Sentinel Lakes Dataset : An Unsupervised Learning Approach

Presenters: Tabassum Nisha and David Josue Molina

Abstract

The overall goal of this project is to explore unsupervised learning methods detecting climate change by looking for increased variability in temperatures. The Trout lakes dataset used is indeed novel but required sufficient pre-processing. Some dates had data missing some hours of the day. We oriented the data in a way that each row represented a date from respective months and each column represented respective years with the cells consisting of average temperature for the day. For statistical analysis we chose to look into the variance in the average temperature of the water for different dates across the years. For visualization we chose regression plots and faceted boxplots. Our baseline model was Lasso Regression Model and the model of our choice was GradientBoost Regressor model.

Background

Many organizations are invested in protecting natural habitats as it not only may have an impact on ecological systems but also humans' daily lives. One such organization includes the Minnesota Pollution Control Agency (MPCA) that collects data on 25 Sentinel Lakes in the state. Through collaborative efforts, they strive to predict the consequences of climate change on lake habitats in a representative subset of their state's lake.

The Minnesota Department of Natural Resources cites as one of its goals to look at new ways to process and analyze data and with an unsupervised learning approach, we can provide such insight on such efforts. Previous efforts from this organization may have suggested that water depth may have been the focus.

Objectives

For the purpose of this report, we have access to a subset of the data that MPCA has specifically on its Trout lakes. Our team hopes to figure out ways to measure variability in temperature. As is often the case of unsupervised learning, we are hoping to use unsupervised learning methods to explore new possible structures or patterns.Specifically, we want to explore how temperature variability might differ from the same depth , hour, month but for different years. Previous efforts from Minnesota Department of Natural Resources may have suggested that water depth may have been the focus. Like previously mentioned, we are hoping to see how temperatures change in response to time. Thus, our hypothesis is that the temperatures of the lake in the dataset is increasing.

Methods

- Data Pre-processing
- Data Exploration/Data Visualisation (regression plots and boxplots)
- Statistical Analysis (mean temperatures, covariance, correlation coefficient)
- Machine Learning Algorithms (Lasso , MLP, XGB,Linear Regression and GradientBoosting Regression)

Dataset created for calculating average temperatures for the day

Date	0	1	2	3	4	5	...	15	16	17	18	19	20	21	22	23	Average_Temperature_Day
20-04-2012	5.04	5.02	4.98	4.98	4.97	4.94	...	4.93	4.92	4.91	4.91	4.93	4.93	4.98	5.05	5.02	113.55
21-04-2012	5.06	5.06	5.06	5.03	5.04	5.02	...	4.99	5.00	4.99	5.05	5.11	5.11	5.12	5.13	5.10	115.73
22-04-2012	5.11	5.12	5.09	5.10	5.06	5.04	...	5.03	5.07	5.11	5.09	5.08	5.18	5.23	5.26	5.24	116.58
23-04-2012	5.29	5.26	5.24	5.23	5.23	5.21	...	5.29	5.29	5.30	5.31	5.35	5.38	5.38	5.38	5.44	121.27
24-04-2012	5.51	5.51	5.50	5.56	5.63	5.63	...	5.57	5.65	5.67	5.71	5.77	5.83	5.86	5.85	5.82	129.21

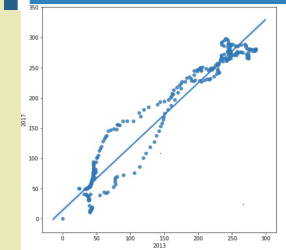
Dataset created for predicting the temperatures of the year 2017 using the data from the year 2012 to 2016

date_month	2012	2013	2014	2015	2016	2017	2018
20-04	113.55	45.73	53.02	75.57	48.90	91.95	0.00
21-04	115.73	45.63	53.34	76.29	50.28	92.84	0.00
22-04	116.58	45.77	53.99	76.93	51.76	94.14	0.00
23-04	121.27	45.83	54.61	77.44	53.89	93.92	0.00
24-04	129.21	45.77	54.91	78.00	54.50	94.35	0.00

Results

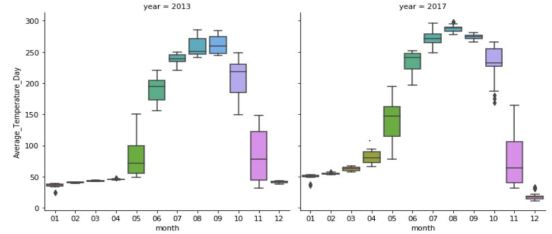
Upon plotting the scatter plot of average temperatures of May 2012 against the average temperatures of May 2016, we observed that there was a positive correlation. We then plotted the linear regression plot of the average temperatures of the year 2013 and 2017 which also showed positive correlation. The faceted boxplot displays variance in the average temperatures of different dates in a month across the years. We can observe that the average temperatures in the month January gradually increased over the years, hence justifying our hypothesis.

Since the covariance between the average temperatures of May 2012 and May 2016 is 1110.54 and 1619.23, which indicates that the relationship between the two variables is positive. The NumPy correlation coefficient and Pearson's correlation coefficient between the same variables is 0.974 which again indicates that the correlation is positive. Thus, this indicates that our hypothesis stands true.



Positive correlation seen between the average temperatures of same month but different years

Increase in the mean temperatures in the same month but different year



Conclusions

Overall, our findings showed that our hypothesis was supported. Additionally, many of our models performed well in training and testing scores but it was the mean squared errors that help differentiate each model from each other.

If we worked on this project further, we would have liked to create predictive models to predict temperature's variance by utilising covariance/standard error rather than average temperatures. Additionally, it would have been fruitful to explore other algorithms as the research suggests that there might be other algorithms that handle time series and hydrology relationships better.