Computer Science Department
CS675 – Introduction to Data Science (CRN: 74028)
Fall 2023
Project #4 / Due 19-Dec-2023

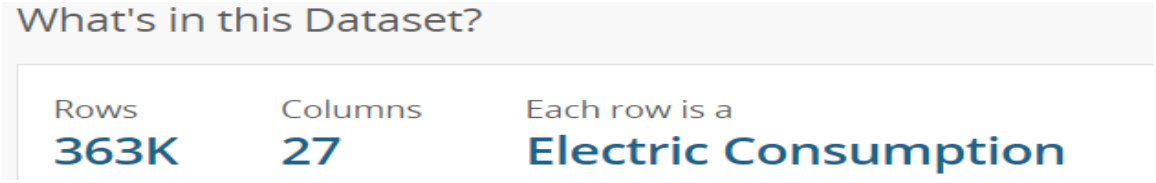
XGBoost is an open-source algorithm often used for many data science use cases. Often the use cases are common classification cases such as fraud detection or regression cases such as house price prediction, but XGBoost can also be extended into time-series forecasting. By using the XGBoost Regressor, we can create a model that can predict future numerical values.

Implement a Time Series Forecasting model in Python/R, by using the XGBoost library.

The forecasting model should be able to predict **New York City's Electricity Consumption** (see below.)

You should test your forecasting model in three (3) distinct datasets. On Daily, Monthly, and Yearly Mean electric consumption.

Daily data:



Dataset contains daily electric consumption for all five (5) boroughs of New York City.
https://data.cityofnewyork.us/Housing-Development/Electric-Consumption-And-Cost-2010-April-2020-/jr24-e7cr

Monthly Mean data:
Take the above (daily dataset) data and average it out based on each month.

Yearly Mean data:
Take the daily data and average it out based on each year.

Write Python/R scripts in order to complete the following tasks along with their output. All work should be done and submitted in a single Notebook file.

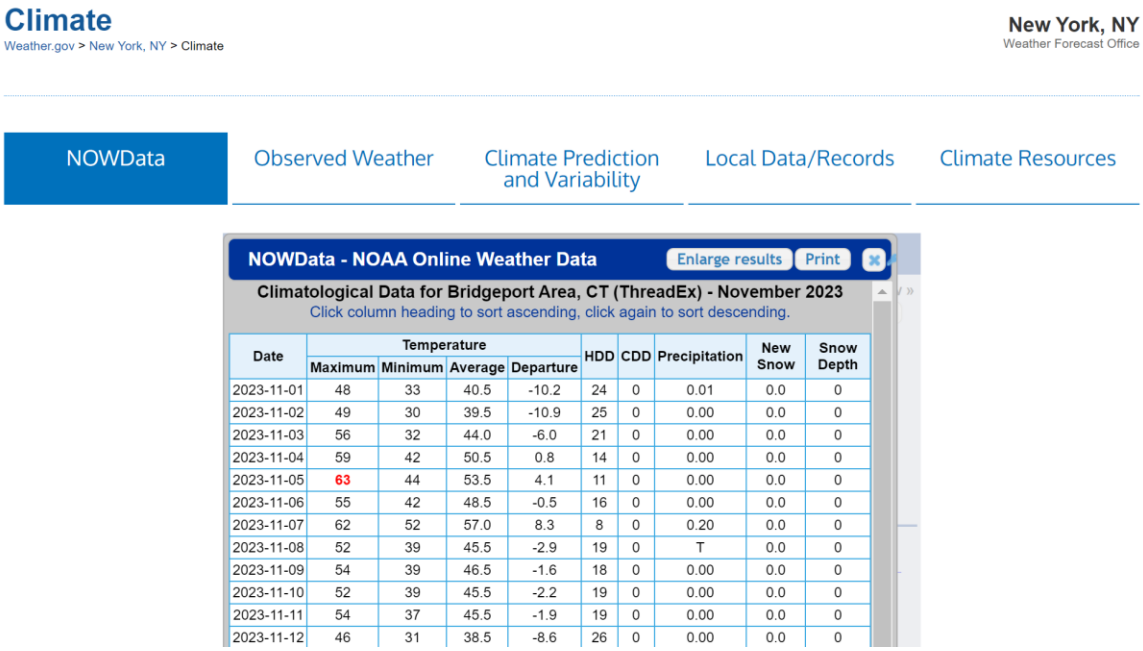
1) Since the time unit (day, month, year) varies from dataset to dataset, make your code agnostic of the input. In other words, have your code to determine the unit of the time series.

2) Then, train your model (on the respective dataset) and predict the Electric Consumption (EC) values from the last date of the dataset into X units of time into the future.

a) Should the unit of time be day, then predict the EC for 100/200/365 days into the future.
b) Should the unit of time be month, then predict the EC for 1/6/9 months into the future.
c) Should the unit of time be year, then predict the EC for 1/10/20 years into the future.

3) Perform **Feature Engineering** tasks by incorporating new time-based features (such day-of-week, month-of-year, week-of-year, etc…) as well as 'external' factors.
For external factors include daily temperature as a new feature and present your findings as far as enhanced model's performance.

To obtain daily temperatures (historical data) for NYC, review this url:
https://www.weather.gov/wrh/climate?wfo=okx



Date	Temperature				HDD	CDD	Precipitation	New Snow	Snow Depth
	Maximum	Minimum	Average	Departure					
2023-11-01	48	33	40.5	-10.2	24	0	0.01	0.0	0
2023-11-02	49	30	39.5	-10.9	25	0	0.00	0.0	0
2023-11-03	56	32	44.0	-6.0	21	0	0.00	0.0	0
2023-11-04	59	42	50.5	0.8	14	0	0.00	0.0	0
2023-11-05	63	44	53.5	4.1	11	0	0.00	0.0	0
2023-11-06	55	42	48.5	-0.5	16	0	0.00	0.0	0
2023-11-07	62	52	57.0	8.3	8	0	0.20	0.0	0
2023-11-08	52	39	45.5	-2.9	19	0	T	0.0	0
2023-11-09	54	39	46.5	-1.6	18	0	0.00	0.0	0
2023-11-10	52	39	45.5	-2.2	19	0	0.00	0.0	0
2023-11-11	54	37	45.5	-1.9	19	0	0.00	0.0	0
2023-11-12	46	31	38.5	-8.6	26	0	0.00	0.0	0

Should you find a 'better' site to retrieve NYC's daily (historical) temperatures, go ahead and use it. Make sure the records retrieved have the same date range as the original time series dataset!

For each model (whatever models you come up with from your Feature Engineering task), print the predicted values in a tabular format and draw a line graph showing both historical data and the future.

4) Evaluate all models by providing their respective **MAE** (Mean Absolute Error) and **MAPE** (Mean Absolute Percentage Error), as well as **R^2** (use Python's sklearn's respective metrics and/or R's).

5) Compare and contrast your findings and model performance from FBProphet forecasting vs XGBoost time series forecasting.

Here are details about the daily dataset (timeseries). You need to manually create the monthly and yearly timeseries.
Daily NYC Electric Consumption: <<NYC Open Data >>
https://data.cityofnewyork.us/Housing-Development/Electric-Consumption-And-Cost-2010-April-2020-/jr24-e7cr

Devel...	Boro...	Acco...	Loca...	Mete...	Mete...	TDS #	EDP	RC C...	Fund...	AMP #	Vend...	UMI
WYCKOFF...	BROOKLYN	WYCKOFF...	BLD 03	AMR		163	272	K016300	FEDERAL	NY00501...	NEW YOR...	9
WYCKOFF...	BROOKLYN	WYCKOFF...	BLD 02	AMR		163	272	K016300	FEDERAL	NY00501...	NEW YOR...	9
WYCKOFF...	BROOKLYN	WYCKOFF...	BLD 03	AMR		163	272	K016300	FEDERAL	NY00501...	NEW YOR...	9
WYCKOFF...	BROOKLYN	WYCKOFF...	BLD 03	INTERVAL		163	272	K016300	FEDERAL	NY00501...	NEW YOR...	9
WYCKOFF...	BROOKLYN	WYCKOFF...	BLD 02	INTERVAL		163	272	K016300	FEDERAL	NY00501...	NEW YOR...	9
WYCKOFF...	BROOKLYN	WYCKOFF...	BLD 03	INTERVAL		163	272	K016300	FEDERAL	NY00501...	NEW YOR...	9
WOODSON	BROOKLYN	WOODSON	BLD 02	NONE		182	285	K018200	FEDERAL	NY00501...	NEW YOR...	9
WOODSON	BROOKLYN	WOODSON	BLD 01	NONE		182	285	K018200	FEDERAL	NY00501...	NEW YOR...	9
WOODSIDE	QUEENS	WOODSIDE	BLD 18	INTERVAL		33	316	Q003300	FEDERAL	NY00500...	NEW YOR...	9
WOODSIDE	QUEENS	WOODSIDE	BLD 19	INTERVAL		33	316	Q003300	FEDERAL	NY00500...	NEW YOR...	9
WOODSIDE	QUEENS	WOODSIDE	BLD 17	AMR		33	316	Q003300	FEDERAL	NY00500...	NEW YOR...	9
WOODSIDE	QUEENS	WOODSIDE	BLD 07	AMR		33	316	Q003300	FEDERAL	NY00500...	NEW YOR...	9
WOODSIDE	QUEENS	WOODSIDE	BLD 15	AMR		33	316	Q003300	FEDERAL	NY00500...	NEW YOR...	9

< Previous Next > Showing Electric Consumptions 1 to 13 out of 362,630

Extra Points: Predict Electric Consumption for each of the 5 Boroughs (independently)!