Drought Prediction using Rudimentary Meteorological & Soil Variables

Presented by: Chad Delany

Data Provided by: Kaggle.org,
North American Drought Monitor (NOAA)

Springboard School of Data:
December 11, 2022

Problem

The United States has an abundance of weather and soil data compared to other countries.

- Using basic weather and rudimentary soil data, can we accurately predict droughts in the United States?
- Can a model be developed that has an accuracy greater than 80% within the next two months?
- Can these results be generalized to other countries with less available data resources?

Key Findings

Table 1. Random Forest model accuracy on training dataset using all variables.

Class	Accuracy
0	79%
1	74%
2	69%
3	69%
4	74%
5	84%
Total	**77%**

Table 2. Random Forest model accuracy on training dataset using select variables.

Class	Accuracy
0	70%
1	62%
2	59%
3	59%
4	66%
5	76%
Total	**69%**

Table 3. Random Forest model accuracy on test dataset using all variables.

Class	Accuracy
0	76%
1	17%
2	15%
3	5%
4	0%
5	0%
Total	**74%**

Key Variables

Meteorological Variables:

- Wind Speed
- Specific Humidity
- Temperature Range
- Surface Pressure
- Dew/Frost Point
- Wet Bulb Temperature
- Precipitation

Soil Variables:

- Latitude/Longitude
- Elevation
- Slope/Aspect
- Natural Cover Type
- Nutrient Availability

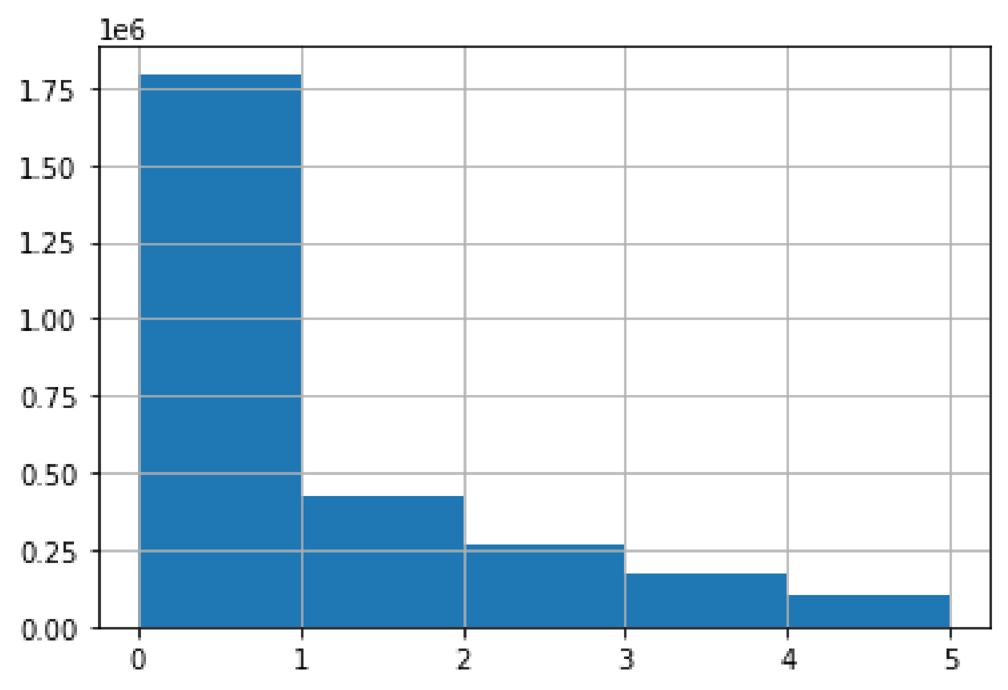
Drought Score:

- Category 0 – No Drought
- Category 1 – Abnormally Dry
- Category 2 – Moderate Drought
- Category 3 – Severe Drought
- Category 4 – Extreme Drought
- Category 5 – Exceptional Drought

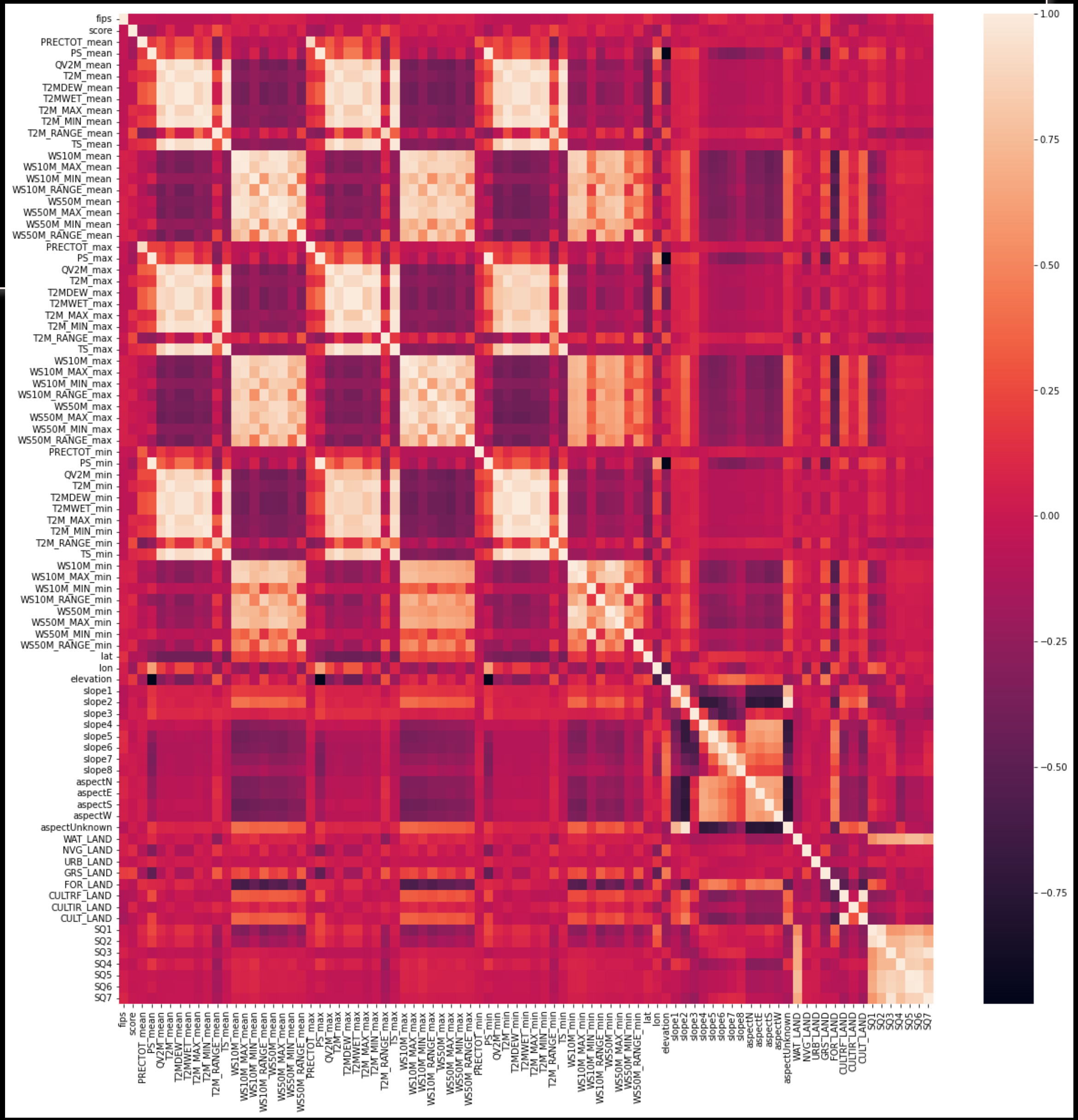
Dataset summary:

- time period 2000 – 2020
- 18 daily meteorological variables
- 1 weekly drought score
- 30 soil variables
- 3,100+ counties in the US
- ~3 million rows of data

Models & Analysis



Skewed distribution of drought scores.



Heatmap of correlation of all soil and derived meteorological variables.

Models & Analysis

Comparison of the regression models' accuracy.

Model & Metric	R**2	MSE	RMSE	MAE
Dummy Regression	0.000	1.497	1.224	0.975
Linear Regression	0.215	1.176	1.084	0.819
Ridge Regression	0.154	1.266	1.125	0.872
ElasticNet Regression	0.074	1.387	1.178	0.929
Nearest Neighbor Regression	0.468	0.796	0.892	0.574
Random Forest Regression	0.714	0.429	0.655	0.434

Models & Analysis

Comparison of the classification models' accuracy.

Model & Metric	ROC AUC	Total Accuracy	Mean Accuracy per Class	Accuracy per Class STD
Logistic Regression	0.844	60%	27%	16%
Nearest Neighbor Classifier	1.472	68%	55%	9%
Random Forest Classifier	1.720	77%	75%	5%

Next Steps

The next steps to improve this model would be:

- Allocate more resources so that the training can be done on the entire training dataset.
- Subset the training dataset and rerun the models to allow for a standard cross validation procedure and allow tools that determine important input variables to be determined.
- Incorporate ordinality information into the classification schema.
- Incorporate a time series analysis that capitalizes on the time nature of the data.
- Use a recurrent neural network to build a time series model.

Conclusions

- Produced a viable model (74% accuracy).
- Demonstrated the usefulness of Random Forest models for this problem.
- It was not able to highlight a few, key variables.
- Given the changing climate and the inherent integration of economies throughout the present-day world, understanding and accurately predicting drought is an important first step in adapting to the current changing conditions of our environment and maintaining a viable global economy.
- Being able to predict drought from simple variables and not overly complex models, would allow them to be applied worldwide.

Thank you.

Questions?