Drought Prediction (CA) Week 10 Check-In

Angela Chang, Nan Li, David Sherman
W210, Section 4
March 12th, 2024

Table of Contents

01
Main Problem

02
MVP

03
Modeling Progress

04
Next Steps

05
Conclusion: Mission

01

Main Problem

Main Problem

- California faces **recurring droughts**, impacting agriculture, water resources, and ecosystem
- Those impacted include farmers, water resource managers, policymakers, and the general public
- Timely and accurate drought predictions would greatly benefit **government agencies** that are responsible for public welfare and resource allocation

02

MVP

MVP: Research Paper

Target Audience:
Local government scientists and decision-makers who guide policy and research

Contribution to Knowledge Base:

- Comparison and evaluation of machine learning techniques in predicting US Drought Monitor (USDM) drought conditions in California
- Verification of ability to predict USDM drought score **2-3 months in advance** leveraging only previous drought scores (univariate) and weekly historical meteorological data (multivariate models)

Key User Questions

1. Does a machine learning approach work well for the task of drought prediction in CA?

2. If a machine learning approach works well, what models work well for the drought prediction task? What characteristics or data does the model need to guide users in implementing similar models in their work?

SME Feedback

- Laurel Larsen, Associate Professor of Geography, UC Berkeley:
    - **LSTM models perform best in her work**
    - Consider memory time scale in our inputs
    - Paper would be useful to for state water agencies to develop ensemble forecasting

- Chris Funk, Director of the Climate Hazards Center, UC Santa Barbara:
    - **Expand data time range due to seasonality**
    - Create dummy variables to indicate seasonality
    - Different counties may be easier to predict drought scores for due to varying conditions

03

Modeling Progress

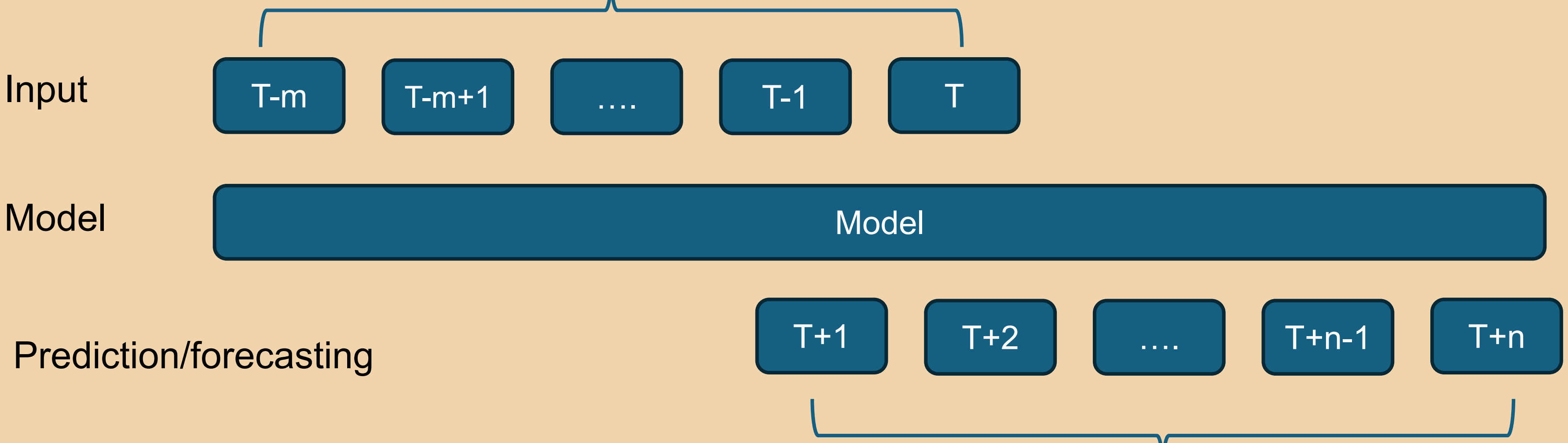
Data & ML Pipelines

- The data pipeline preprocesses weekly meteorological data and drought score data through
  - Train-validation-test splitting
  - Windowing (Transform time series forecasting problem into supervised machine learning)
  - Normalization
- ML architecture: sequence-to-sequence forecasting

ML Architecture

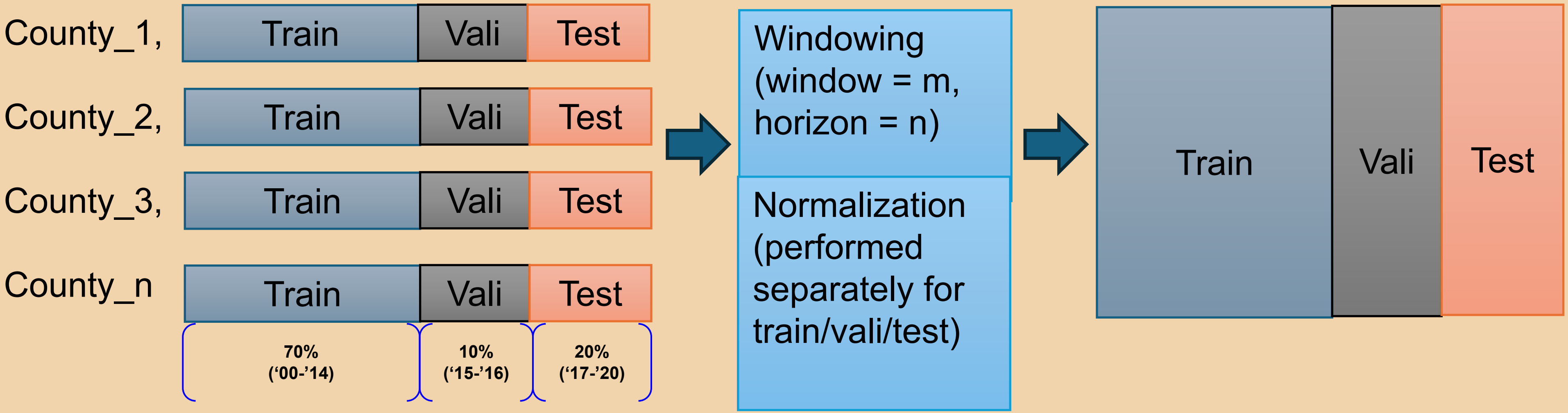
m weeks of historical data
(window = m)

m ranges from 8-52 weeks (30 weeks)
n ranges from 4-12 weeks (12 weeks)



n weeks of forecasting output
(horizon = n)

Data Engineering Pipeline



Data Engineering – Windowing

Transform time series forecasting problem into supervised machine learning

Features

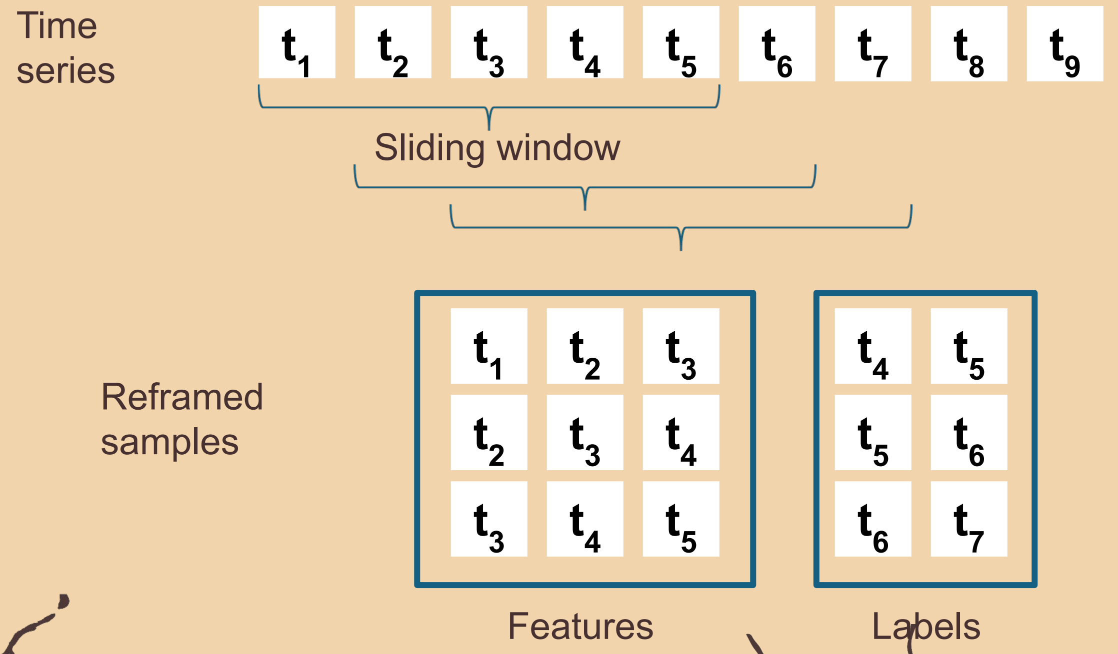
	Var1	Var2	Var3	…	Varm
T-m,	[[Var1_t-m,	Var2_t-m,	Var3_t-m, ...,		Varn_t-m],
T-m+1,	[Var1_t-m+1,	Var2_t-m+1,	Var3_t-m+1,...,		Varn_t-m+1],
…	…				
T-1,	[Var1_t-1,	Var2_t-1,	Var3_t-1,…,		Varn_t-1],
T	[Var1_t,	Var2_t,	Var3_t, ,…,		Varn_t]]

Labels

T+1,	[Score_t+1,
T+2,	Score_t+2,
…	…
T+n-1,	Score_t+n-1,
T+n	Score_t+n]

Data Engineering – Windowing

Transform time series forecasting problem into supervised machine learning



Modeling Approaches

Persistence

Assumes the future values results from persisting past trends

ARIMA

AutoRegressive Integrated Moving Average

CNN

Convolutional neural network

Random Forest

A “forest” of randomly generated decision trees

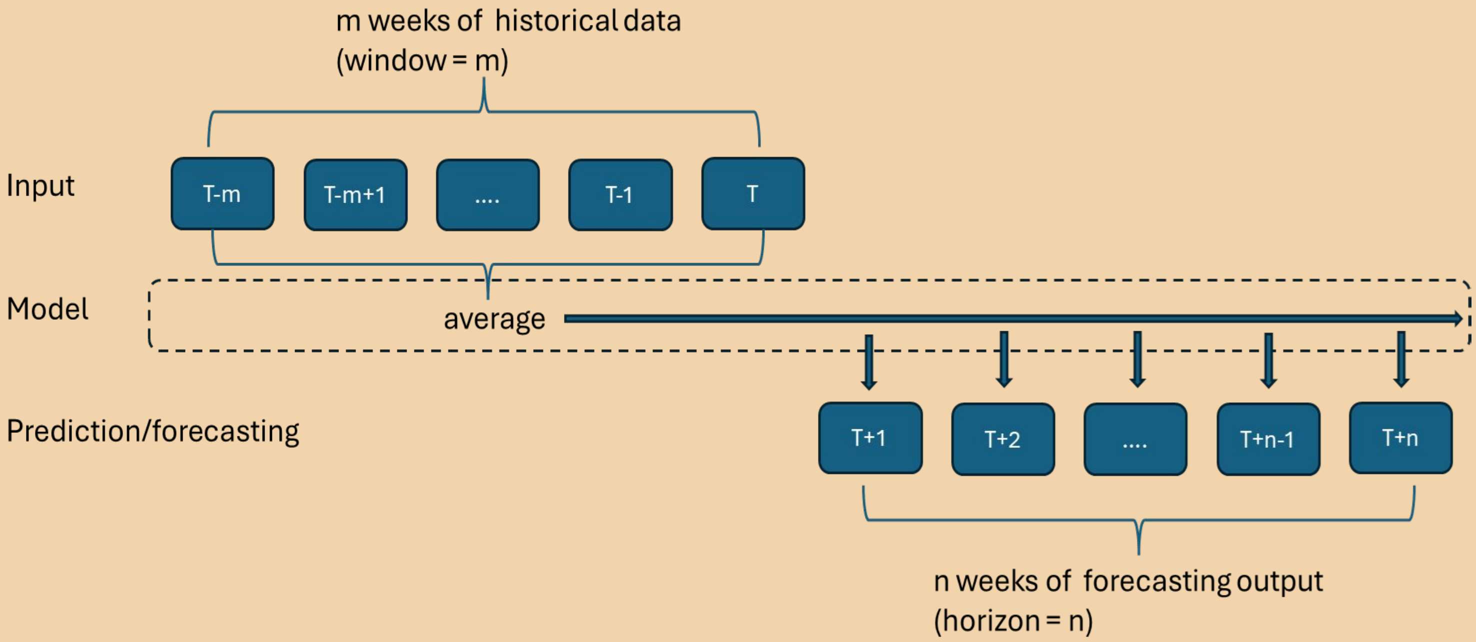
XGBoost

Extreme Gradient Boosting of decision tree ensemble

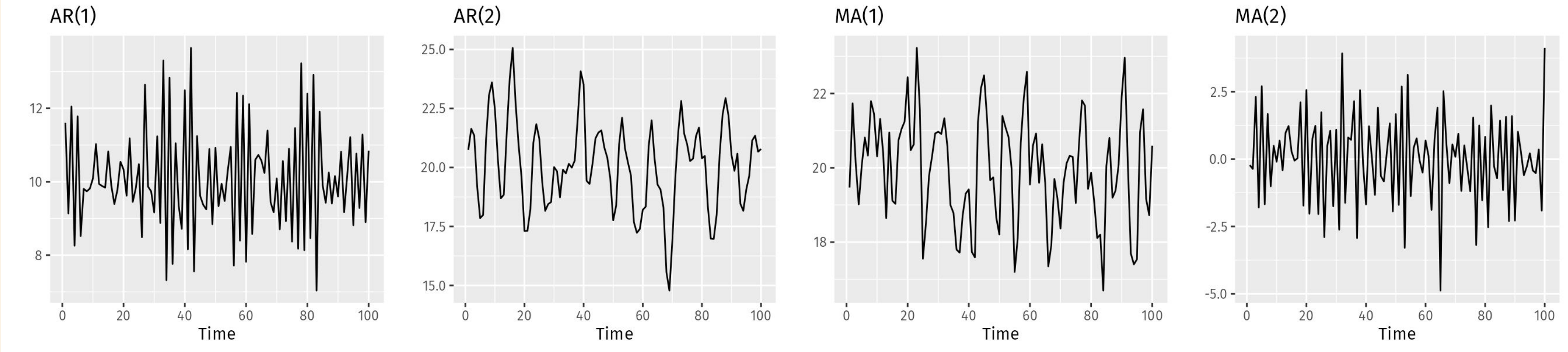
LSTM

Long short-term memory network or a variation of a recurrent neural network

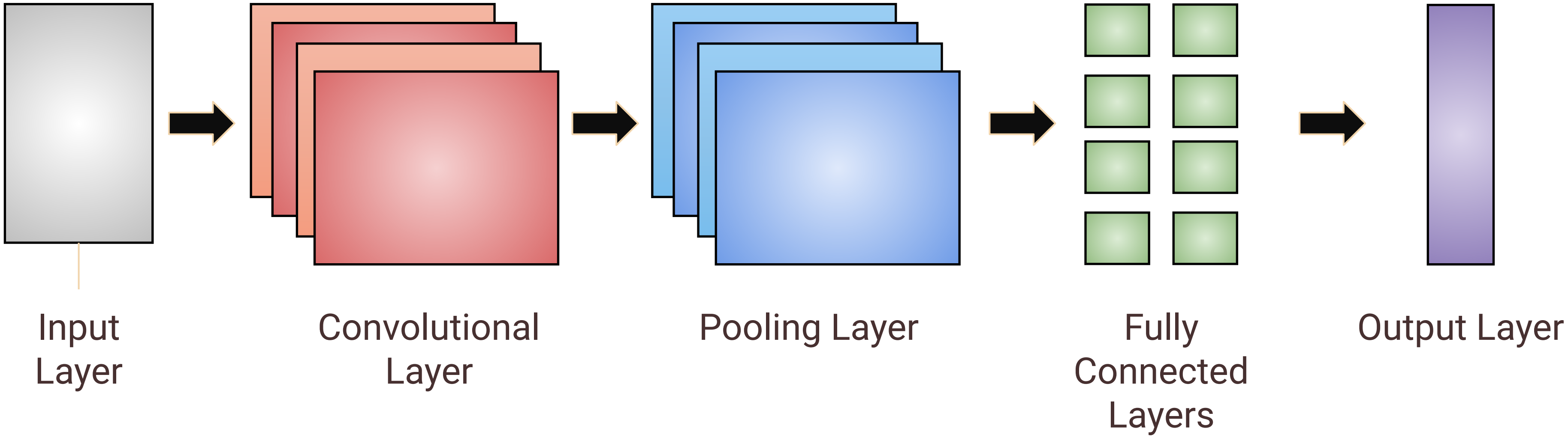
Baseline (Persistence Model)



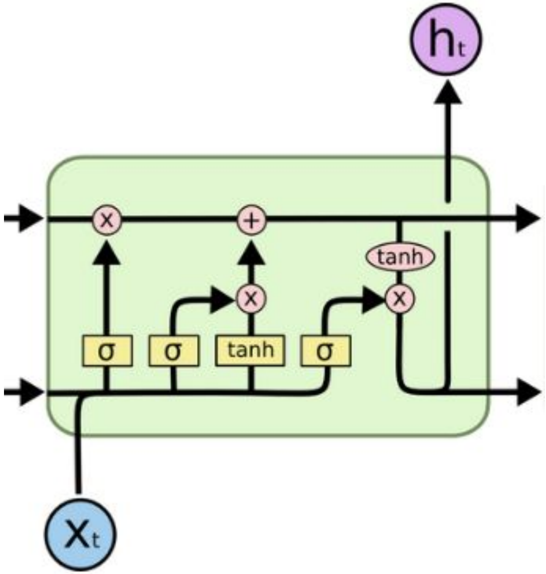
ARIMA:
AutoRegressive Integrated Moving Average



CNNs:
Convolutional Neural Networks



LSTM:
Long Short-Term Memory



$$i_t = \sigma\,(x_t W_{xi} + h_{t-1} W_{hi} + b_i) \qquad \textbf{(1)}$$

$$f_t = \sigma\,(x_t W_{xf} + h_{t-1} W_{hf} + b_f) \qquad \textbf{(2)}$$

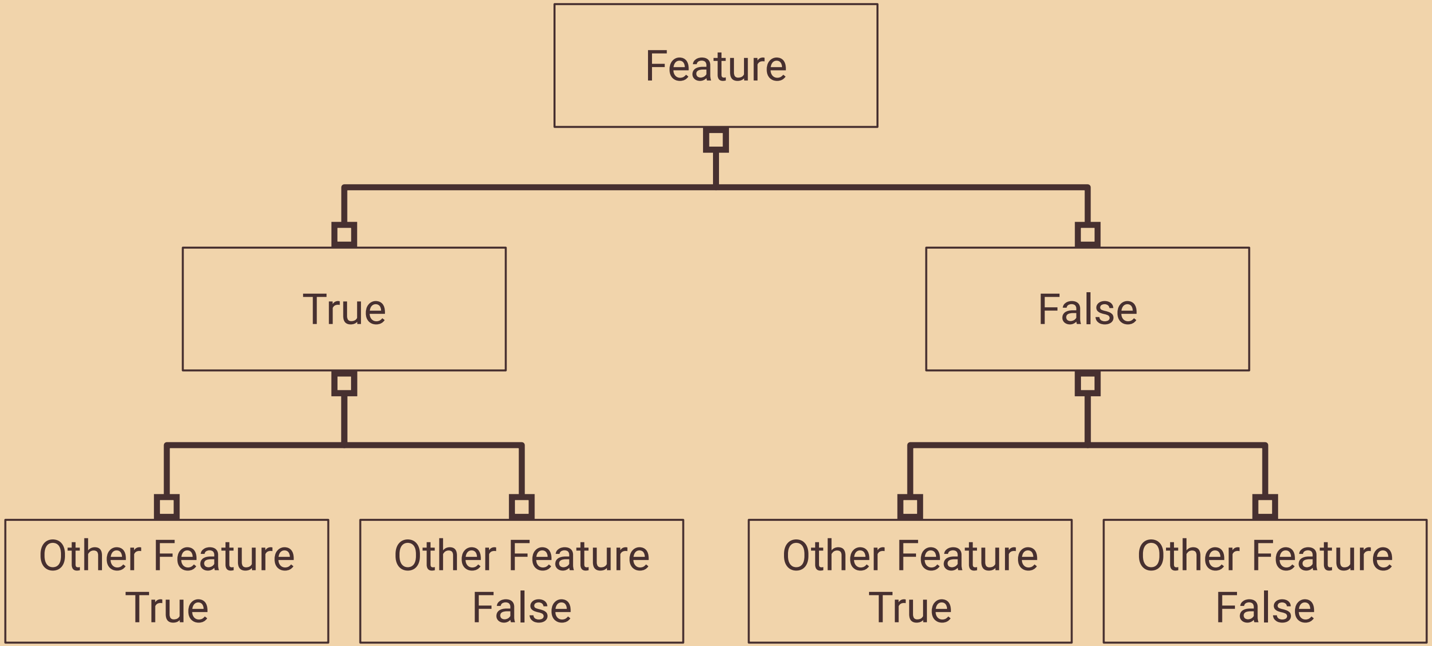
$$o_t = \sigma\,(x_t W_{xo} + h_{t-1} W_{ho} + b_o) \qquad \textbf{(3)}$$

$$c_t = f_t \cdot c_{t-1} + i_t \cdot tanh(x_t W_{xc} + h_{t-1} W_{hc} + b_c) \qquad \textbf{(4)}$$

$$h_t = o_t \cdot tanh(c_t) \qquad \textbf{(5)}$$

Image Source for LSTM Structure: Chris Olah, *Understanding LSTM Networks*, https://colah.github.io/posts/2015-08-Understanding-LSTMs

Random Forest and XGBoost



Univariate Test Model Results – Fresno County

Model	MSE	MAE	F1 for Severe Drought*
Persistence (Baseline)	0.51	0.45	
CNN	0.84	0.53	
Random Forest	0.35	0.45	
XGBoost	**0.35**	**0.45**	**0.56**
LSTM	0.39	0.41	

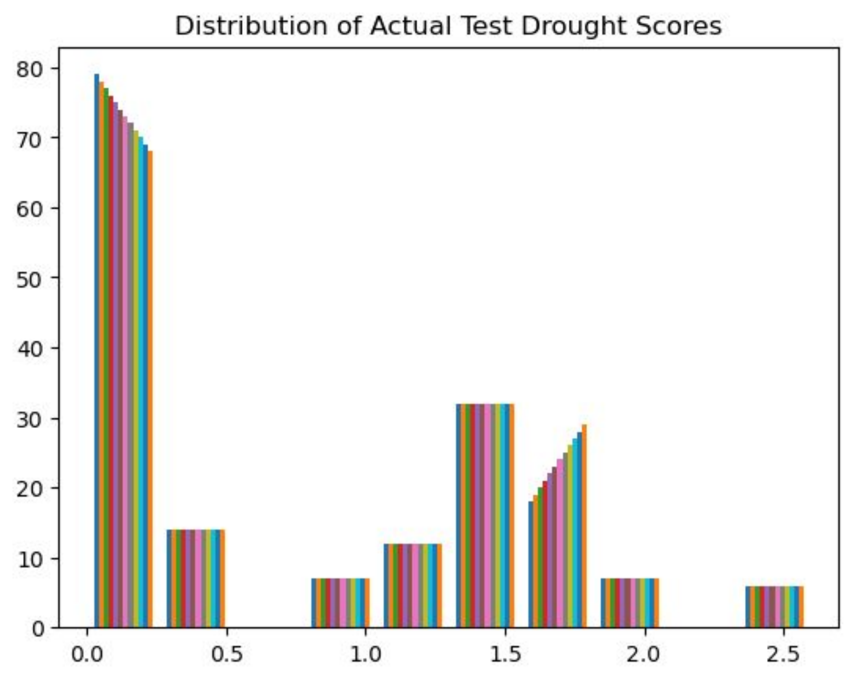
*We set a binary classification of severe drought at a score of 2.5 or higher.

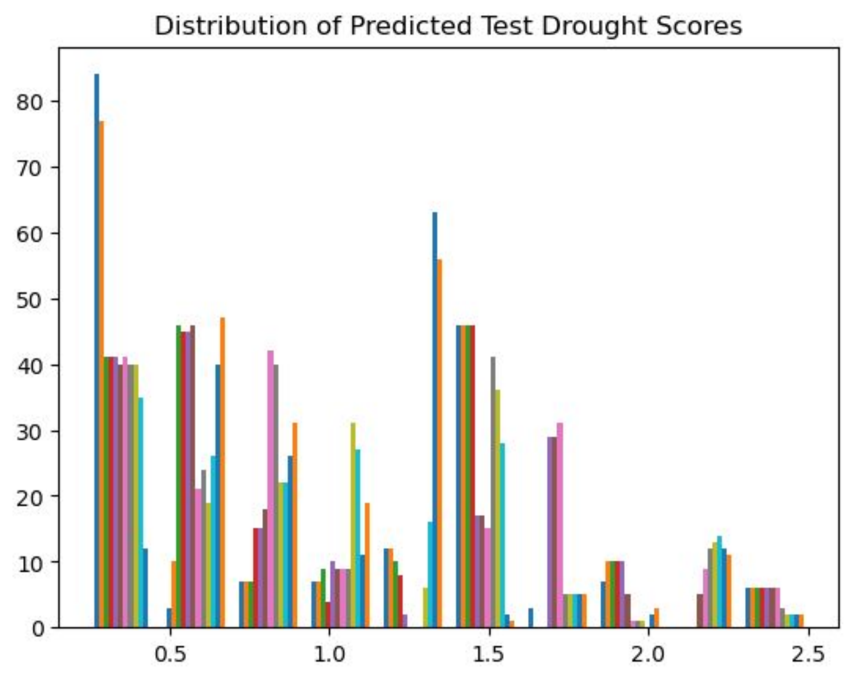
Multivariate Test Model Results – Fresno County

Model	MSE	MAE	F1 for Severe Drought+
Persistence (Baseline)	0.51	0.45	
CNN	0.89	0.74	
Random Forest	0.29	0.38	
XGBoost	0.33	0.44	0.56
LSTM	0.39	0.37	

*We set a binary classification of severe drought at a score of 2.5 or higher.

Fresno Multivariate XGBoost

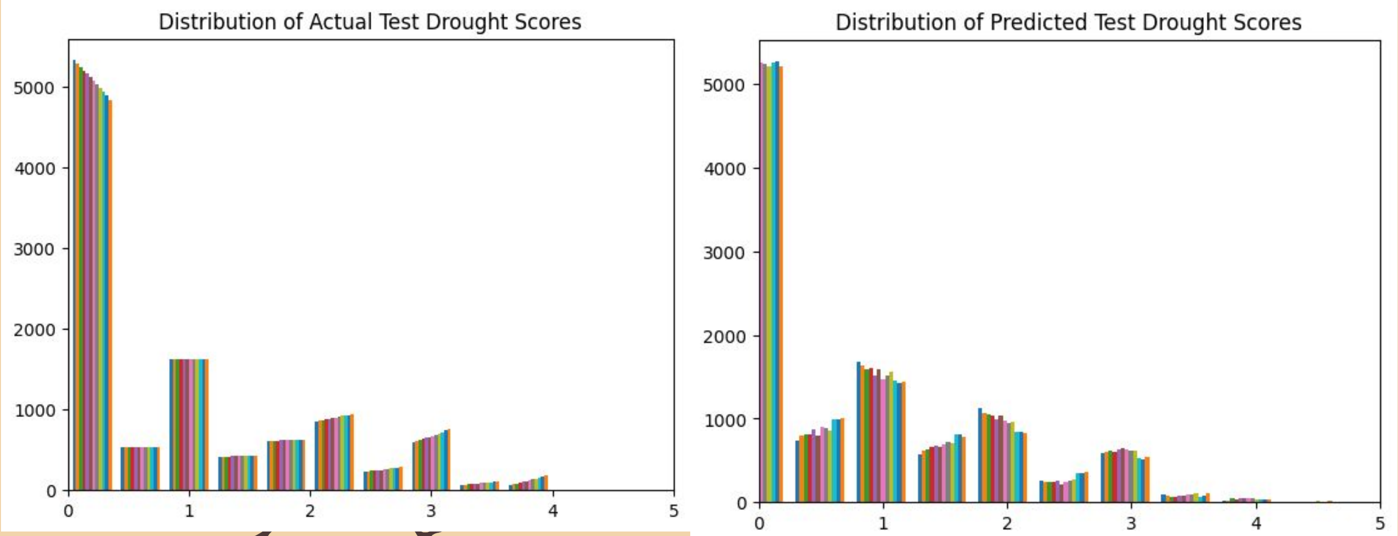




State-Wide LSTM

Model	MSE	MAE	F1 For Severe Drought+
LSTM	0.336	0.340	0.89

	precision	recall	f1-score	support
0	0.97	0.99	0.98	110778
1	0.92	0.72	0.81	13110
accuracy			0.96	123888
macro avg	0.94	0.86	0.89	123888
weighted avg	0.96	0.96	0.96	123888



04

Next Steps

Next Steps

Continue to improve, refine, and finalize our models

Identify and reach out to a local government scientist or policymaker for stakeholder feedback

Begin working on the MVP (research paper), web-based version, and final presentation

05

Conclusion: Mission

Mission Statement

"To **empower** governments with accurate **drought intensity predictions** using time series machine learning models, facilitating proactive planning and water management for California"

Thanks!

Do you have any questions?

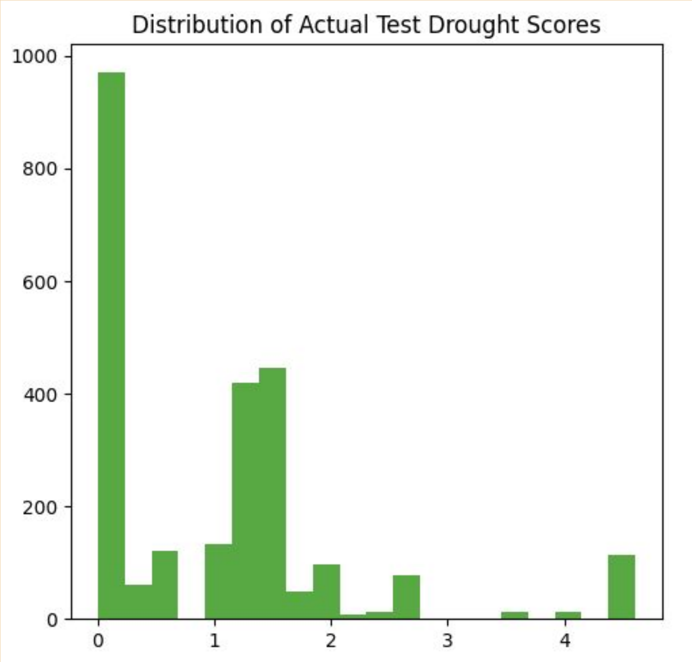
am_chang@berkeley.edu
nankli@berkeley.edu
davidsherman@berkeley.edu

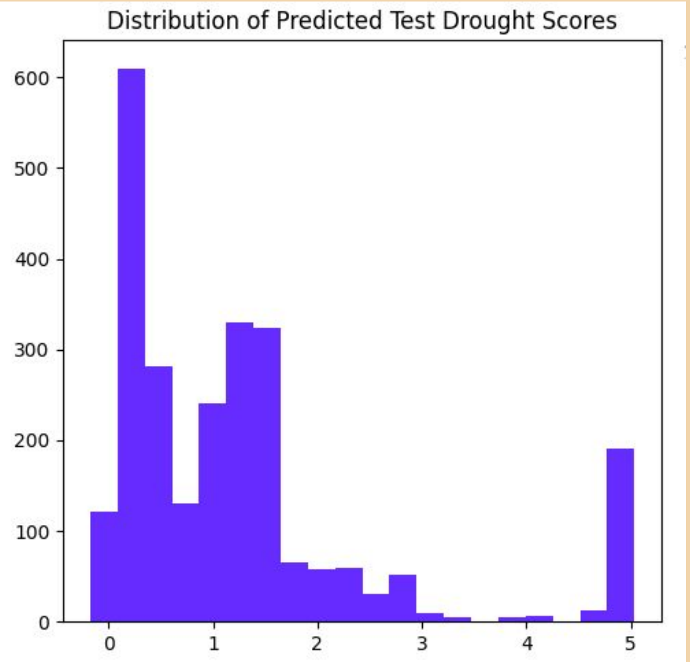
References

- Drought Prediction: A Comprehensive Review of Different Drought Prediction Models and Adopted Technologies
- Frontiers | DroughtCast: A Machine Learning Forecast of the United States Drought Monitor (frontiersin.org)
- Forecasting drought using neural network approaches with transformed time series data - PMC (nih.gov)
- https://www.scientificamerican.com/article/drought-takes-2-7-billion-toll-on-california-agriculture/
- https://www.mercurynews.com/2024/01/02/sierra-nevada-snowpack-at-lowest-level-in-10-years-what-it-means-for-californias-water-supply/
- https://droughtmonitor.unl.edu/CurrentMap/StateDroughtMonitor.aspx?CA
- https://www.drought.gov/drought-status-updates/california-nevada-drought-status-update-2023-10-19

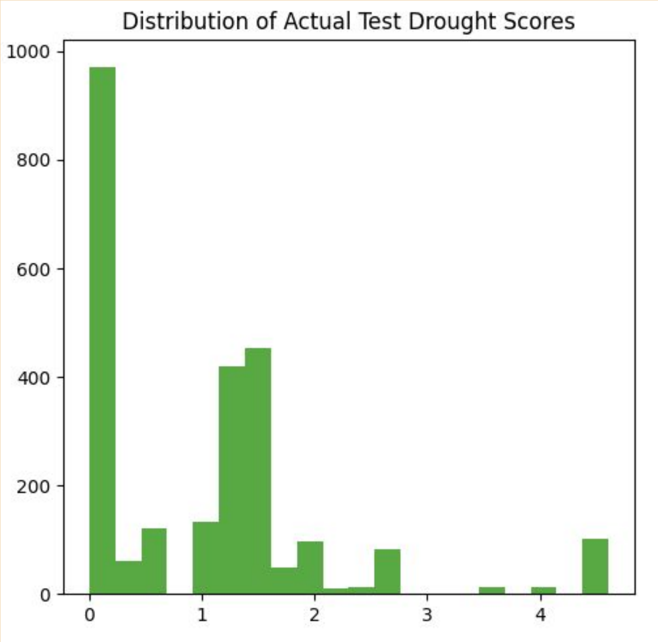
Appendix

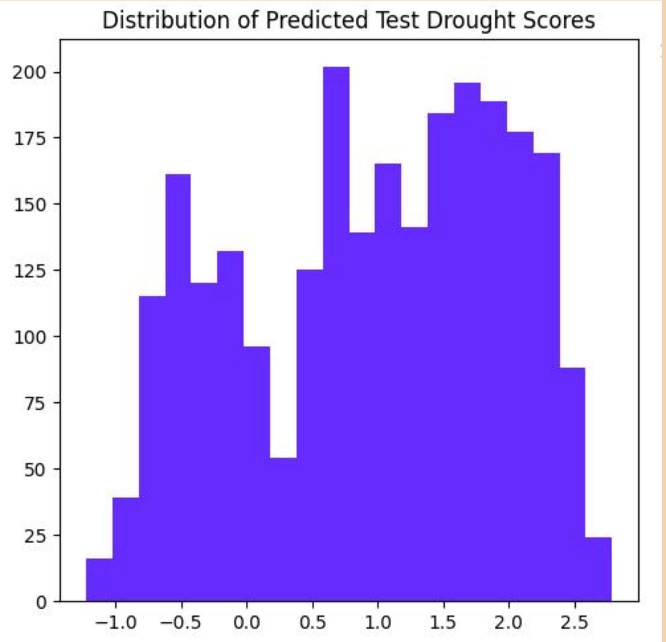
Fresno Univariate CNN



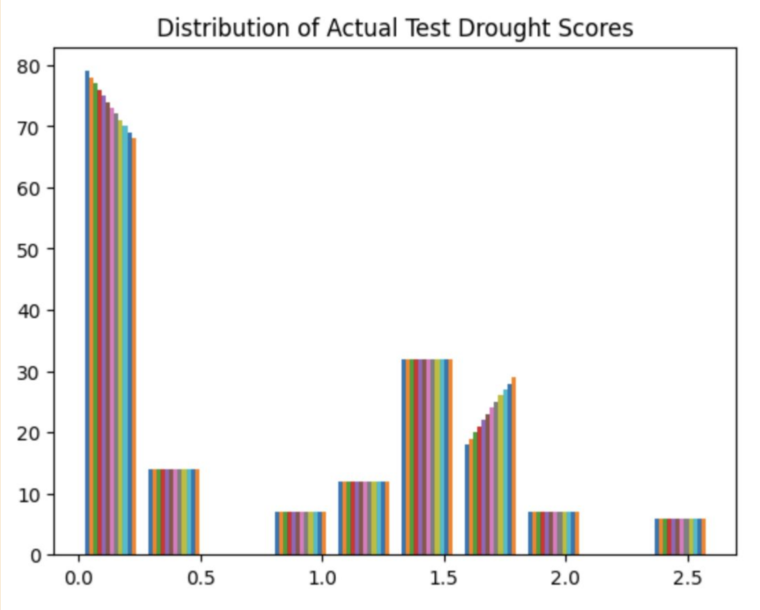


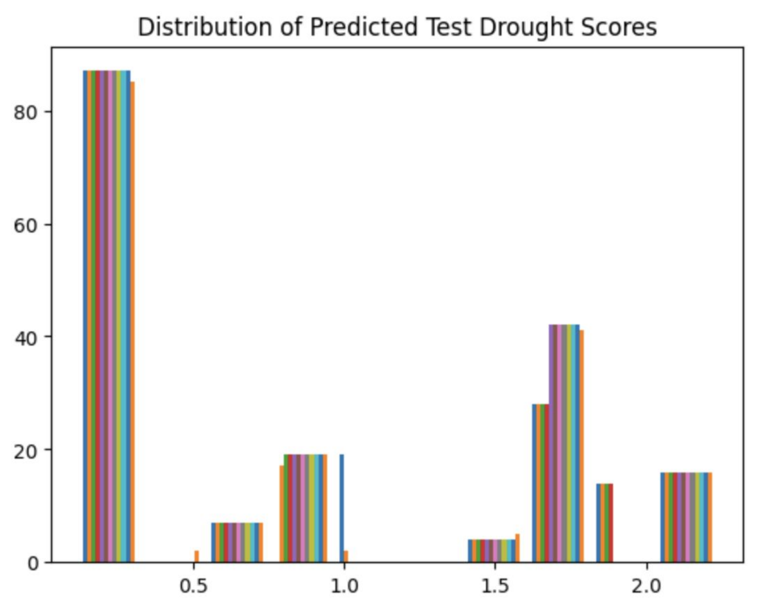
Fresno Multivariate CNN





Fresno Random Forest





LSTM – Fresno County

	precision	recall	f1-score	support
0	0.97	1.00	0.98	2028
1	0.00	0.00	0.00	72
accuracy			0.96	2100
macro avg	0.48	0.50	0.49	2100
weighted avg	0.93	0.96	0.95	2100

Multivariate Test Model Results – CA

Model	MSE	MAE	Macro F1
Persistence (Baseline)	0.50	0.43	0.81
CNN	0.36	0.37	0.87
Random Forest	0.33	0.39	0.86
XGBoost	0.30	0.35	0.88
LSTM	0.32	0.33	0.90