**storeLSTM\_Learning**

This folder contains all the code from trying to learn how to use LSTM. I followed a few youtube tutorials and was testing out various prediction methods.

Tutorial 1: <https://www.youtube.com/watch?v=c0k-YLQGKjY> - LSTM and how to do prediction

Tutorial 2: <https://www.youtube.com/watch?v=H6du_pfuznE> - LSTM and how to do future prediction

**LSTM-deadExperiments**

This stores most of all the random LSTM/prediction attempts. This includes prediction by season, randomization, and June/July/August analysis. It’s not really worth looking in here unless you know the exact snippet of code you want to access or be inspired from.

**Experiment 1**

Experiment 1 is about discovering whether we can do prediction by month, so we look at the energy demand for the season of summer: June, July, and August. It also has the per day energy demand if you want to see that.

**Experiment 2**

Experiment 2 develops the baseline of performance. LSTM - 2 hidden layers, 20 epochs, etc. 24 hours of previous data (energy data and weather variable data) to make one hour of prediction, one month of data for training. The performance baseline is established for the month of July in this experiment.

**Experiment 3**

Experiment 3 explores changing the number of hours of data used to make one hour of prediction. So previously we used 24 previous hours of data to make one hour of prediction. Here, we are varying the number of hours to use 48, 72, and 96 hours of data. This is again, for the month of July.

**Experiment 4**

Experiment 4 explores changing the number of months of data used for training the model. First we use only one, the month of July. Now, we can trying more months.

**Experiment 5**

Experiment 5 is the synthesis of everything learned from experiment 3-4-5. We are making the baseline performance on the month of August, with random testing days. Then we use 72 hours of previous data for prediction and 3 months of data.

Datasets & DataProcess

Below are all the files associated with processing data that will be used in the LSTM modeling. The overall goal of this code is to take the weather data from NOAA (it is an average of several years data in the format of one year) and put the energy consumption data from CAISO alongside.

* Process CAISO data.ipynb
* Process NOAA data.ipynb
* NOAA and CAISO data together.ipynb

Their collective output is stored in the Datasets folder.

### **Process CAISO data.ipynb**

Some important notes:

* February 29, 2020 data has been removed from the dataset because it is a leap year
* NOAA data does not consider February 29th
* CAISO data from 2018 and 2022 are not complete
  + 2018 starts from 04-10
  + 2022 ends at 02-14 (the last time I collected data)
* if you get an error running, it's probably the date/time format is wrong (see notes below) or .DS\_Store hasn't been removed

Goal:

* turn CAISO data into the format with 2 columns: date and megawatts

Inputs:

* data from CAISO (2018, 2019, 2020, 2021, 2022)
* from directory = '/Users/yukahatori/A\_Fairness/FairnessML\_git/Datasets/CAISO\_NetDemand\_Clean/'
  + had to clean the data because some of the dates became wonky
  + ex: hour 00:00 sometimes was in the format of 0:00
    - so I manually cleaned this by pulling the csv into google docs, changing the number format, and then putting it back in the directory
    - thus the source for the data is CAISO\_NetDemand\_Clean, but in my google drive is a CAISO data folder called CAISO\_NetDemand\_Raw, in which I didn't fix the formatting

Outputs:

* files in the format and with the name CAISO\_NetDemand\_\_Megawatts.csv
* to directory = '/Users/yukahatori/A\_Fairness/FairnessML\_git/Datasets/CAISO\_NetDemand\_Megawatts/'

### **Process NOAA data.ipynb**

Goal:

* turn NOAA data with just the hourly data and remove the flags and years data
* there are two types of flags: completeness and measurement
  + measurement:
    - says whether it is rounded down (X) or missing (M)
    - rounded down is fine but missing is not
    - replace the missing value with an average of the a window of values from 1-5
    - have function to determine the optimal values of prev
  + completeness:
    - replacing the 'years\_' version column
    - same process as measurement

Inputs:

* data from 4 NOAA locations: San Diego (SD), Sacramento (SAC), San Francisco (SF), Los Angeles(LA)
* from directory = '/Users/yukahatori/A\_Fairness/FairnessML\_git/Datasets/NOAA\_Data/'

Outputs:

* files in the format and with the name of their location.csv
* to directory = '/Users/yukahatori/A\_Fairness/FairnessML\_git/Datasets/NOAA\_Data\_Clean/'

### **NOAA and CAISO data together.ipynb**

Some important notes:

* I am only using the data from 2019, 2020, 2021 currently
  + because 2018 and 2022 don't have a full year of data

Inputs:

* data from CAISO (2018, 2019, 2020, 2021, 2022)
* data from NOAA (SF, SD, SAC, LA)
* from directory = '/Users/yukahatori/A\_Fairness/FairnessML\_git/Datasets/CAISO\_NetDemand\_Clean/'

Outputs:

* files in the format and with the name CAISO\_NetDemand\_\_Megawatts.csv
* to directory = '/Users/yukahatori/A\_Fairness/FairnessML\_git/Datasets/CAISO\_NetDemand\_Megawatts/'

## **Extra**

Some extra folders I have are:

* ARMA
  + this is from when I was attempting an ARMA model
* storeLSTM\_Learning
  + this stores the files I used to learn about LSTM and time series modeling
  + some are using Youtube tutorials
  + more documentation in each individual file
* CAISOprices
  + I was working on using CAISO's API to get pricing
  + this might be helpful later