg Pred Value : 0.7579
             Prevalence : 0.5195
         Detection Rate : 0.4012
   Detection Prevalence : 0.5114
      Balanced Accuracy : 0.7715

       'Positive' Class : 2
```

DataRobot's Decision Trees:



Selection Summary Export

F1 Score	True Positive Rate (Sensitivity)	False Positive Rate (Fallout)	True Negative Rate (Specificity)	Positive Predictive Value (Precision)	Negative Predictive Value	Accuracy	Matthews Correlation Coefficient
0.7753	0.9	0.433	0.567	0.6809	0.8467	0.7357	0.4964

		Predicted -	Predicted +	
Actual	-	254 (TN)	194 (FP)	448
	+	46 (FN)	414 (TP)	460
		300	608	908

ROC Curve

Data Selection: Holdout

Prediction Distribution

Threshold (0-1) 0.3188

Threshold set for Prediction Output: 0.5

Y-Axis Density

Cumulative Charts

Chart Type: Cumulative Gain

Class: Positive

The Accuracy of our Decision Trees is 0.7715, which means that our Decision Trees can predict 77.15% of the whole dataset correctly. The Accuracy of DataRobot's Decision Trees is 0.7357. DataRobot's model is slightly less accurate than ours.

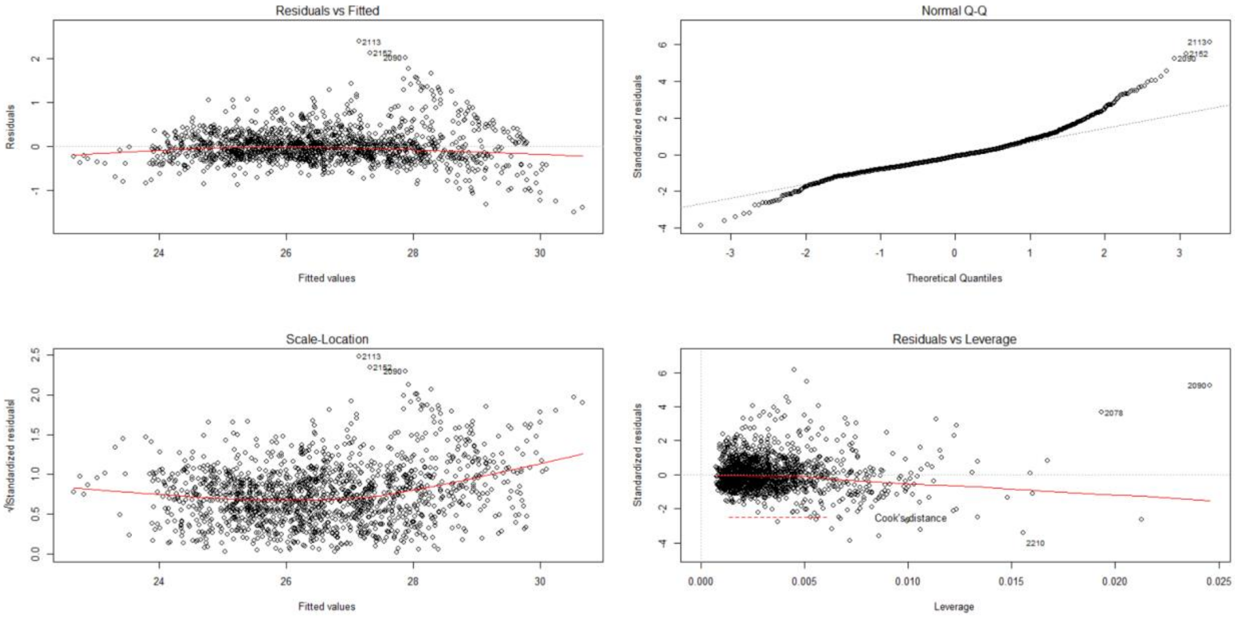
One thing worth noting when comparing Decision Tree model and KNN model generated by data robot is that decision tree produced much more FP (false positive) prediction than KNN, i.e. decision tree is an "over-optimistic" in this case that it wrongly classified data point from buoy 48 as that from buoy 2.

Management of Outliers, Overfitting and Target Leakage

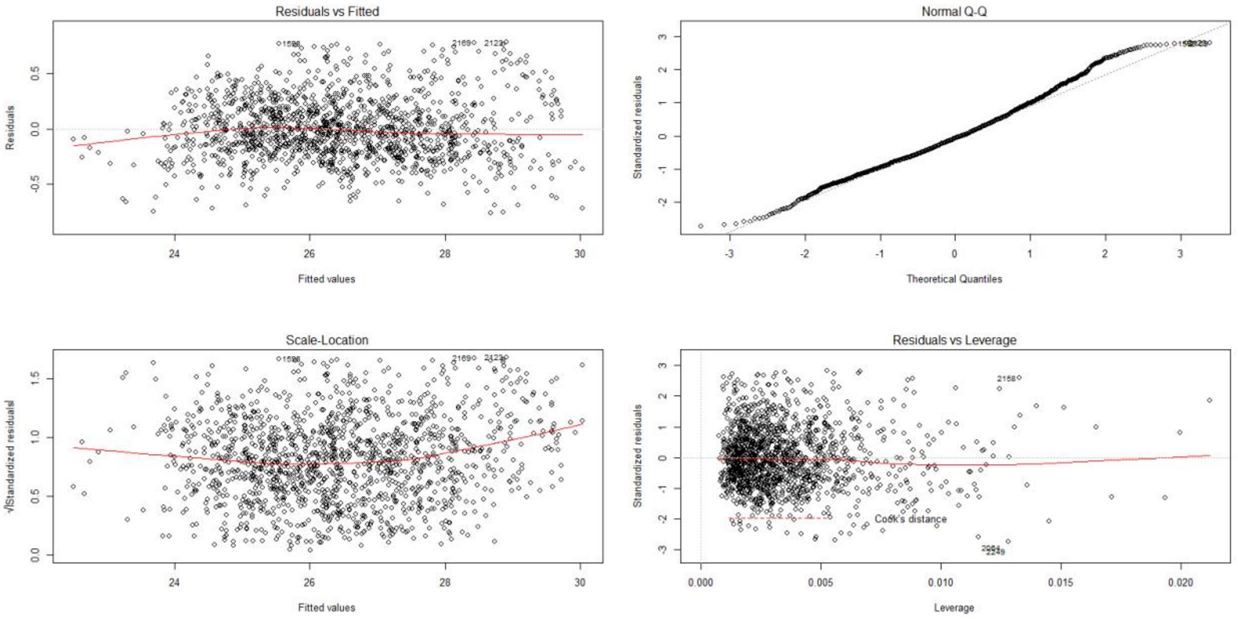
In regression analysis, we discovered numerous outliers by plotting the normal Q-Q plot and residuals and eventually removed 87 data points one by one from the training set.

Two plots of before and after removing outliers are presented below for comparison:

Residual Plotting (Original)



Residual Plotting (After Removing Outliers)



As we can see from the second plot, the Normal Q-Q plot aligns more with the straight line at two ends, while in all other three subplots, data points stay around the horizontal line: x = 0. Though the plot says that the dataset is far from perfect, we usually do not wish to eliminate all outliers. There is a trade-off of removing outliers and including information conveyed by them. The plots above are well enough and justify the usage of linear regression.

Also, in our project, we checked the Accuracy, Sensitivity, and Specificity of our models. We kept adjusting our models until the Accuracy, Sensitivity, and Specificity are all high, but not close to 1.

In Datarobot, cross-validation is conducted for each model in order to reduce bias introduced by training/testing data split. Cross-validation is crucial in the case of decision tree modelling. Decision tree, with its ‘greedy’ nature, always tends to overfit the training data, and thus do not predict well on testing data. Cross-validation mitigates this problem by running several decision trees models on different splits.

There were several times when we carefully tuned the dataset and model to avoid target leakage. Firstly, we excluded geographical information from the dataset and refrained from using it during the whole modelling process. A buoy tended to moor at a fixed position, though sometimes it traveled along the equatorial line. With the help of geographical information, it is way too easy to infer the index of a buoy. However, in the real case scenario, we can lose geographical information and index of a buoy at the same time, as a result, the classification model should be able to identify the index of a buoy without geographical information.

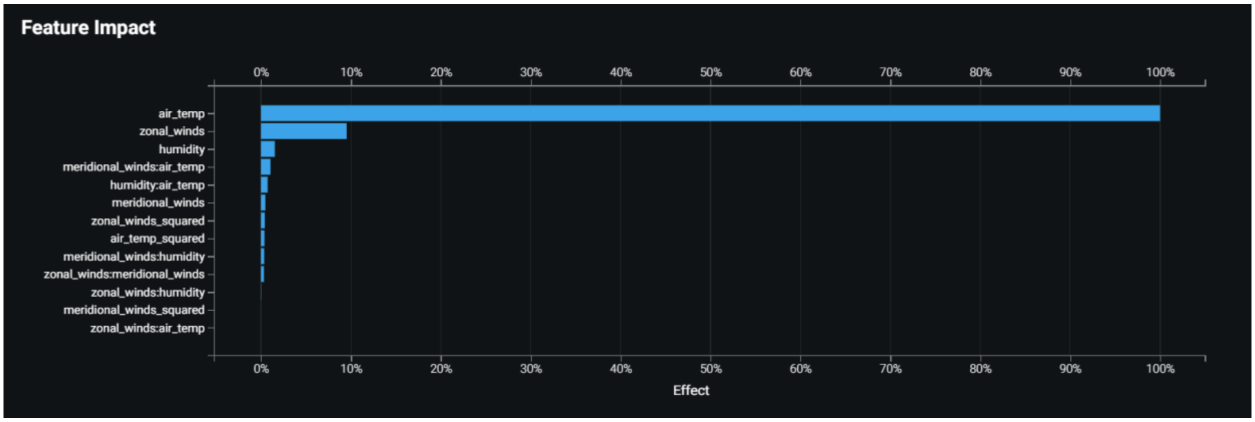
Secondly, on DataRobot, we used the feature list with leakage removed to run the models. We don’t know what runs behind the leakage removal algorithm, but the resulting accuracy suggests that models generated by Datarobot did not encounter the problem of target leakage. Results from Datarobot and from our models align well, which is also a positive sign for the robustness of the models.

Overall, we were careful at handling the dataset, and feeding in only data recorded by the buoy at that time. We tuned the models to get rid of outliers as well. As a result, models do a good job, but not ‘too good’ either.

Insights

Linear Models

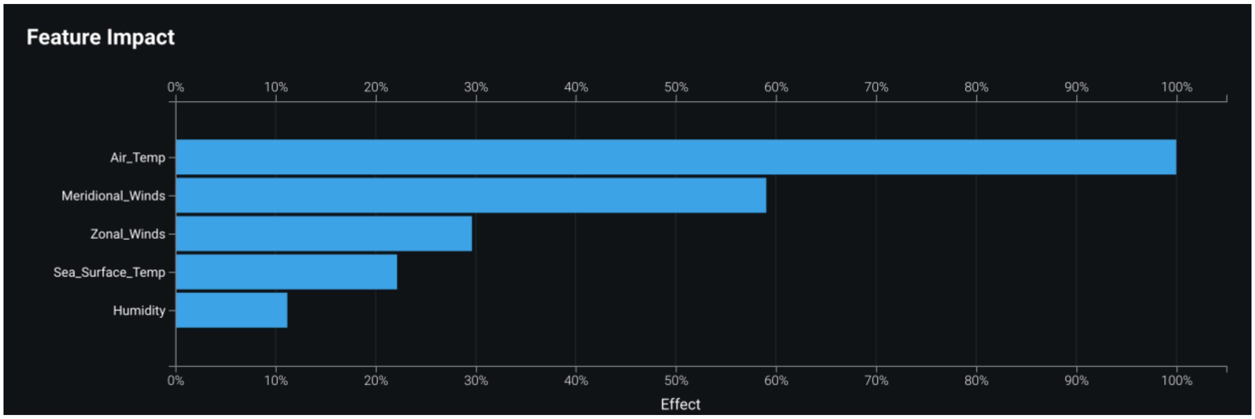
Both our and DataRobot linear models show us that air temperature is critical for forecasting sea surface temperature, which makes sense as they are both temperature measurements. However, analysts should note that they differ in timing because air temperature heats up faster than water temperature. Zonal_winds also show a good impact on forecasting sea surface temperature, but its feature impact is much less than the one of air_temp. Although the higher power and interaction terms in our linear model are statistically significant and do increase the model predictability, their feature impact are very low in DataRobot, shown by the graph below.



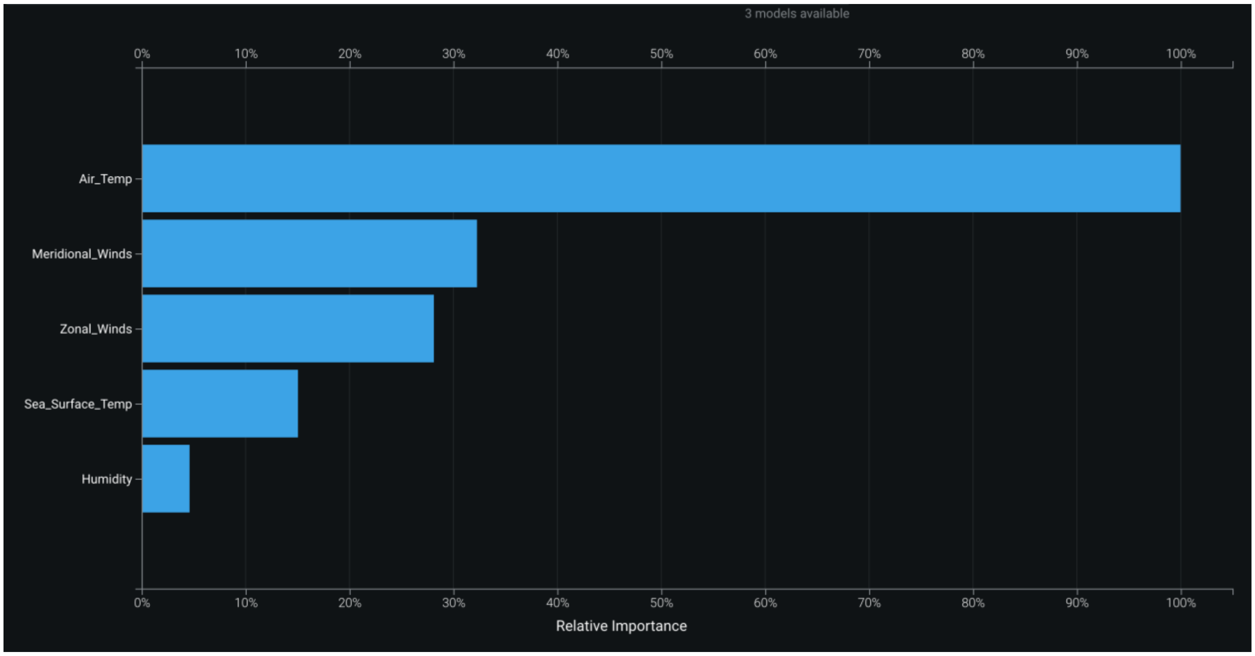
Predicting temperature is only one element for understanding the El Niño effects in the tropical Pacific. All of the original five factors should all play an important role in studying climate activities and unusual changes. We are limited by our scope to expand further research on other measures. Keeping those measures tracked is necessary for getting a holistic picture of the Pacific oceanography.

Classification Models

The comparison section above demonstrates that K-Nearest Neighbors and Decision Trees are efficient models for classifying which buoy measures the data. DataRobot indicates that air temperature is the most important variable in the KNN model to classify the buoy. Please refer to the graph below.



DataRobot indicates that air temperature is the most important variable in the Decision Tree model to classify the buoy. Importance at the tree node is computed by measuring the forecast accuracy improvement from that node weighted by the number of samples coming through that node. This suggests the two buoys were located further away from each other (by latitude and longitude) since temperature variations across the equator are very minimal.



The two graphs above have the same pattern, both suggest that all 5 variables are important for classification. Among them, air temperature is the most significant variable to classify the buoys.

Conclusion & Reflection

We wrangled the data using Trifacta and used Python to label them into 72 buoys. Because the buoys had been placed in the ocean for years, some older buoys might have technical issues not measuring those variables accurately or losing tracks. The dataset had numerous missing values. As a result, simply removing variables due to missing values may cause omitted variable bias when conducting modeling. Because the given dataset does not label buoys, we thought that it is important to classify the buoys and find out which buoy collected relatively complete data. Further analysis should be operated on the data from the one or two reliable buoys.

The regression analysis suggests that sea surface temperature can be predicted with good accuracy. Certain variables such as meridional wind and humidity have coefficients that are statistically significant while regressing the sea surface temperature on them. Moreover, significant interaction terms indicate that interacting variables may play a role on the sea surface temperature as well. Afterall, it is highly probable that we detect an unusual sea surface temperature data point when comparing our predictive results to the actual temperature, and the abnormal sea surface temperature may serve as a prophetic sign for the El Nino Effect.

Our preliminary classification gives 72 distinct buoys. By later modeling as we already discussed earlier in the report, we were able to classify the buoys more scientifically by automating KNN and decision trees. These models further help distinguish the data source and improve the data reliability.

Clustering is not a focus in our report, but in our A4 assignment, we used the K-Means clustering to divide the data into two clusters: the wet season and the dry season. These clusters are useful because they provide information beyond the buoy index. Readers are able to have an insightful first impression by learning that the data contain seasonal characteristics.

Our study on the El Nino effect in the equatorial Pacific may not provide insights on climate change in a way that an environmental science study does, but we investigated how we should treat the measurement from this dataset and generated statistical insights of those measurements.

Last but not least, we found some other datasets that look very promising to generate insights into the El Nino Effect. For example, a dataset from *www.ncdc.noaa.gov* contains information on storm data. The data includes hurricanes, tornadoes, thunderstorms, hail, floods, drought conditions, lightning, high winds, snow, and temperature extremes. The dataset, once connected with tropical atmosphere ocean data we are using, may shine light on a predictive model for El Nino Effects based on sea surface temperature.

Another site of interest related to the ENSO cycles is available here: https://www.cpc.ncep.noaa.gov/products/precip/CWlink/MJO/enso.shtml This site contains information on twelve areas of the world that have demonstrated ENSO-precipitation relationships. We may be able to closely observe how climate variations in equatorial oceanic areas impact precipitations through the ENSO cycles.

Methodology we have introduced in this report can well be applied to these scenarios and hopefully, El Nino effects can be closely predicted and monitored, so that they do less harm to people’s normal lives and the earth environment as a whole.