Using a KNN Classifier to Predict Beach Closings

Tamara Hadzic (hadzic.t@northeastern.edu)

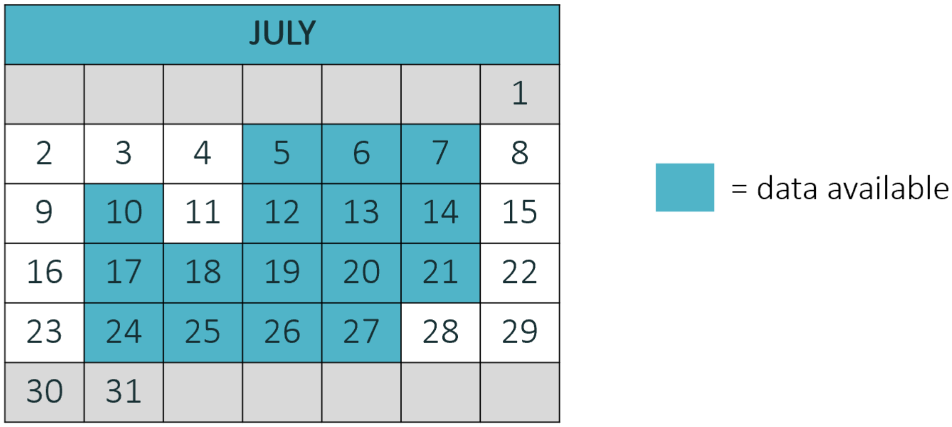
Problem Statement and Background

Last summer, due to the heavy rainfall Massachusetts had experienced, many beaches were shut down due to overgrowth of bacteria. At the time, the data on the Massachusetts government website was exclusively available as a downloadable CSV file for the current week. Since then, the website has added an interactive dashboard that displays the current beach postings on a map. However, as beaches are currently off season and past data is not displayed on the dashboard, users cannot easily look for trends in the data.

This project aims to use past data from the summer to predict beach closings by using a KNN algorithm. Specifically, the main question under investigation is whether data from three weeks of beach closures, weather, and location data can be used to predict which beaches will close in the following week. Being able to predict closures from bacterial exceedance will benefit the general public and will allow them to make more informed decisions about which beaches to visit. In addition, beach owners, economists, and public policy makers could benefit from knowing which beaches will suffer most heavily during waves of bacterial exceedance.

Introduction to your Data

The first piece of data used was daily beach closure records from the Massachusetts State Online Archives (July 2023).[1] It was collected by the Bureau of Climate and Environmental Health and the Department of Public Health and comes from constant monitoring of beach water quality. This is the main data that was used in both the features and the labels, meaning that the introduction of bias into this dataset would make our results less valid. One of the main issues with this dataset is the availability of data. As shown in the figure below, data was not available for every day in July, meaning that the data used for this project was incomplete.



JULY						
						1
2	3	4	🟦 5	🟦 6	🟦 7	8
9	🟦 10	11	🟦 12	🟦 13	🟦 14	15
16	🟦 17	🟦 18	🟦 19	🟦 20	🟦 21	22
23	🟦 24	🟦 25	🟦 26	🟦 27	28	29
30	31					

🟦 = data available

Figure 1: Visualization of available beach closing data for July, created with Microsoft PowerPoint

The same organizations also made current postings available on the Massachusetts Beach Water Quality Dashboard, which was used to get coordinates for beaches. This website was accessed in November 2023.[2] The third piece of data used was a shapefile from the MassGIS database created by The Massachusetts Department of Public Health Center for Environmental Health Environmental Toxicology Program and represents point coordinates and markers for the extent of each beach.[3] Finally, the project also used precipitation data accessed from stormglass.io via their API.[4] This data was collected from NOAA at precipitation measuring stations across Massachusetts.

Data Science Approaches

The main data science algorithm used was a K-nearest neighbors classification, which predicts labels for unlabeled points based on their distance to labeled points. The features chosen for this algorithm were (1) fraction of days closed during week 1, (2) fraction of days closed during week 2, (3) fraction of days closed during week 3, (4) latitude, (5) longitude, (6) total precipitation during the first 3 weeks of July. The labels were whether the beach was closed during any of the days in the fourth week of July (True) or whether it remained open during all of the days (False). We tested k's ranging from 4 to 19 in cross-fold validation and evaluated the choice of k by maximizing accuracy, precision, and recall. After the best k was chosen, we evaluated the performance of the model with the F1 score.

The other data science algorithm used was Euclidean distance. Since the free version of the API only allowed 10 calls per day, we could not retrieve precipitation data for every single coordinate in our dataset. Therefore, we chose 8 cities to represent the different coastal regions in Massachusetts and assigned precipitation data to the coordinates in our dataset based on the Euclidean distance to each of the cities.

Results and Conclusions

After performing cross-fold validation, the optimal k was 6. At this k, the accuracy was 95.84%, the precision was 87.5%, and the recall was 32.92%. A figure of the changes in the three metrics as k varied is shown below. All three metrics have a maximum at k = 6. At this value of k, the F1 score was 51.47%, meaning that the overall performance of our model was adequate.

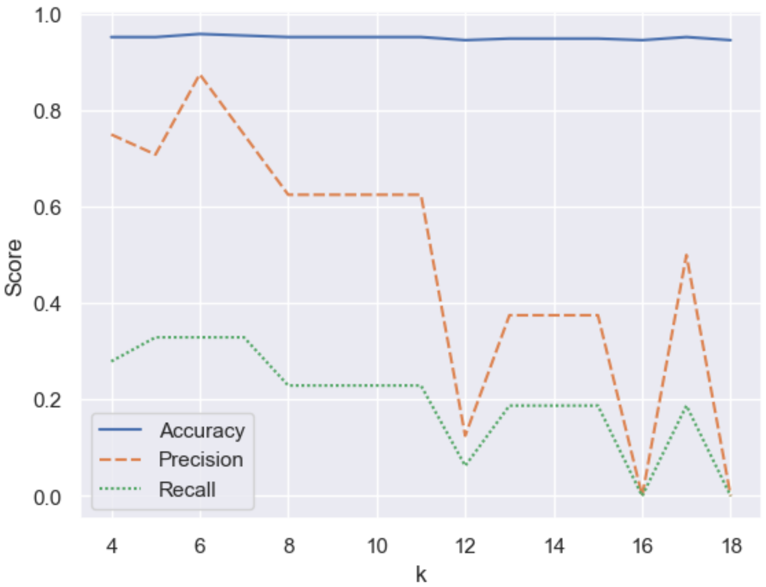


Figure 4: Change of Accuracy, Precision, and Recall with different values for k in the KNN, created with seaborn

In the testing dataset, the model was able to predict the closing of Malibu Beach in Boston (which was closed on July 24th) and King's Beach in Swampscott (which was closed on July 24th, 25th, 26th, and 27th). However, it was unable to predict the closing of Tenean Beach in Boston (which was closed on July 24th, 25th, 26th, and 27th), Constitution Beach in Boston (which was closed on July 24th), and Sandy Point Beach in Ipswich (which was closed on July 27th). In figure 3A and 3B below, one can see which beaches were predicted to close during the last week of July and which ones actually closed during the last week of July.

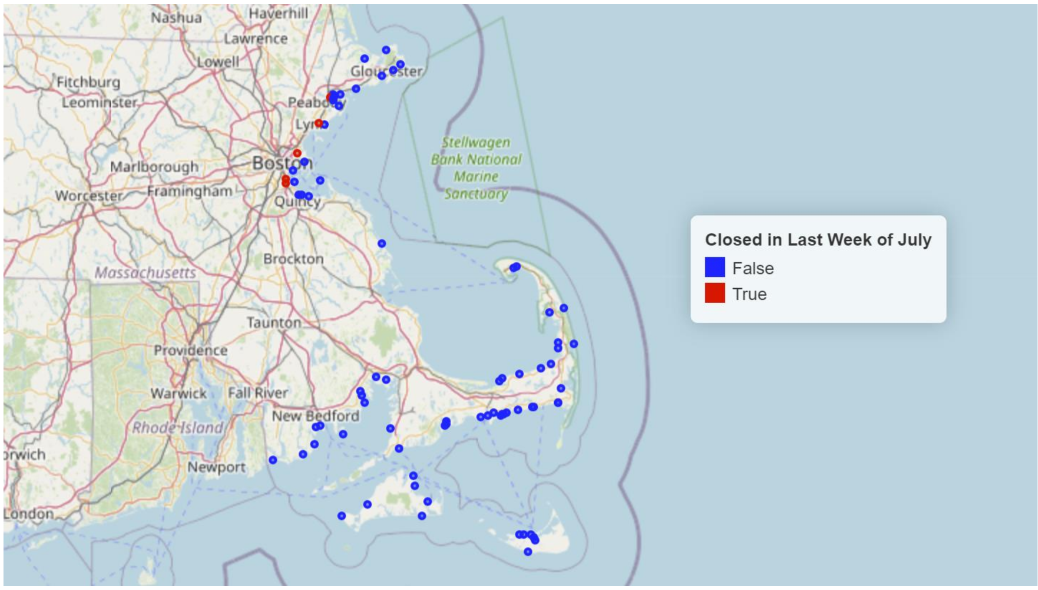


Figure 3A: Beaches in the testing set labeled by whether they were actually closed in the last week of July (True) or not (False), created with geopandas

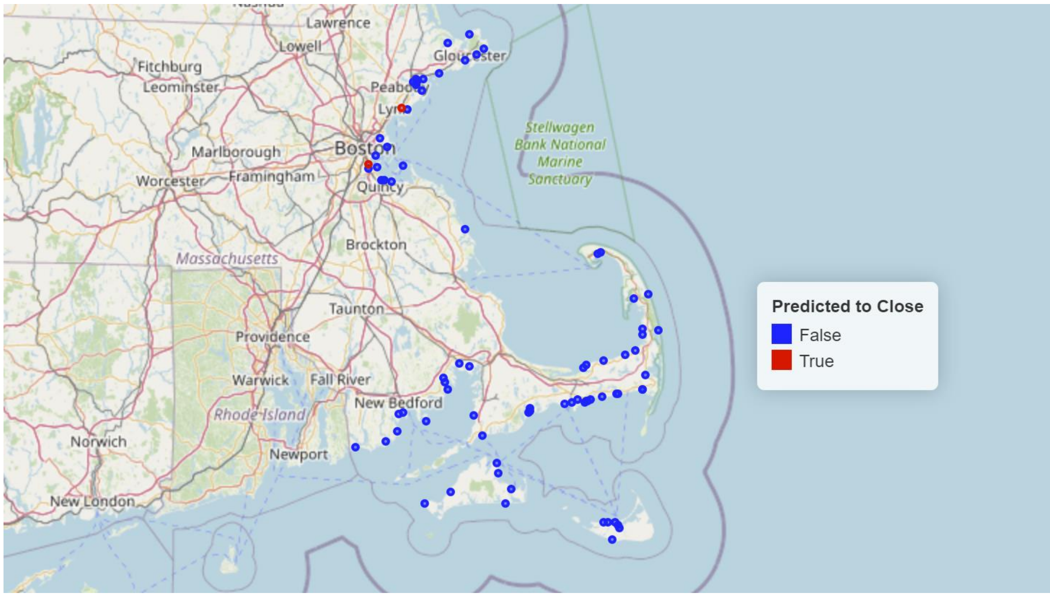


Figure 3B: Beaches in the testing set labeled by whether they were predicted by the KNN to close in the last week of July (True) or not (False), created with geopandas

One can see from the map that the model had a relatively high false negative rate. This is also apparent in the confusion matrix, where there were 74 true negatives, 0 false positives, 3 false negatives, and 2 true positives.

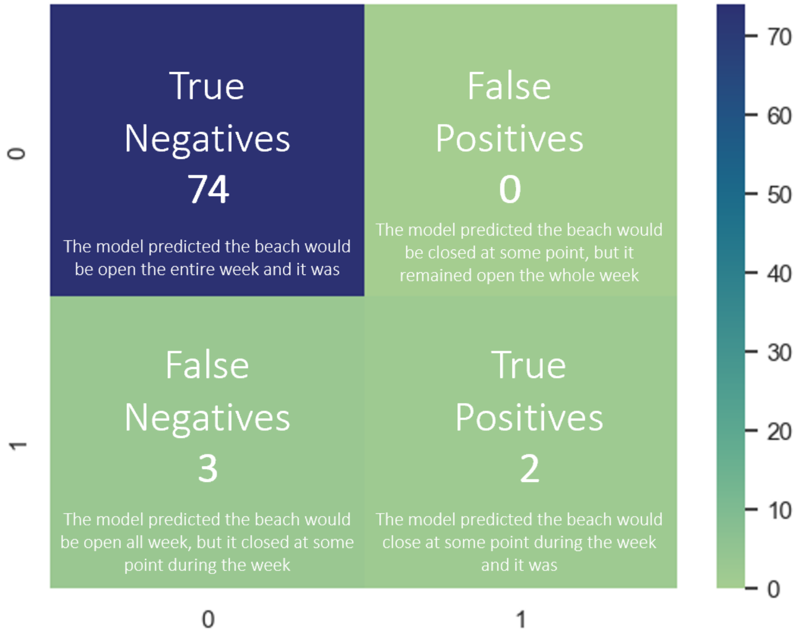


Figure 4: Confusion matrix summarizing the number of true negatives, false positives, false negatives, and true positives in the dataset, created with seaborn

Future Work

If we want to optimize this model for consumers, in the future we would want to use a model that minimizes the false negative rate, even if that means increasing the false positive rate. This would provide beachgowers with an abundance of caution in deciding which beach to go to, so they can be more certain that the beaches predicted by the model to stay open actually stayed open.

To create a better model, we must also recognize the shortcomings of this dataset, specifically concerning the ratio of open beaches to closed beaches. In the figure below, one can see that most beaches were open 100% of the time during each of the weeks. The small amount of beaches that were closed raises concern over the model overfitting to those beaches, leading poor performance on the testing data.

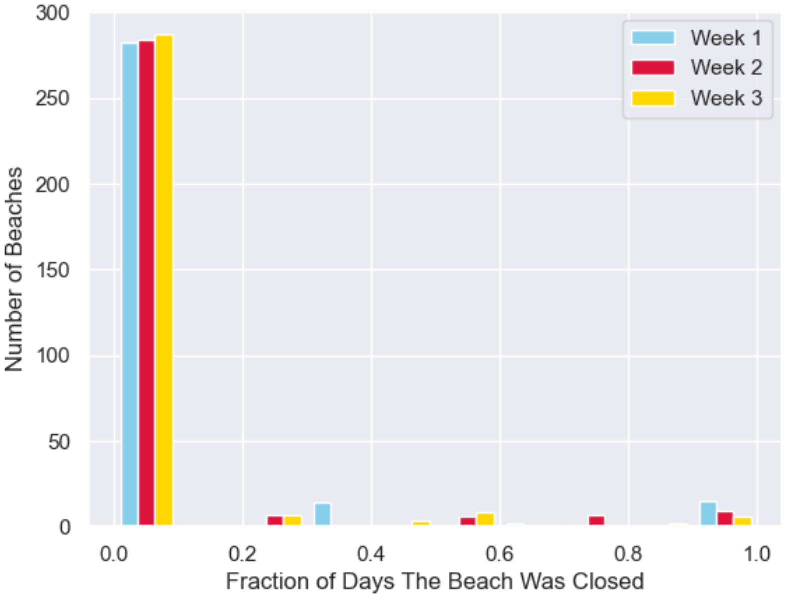


Figure 5: Distribution of beach closing data by week for the first 3 weeks of July, showing the fraction of days that the beach was closed during that week on the x-axis and the number of beaches on the y-axis, created with matplotlib

Another concern was with data availability and matching. Specifically, since the data all came from different sources (ie, the shapefile was put out by a different organization than the beach closing data), there were issues with merging these sources together. Most notably, Sandy Point Beach in Ipswich was closed in the last week of July, but Figure 3A places this beach closer to Peabody. This is because there are actually two Massachusetts beaches named Sandy

Point Beach – one in Ipswich (which was in the beach closing file) and one in Beverly (which was in the shapefile). When the two files were merged by beach name, these beaches were consolidated. The only way to truly avoid issues like this would be to check each beach by hand, which is difficult with over 300 beaches in the dataset. If this work were to be expanded, it would be useful to create a naming system for each beach so that beaches with multiple names or multiple beaches with the same name do not get consolidated together in analysis. Namely, a standardized four-digit code corresponding to each beach would eliminate confusion in consolidating data from multiple sources.

Finally, if this project were to be expanded, it would be useful to pay for more calls on the stormglass.io API so that precipitation data can be more accurately matched to beaches. Data used for this project and corresponding code and package requirements can be accessed at https://github.com/tamarahadz/ds500_proj.

Works Cited

1. Current public beach postings | Mass.gov. (n.d.). Www.mass.gov. https://www.mass.gov/info-details/current-public-beach-postings
2. Interactive Beach Water Quality Dashboard | Mass.gov. (n.d.). Www.mass.gov. Retrieved December 5, 2023, from https://www.mass.gov/info-details/interactive-beach-water-quality-dashboard
3. MassGIS Data: Marine Beaches | Mass.gov. (n.d.). Www.mass.gov. Retrieved December 5, 2023, from https://www.mass.gov/info-details/massgis-data-marine-beaches#downloads-
4. stormglass.io Weather API. (n.d.). Stormglass.io. Retrieved December 5, 2023, from https://stormglass.io/