Waimea Bay Waves: Initial Assessment – June 2023

Adi Srikanth, Strong Analytics

Project Code: GitHub Repo

Framing the Problem

→ *What questions would you ask the WSL on a kick-off call before beginning on the project?*

Overview

Prior to playing any game, it's important to understand two things: the object of the game and the rules of the game. Here, we take a similar approach to understanding the given problem by looking to clearly define the goal and the parameters of the project. We are unlikely to get full clarity on all aspects of the problem from a single kickoff call. So, we prioritize questions that can narrow down the project scope as much as possible.

Project Goal

What we know:

- WSL wants to host a surf contest at Waimea Bay
- Proper wave conditions at Waimea Bay are critical requirements for a surf contest
- One of the aforementioned conditions is that waves must be at least 3m in height

What we are assuming:

- The surf contest should be, if possible, in Waimea Bay
- Wave conditions cannot be artificially adjusted

High Priority Knowledge Gaps

- More clarity on wave conditions (this will inform what we aim to predict)
    - Are there any other measurable requirements for wave conditions?
    - How long do wave conditions have to hold to allow for a contest?
    - What factors typically impact wave conditions?
- More clarity on the surf contest
    - When are eligible dates and times for the surf contest?
    - How far in advance does planning begin?
    - What does a typical surf contest look like (schedule, scale, etc)?

Project Parameters

What we know:

- We have some data on wave conditions from five distinct buoys

High Priority Knowledge Gaps:

- What other data and information sources are readily available?
- What data sources are known to be off-limits (because of privacy violations, inconsistent availability, etc)?
- Does our solution have to check any specific boxes beyond informing our broader objective?
- Is there a catalog of previous attempts or known strategies?

Questions for WSL

Ultimately, we would like to fill in as many high priority knowledge gaps as possible during our kickoff call. However, instead of asking questions to specifically fill the aforementioned gaps, it is oftentimes a better strategy to to keep questions open-ended. This allows the customer to implicitly identify what they value most instead of relying on our imposed prioritizations. It also allows the customer to provide tangential, but equally useful information that we did not identify ourselves.

In order to allow for open-endedness while still respecting the knowledge gaps we identified, we should ask broad questions that center around our knowledge gaps before drilling down into details. For example, instead of asking *"Are there any other measurable requirements for wave conditions?"* we can ask "*what sorts of environmental conditions can make or break a surf contest?"* in order to fill in our knowledge gaps without limiting conversation.

Enrichment

→ *Before diving into the analysis and forecasting, which new data sources would you consider to complement the buoy data? How would you consider using these data on an ongoing basis?*

Data Sources

The data required for this project is heavily dependent on the various details that we would expect to learn from our kickoff call and follow-up communications. However, at first glance, there are a few data sources that seem valuable in this space.

Specifically, historical data on past surf competitions is incredibly valuable as it can help establish a relationship between the environmental conditions and event success for a competition (can be defined via viewership metrics, contestant performance metrics, etc).

Beyond this, compiling a comprehensive dataset to describe environmental conditions can help us predict what the wave conditions will be on a given date. Such data is available from trustworthy organizations such as the National Oceanic and Atmospheric Administration (NOAA) along with many commercial enterprises.

As a third area to source data, it would be worthwhile to investigate other data sources closer to the intended event location. We currently have various buoy data; but, this data can be augmented, improved, or even replaced if a better data source is available.

Ongoing Data Collection

In addition to rounding out our dataset to be as robust as possible, it is important that our datasets are consistent over time. Given that the use case is designed for predicting future conditions without a clear timebox, our data sources must be generally available at any given time in order to ensure that we don't miss a window of opportunities. Considerations to ensure this would include both a thorough evaluation of the data source (its track record of availability, the reliability of its data measurement, etc) and a meticulous plan to stand up fault tolerant data architecture to ingest data.

This could involve using a reliable cloud provider to host data jobs and should involve logging to track any missed data points.

Data Integrity

→ *How would you handle missing data in this project?*

Overview

Handling missing data is a crucial, yet complex task that is highly dependent on a given data set and project goal. For the purposes of this discussion, we will ignore additional data sources that can be pulled into the fold and will instead focus on handling missing data if the given data set is our primary dataset.

There are several strategies for handling missing data. For this particular use case, we focus on three options: dropping data, imputing data, and relabeling data. Dropping data simply removes any rows of data with missing values. Imputing data replaces missing values with some aggregate value. Relabeling data takes missing data and treats it as a class of its own while categorizing the rest of the existing data.

Buoy Categories

The buoys included in our dataset are located near Tokyo, Alaska, and Hawaii (three). We are interested in checking to see if any of the buoys are missing data with some sort of noticeable trend. As in, we want to see whether or not a buoy missing data is indicative or correlated with a certain trend. If so, we should treat missing data as its own class.

After an analysis, we find that the Tokyo and Alaska buoys do not exhibit an evident trend with their missing data. The distributions of our target variable and our other features are largely the same regardless of whether the Tokyo/Alaska buoy data is missing or not. On the other hand, the Hawaii buoys do exhibit trends with respect to missing data.

Missing Data for Tokyo and Alaska Buoys

Because we did not find a clear trend amongst the missing data for these buoys, we choose to address the missing data here by imputing missing values with the median value of the feature. This allows us to retain rows with missing data while avoiding prescribing significance to missing data. We use the median specifically in order to account for possible outliers in our dataset.

Missing Data for Hawaii Buoys

The Hawaii buoys did appear to exhibit missing data with some trend pertaining to environmental factors. So, instead of imputing missing values, we choose to represent missing values as their own categorical class. This allows any future model to learn potential relationships between the lack of data and our target variable.

In order to do this, we take our numeric buoy data and convert them into categorical classes. Specifically, we take each feature and convert the numeric values to their respective quantiles and treat each quantile as a category. Missing data points then become their own, sixth category.

Approaches

→ *Which approaches would you consider for this project? Explain your recommendations as if you were communicating directly to the CEO of WSL.*

Overview

Ultimately, the ideal approach for this problem is the solution that is associated with a tradeoff that targets benefits that are highly meaningful to the customer and allows for costs that are less significant to the customer. For example, in the case of a surf competition, it is likely incredibly important to avoid scheduling a competition for subpar wave conditions. On the other hand, while missing out on an opportunity to host a competition is certainly a cost, not hosting a competition is far less damaging than hosting a competition that is deemed a failure. As such, we would look for a solution that may not identify every single opportunity to host a competition, but does not erroneously identify an opportunity.

Possible Approaches

There are two major tradeoffs for this type of project. The first is the tradeoff between technical complexity and ease of implementation. A technically complex solution would likely be more accurate and precise, but would incur a higher cost (time and money) and could be less interpretable to nontechnical stakeholders. Understanding the priorities for this project will inform where the best balance is for this tradeoff.

The second major tradeoff for this type of project lies in further specifying the goal of the project. What should the solution optimize? The solution can be designed to mitigate the worst case scenario, maximize the best case scenario, or fall somewhere in between.

Technically Complex	Predictive Model (ML), Recommender System, Randomized Experiment
Technically Simple	Similarity Matching, Predictive Regression, Data-Driven Research

As far as what solution to optimize, this can be controlled by how the objective is set for any solution. For example, we can choose to optimize for precision, recall, accuracy, etc to favor a specific outcome.

Recommended Approach

We recommend a ML-based, predictive model for this particular problem. While this approach is technically complex, the problem at hand is reasonably complex itself. Additionally, accuracy and precision are of paramount importance in this use case given potential risks to safety, risks to competition quality, and various economic and social externalities that are at play as a result of the surf competition. A ML-based predictive model would provide a sophisticated understanding of future wave conditions. The model can be made interpretable with proper care and attention to explainability.

Communication

→ *How would you recommend that these results be communicated to the WSL on an ongoing basis?*

Overview

Simply reporting a slew of accuracy metrics is insufficient to support the actual goal of the WSL. Instead, we should focus on using our expertise to interpret the results of analyses and provide reasonable expectations for the WSL.

A good high-level solution could be a brief report that identifies the best opportunities to host an event in the near future. For each opportunity, we can list a worst-case, best-case, and most-likely case scenario. The focus here is to give the WSL a concrete outcome that they can expect; using this, they can choose which option fits their needs best.

A more detailed solution, if desired, could be a similar report that contains a predicted outcome set for any date in the near future. For each date, we can include worst-case, best-case, and most-likely case scenarios. This solution would be more helpful if instead of being prescribed possible dates for an event, the WSL would rather choose a potential date themselves and then evaluate the date for feasibility.

In both cases, we can make the raw data available. However, it is important for us to remember that we were hired to sift through the data – we should not expect the WSL to have to do this themselves.

Ongoing Communication

While we can continually update the WSL, it is best to empower them and give them the agency to use our solution how they please. So, we should look to make our reports available whenever the WSL would like to access it. This can be something simple, such as an automated upload to a Google Drive. Or, it can be something more complex, such as making an API endpoint available to the WSL.

Reports can be as sophisticated as an interactive dashboard with live updates and additional functionality to support further analysis. Or, reports can be as simple as an automated one-page report that is easily digestible. The report format should be tailored to the end user.