Lake Michigan Influences

Blake Wallace
Capstone Technical Report

May 14, 2019

Data science objectives:

i. Does Rain have an effect on the average daily water temperature?

ii. What effect does rain have on the average daily temperature near the water?

iii. What effect does rain have on the average daily temperature far from the water?

iv. When it rains, is there a statistically significant difference between the amount of rain that falls in downtown Chicago compared to the Ohare airport?

v. How much correlation exists between the average daily temperature of Lake Michigan and the temperature difference between the downtown Chicago area and the Ohare airport?

vi. Can we build a model that predicts with at least 80% accuracy the difference in total precipitation between Ohare airport and the the Botanical gardens?

vii. Is there any statistically significant difference between the daily temperature near the water as apposed to far from the water?

Data:

Links

Data Dictionary

Bouy Data Dictionary

O'Hare Airport Data Dictionary

Botanical Garden Data Dictionary

"The five core values are:"

ohare_prcp - Precipitation (PRCP) (inches)

ohare_snfall - Snowfall (SNOW) (inches)

ohare_sndpth - Snow depth (SNWD) (inches)

ohare_maxtmp - Maximum temperature (TMAX) (Fahrenheit)

ohare_mintmp - Minimum temperature (TMIN) (Fahrenheit)

Other Features

lake-temp - Average Daily Surface Water Temperature for Lake Michigan (Fahrenheit)

garden_prcp - Precipitation (PRCP) (inches)

garden_maxtmp - Maximum temperature (TMAX) (Fahrenheit)

garden_mintmp - Minimum temperature (TMIN) (Fahrenheit)

garden_tobs - Temperature at time of observation (TOBS) (Fahrenheit)

ohare_wspd - Average daily wind speed (AWND) (miles per hour)

ohare_atmp - Average Temperature (TAVG) (Fahrenheit)

ohare_w2dir - Direction of fastest 2-minute wind (WDF2) (the direction the wind is coming from in degrees clockwise from true N)

ohare_w2spd - Fastest 2-minute wind speed (WSF2) (miles per hour)

Feature Engineering

target - absolute difference between the precipitation measurements at Ohare and the garden (ohare_prcp - garden_prcp)

garden_didrain - categorical, 1 for yes, 0 for no

ohare_didrain - categorical, 1 for yes, 0 for no

garden_medtmp - Median daily temperature at the Garden/ midpoint between the max and min temperatures ((garden_maxtmp + garden_mintmp)/2)

ohare_medtmp - Median daily temperature at ohare/ midpoint between the max and min temperatures ((ohare_maxtmp + ohare_mintmp)/2)

tmpdiff - difference between the median temperatures at ohare and the garden (ohare_medtmp - garden_medtmp)

Data Cleaning/Data Manipulation/EDA:

Models and Evaluation:

Model	Training score	Testing score	Training MSE	Testing MSE	cross validation
Linear no poly	0.0825	0.1052	0.0933	0.0683	0.0785
Linear gs	0.1222	0.1329	0.0893	0.0662	0.0984
Decision Tree	0.1139	0.0691	0.0901	0.0711	0.0429
Decision Tree gs	0.0937	0.0584	0.0922	0.0719	0.0450
Random Forest	0.8614	0.0517	0.0134	0.0724	0.0554
Random Forest	0.8711	0.1078	0.0131	0.0681	0.0770
Random Forest gs	0.8658	0.0905	0.0136	0.0694	0.0651
Random Forest	0.8677	0.0682	0.0135	0.0711	0.0660
Random Forest	0.8704	0.1153	0.0132	0.0676	0.0767
Random Forest	0.8080	0.0957	0.0195	0.0690	0.0787
Random Forest ada	0.9547	0.0549	0.0331	0.0722	0.0331
Random Forest ada	0.9445	0.0525	0.0056	0.0723	0.0283
Random Forest bag	0.6735	0.1130	0.0332	0.0677	0.0928
Random Forest bag	0.6705	0.1239	0.0335	0.0669	0.0943

Tests and Evaluation:

Data	t-score	p-value	significance	Gardens Avg (F)	Ohare Avg (F)
All Data	0.5876	0.5568	None	59.24	59.43
No Rain	3.285	0.0010	Yes	58.99	60.57
Both Rain	-2.629	0.0086	Yes	59.48	57.7
ohareRain	-1.9557	0.0506	None	59.06	57.43
gardensRain	0.0904	0.9280	None	59.99	60.07

Resources:

1. An executive summary: What is your goal? Where did you get your data? What are your metrics? What were your findings? What risks/limitations/assumptions affect these findings? 2. Summarize your statistical analysis, including: implementation evaluation inference 3. Clearly document and label each section of your notebook(s) Logically organize your information in a persuasive, informative manner. Include notebook headers and subheaders, as well as clearly formatted markdown for all written components. Include graphs/plots/visualizations with clear labels. Comment and explain the purpose of each major section/subsection of your code. Document your code for your future self, as if another person needed to replicate your approach. 4. Clearly document all of your

decision points in the relevant sections How did you acquire your data? How did you transform or engineer your data? Why? How did you select your model? How did you optimize hyperparameters? 5. Host your notebook and any other materials in your own public Github Repository. You repo should have README file that guides us through the repository and links to important files. Include links and explanations to any outside libraries or source code used. Host a copy of your dataset or include a link to a remotely hosted version.