The Greatest Lakes: Using Data Mining Techniques to Analyze Physical Properties of the Great Lakes

Nongjie Feng, Rhys Leahy, and Carter Rogers

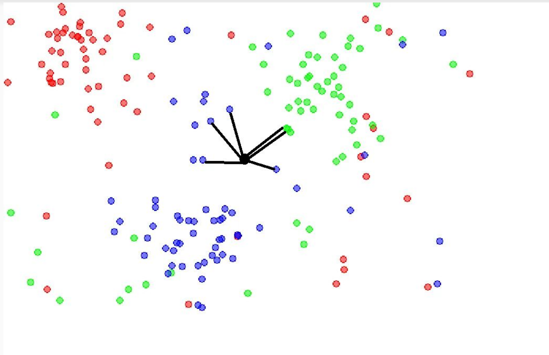
https://northernmichiganlandbrokers.com/page/the-great-lakes/

Overview: Key Questions

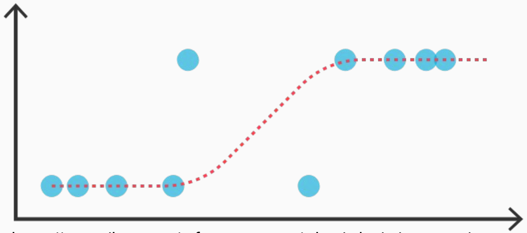
- Based on surface temperature and physical properties, can you predict the percentage ice concentration and whether that percentage will exceed a given threshold across the Great Lakes?
- Can you predict which lake a set of characteristics (surface temperature, ice concentration, and physical properties) most likely belongs to?

Algorithms used

- K-Nearest Neighbors for classification
- Logistic Regression for surface temperature and ice concentration



https://www.unite.ai/what-is-k-nearest-neighbors/



https://www.tibco.com/reference-center/what-is-logistic-regression

Description of data set

- Public datasets collected and published by NOAA's CoastWatch program and The Great Lakes Environmental Research Laboratory.
- We queried and cleaned 26 .dat files containing observed surface temp and ice concentration with Requests and Pandas.
- We also merged a table with the physical characteristics of each of the Great Lakes.

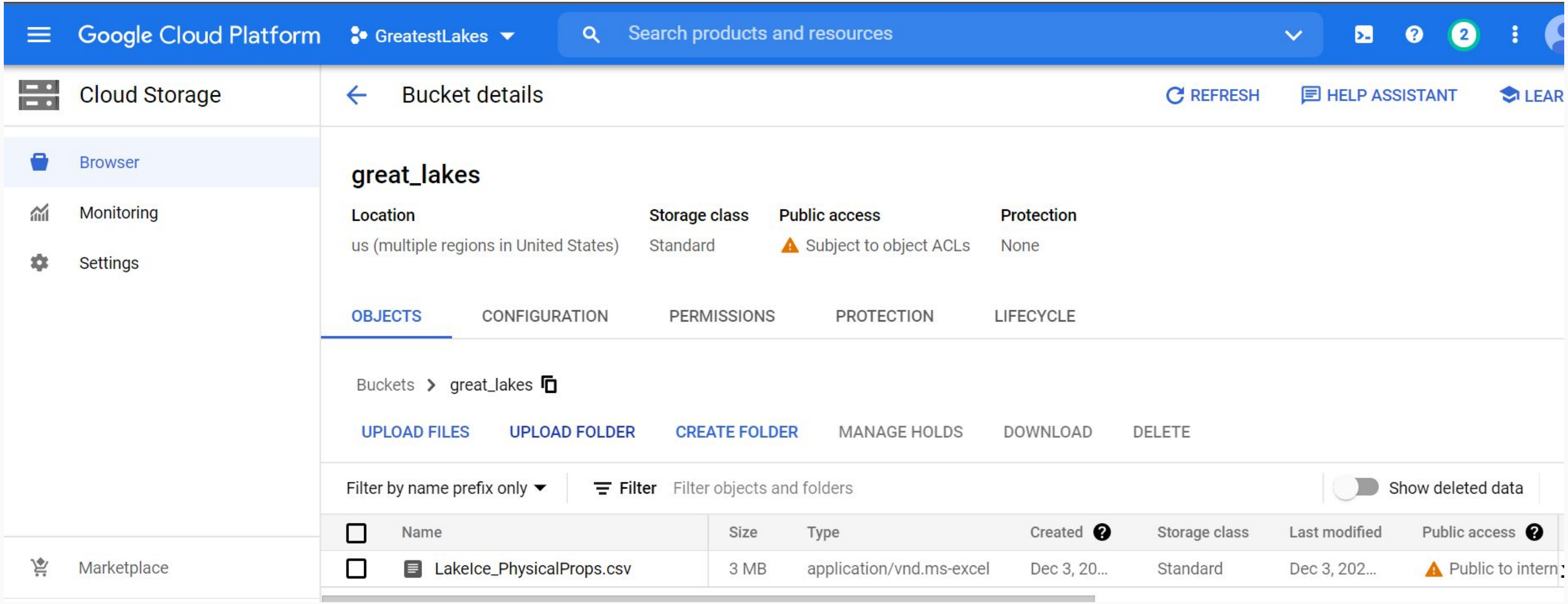
Great Lakes Average Ice Concentration

		Ice Concentration (%)						
Year	Day	Sup.	Mich.	Huron	Erie	Ont.	St.Clr	GL Total
2008	344	2.10	2.12	5.58	0.42	0.24	34.56	2.76
2008	346	2.08	2.29	6.24	0.63	0.27	15.33	2.90
2008	350	3.65	4.24	8.64	7.76	1.05	24.88	5.25
2008	353	4.94	7.66	10.39	6.93	1.40	53.04	6.97
2008	357	5.34	15.50	18.13	13.43	3.06	92.41	11.61
2008	360	7.51	18.18	16.52	16.57	2.63	91.41	12.88
2008	364	4.26	9.29	11.44	9.68	1.65	60.36	7.64
2009	001	6.13	13.52	15.76	9.57	2.85	60.41	10.38

Data cleaning and preprocessing

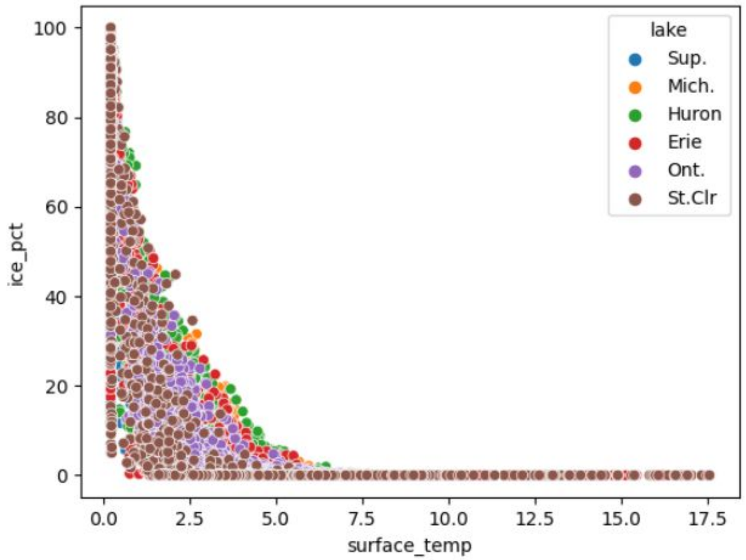
- Resulting dataset contained a total of 31,126 rows and 16 variables.
- Most important features were surface temperature and ice concentration.
  - Only 8, 855 rows contained ice data.
- To circumvent connection errors across machines, we hosted a copy of our full dataset on GCP

Data Warehousing



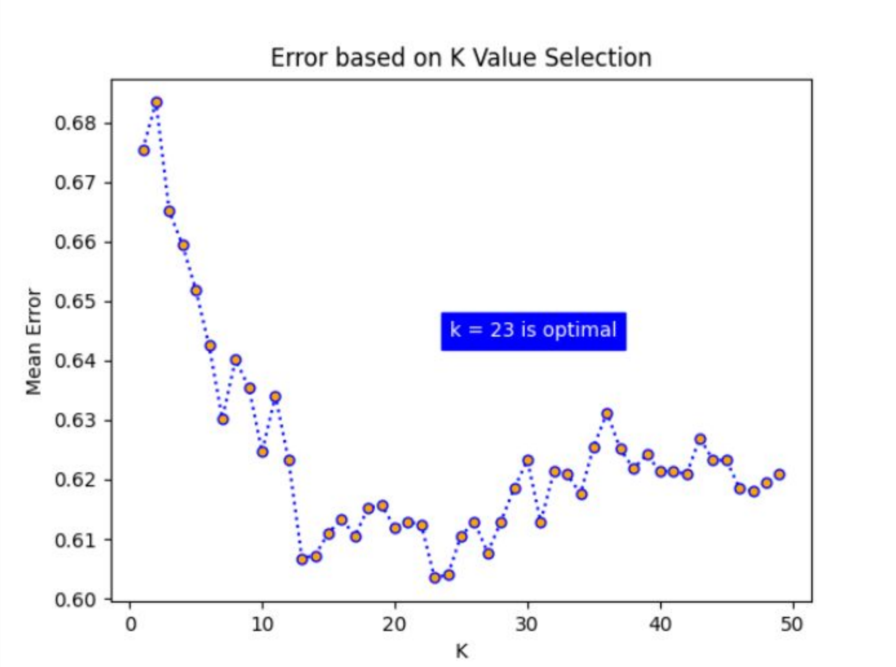
Implementation of K-Nearest Neighbors

- Classifying data points by lake
- Use surface temperature and ice concentration %
- Naive baseline model: 16.7% chance of correctly guessing lake



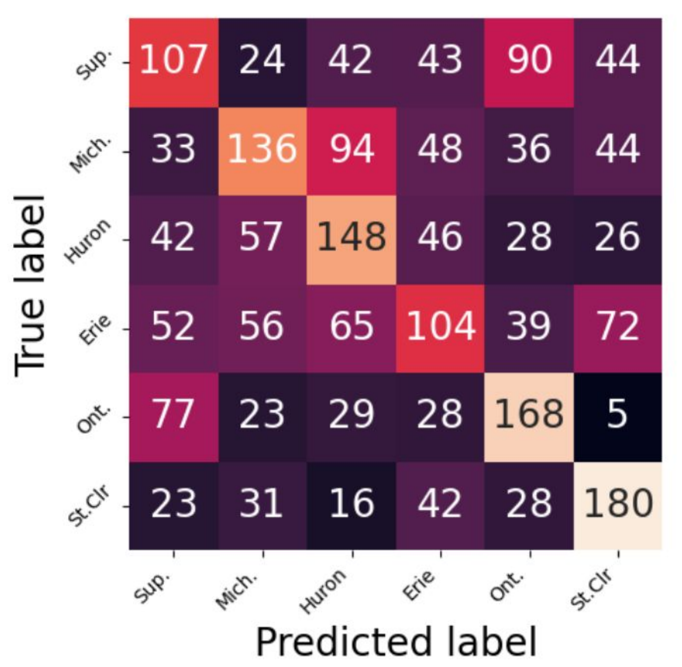
Implementation of K-Nearest Neighbors

- Tuning k-value to maximize accuracy of classification
- Loop through k-vals from 1-50
- Optimal k = 23
- Final model accuracy score: 39.7%



Results

- Final KNN model (k = 23) classified observations with 39% accuracy
- Best accuracy: Lake St. Clair
- Distinguishing between Lake Superior vs. Lake Ontario

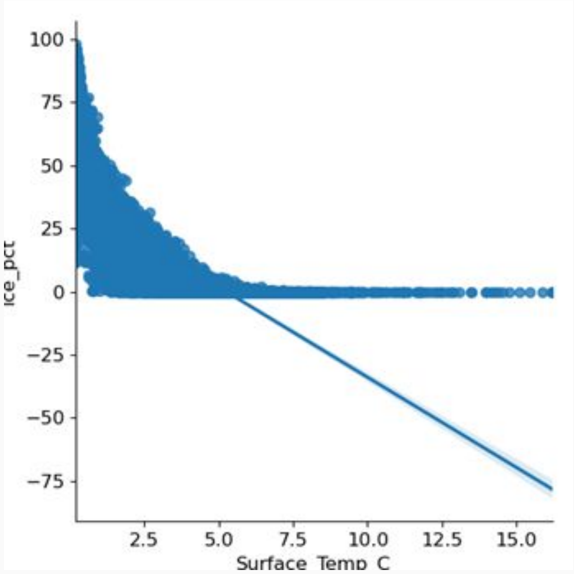


Implementation of Linear regression

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \epsilon$$

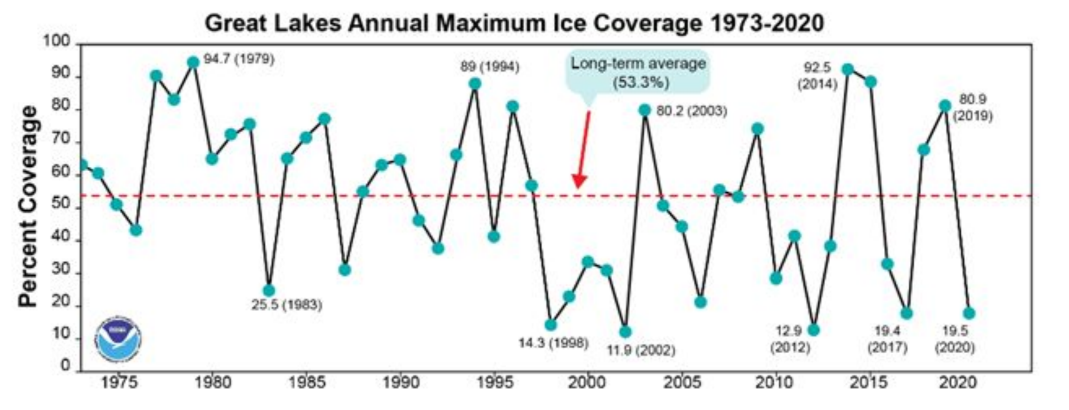
- Lake Erie:0, Lake Huron:1, Lake Michigan:2, Lake Ontario:3, Lake Superior:4.
- Test set MSE: 324.3781 Train set MSE:323.9782

	Coefficient
Elevation_meters	1.531980e+14
Length_km	-2.505699e+13
Breadth_km	8.195978e+13
Avg_Depth_meters	2.239198e+13
Max_Depth_meters	-5.039058e+13
Volume_km3	-2.284854e+14
Water_Area_km2	2.197783e+14
Land_Drain_Area_km2	-7.097686e+13
Total_Area_km2	-1.567553e+14
Shore_Length_km	-1.610179e+12
Retention_Time_years	-3.299119e+13
Lake	2.036360e+14
Surface_Temp_C	-1.574068e+01



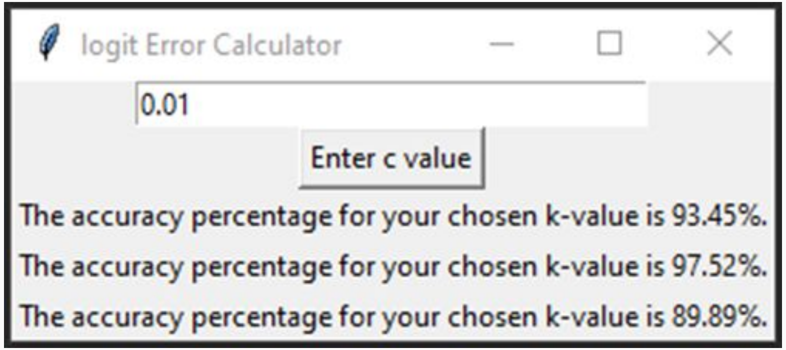
Implementation of Logistic regression

- Threshold: 0.533
- Model accuracy score: 0.9345
- Training-set accuracy: 0.9365
- Test-set accuracy: 0.9345
- No overfitting



Implementation of Logistic regression

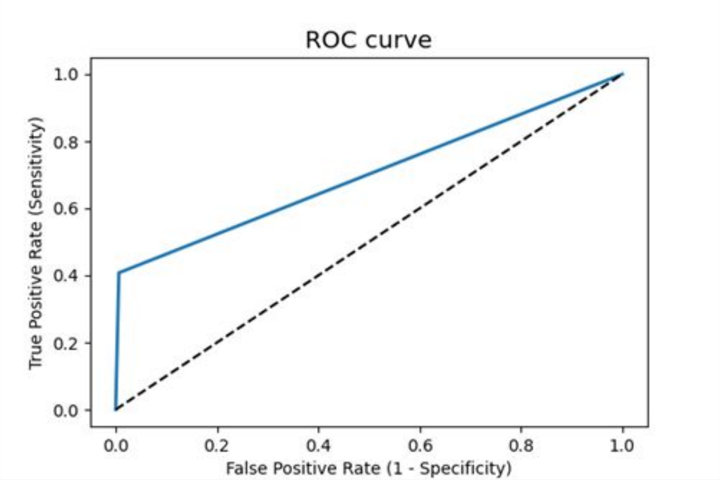
- Increase C to 100
  - Training-set accuracy: 0.9771; Test-set accuracy is 0.9752
- Setting C to 0.01
  - Training-set accuracy: 0.8964; Test-set accuracy: 0.8989
- More complex model should perform better



Implementation of Logistic regression

	Actual Positive:1	Actual Negative:0
Predict Positive:1	1582	10
Predict Negative:0	106	73

- Precision: 0.9937
- Sensitivity: 0.9372
- Specificity: 0.8795

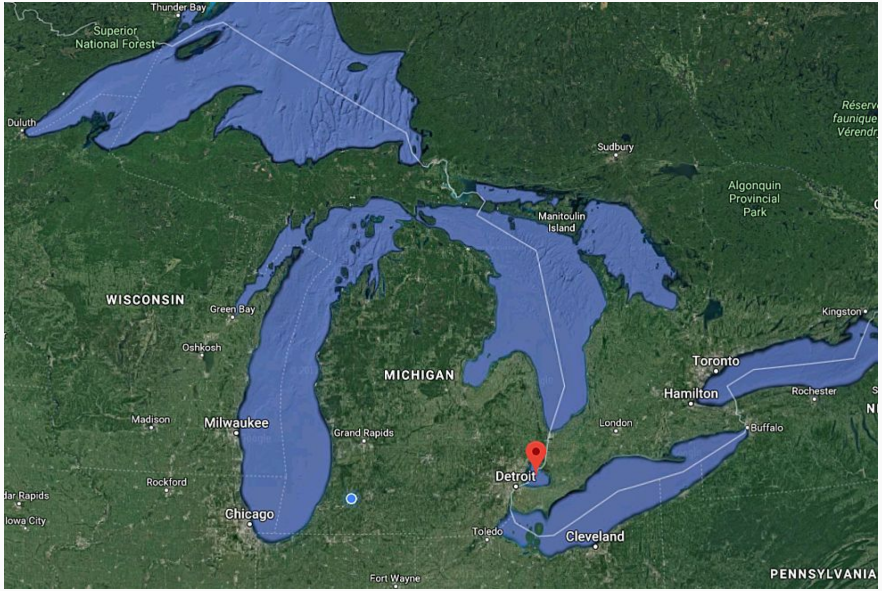


Model Summary

- Majority of observations predict that there will be no maximum ice coverage
- Logistic regression is safe and robust
- Use different threshold
- Lasso regression and elastic net

Conclusion

- Classification: Lake Ontario and Lake Superior similarities
- Lake St. Clair observations easiest to classify



https://k1025.com/should-lake-st-clair-be-the-sixth-great-lake/

References

Beeton, Alfred M.. "Great Lakes". Encyclopedia Britannica, Nov 23rd, 2020, https://www.britannica.com/place/Great-Lakes. Accessed 3 December 2021.

Chandra, R. V., & Varanasi, B. S. Python requests essentials. Packt Publishing Ltd, 2015.

Great Lakes CoastWatch Program Overview. https://coastwatch.glerl.noaa.gov/overview/cw-overview.html. Accessed 5 Dec. 2021.

Great Lakes Physical Characteristics. https://coastwatch.glerl.noaa.gov/statistic/physical.html. Accessed 5 Dec. 2021.

Great Lakes Statistics. https://coastwatch.glerl.noaa.gov/statistic/statistic.html. Accessed 5 Dec. 2021.

Joby, Amal. “What Is K-Nearest Neighbor? A ML Algorithm to Classify Data.” *Learn Hub*, 19 July 2021, https://learn.g2.com/k-nearest-neighbor.

McKinney, W., et al. Data structures for statistical computing in python. In Proceedings of the 9th Python in Science Conference (Vol. 445, pp. 51–56), 2010.

References

scikit-learn. 2021. *sklearn.linear_model.LinearRegression*. [online] Available at: <https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LinearRegression.html> [Accessed 6 December 2021].

scikit-learn. 2021. *sklearn.linear_model.LinearRegression*. [online] Available at: <https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LinearRegression.html> [Accessed 6 December 2021].

En.wikipedia.org. 2021. *Logit - Wikipedia*. [online] Available at: https://en.wikipedia.org/wiki/Logit [Accessed 6 December 2021].

Hastie, T., Tibshirani, R. and Friedman, J., n.d. *The elements of statistical learning*.

James, G., Witten, D., Hastie, T. and Tibshirani, R., n.d. *An introduction to statistical learning*.

Kaggle.com. 2021. *Linear Regression House price prediction*. [online] Available at: <https://www.kaggle.com/faressayah/linear-regression-house-price-prediction> [Accessed 6 December 2021].

Kaggle.com. 2021. *Logistic Regression Classifier Tutorial*. [online] Available at: <https://www.kaggle.com/prashant111/logistic-regression-classifier-tutorial> [Accessed 6 December 2021].

References

Kaggle.com. 2021. *Heart Disease Prediction using Logistic Regression*. [online] Available at: <https://www.kaggle.com/neisha/heart-disease-prediction-using-logistic-regression> [Accessed 6 December 2021].

pandas?, A., Troy, A. and Zwinck, J., 2021. *Any way to get mappings of a label encoder in Python pandas?*. [online] Stack Overflow. Available at: <https://stackoverflow.com/questions/42196589/any-way-to-get-mappings-of-a-label-encoder-in-python-pandas> [Accessed 6 December 2021].

Research.noaa.gov. 2021. *NOAA projects 30-percent maximum Great Lakes ice cover for 2021 winter - Welcome to NOAA Research*. [online] Available at: <https://research.noaa.gov/article/ArtMID/587/ArticleID/2706/NOAA-projects-30-percent-average-Great-Lakes-ice-cover-for-2021-winter> [Accessed 6 December 2021].

Brownlee, J., 2021. *A Gentle Introduction to Threshold-Moving for Imbalanced Classification*. [online] Machine Learning Mastery. Available at: <https://machinelearningmastery.com/threshold-moving-for-imbalanced-classification/> [Accessed 6 December 2021].