Drought Prediction in the American West

Tyler Meester

Problem Statement

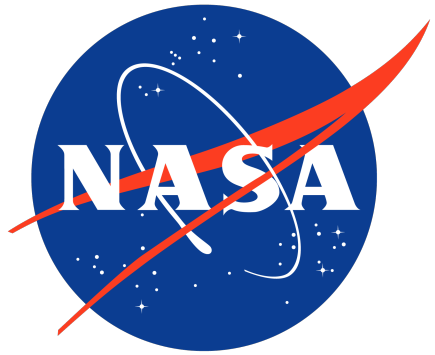
- All around the world wildfires are damaging the lives and property of many people.
- Due to a multitude of factors such as drought, bark-beetle infestation, bad fire management practices, increases in outdoor recreation, degrading energy infrastructure, and much more, the western United States is now host to catastrophic wildfires that haven’t been witnessed in recorded history.
- Unfortunately, climate scientists warn that this is the “new normal”.

Problem Statement

- Being able to predict wildfire risk could be extremely helpful for communities, businesses, fire management crews, and many others to help plan and prepare as we continue into the future. One part of the equation in preparing for wildfires is being able to predict future drought levels. **The aim of this project is to predict the drought level of a given county 1 month into the future.**
- This information could prove invaluable for wildfire fighting preparations. Being able to quickly fight a growing fire is essential, and having an idea of the most at risk areas could make all the difference in succesfully fighting a fire. **More specifically, a model such as this could be used by fire management crews for staffing forecasting and fire advisory/restriction implementation.**
- This project can be considered a success if the model is able to correctly identify 75% or more of the high drought counties in the dataset.

Data - Meteorological Timeseries

- This dataset includes data from the NASA Langley Research Center (LaRC) POWER Project which is funded through the NASA Earth Science/Applied Science Program. The meteorological timeseries dataset was compiled by Christoph Minixhofer and uploaded to Kaggle.

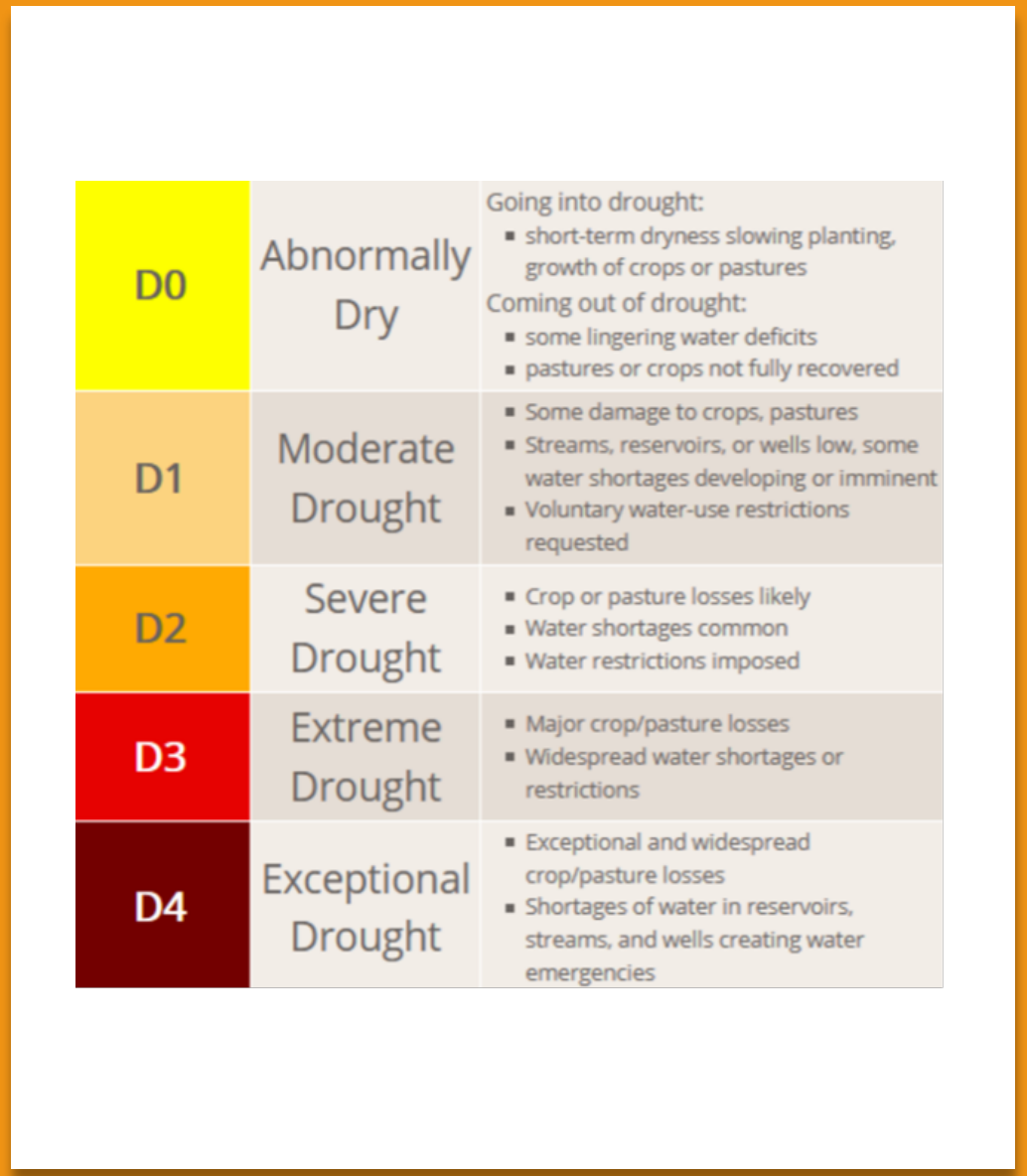


Data – Soil

- The features in the soil dataset are from the Harmonized World Soil Database, which contains data generated by the NASA Shuttle Radar Topographic Mission (SRTM). This includes digital elevation data (DEMs) for over 80% of the globe with 3 arc second (approximately 90 meter) resolution at the equator. This soil datset was compiled by Christoph Minixhofer and uploaded to Kaggle.

Data – Drought Scores (Target Feature)

- The US Drought Monitor collects drought scores for counties throughout the US on a scale from No Drought - D4. In the dataset these values are floating point values ranging from 0-5, 0 corresponding to No Drought and 5 corresponding to D4. These values are the county (FIPS Code) average and are collected weekly.



D0	Abnormally Dry	Going into drought: ▪ short-term dryness slowing planting, growth of crops or pastures Coming out of drought: ▪ some lingering water deficits ▪ pastures or crops not fully recovered
D1	Moderate Drought	▪ Some damage to crops, pastures ▪ Streams, reservoirs, or wells low, some water shortages developing or imminent ▪ Voluntary water-use restrictions requested
D2	Severe Drought	▪ Crop or pasture losses likely ▪ Water shortages common ▪ Water restrictions imposed
D3	Extreme Drought	▪ Major crop/pasture losses ▪ Widespread water shortages or restrictions
D4	Exceptional Drought	▪ Exceptional and widespread crop/pasture losses ▪ Shortages of water in reservoirs, streams, and wells creating water emergencies

Methods

The data goes back 20 years and contains numerous features, some of which have different time frequencies (daily, weekly, and some with no time series aspect at all).

Initially this was approached as a multi-class classification problem, the time series information was to be captured through creative feature engineering.

After further consideration I decided to simplify the problem even further by turning it into a binary classification problem, splitting the target feature "drought score" into two categories:

- Low Drought (No Drought, D0, D1)
- High Drought (D2, D3, D4).

The full dataset included millions of daily entries for thousands of individual counties in the United States.

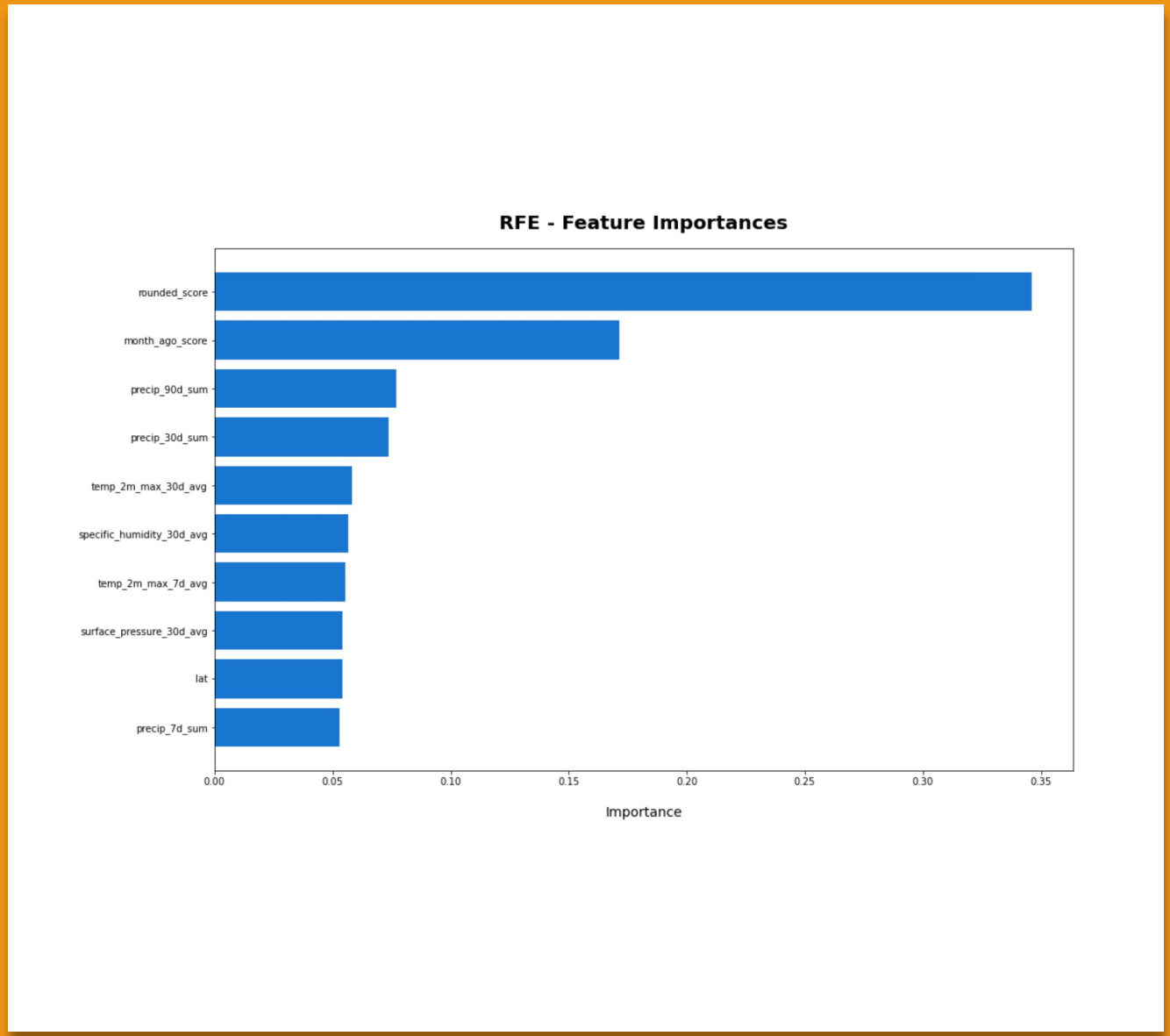
A smaller subset of this data was created that focused on the West Coast States (California, Oregon, and Washington). These particular states were selected as they have experienced the most destructive wildfires over the past decade.

Data Cleaning and Preparation

- The data cleaning and preparation was by far the most complex stage of this project, as it required a great deal of feature engineering and involved data with different collection frequencies. I developed functions that would create new datasets with the desired features for each county (FIPS Code) and then combine them into one large dataset.
- **The developed functions performed the following operations:**
  - **Aggregated the sums and averages of the daily meteorological features**
  - **Extracted the month from the date of each entry**
  - **Dropped all NaN values, to make the time frequency of the data uniform (weekly)**
  - **Rounded the drought scores to change from floating point numbers to integers**
  - **Created new features that showed the previous and following month's drought scores**
  - **Combined county datasets into one large dataset**
- After these datasets were compiled I added the corresponding soil data for each county.
- **By the end of this process the final datasets had 56 features. The full training dataset had 3,014,760 entries and the west coast training dataset had 129,010 entries.**

Exploratory Data Analysis

- From this visualization of feature importances in the RFE model, we can see the most important features in predicting the target feature.
- I decided to include all of these features in the final dataset, greatly reducing the number of features from 56 to 10.

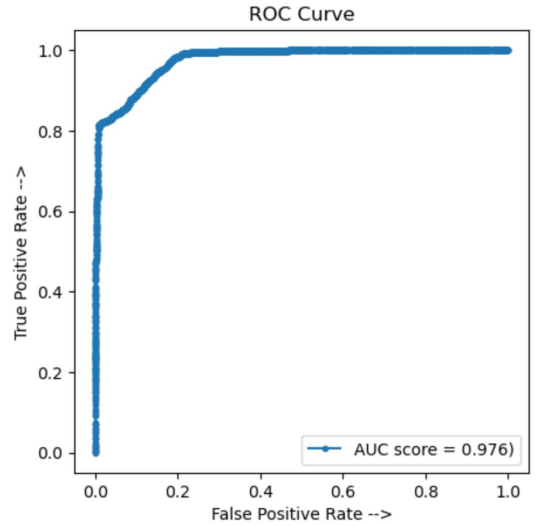


Final Model Performance

- The final model has an ROC AUC Score of ~0.98 and a Recall Score of 80%. This indicates that **the model was able to correctly predict 80% of the actual high drought counties in the testing dataset.**

Testing Data Classification Report:

	precision	recall	f1-score	support
0	0.96	0.99	0.98	9319
1	0.96	0.80	0.87	1853
accuracy			0.96	11172
macro avg	0.96	0.90	0.93	11172
weighted avg	0.96	0.96	0.96	11172



Future Improvements

- There is undoubtedly room for improvement on this project, I think the below changes would be worth implementing in the effort to boost model performance:
  - The model might be limited in its predictive power by the way the training and testing data is split up (chronologically). Since I have removed the time-series component, it could be beneficial to combine the training and testing datasets (which are currently completely seperate) and do a train_test_split on the full dataset so that some of the later observations, with higher drought scores, are included in the training data.
  - it is apparent that the dataset is quite umbalanced, with a majority being 'Low Drought', more time could be spent trying to work around this limitation in future iterations of the project.

Final Thoughts

- Overall this project was very fun and I learned a great deal. I enjoyed developing my feature engineering skills and implementing more custom functions to limit copy/pasting and repetitive actions. I found it fascinating that we could take a time-series problem and approach it as a classification problem with creative feature engineering.
- **I believe that the level of accuracy of this model makes it usable by fire management crews for staffing forecasting and introducing fire advisories/restrictions in high risk areas ahead of time.**