Predicting Best Time to Shoot Some Curls

Matias Gonzalez

About Me

- I love surfing, skating and snowboarding
- I can solve a Rubik’s cube in less than 2 minutes
- Originally from El Salvador
- Went to Florida State

Motivation

Have you ever been next to a beautiful beach but as soon as you take out your phone and you notice your subscription to Surfline has expired? How are you going to know when it's the best time to surf? The goal of this project is very simple, to predict at what time will waves be at their highest. We are going to be taking a look at an oceanic dataset to see if we can create a model capable of predicting wave height.

EDA

The dataset contains 75 individual features and has close to a billion entries. This is a huge dataset that is feature-heavy and has a lot of data points. To simplify the project and make it feasible to run on a common machine we had to cut the dataset by a great amount. My next idea was to look at the NaN's since maybe dropping those could provide a workable dataset. This is what some key feature percent of NaN's per columns looks like.

1. I will be favoring using latitude and longitude over country code since it complete

2. I will have to drop wind_speed, swell_height, wave_direction and swell_speed, from the data, since there are way too many NaNs meaning it is mostly incomplete or not precise data

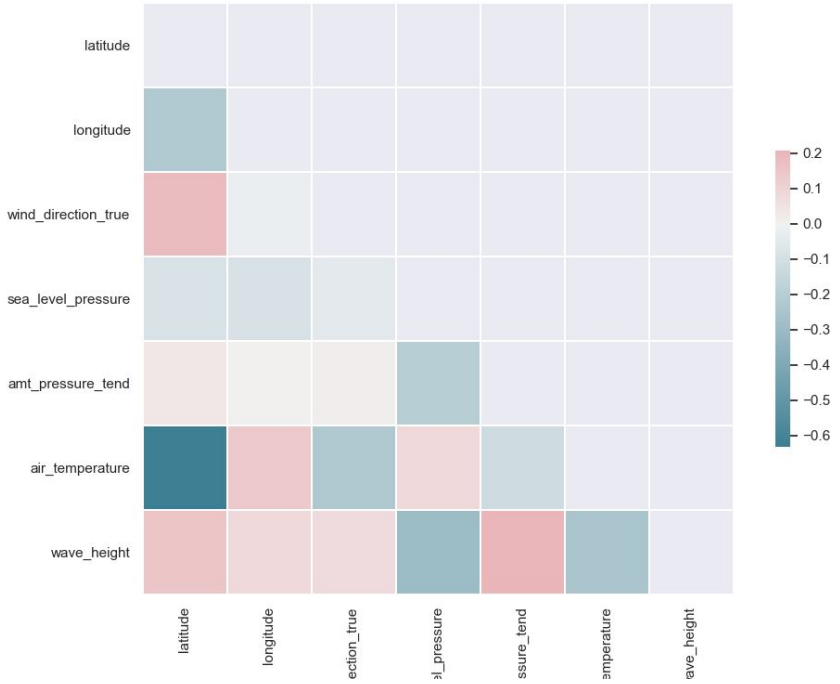
3. I can ignore the other date columns and only use the timestamp

4. Since I am trying to predict wave_height this will be the biggest contributor to row dropping.

Feature Name	% NaN's per column
latitude	0
longitude	0
sea_surface_temp	39.6
country_code	93.7
wind_speed	95.9
wind_direction_true	0
amt_pressure_tend	74.1
air_temp	5.5
sea_level_pressure	7.3
wave_direcion	100
wave_height	89.9
wave_speed	97.5
swell_height	98.8
swell_speed	100
timestamp	0

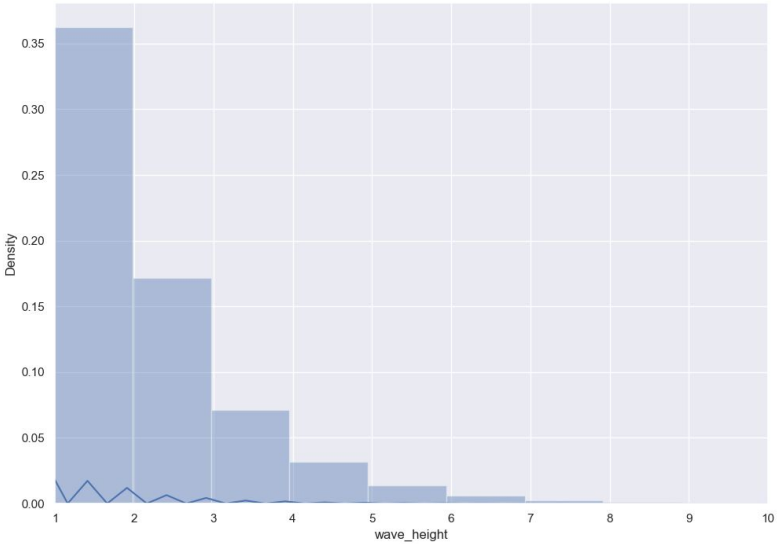
EDA

As we can see from out correlation matrix our variables are correlated but the strength is not too high. The only two variables that have a high positive correlation of around .60 are air temperature and latitude. This makes sense since the further away you are from the equator the more extreme temperatures get.



EDA

The graph is very left-skewed, meaning the distribution of our waves is mostly toward the smaller side. After closer inspection, I can see that a little above 35% of our waves are less than 1 meter, and around 17% are around 2 meters. This means that our target of big waves is the minority of the distribution making this even harder.



EDA

We can gather from the scatter plot that there are a lot of errors in this dataset. We can see where the normal distribution was starting to form underneath all the noise. This is probably going to be very impactful when creating our model.

Results

I was very determined to use a neural network to make my predictions due to the sheer amount of data I was working with. I decided to use a sequential model with the basic 3 layers. Unfortunately due to time constraints I only had time for one model run with my complete dataset which took a whopping 8 hours to run. This means I had no time to tune my model or try to make any adjustments to it.

Since my data is time-based we had to apply a time series train test split in order to respect time. this was done easily since we had a timestamp column. My model was not the best at predicting when the best surf time is going to be. I used kfold and mean squared error as a means of cross-validation. I got a score of 0.185, which s pretty high. this means our model is usually around 18.5 percent incorrect.

I had a pretty low dropout rate for regularization which could have been tuned higher in order to account for specialization. In general, I would have liked a bit more time in order to work more on the model and get a better result from it. It was pretty hard to work on something that took 8 hours to run when the project had a one-week time constraint.

Conclusions and Next Steps

One major stepback in this project was the amount of error present in the dataset. Upon further inspection, it seems that all of the indicators for accuracy have very high numbers meaning our data is not that accurate. Looking back at my steps I would have liked ot keep the wind_speed column. Even though we were going to lose a lot of data we had enough data to simply drop around 96 percent of it. It would have probably been beneficial to the project as a whole as it would have probably diminished our run times.

Questions?

Matias Gonzalez

matiasgonzalezrivera@gmail.com
https://www.linkedin.com/in/matiasgonzalezrivera/
https://github.com/matiasgonz

Appendix

My biggest note is Chris suggested I change datasets and I should have listened, this was very difficult to work with due to its sheer size and our time limitation when running it.

Look into monte Carlo augmentation in order to work with wind_speed.

Appendix

Feature Name	Data Type
year	INT
month	INT
day	INT
hour	FLOAT
latitude	FLOAT
longitude	FLOAT
imma_version	INT
attm_count	INT
time_indicator	INT
latlong_indicator	INT
ship_course	INT
ship_speed	INT
national_source_indicator	INT
id_indicator	INT
callsign	STR
country_code	STR
wind_direction_indicator	INT
wind_direction_true	INT
wind_speed_indicator	INT
wind_speed	FLOAT
visibility_indicator	INT
visibility	INT
present_weather	INT
past_weather	INT
sea_level_pressure	FLOAT
characteristic_of_ppp	INT
amt_pressure_tend	FLOAT
wbt_indicator	INT
wetbulb_temp	FLOAT
dpt_indicator	INT
dewpoint_temp	FLOAT
sst_measurement_method	INT
sea_surface_temp	FLOAT
total_cloud_amount	INT
lower_cloud_amount	INT
lower_cloud_type	STR
cloud_height_indicator	INT
cloud_height	STR
middle_cloud_type	STR
high_cloud_type	STR
wave_direction	INT
wave_period	INT
wave_height	FLOAT
swell_direction	INT
swell_period	INT
swell_height	FLOAT
box_system_indicator	STR
ten_degree_box_number	INT
one_degree_box_number	INT
deck	INT
source_id	INT
platform_type	INT
dup_status	INT
dup_check	INT
track_check	INT
pressure_bias	INT
wave_period_indicator	INT
swell_period_indicator	INT
second_country_code	INT
adaptive_qc_flags	STR
nightday_flag	INT
trimming_flags	STR
ncdc_qc_flags	STR
external	INT
landlocked_flag	INT
source_exclusion_flags	INT
unique_report_id	STR
release_no_primary	INT
release_no_secondary	INT
release_no_tertiary	INT
intermediate_reject_flag	INT
timestamp	OBJ