Surfability Prediction Using Machine Learning

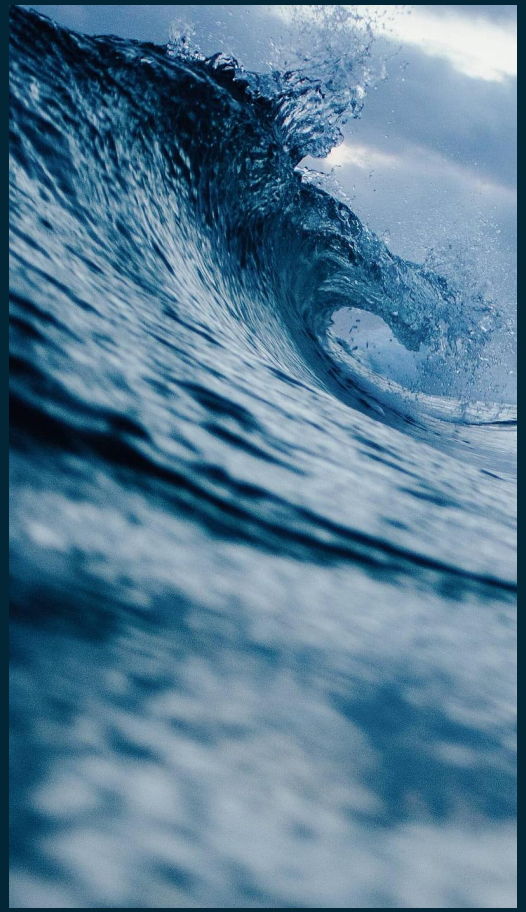
Dylan Hernandez, Micah Chen, Bryce Uota, Michael Man

How Do We Know If It's a Good Day to Surf?

- Hard to combine all variables mentally

- Surf quality depends on many changing factors

- Goal: Build a model to classify surfing conditions



How We Defined “Surfable”

Surfability = 1 iff 3/4 principle conditions met:

Wave height **0.9–2.5 m** (~3–8 ft)

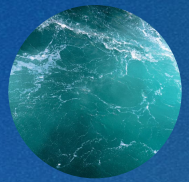
Swell period >= **9 s**

Wind <= **6 m/s**

Tide **-0.3-1m**

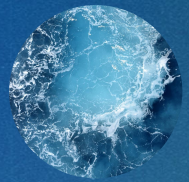
Surfability = 0 otherwise

Data Sources



NDBC

National Data Buoy Center



CO-OPS

Center for Operational Oceanographic Products and Services

ERDDAP

Environmental Research Division's Data Access Program

Surf & Weather Variables

Ocean

WVHT_M — Wave height (m)

DPD_S — Wave period (s)

MWD_DEG — Wave direction

TIDE_M — Tide level (m)

Atmosphere

WSPD_MPS — Wind speed (m/s)

WD_DEG — Wind direction

GST_MPS — Gust speed

ATMP_C — Air temp

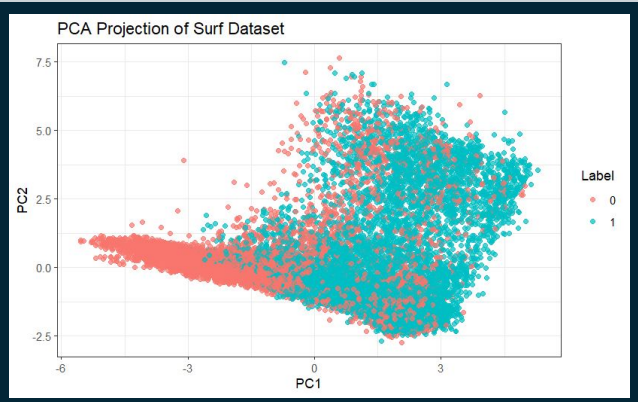
WTMP_C — Water temp

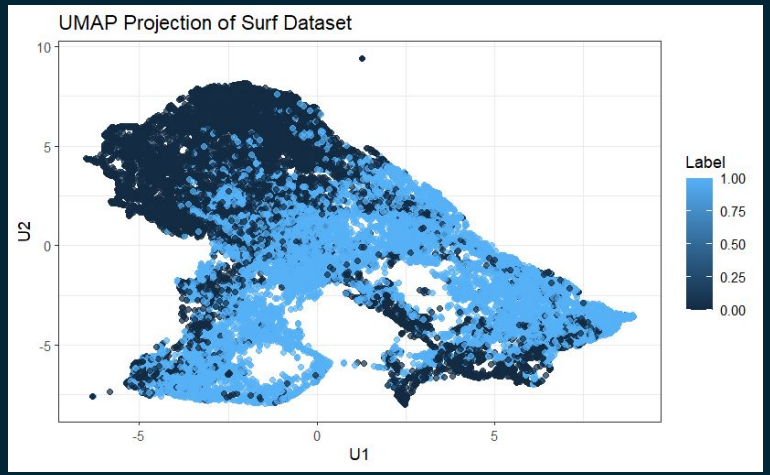
BAR_HPA — Pressure

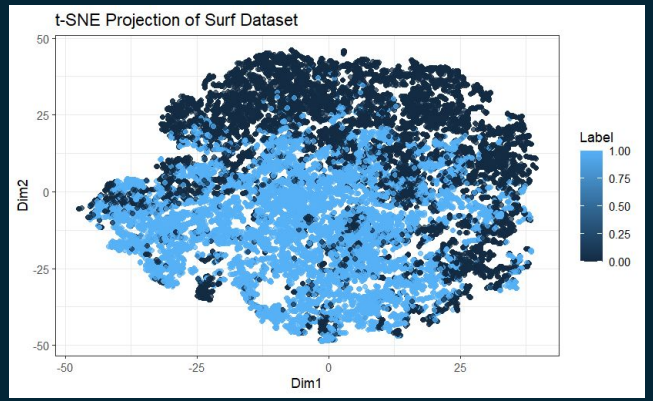
How the Data Was Prepared

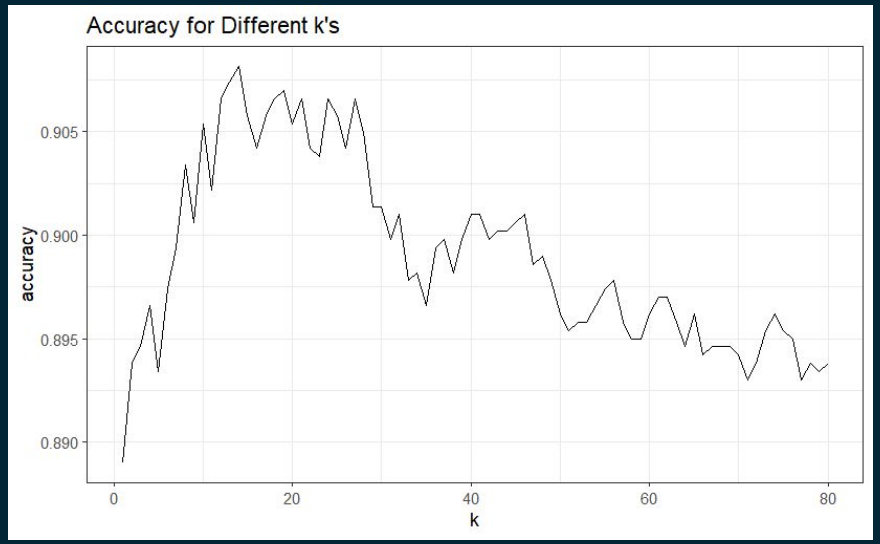
- Cleaned missing / invalid values (removed rows with any missing values, weren't many)
- Applied surfability rule to create labels
- 2022-2023 is training, 30% of 2024 data was validation set (assigned randomly), other 70% testing set

Visuals


PCA Projection of Surf Dataset
7.5
5.0
2.5
0.0
-2.5
PC2
-6
-3
0
3
PC1
Label
0
1


UMAP Projection of Surf Dataset
10
5
0
-5
U2
-5
0
5
U1
Label
1.00
0.75
0.50
0.25
0.00


t-SNE Projection of Surf Dataset
50
25
0
-25
-50
Dim2
-50
-25
0
25
Dim1
Label
1.00
0.75
0.50
0.25
0.00



Model Choice: K-Nearest Neighbors

- Uses euclidean distance between positions
- Works well with overlapping clusters
- Can handle wide variety of data types
- Easy to interpret

Results

Accuracies

Test Accuracy: about 91%

Train Accuracy: about 93%

F1 Score

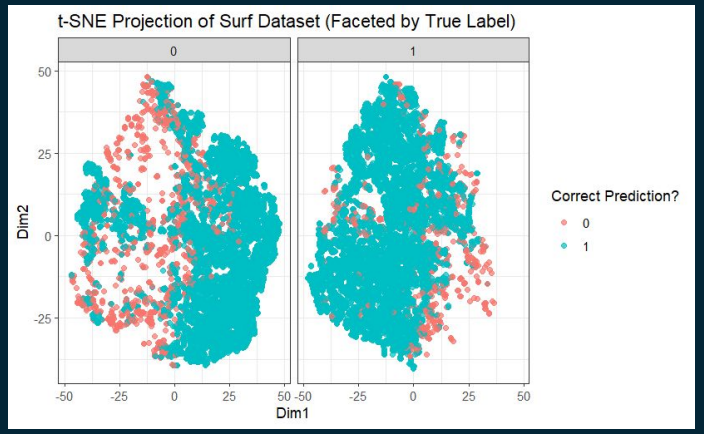
F1 Score (positive): about 0.91

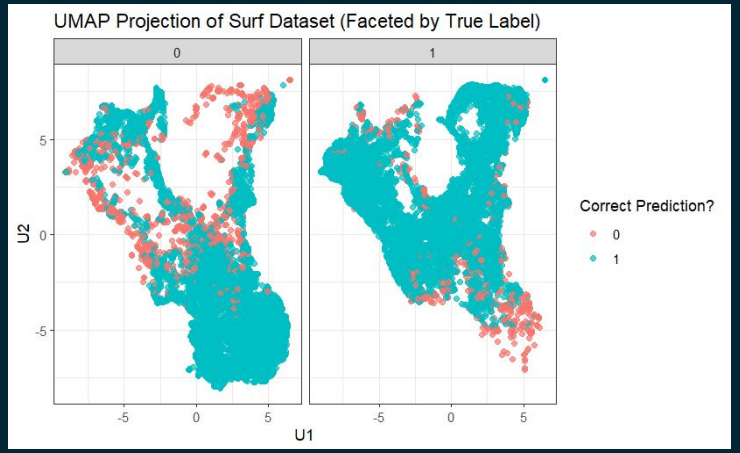
F1 Score (negative): about 0.91

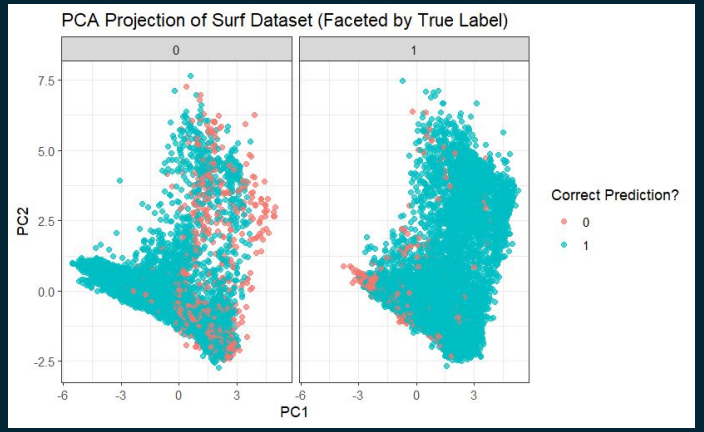
Macro F1 Score: about 0.91

* random output due to random validation set. Accuracies/scores varied between 0.91-0.93

Results Cont.







What We Learned

- Mapping multi-dimensional data into 2d visualization
- Weaknesses of k-NN due to cluster overlap
- ETL of public data over APIs
- Results can be vague due to its binary nature
- Algorithm selection is important

Difficulties

- We had to obtain additional data for more training data
- This made our training data to be 70%, validation data to be 10% and testing data to be 20%.
- Increasing from our original 50% training data
- It increased our accuracy to around 90%
- Running training algorithm multiple times due to random output