

**Predictive Analytics**

***Identifying Key Factors in***

***Drought Prediction***

*Nidhi Danayak*

*Luke Jackson*

*Jason Winn*

**School of Graduate Professional Studies**

Data Analytics

IE-575 – Predictive Analytics

Spring 2021

# Document Control

## Work carried out by:

|  |  |  |
| --- | --- | --- |
| **Name** | **Email Address** | **Task description** |
| Nidhi Danayak | nmd5605@psu.edu | Research, coding, documentation |
| Luke Jackson | lmj5310@psu.edu | Research, coding, documentation |
| Jason Winn | jew5184@psu.edu | Research, coding, documentation |

## Revision Sheet

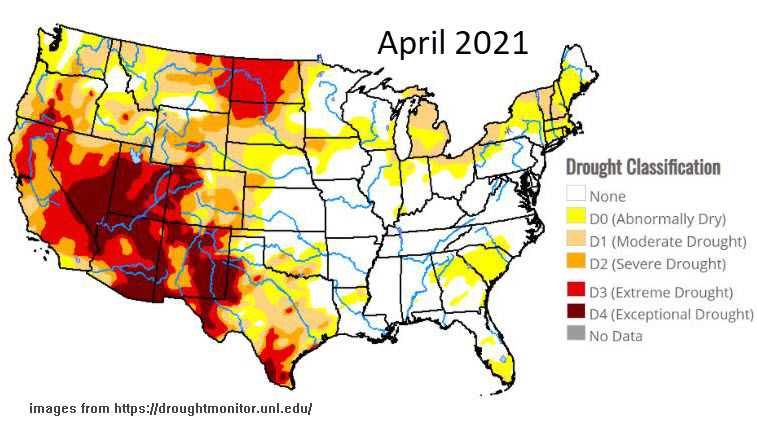
|  |  |
| --- | --- |
| **Date** | **Revision Description** |
| 27-Apr-2021 | Initial draft |
| 29-Apr-2021 | Final submission |

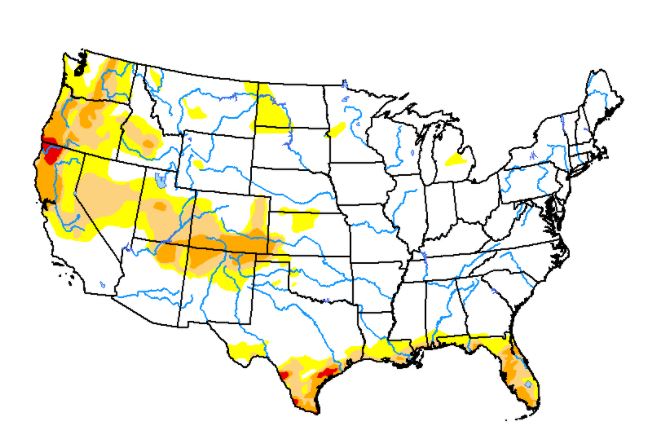
**TABLE OF CONTENTS**

1. Introduction
2. Problem Statement / Research Question
   1. Importance
3. Data
   1. Exploratory Data Analysis
   2. Data Preprocessing
4. Methodology
5. Results
   1. Discussion of Results
6. References
7. Appendix

##### 1. Introduction

The United States is currently in a megadrought, reported by various news agencies as the most severe drought in modern US history. Many news media outlets display a map from the United States Drought Monitor, a joint production of various government agencies and educational institutions.



Above is the map for April 20, 2021. The drought classification score is a measure of relative draught intensity, and for comparison the map for April 21, 2020 is shown here on the left. This information is made publicly available and is used primarily for public policy decision making. Additionally, the authors of the monitor encourage interaction with any interested parties, declaring on their website "let's take a look at the data and see what it says." Inspired by this maxim, our team found a cleaned-up version of this dataset on Kaggle, and gave it a try.

April 2020

##### 2. Problem Statement / Research Question

Drought response is difficult, when compared to other natural disasters such as tornados or hurricanes, because the effects come on slowly and they last a long time. Additionally, drought response is very decentralized - no single federal agency is responsible for US drought policy. One of the goals of the US drought monitor is to unify federal drought response and policy. And one way they do this is to investigate if droughts could be predicted using weather and soil data. Today, the tool is only able to look backwards – to provide the data about what already happened. But what if we could use that same historical data to predict what is likely to happen?

Our team would like to analyze the data behind the US Drought Monitor using R and the predictive analytics tools that we've acquired in this class. Using various modelling techniques, we would like to learn which factors are most important in determining drought severity. We have included some obvious factors in our analysis, such as precipitation. Hight levels of precipitation will predict low drought levels. However, most of our factors involve measures of wind speed and temperature above and below ground. These and other data points are available to farmers and other land developers, and our group hopes to determine whether this data is useful in predicting drought severity.

##### 2a. Importance

Drought is a regular part of a normal climate cycle. However, drought can cause destruction and losses to agriculture, energy production, public health, and the public water supply. Current conditions in the United States are historically dry and look as though they will only get worse. The federal government is preparing to issue the first ever water shortage declaration, as they anticipate that for the first time ever, the water level at Lake Mead will fall below 1,075 ft. In June, 2021. It also threatens to fall to such lows that Hoover Dam will no longer be able to generate electricity. Long-term development plans and short-term responses can all be improved with improved drought prediction. Our group hopes to help in this effort by identifying drought severity factors, especially those that might be less obvious.

##### 3. Data

Our data was downloaded from a Kaggle dataset called "Predict Droughts using Weather & Soil Data". <https://www.kaggle.com/cdminix/us-drought-meteorological-data>

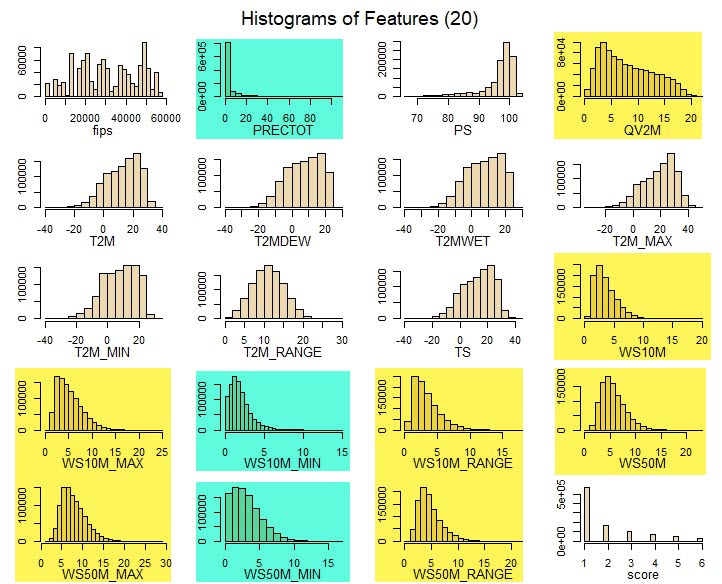
This dataset is an amalgamation of weather and soil data obtained from:

* [The US Drought Monitor](https://droughtmonitor.unl.edu/About/WhatistheUSDM.aspx)
* [NASA POWER Project](https://power.larc.nasa.gov/)
* [Harmonized World Soil Database](http://www.fao.org/soils-portal/data-hub/soil-maps-and-databases/harmonized-world-soil-database-v12/en/)

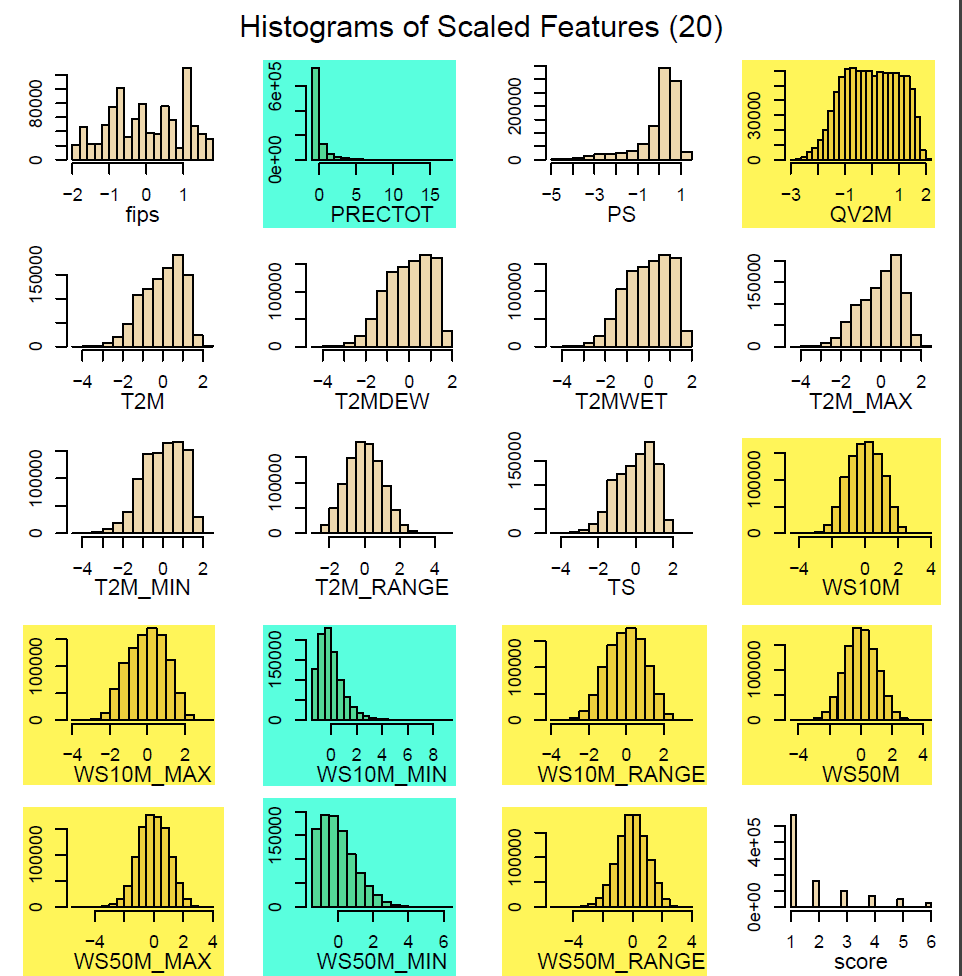
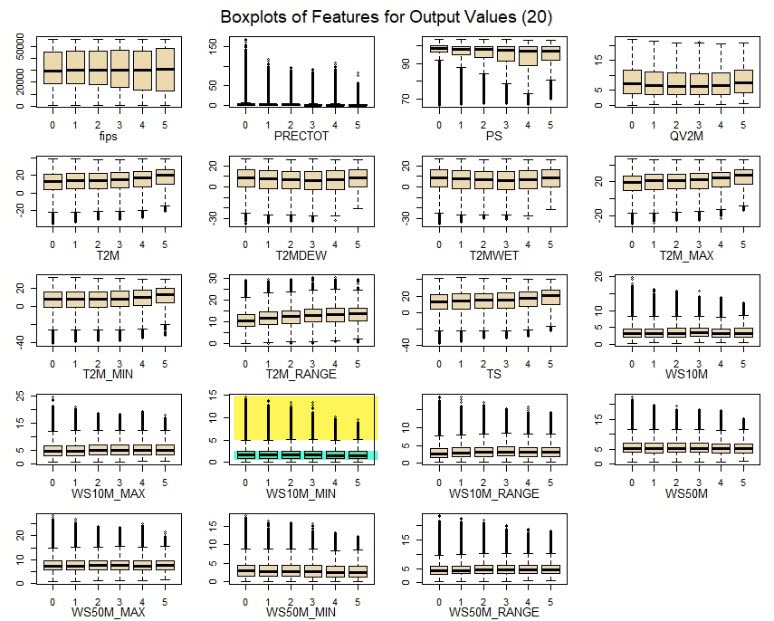
There are 18 scientific measurements available, consisting primarily of temperature and wind speed measurements. The date and geographic location of each measurement is provided, as well as the associated drought score.

|  |  |  |
| --- | --- | --- |
| **Column Name** | **Description** | **Data Type** |
| fips | 5-digit FIPS code used to identify geographic location (e.g. US counties)  \*we will read codes with leading zeros as 4-digits | numeric (categorical, 4 or 5-digit non-ordinal integers) |
| date | measurement observation date | date (yyyy-mm-dd format) |
| PRECTOT | Precipitation (mm day-1) | numeric (continuous, 2-decimal scientific measurements) |
| PS | Surface Pressure (kPa) |
| QV2M | Specific Humidity at 2 Meters (g/kg) |
| T2M | Temperature at 2 Meters (C) |
| T2MDEW | Dew/Frost Point at 2 Meters (C) |
| T2MWET | Wet Bulb Temperature at 2 Meters (C) |
| T2M\_MAX | Maximum Temperature at 2 Meters (C) |
| T2M\_MIN | Minimum Temperature at 2 Meters (C) |
| T2M\_RANGE | Temperature Range at 2 Meters (C) |
| TS | Earth Skin Temperature (C) |
| WS10M | Wind Speed at 10 Meters (m/s) |
| WS10M\_MAX | Maximum Wind Speed at 10 Meters (m/s) |
| WS10M\_MIN | Minimum Wind Speed at 10 Meters (m/s) |
| WS10M\_RANGE | Wind Speed Range at 10 Meters (m/s) |
| WS50M | Wind Speed at 50 Meters (m/s) |
| WS50M\_MAX | Maximum Wind Speed at 50 Meters (m/s) |
| WS50M\_MIN | Minimum Wind Speed at 50 Meters (m/s) |
| WS50M\_RANGE | Wind Speed Range at 50 Meters (m/s) |
| score | Measure of drought ranging from 0 (no drought) to 5 (exceptional drought) | numeric (continuous) |

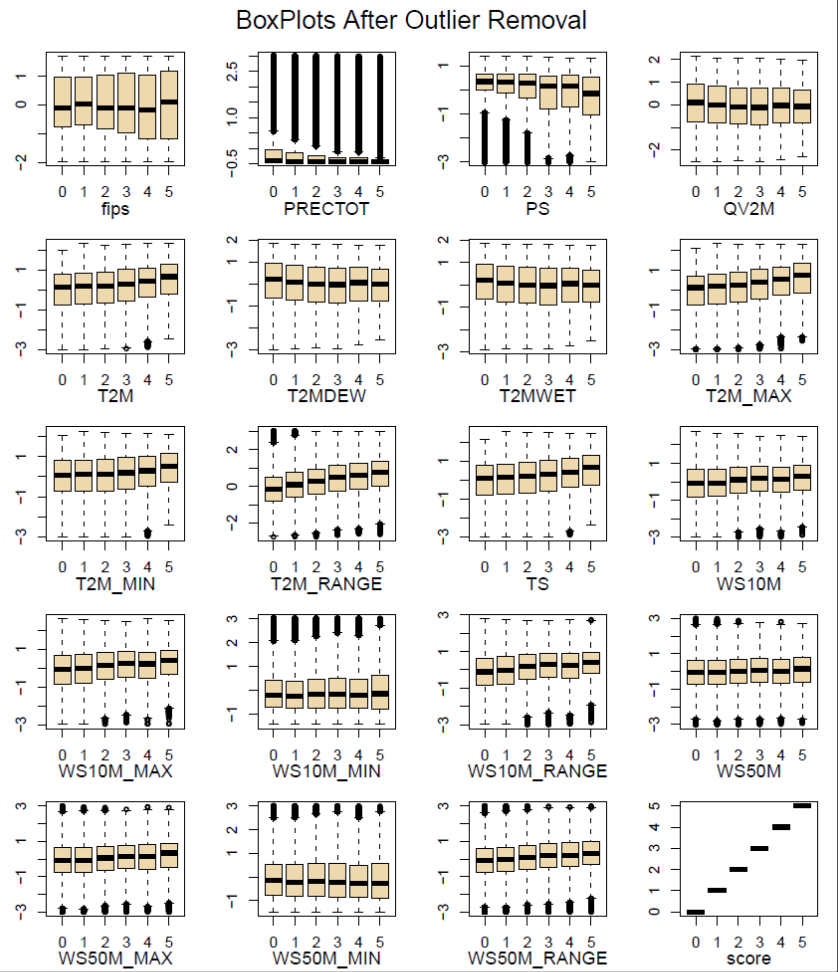
##### 3b. Exploratory Data Analysis

We built histograms of all of the features and realized that some columns are not normally distributed. To solve this issue, we pre-processed our data using the preProcess function's 'BoxCox' method, along with center scaling. The figures below show the distributions of our features before and after scaling.

* The distribution normalization was most effective for the yellow highlighted factors.
* The x-axis scale was normalized for all factors.
* Notice the right-tail reduction the blue-highlighted factors, which reduces outlier impact.

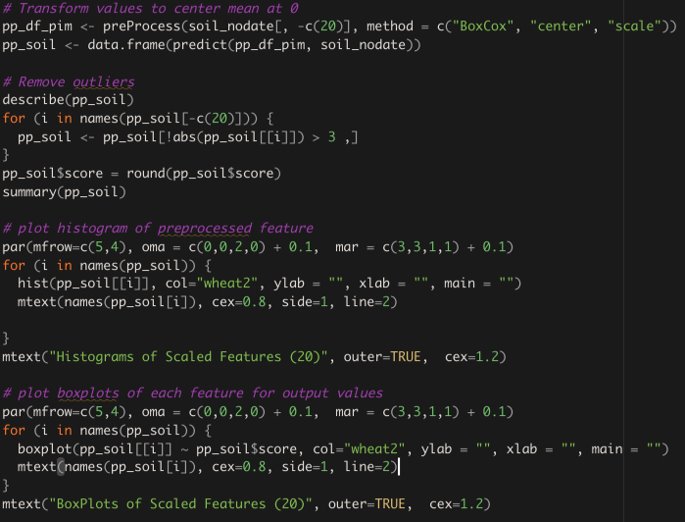
We next created boxplots of each feature for score values. We noticed outliers spread across each feature. For example, if you look at WS10M\_MIN highlighted below, you'll see how the "outliers" outnumber the data points in the center of the boxplot.

We decided to remove these outliers to get better accuracy

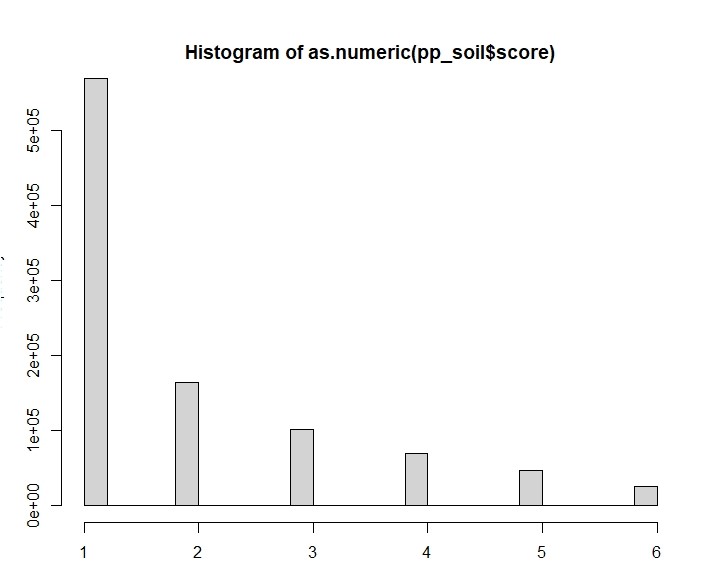


##### 3b. Data Preprocessing

We experimented with many different data preprocessing techniques in order to end up with the best data set for our models. We spent a significant amount of time "thinking" through our dataset and how we could compensate for the significant non-normalized (high skew) nature of the data set. For our final process, we:

* imported the data and removed any duplicates and any NAs.
* centered all features to have a mean of 0.
* removed any outliers greater than 3 standard deviations from the mean.
* rounded off the output-y (drought scores) to get 6 distinct, discrete drought scores
* plotted the histograms of each feature
* plotted the boxplots of each feature for each score
* reduced the size by limiting it to 5 years of data: 2011-2016.

We began with a very large and imbalanced data set.



* There are significantly more samples of level 1 - 'No Drought', than of any other sample. the number of samples then decreases as the severity level gets higher.

We dealt with this imbalance by using downsampling. We decided that the category with the fewest samples, drought score 6, contained enough, so we reduced all of the category sample sizes to 24,665.

* The data set was too large for R and for our machines. The Original Files (1 Record per Day per County), were over 2G.

|  |  |  |
| --- | --- | --- |
| **Split** | **Records** | **Size (G)** |
| Train | 19,300,681 | 2.1G |
| Validation | 2,268,841 | 246M |
| Test | 2,271,949 | 246M |



* After running the above Dplyr Group by Year, and Month (2020-04), the reduced data set is shown below.

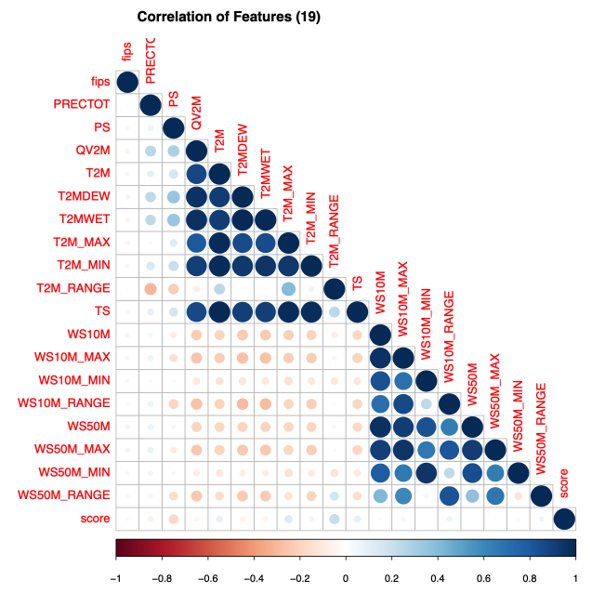
|  |  |  |
| --- | --- | --- |
| **Split** | **Records** | **Size (G)** |
| Train | 634,033 | 83M |
| Validation | 74,593 | 9.8M |
| Test | 74,593 | 9.7M |

* We consciously excluded date from our analysis due to the large size of the dataset.

##### 4. Methodology

After we were familiar with the data, we decided on an objective of finding which factors "most" influence drought score. Ordinal regression and multiple classification models provided the best means to determine possible relationships between weather data features and the severity score of drought.

Unfortunately, as visible from our Corrplot only a few features had any correlation to drought score. We expand on this in our findings.



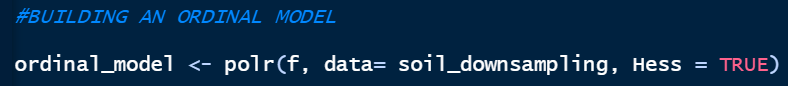
According to these correlation matrix results, only the 5 variables highlighted yellow are correlated to our dependent variable score.

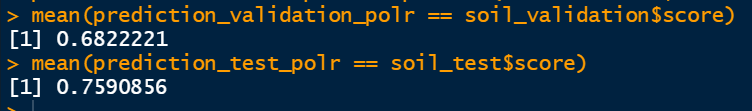
|  |  |
| --- | --- |
| **FEATURES​** | **CORRELATION​** |
| fips​ | -0.05160543​ |
| Surface Pressure (kPa)​, PS | -0.16924778​ |
| Temperature at 2M​ | 0.1070144​ |
| Maximum Temperature at 2M​, T2M\_MAX | 0.15054281​ |
| T2M\_RANGE​ | 0.24948646​ |
| Wind Speed Range at 10 Meters (m/s)​, WS10M\_RANGE | 0.11101822​ |
| Precipitation ​ | -0.07166571​ |
| Specific Humidity 2m​ | -0.06784445​ |
| T2MDEW​ | -0.06591331​ |
| T2MWET​ | -0.0651899​ |
| T2M\_MIN​ | 0.0716427​ |
| Earth Skin Temp (C)​ | 0.1175141​ |
| WS10M​ | 0.06639164​ |
| WS10M\_MAX​ | 0.08847621​ |
| WS10M\_MIN​ | 0.01140501​ |
| WS50M​ | 0.03063234​ |
| WS50M\_MAX​ | 0.0577941​ |
| WS50M\_MIN​ | -0.01900761​ |
| WS50M\_RANGE​ | 0.09391972​ |
| score​ | 1​ |

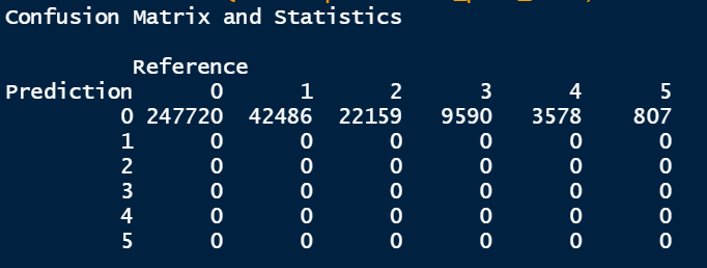
Using these features, we attempted to construct most of the models we learned about in this class. In the end we were able to successfully construct the following 4 types of models:



* Ordinal



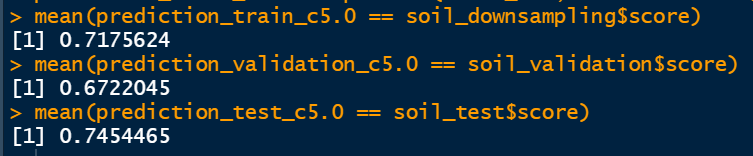


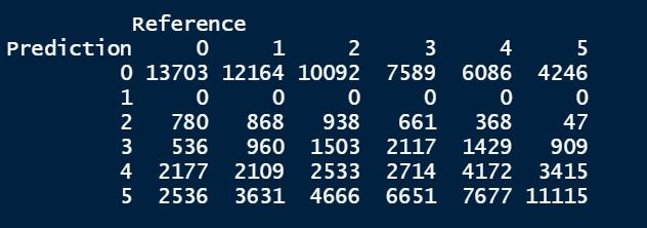


This model gives good accuracy, but on further analyzing the confusion matrix we can conclude that it underperformed

* C5.0

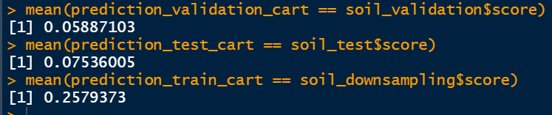


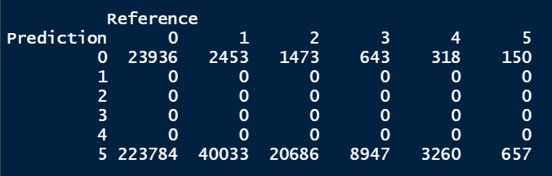


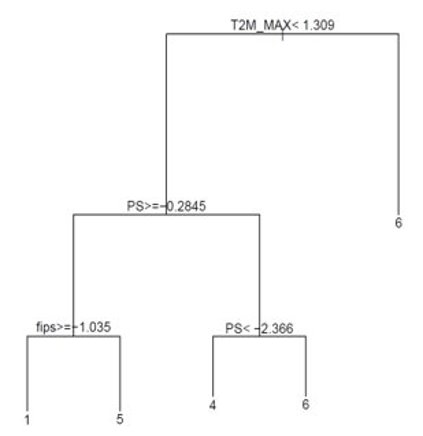


* CART



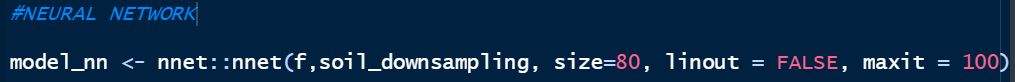


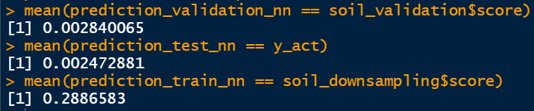


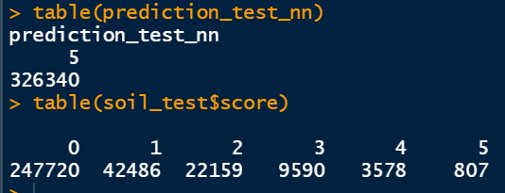


CART underperformed tremendously in both accuracy and confusion matrix. But it gave us an in site on the major features driving the model's decision making

* Neural Network



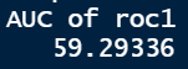
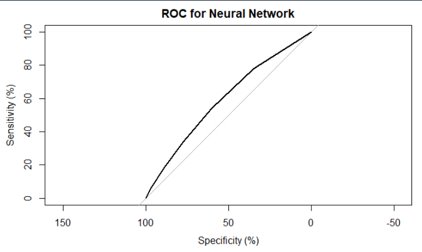


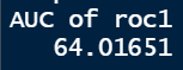
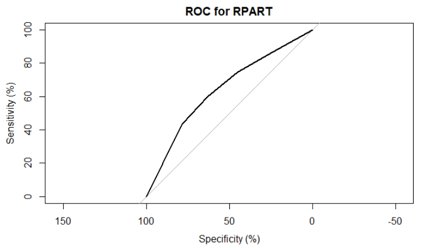


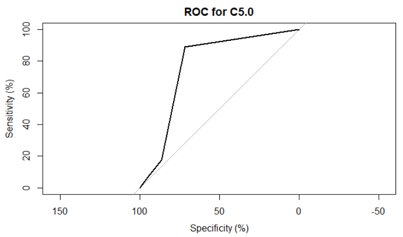
* Neural Network predicts everything in class 5 which show its underperformance despite its high power
* We tried running the model with maximum iteration>=1000, but the high computational load of this model crashed our systems

5. Results

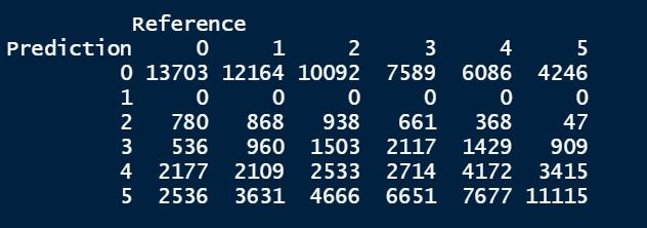
None of the models performed particularly well, but the C5.0 was the best.







Below is the confusion matrix for the C5.0 model.



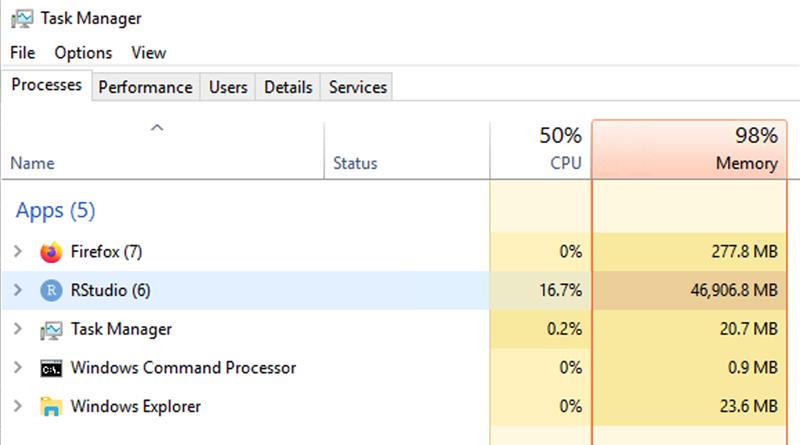
5a. Discussion of Results

Some of the things that we learned about this data set were:

* We discovered that there were no features that were meaningfully correlated with our desired output variable, drought. Therefore, no amount of data could have delivered a reliably accurate model.
* Of the features that we did eventually include in the model, drought score could be seen dependent on Surface Pressure, Temperature, Wind Speed Range.

Most of the lessons that we learned related to the modelling process, rather than the subject matter of drought prediction. Some of our the primary takeaways are:

* Data Preprocessing is a quintessential
* Down sampling led to better model evaluation
* We had difficulty using RStudio for some of our models. As the graphic below illustrates, the size of the data set and complexity of the models often crashed our systems.



We were never able to fully remediate this issue, but we did several things to account for it:

* + We continued to downsample the data as much as possible while still maintaining accuracy ratings
  + We used high-CPU virtual machines available to us through Penn State.

6. References

United States Drought Monitor

<https://droughtmonitor.unl.edu/>

Predict Droughts using Weather & Soil Data

Kaggle data set by Christopher Minixhofer

<https://www.kaggle.com/cdminix/us-drought-meteorological-data>

7. Appendix

