CLIMATE CHANGE: XTREME WEATHER FORECASTING

NISHRIN KACHWALA

OMDENA SILICON VALLEY, CHAPTER LEAD

DATA SCIENCE FOR GOOD

JAN 28, 2023

AGENDA

- Welcome & Introduction (5 min)
- Climate Change Challenge & Objectives (20min)
- SV Omdena Workflow & Process (5)
- Q&A — Collaborators ask questions (10min)

INTRODUCTION

Nishrin Kachwala
Omdena Silicon Valley Chapter Lead
WIDS Ambassador
AI for Good Enthusiast
PM, Mentor

THE CONSEQUENCE

Typical climate change related events: floods, heatwaves, drought, hurricanes, wildfires and loss of glacial ice. (NOAA)

THE PROBLEM TO BE SOLVED

- Extreme swings in temperatures and precipitation drive *the need for accurate long-term forecasts.*
- Physics-based models dominate short-term weather forecasting. But these models have a limited forecast horizon.
- With the availability of meteorological data, the data scientist can improve weather forecasting by blending it with physics-based forecasting.

WEEKLY PLAN

- Semi-self driven project.
- You can go as far as you want.
- Anyone can be a Task Lead.
- Task Lead has Responsibility & Visibility

Week 1	Week 2	Week 3	Week 4
– Data Pre-processing -Data Insights - **Research**	– Exploratory Data Analysis (EDA) - Plots -Feature Engineering	-Model Development -Model Evaluation - Initiate deployment	-Submit to Datathon (Optional) - Deployment Build a Dashboard or App

THE DATA & OBJECTIVE

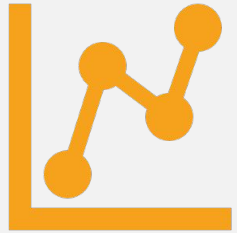
- Provided by WIDS, in collaboration with Climate Change AI (CCAI). Train/Test sets.
- Our Xtreme Weather Project g-drive in Data Folder has Train/Test data sets.

What is my Objective?

- *Predict the maximum and minimum temperature over the next 14 days for each location and start date.*

EXPLORE THE DATA

What is the data about?



How does the data look? Good/Bad? How do we identify Bad?

Weather Data for each row in the data corresponds to a single location and a single start date for the two weeks.

400K points with 246 variables

Train data from 2014-2017

Test data from 2022

Quantitative analysis (descriptive statistics)

Gaps (Missing, incomplete entries)

Erroneous data(outliers, human errors)

Visualize. '*A Picture can speak a 1000 words*'

https://extremepresentation.com/wp-content/uploads/choosing-a-good-chart-09-1.pdf

TRANSFORM THE DATA

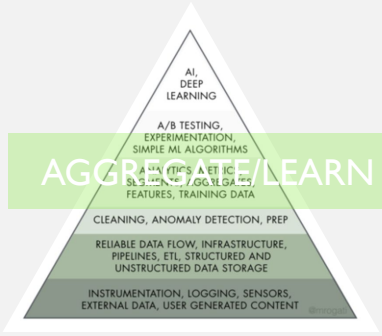
TRAIN

- **IDENTIFY (TRAIN)**
  - missing or erroneous values
  - outliers(box plot, z-score, typical ranges, domain expertise)
  - incorrect data types

TEST

- **FIX (TRAIN, TEST, VALIDATION)**
- Homogenize your data(dates in different formats Jan versus January, temp in different units)
- Convert to numeric.

VALIDATION



AGGREGATE/LEARN

THE GOAL: FORECAST WEATHER

Develop machine learning models to forecast long term weather

What kind of a problem is it? Time Series, Regression

Features contain predictive power – Engineer the heck out of it!

Start simple – Baseline Model first

RMSE is the chosen evaluation metric

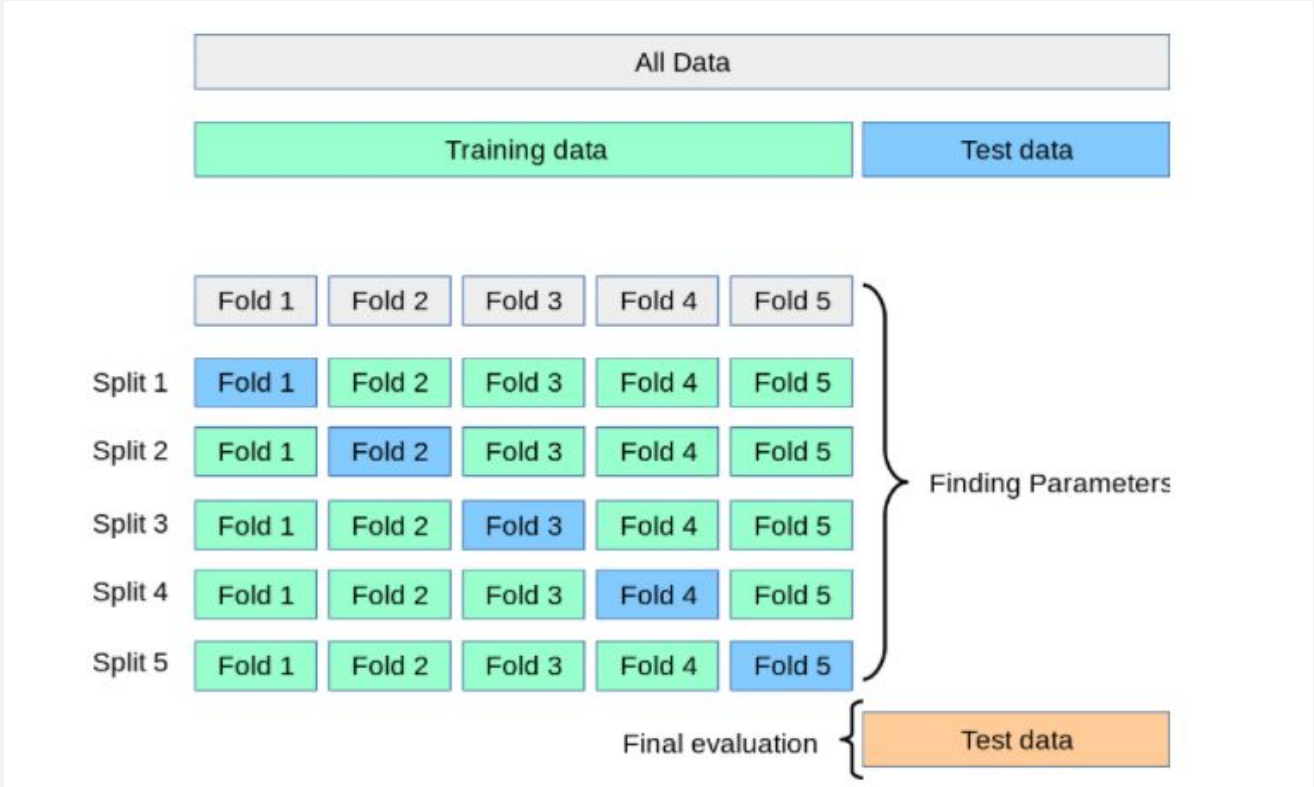
Explain away! Does the model result match intuition?

VALIDATION & TEST SETS

- Tests your Model outputs(several times) – Validation Set
- You can improve your model using the validation set.
- The Test set is 'NOT' for learning and improving.
- Test set is a final go for deployment or the next phase of your ML

CROSS-VALIDATION

- Used for Model Tuning
- Divy up the data, use part for Modeling and Part for Evaluation
- Example shown from ScikitLearn is 5-fold cross validation

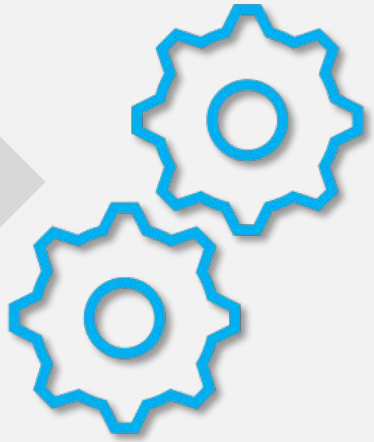


TUNING

- Hyperparameters(numerical setting of your algorithms, that you can change to tune your model to your data)
- This is a setting that is set even before your data hits your model. Default settings. Which you may fine tune and dial-in after you have gone through some iterations of your model.
- FOR Loop through the hyperparameters-> come to the best performing model.
- Cross-validation is used only for ‘TUNING’ not validation!!

ALGORITHM SELECTION

DATA

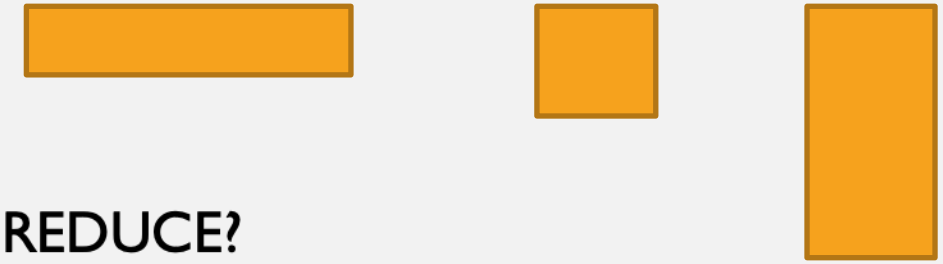


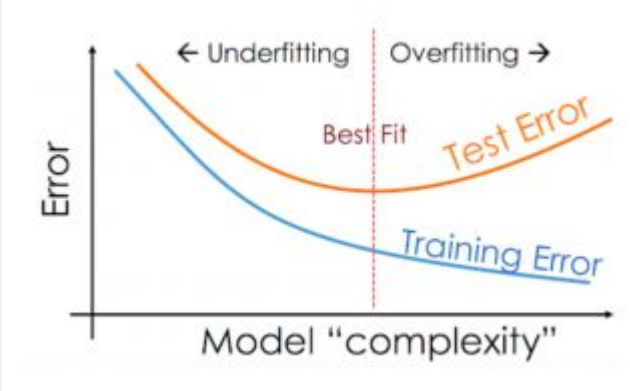
ALGORITHM(S)

MODEL

SCORE

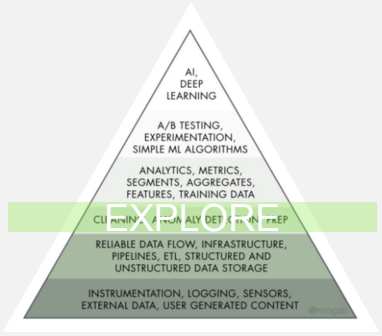
FEATURES

- WHAT SHOULD YOU START WITH?
  - ANSWER IS: NOT ALL FEATURES* and NOT NN
  - DEPENDS 
- HOW CAN I REDUCE?
- Consult experts, Analytics, Algorithm



- Model complexity and the number of instances – constrain the number of feature you can use
- Model Complexity ∝ Overfitting


REGULARIZATION
Balances Complexity and Errors



LET'S GET TO THE CODING

- Develop a starter notebook with empty cells for uploading the data. Colab is easy but, use tools you are comfortable with.
- Gathering info about your data, visuals etc, - univariate, bivariate, correlation etc
- Clean data – missing values, formats, outliers
- Start with a simple model, get to the important features
- Validate and tune your model
- Final- test with never seen data

Silicon Valley Omdena – How do we work together

SLACK COMMs

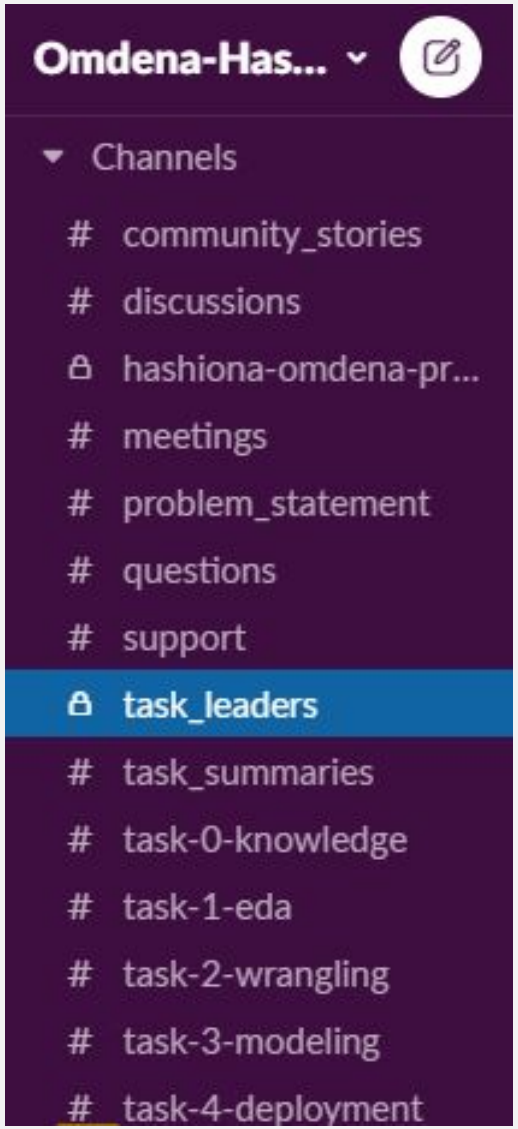
- SLACK – PRIMARY communication channel
- Weekly Meeting – Time/Day based on Poll in Slack # meetings. Task Leads responsible for meetings/notice/agenda
- Various Slack channels

  # questions for any question for our PO

  # support for technical questions for experienced colleagues

  # meeting to note or announce upcoming meetings

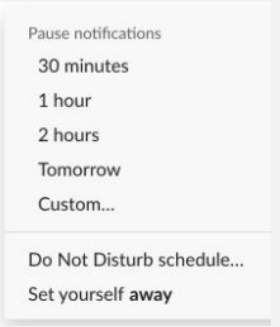
  - # task-1-exploratory-data-analysis – Leads: Vishu & Pooja
  - # task-2-model – Leads: Tekle & Kelly
  - # task-x? Go for it, subtask or another task



Be respectful

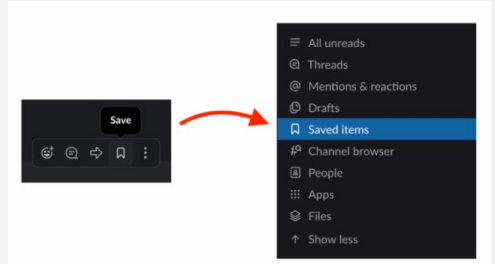
Use threads. Seriously.

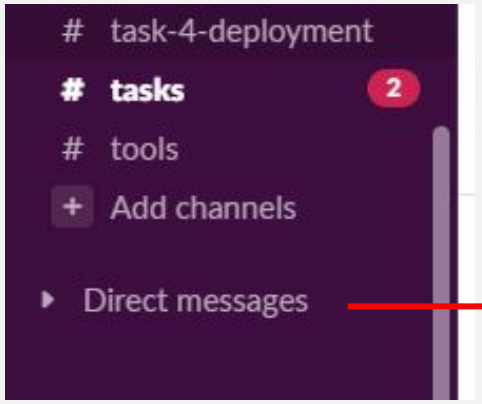
Reduce off-hours pings with Do Not Disturb



Add locations to your profile

Pin important messages in Channels for quick access. Messages go by quickly, save them





Limit DM's to personal questions – otherwise open communication

NOTE: USE TAGS FOR ATTENTION IN YOUR MESSAGES E.g. @Nishrin

OTHER CHANNELS OF COMMUNICATION

- GitHub – Post your clean code with comments, functions, data links, requirements etc. Follow Best software practices so that others can understand your code. *NOT PROVIDED ACCESS YET. I NEED TO SEE SOME CODE WITH GOOD PRACTICES.*
- Official Challenge Google Drive – Stores data, research papers, task presentations, and other files, task plans and tracking documents, recordings
- Online Video – Use free tools, 45min Zoom, gmeet(cannot record), GoToMeeting, RingCentral.

TASK & TASK CHANNELS

TASK #	TASK LEADER
TASK 0: Knowledge	
TASK 1: EDA sub Task 1A: X sub Task 1B: Y sub Task 1C: Z	**Vishu & Pooja**
TASK 3: Feature Engineering	
TASK 4: Predictive Modeling #task-4_6-a-linear_regression #task-4_6-b-ann-rnn #task-4_6-c-alternate_ls-knowledge_graph #task-4_6-d-rf-svm	**Tekle & Kelly**
TASK 5: Deployment	

Meetings

BRAINSTORMING SESSION

TEAM WEEKLY MEETINGS

TASK MEETINGS

TUTORIALS

Meeting Cadence

Meeting	Day and time
Weekly Team Meeting	TDB 8 AM PST | 16 UTC | 9:30 PM IST
TASK 0: Knowledge	TDB
TASK 1: EDA	14 UTC SUNDAYS
TASK 2 : DATA WRANGLING	TDB
TASK 4: MODELING	16 UTC SUNDAYS
TASK 5: DEPLOYMENT	TBD
SPRINTS	TDB
*Learning & Education Series/Workshops	Every Saturday, based on volunteers

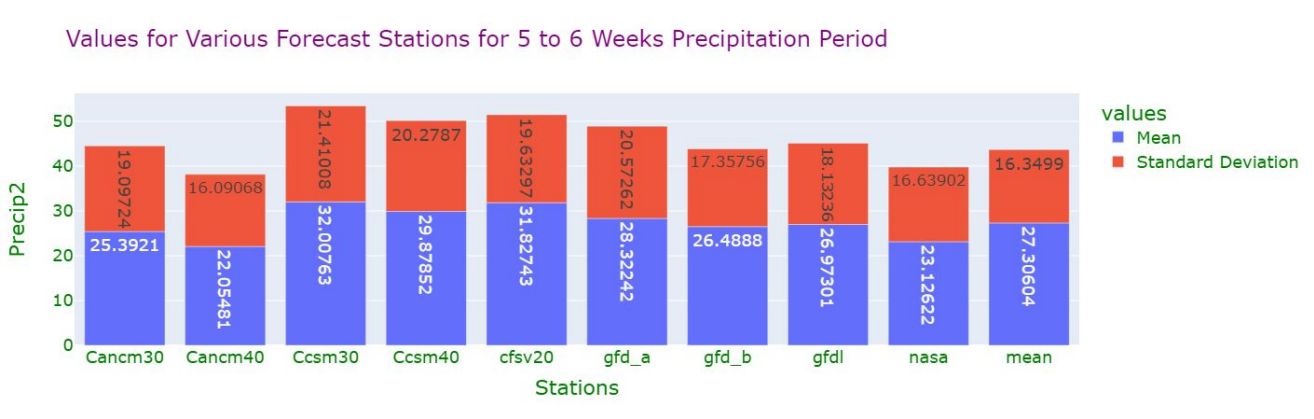
WHAT HAS BEEN DONE SO FAR

- Some EDA and Visualizations
- Your colleagues have posted notebooks in Slack (you can start from there). Or head to Kaggle WIDS 2023 Challenge, and there are lots of resources there.
- Ask your colleagues to explain their work in Task meetings or Slack.

Some EDA Findings…

- There are eight columns with missing values.
- On the missing values. Using the neighbors' average values of each area's missing value for each day. On the column that is without outliers.
- IQR to find outliers
- Looked at the bar plot distribution of precipitation rates from different weather stations, it showed the same mean and deviation values, so we need not look at data from all stations but just one (reducing highly correlated columns that don't add information)
- Columns with El Nino data are compressed into 10 using their means
- After some assessment, the dataset has 44 columns with at least some significance in the target temperature computation.

Highly correlated values from various Weather stations



Missing values-8 Cols, 2 groups are seen