

# **Prediction of Droughts using Weather and Soil Data**

Sahana Pandurangi Raghavendra (2176508)

Pooja Tank(2176991)

Pooja Nadagouda (2173740)

submitted in partial fulfillment of the  
requirements for the completion of  
CSS 581 Introduction to Machine learning

Master of Science  
University of Washington

December 2021

Professor and Guide  
Dr. Muhammad Aurangzeb Ahmad

University of Washington

## **Abstract**

Global warming and climate changes have increased the probability of natural disasters like drought, hurricanes, floods etc. Natural disasters like droughts have intense adverse impacts on human and animal lives. Early prediction of droughts can help save lives and facilitate better resource allocation. Here in our project, we have considered common weather and soil parameters not specific to a geographical region and devised a machine learning approach where we apply LSTM, ARIMA and XGBoost machine learning models on multivariate time series data to predict droughts over the period of years 2012-2020. We observed that XGBoost and LSTM models provided promising results and ARIMA model can be used as a baseline model. This machine learning model can be used to predict droughts in any geographical region. Some of the challenges and future work in this problem is concept drift due to the ever-changing climatic conditions caused by global warming.

# TABLE OF CONTENTS

Chapter 1. Introduction .....	4
1.1 Importance of Drought Prediction .....	4
1.2 How is the problem being addressed? .....	4
1.3 What can one learn from the proposed solution and limitations .....	4
1.4 Machine learning problem .....	5
Chapter 2. Related work .....	6
Chapter 3. Data .....	7
3.1 Feature Description .....	7
3.1 Data Distribution And Exploration .....	9
3.1.1 Distribution of data for each of the features: .....	9
3.1.2 Relation of each feature with respect to target variable:.....	11
3.2 Feature Analysis and Data Transformation .....	13
3.2.1 Handling Missing values in the data .....	13
3.2.2 Preparing data for Multi-class Classification.....	14
3.3 FEATURE ENGINEERING: .....	14
3.3.1 RELATION WITH TARGET VARIABLE: .....	14
3.3.2 CORRELATION MATRIX: .....	14
3.4 Handling Data Imbalancing by converting to a binary classification problem .....	15
3.5 Time Series Data pre-processing .....	16
Chapter 4. Experiments.....	17

4.1	ARIMA Model – Baseline Model.....	17
4.1.1	Model Details:.....	18
4.1.2	Results:.....	19
4.2	XGBoost Model.....	19
4.2.1	Implementation details.....	19
4.2.2	Results.....	20
4.3	LSTM.....	20
4.3.1	Model Details.....	21
4.3.2	Results.....	23
4.4	Bidirectional LSTM.....	24
Bibliography .....		<b>Error! Bookmark not defined.</b>
Appendix A.....		<b>Error! Bookmark not defined.</b>

## LIST OF FIGURES

Figure 1 Category distribution provided in the data .....	5
Figure 2: Distribution of data into train, test and validation sets.....	7
Figure 3: Feature Descriptions.....	8
Figure 4: data distribution graph for each variable .....	11
Figure 5 Relation of each feature with respect to target variable .....	12
Figure 6 Correlation metrics for each variable with every other variable. ....	15
Figure 7 Data Distribution after converting to binary classification .....	16
Figure 8 preparing data to include past knowledge. ....	16
Figure 9 transformed data in left and residual graph on right.....	17
Figure 10: Arima model details .....	18
Figure 11: Arima model results .....	19
Figure 12: XGBoost model results and comparison of Multiclass classification, binary classification and XGBoost regression .....	20
Figure 13: Prediction results of XGBoost models .....	21
Figure 14: LSTM model summary .....	22
Figure 15: Training of LSTM model .....	23
Figure 16: Prediction and actual results mapping of LSTM model.....	23
Figure 17: bidirectional LSTM binary classification and multi-classification results.....	24
Figure 18: Comparison between results of all the models implemented .....	25

## Chapter 1. INTRODUCTION

Droughts are the second most costly weather events causing large amounts of economic and social losses. A drought can be defined as a period when certain geographical regions experience below-normal precipitation levels. This condition leads to reduced soil moisture or groundwater, diminished stream flow, crop damage, shortage of food and water. Drought conditions are known to last for weeks, months or years together.

### 1.1 IMPORTANCE OF DROUGHT PREDICTION

An estimated 55 million people globally are affected by droughts every year, and they are the most serious hazard to livestock and crops in nearly every part of the world. Drought threatens people's livelihoods, increases the risk of disease and death, and fuels mass migration. Water scarcity impacts 40% of the world's population, and as many as 700 million people are at-risk of being displaced because of drought by 2030. The prediction of droughts may provide important information for drought preparedness and farm irrigation. Thus, if these predictions are done well in advance, then it would facilitate timely resource allocation to minimize loss. Especially, save lives of plants and animals, reduce the impact of drought on human health that would in turn result in low morbidity and deaths due to droughts.

### 1.2 HOW IS THE PROBLEM BEING ADDRESSED?

The drought prediction machine learning model uses a dataset of 19M instances to train and test the model. The machine learning problem is a classification problem consisting of two classes being no drought and drought. We aim to solve the problem using time series data and build LSTM, XGBoost and ARIMA models to predict drought conditions.

### 1.3 WHAT CAN ONE LEARN FROM THE PROPOSED SOLUTION AND LIMITATIONS

We learn how the different soil and weather parameters irrespective of a geographical region affect drought conditions. The data is collected from regions all over US. This model can be extended to other geographical areas. The limitation of this approach is concept drift is not taken into account due to climatic changes. Another limitation is that more soil data can be included in the data to build the prediction model.

## 1.4 MACHINE LEARNING PROBLEM

The Machine Learning Problem is a multi-class classification problem with target variables divided into 6 categories which are None, D0, D1, D2, D3, D4. The encoding for these variables is as shown below in Fig 1.

Category	Description	Possible Impacts
D0	Abnormally Dry	Going into drought: <ul style="list-style-type: none"><li>■ short-term dryness slowing planting, growth of crops or pastures</li></ul> Coming out of drought: <ul style="list-style-type: none"><li>■ some lingering water deficits</li><li>■ pastures or crops not fully recovered</li></ul>
D1	Moderate Drought	<ul style="list-style-type: none"><li>■ Some damage to crops, pastures</li><li>■ Streams, reservoirs, or wells low, some water shortages developing or imminent</li><li>■ Voluntary water-use restrictions requested</li></ul>
D2	Severe Drought	<ul style="list-style-type: none"><li>■ Crop or pasture losses likely</li><li>■ Water shortages common</li><li>■ Water restrictions imposed</li></ul>
D3	Extreme Drought	<ul style="list-style-type: none"><li>■ Major crop/pasture losses</li><li>■ Widespread water shortages or restrictions</li></ul>
D4	Exceptional Drought	<ul style="list-style-type: none"><li>■ Exceptional and widespread crop/pasture losses</li><li>■ Shortages of water in reservoirs, streams, and wells creating water emergencies</li></ul>

Figure 1 Category distribution provided in the data

We aim to implement different classification models to predict drought conditions using a large-scale dataset of 19 million rows and 21 features, analyze them with different performance metrics like precision, recall, accuracy, and F-1 score, using the dataset that contains features common to different climatic or geographical conditions. Thus, the model can be extended to predict drought conditions in different climatic conditions. We also aim to reduce the number of false negatives in comparison to false positives as having a false alarm about the drought is better than no alarm when it is required. We lay more emphasis on optimizing the performance metric recall to get smaller number of false negatives.

## Chapter 2. RELATED WORK

The existing drought prediction models are mainly based on a single weather station. Efforts need to be taken to develop a new multi-station-based/generic prediction model to extend their applications to new heterogeneous geographical locations. There are severe drought area prediction models that have been implemented using random forest classifiers on satellite images and topography data. Some drought prediction models implement XGBoost machine learning algorithms and have provided promising results, while some models have focused on training data with specific meteorological data that are drought indicators like the standardized precipitation index (SPI) and the standardized soil moisture Index (SSMI).

Studies over the past century have shown that meteorological drought is never the result of a single cause. It is the consequence of complex individual interactions. [1] Some of the following may be components of a drought event like Global weather pattern, High Pressure, The tropical outlook, temperate zone outlook and too many other variables.. The various work has been done in this field, most of the work is done with the statistical predictions and not involving machine learning. The Global Integrated Drought Monitoring and Prediction System (GIDMaPS) [2] is one such tool that uses SPI values to predict droughts.

Some recent work shows the improvement in the accuracy and rmse score using the random forest prediction. Morteza et al [3] showed the comparison of drought prediction result using SPI, SPEI and random forest, and concluded that Random Forest gives the best results for areas in Iran. Felsche et al. [4] implemented explainable AI to understand the relation between various features and the drought condition. The paper concludes that The best-performing models obtain accuracies of 57 % for the Lisbon domain and 55 % for the Munich domain.



## Chapter 3. DATA

The dataset is provided by NASA Langley Research Center (LaRC) Power Project funded through the NASA Earth Science/Applied Science Program. [5] It consists of weather data and soil data with 21 features covering parameters like temperature, wind speed, humidity, precipitation, surface pressure, earth skin temperature, frost point, wet bulb temperature that contribute to indicating drought conditions. The training set consists of data collected from different regions of United States consisting of 19 million instances with temporal data from year 2000 to 2020 while the validation and testing set consists of temporal data from the years 2010 – 2011 and 2012 -2020 respectively. Table 1 represents the division of train, validation, and test sets across the dataset.

Split	Year Range (inclusive)	Percentage (approximate)
Train	2000 - 2009	47%
Validation	2010-2011	10%
Test	2012-2020	43%

Figure 2: Distribution of data into train, test and validation sets

### 3.1 FEATURE DESCRIPTION

The dataset includes various features related to weather and soil. Some of them are listed below.

ATTRIBUTE	DESCRIPTION
PRECTOT	Precipitation (mm day-1)
PS	Surface pressure (kPa)
TS	Earth skin Temperature (C)
QV2M	Specific Humidity at 2 meters (g/kg)
T2MDEW	Dew / Frost Point at 2 meters (C)
T2MWET	Wet Bulb temperature at 2 meters (C)

T2M	Temperature at 2 meters (C)
T2M_MIN	Minimum temperature at 2 meters (C)
T2M_MAX	Maximum temperature at 2 meters (C)
T2M_RANGE	Temperature range at 2 meters (C)
WS10M	Wind Speed at 10 Meters (m/s)
WS10M_MIN	Minimum Wind Speed at 10 Meters (m/s)
WS10M_MAX	Maximum 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_MIN	Minimum Wind Speed at 50 Meters (m/s)
WS50M_MAX	Maximum Wind Speed at 50 Meters (m/s)
WS50M_RANGE	Wind Speed range at 50 Meters (m/s)
DATE	Date on which the parameters are measured and collected

Figure 3: Feature Descriptions

The above set of attributes are studied in detail to understand their influence on droughts. We observe that these are generic attributes that are applicable to heterogeneous geographical locations. A detailed report on these attributes are as follows.

- *Precipitation (PRECTOT)*: To evaluate degree of drought conditions we can measure the duration of dry period and access the degree of dryness. Normal range is 2.5 mm – 7.6 mm/hour.
- *Skin temperature (TS)*: It is the physical temperature of the Earth's surface.
- *Surface pressure (PS)*: It is the pressure that the air exerts on the surface of the earth. High-pressure regions reduce evaporation and moisture in the atmosphere causing dryness.
- *Specific humidity at 2m (QV2M)*: It is defined as the mass of water vapour in a unit mass of moist air, usually expressed as grams of vapour per kilogram of air. The specific humidity is an extremely useful quantity in meteorology. Low humidity is an indication of drought.
- *Dew/Frost Point at 2 Meters (T2MDEW)*: It is the point when air cannot hold water in gas form. The dew point is the temperature at which the air is saturated with respect to water vapor over a liquid surface. The frost point is the temperature at which the air is saturated with respect to water vapor over an ice surface. The higher the dew point rises, greater the amount of moisture in the air.

- *Wet bulb temperature (T2MWET)*: It essentially measures how much water vapor the atmosphere can hold at current weather conditions. The wet-bulb temperature is the lowest temperature that can be reached under current ambient conditions by the evaporation of water only.
- *Temperature at 2 meters (T2M)*: High temperatures enhance evaporation, decreasing the surface water causing increased dryness. The regions with low precipitation are even more affected at high temperatures than when in low temperatures.
- *Wind Speed (WS10 / WS50M)*: Wind or the atmospheric circulations in large scales, control the ocean evaporation and the transfer of evaporated moisture thus indirectly controlling the degree of drought.

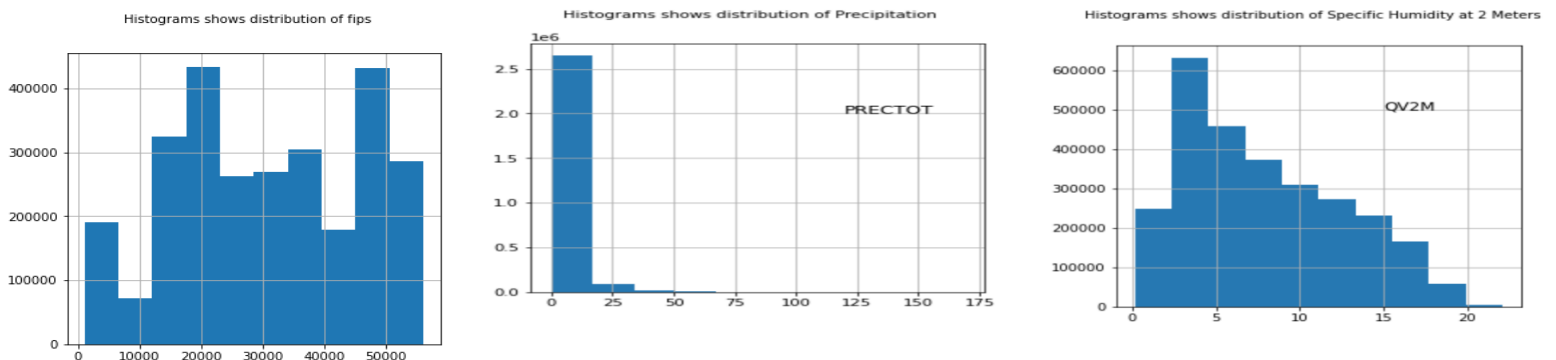
It can be noticed that all the above factors are not independently controlling drought, these factors are inter-related and influence each other thus influencing the drought conditions.

### 3.1 DATA DISTRIBUTION AND EXPLORATION

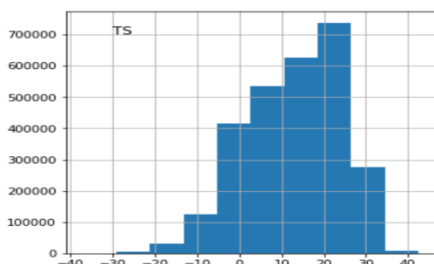
The attributes in the dataset follow different metrics. Hence, they must be analyzed to determine if they should be scaled to achieve uniformity in the data. Also, the outliers present in the data should be detected and eliminated. To achieve this, we need to understand the distribution of data in every attribute.

#### 3.1.1 Distribution of data for each of the features:

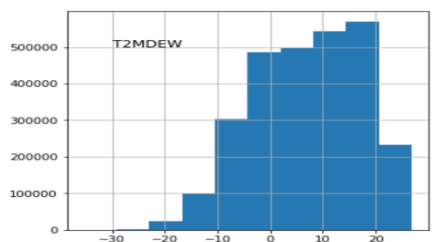
Histograms for all attributes are as follows:



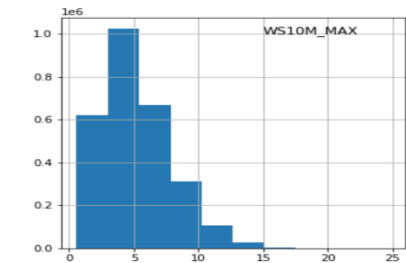
Histograms shows distribution of Earth Skin Temperature



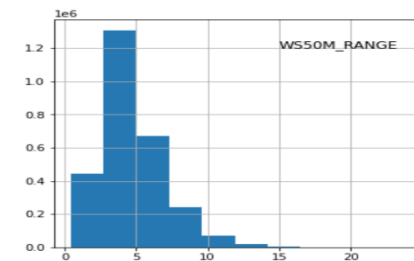
Histograms shows distribution of Dew/Frost Point at 2 Meters



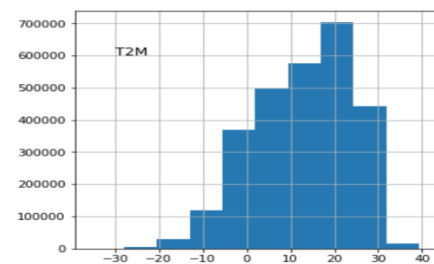
Histograms shows distribution of Maximum Wind Speed at 10 Meters



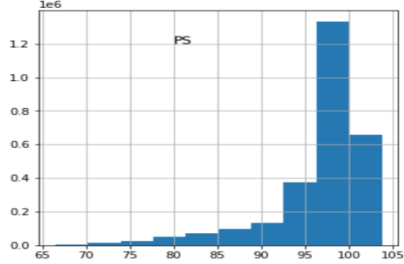
Histograms shows distribution of Wind Speed Range at 50 Meters



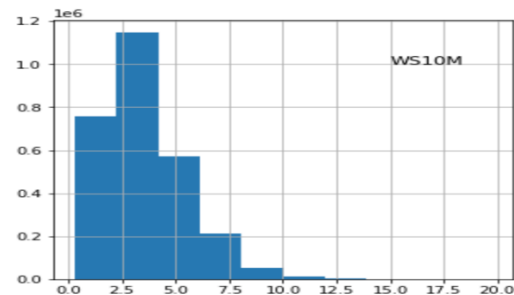
Histograms shows distribution of Temperature at 2 Meters



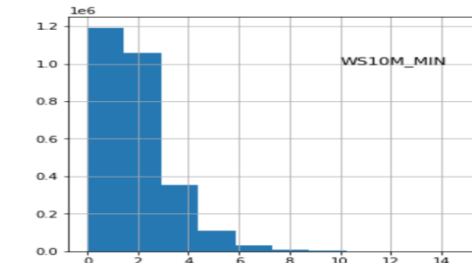
Histograms shows distribution of Surface Pressure



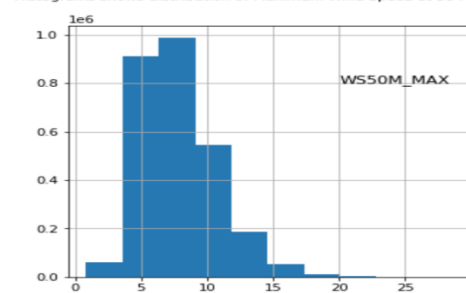
Histograms shows distribution of Wind Speed at 10 Meters



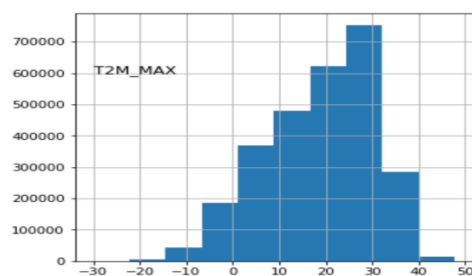
Histograms shows distribution of Minimum Wind Speed at 10 Meters



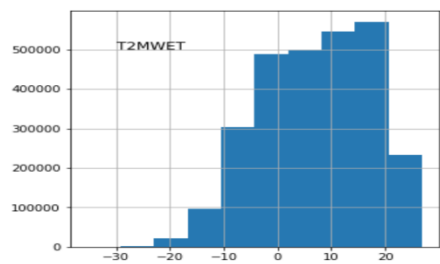
Histograms shows distribution of Maximum Wind Speed at 50 Meters



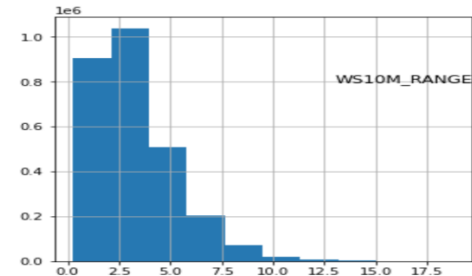
Histograms shows distribution of Maximum Temperature at 2 Meters



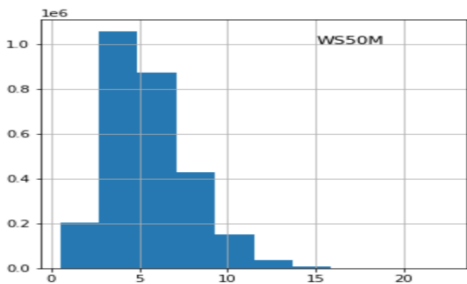
Histograms shows distribution of Wet Bulb Temperature at 2 Meters



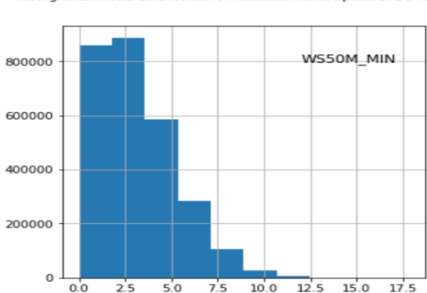
Histograms shows distribution of Wind Speed Range at 10 Meters



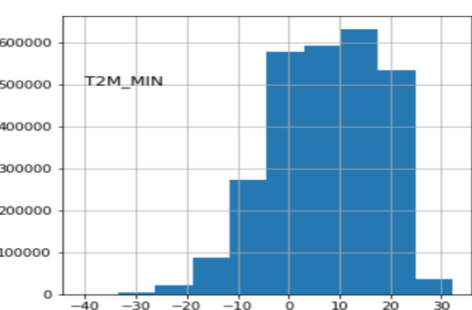
Histograms shows distribution of Wind Speed at 50 Meters



Histograms shows distribution of Minimum Wind Speed at 50 Meters



Histograms shows distribution of Minimum Temperature at 2 Meters



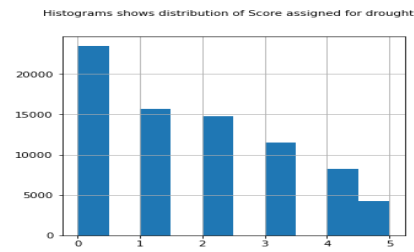


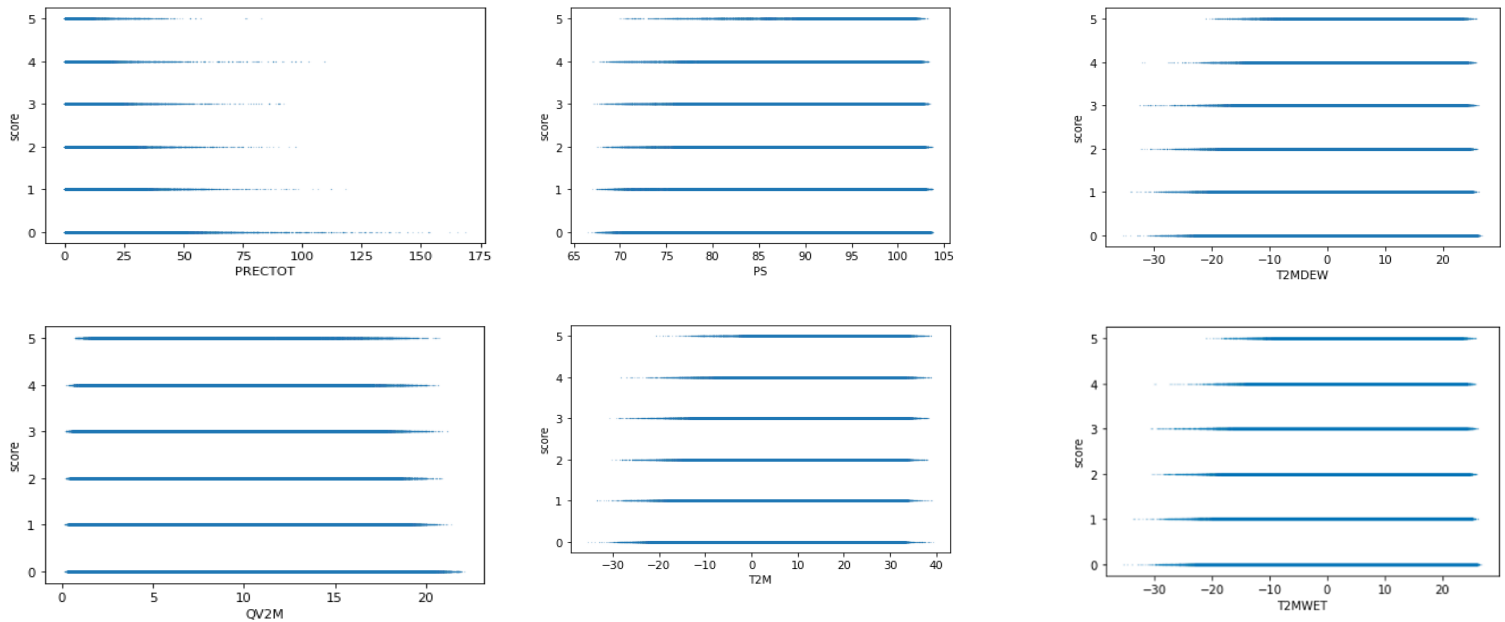
Figure 4: data distribution graph for each variable

### 3.1.1.1 Inferences from the above histograms

For the target variable the data is imbalanced as most of the data maps to no drought condition, hence we need to handle imbalance data so that the model isn't biased. All other attributes have uniform distribution, hence no scaling required.

### 3.1.2 *Relation of each feature with respect to target variable:*

In order to understand the influence of each attribute on the target variable we have done a detailed data exploration on the relationship of each feature with respect to target feature and drawn conclusions based on the data represented by these scatter plots.



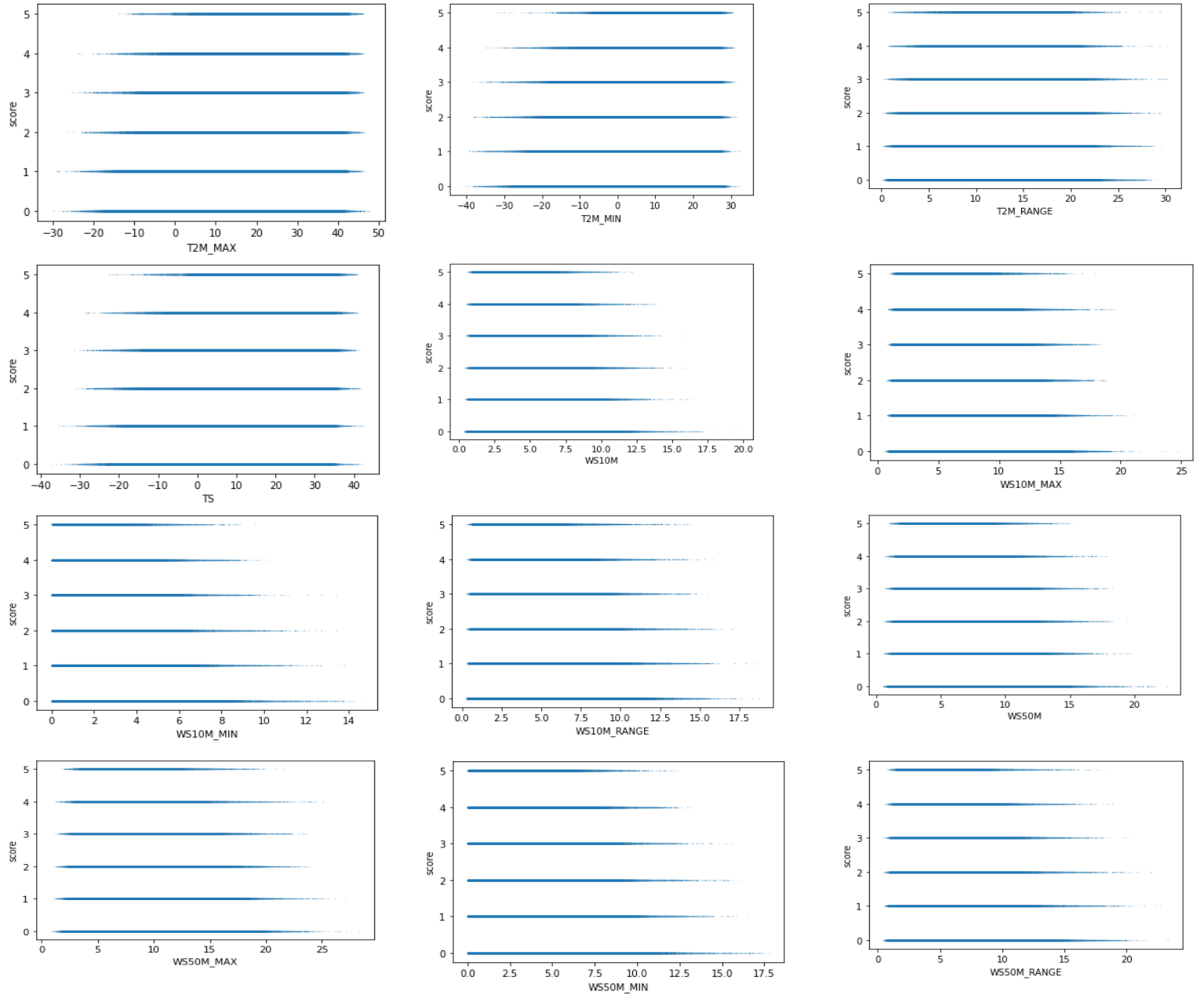


Figure 5 Relation of each feature with respect to target variable

### 3.1.2.1 Inferences drawn from data exploration:

- Precipitation (mm day-1): A precipitation of below 100 millimeter per day is observed in drought hit areas.
- Severe drought conditions are observed for geographical areas with precipitation of below 50 mm per day.
- Severe drought conditions are not observed when maximum wind speed at 10 Meters (m/s) is above 20 m/s
- Severe drought conditions are not observed when maximum wind speed at 50 Meters (m/s) is above 22 m/s

- Severe drought condition is not observed when minimum temperature at 2 Meters from earth's surface is between 25C to 30C
- All the parameters like precipitation, wind speed, temperature at 2 meters from earth's surface, humidity together contributes as a drought indicators at a specific geographical location.
- Humidity alone at a specific location doesn't play a significant role in indicating droughts.

## 3.2 FEATURE ANALYSIS AND DATA TRANSFORMATION

This section provides the details on exploring the data and using various techniques used to handle missingness in the data, validating the data for correctness, exploring the relationship between various features and target variable, detecting outliers, normalizing and transforming the attributes and handling imbalanced classes in the features.

### 3.2.1 *Handling Missing values in the data*

The training dataset has 19300680 instances and includes 21 features. On further analysis, the target variable “**score**” has 16543884 has null values. This is because the drought scores are available weekly while the meteorological data points are available daily. We use the following techniques to handle missing data in the target variable “**score**”. Below are some methods enlisted to handle missing values in the dataset.

#### 3.2.1.1 Copying same values for 1 week

The data is interpreted as that for the entire week, the drought condition remains same for the area and hence that value is used for each day of that week.

#### 3.2.1.2 Eliminating the rows without target value:

It is assumed that the other feature values remain constant for the entire week and only one day's values are taken as the representative data of the entire week. All the other data points of that week are ignored.

#### 3.2.1.3 Averaging 7 rows to 1 row

We take the average of all the feature values of the week and keep only one row with the target value specified in that week.

#### 3.2.1.4 Used KNN imputation for missing values in the target variable:

For the target column which has the missing values, KNN method is used to check the 7 nearest values of the drought condition and use that as a value for that day.

### 3.2.2 *Preparing data for Multi-class Classification*

The target variable score has continuous values. For a multi-class classification problem, we need to convert them to integer values. We achieve this by using the `ceil()` function over the range of continuous values. `Ceil()` and `floor()` function over the target variable provide the same data distribution for the score variable. We have chosen the `ceil()` value here.

## 3.3 FEATURE ENGINEERING:

Selecting feature using following methods.


1. Analyzing the graph plots to understand relation of all other variables with target variable.
2. Using correlation matrix.
3. Using scikit tool `SelecKBest`.

### 3.3.1 *RELATION WITH TARGET VARIABLE:*

By analyzing relation of all attributes with target variable using graph plots we could notice that following attributes influence target variable more in comparison to other attributes.

['PRECTOT', 'WS50M\_RANGE', 'PS', 'T2M', 'QV2M', 'TS']

### 3.3.2 *CORRELATION MATRIX:*

```
 # Correlation matrix plot  
corealtion_matrix = data.corr()  
corealtion_matrix
```



	fips	PRECTOT	PS	QV2M	T2M	T2MDEW	T2MWET	T2M_MAX	T2M_MIN	T2M_RANGE	TS	WS10M	WS10M_MAX	WS10M_MIN	WS10M_RANGE	WS50M	WS50M_MAX	WS50M_MIN	WS50M_RANGE	score
fips	1.000000	-0.017089	-0.035969	-0.055478	-0.048457	-0.050700	-0.050886	-0.050303	-0.046373	-0.022537	-0.044336	0.046493	0.041321	0.035296	0.031267	0.051473	0.053376	0.036962	0.034202	-0.030001
PRECTOT	-0.017089	1.000000	0.067633	0.249676	0.100949	0.234136	0.234132	0.036296	0.151195	-0.294651	0.097151	0.063153	0.072625	0.034942	0.073766	0.079093	0.087746	0.064506	0.052563	-0.064059
PS	-0.035969	0.067633	1.000000	0.278020	0.163697	0.338053	0.338135	0.112778	0.206536	-0.220336	0.163470	-0.076628	-0.131815	0.026716	-0.196463	-0.039484	-0.086798	0.038759	-0.152273	-0.185663
QV2M	-0.055478	0.249676	0.278020	1.000000	0.872572	0.959707	0.960798	0.807513	0.907267	-0.068247	0.864949	-0.220085	-0.248944	-0.112763	-0.257725	-0.203760	-0.243984	-0.088115	-0.235367	-0.049752
T2M	-0.048457	0.100949	0.163697	0.872572	1.000000	0.915029	0.915711	0.983352	0.981779	0.241837	0.997505	-0.206279	-0.216074	-0.132319	-0.199782	-0.194210	-0.203542	-0.121098	-0.149828	0.088532
T2MDEW	-0.050700	0.234136	0.338053	0.959707	0.915029	1.000000	0.999968	0.857073	0.940375	-0.012822	0.906728	-0.233570	-0.261717	-0.120140	-0.269843	-0.202789	-0.239970	-0.089294	-0.228925	-0.056618
T2MWET	-0.050886	0.234132	0.338135	0.960798	0.915711	0.999968	1.000000	0.857745	0.941063	-0.012703	0.907454	-0.233197	-0.261304	-0.120055	-0.269344	-0.202672	-0.239790	-0.089339	-0.228644	-0.055894
T2M_MAX	-0.050303	0.036296	0.112778	0.807513	0.983352	0.857073	0.857745	1.000000	0.938152	0.405033	0.980068	-0.215848	-0.218225	-0.149195	-0.190969	-0.197144	-0.193713	-0.141922	-0.116516	0.126252
T2M_MIN	-0.046373	0.151195	0.206536	0.907267	0.981779	0.940375	0.941063	0.938152	1.000000	0.063431	0.979385	-0.204707	-0.221708	-0.119467	-0.216308	-0.198914	-0.222920	-0.104897	-0.191203	0.057611
T2M_RANGE	-0.022537	-0.294651	-0.220336	-0.068247	0.241837	-0.012822	-0.012703	0.405033	0.063431	1.000000	0.238746	-0.081594	-0.043549	-0.114567	0.020758	-0.042976	0.030309	-0.132080	0.169069	0.211284
TS	-0.044336	0.097151	0.163470	0.864949	0.997505	0.906728	0.907454	0.980068	0.979385	0.238746	1.000000	-0.188707	-0.199017	-0.117590	-0.186984	-0.182234	-0.190985	-0.111342	-0.142687	0.096073
WS10M	0.046493	0.063153	-0.076628	-0.220085	-0.206279	-0.233570	-0.233197	-0.215848	-0.204707	-0.081594	-0.188707	1.000000	0.953171	0.836524	0.705776	0.967086	0.912131	0.800819	0.419047	0.021216
WS10M_MAX	0.041321	0.072625	-0.131815	-0.248944	-0.216074	-0.261717	-0.261304	-0.218225	-0.221708	-0.043549	-0.199017	0.953171	1.000000	0.695674	0.866766	0.911921	0.948009	0.667891	0.596283	0.040860
WS10M_MIN	0.035296	0.034942	0.026716	-0.112763	-0.132319	-0.120140	-0.120055	-0.149195	-0.119467	-0.114567	-0.117590	0.836524	0.695674	1.000000	0.244735	0.842330	0.675113	0.944406	-0.033666	-0.014314
WS10M_RANGE	0.031267	0.073766	-0.196463	-0.257725	-0.199782	-0.269843	-0.269344	-0.190969	-0.216308	0.020758	-0.186984	0.705776	0.866766	0.244735	1.000000	0.646070	0.810867	0.245828	0.828197	0.064752
WS50M	0.051473	0.079093	-0.039484	-0.203760	-0.194210	-0.202789	-0.202672	-0.197144	-0.198914	-0.042976	-0.182234	0.967086	0.911921	0.842330	0.646070	1.000000	0.920635	0.852307	0.379915	-0.002276
WS50M_MAX	0.053376	0.087746	-0.086798	-0.243984	-0.203542	-0.239970	-0.239790	-0.193713	-0.222920	0.030309	-0.190985	0.912131	0.948009	0.675113	0.810867	0.920635	1.000000	0.656790	0.675647	0.017255
WS50M_MIN	0.036962	0.064506	0.038759	-0.088115	-0.121098	-0.089294	-0.089339	-0.141922	-0.104897	-0.132080	-0.111342	0.800819	0.667891	0.944406	0.245828	0.852307	0.656790	1.000000	-0.112161	-0.030854
WS50M_RANGE	0.034202	0.052563	-0.152273	-0.235367	-0.149828	-0.228925	-0.228644	-0.116516	-0.191203	0.169069	-0.142687	0.419047	0.596283	-0.033666	0.828197	0.379915	0.675647	-0.112161	1.000000	0.052290
score	-0.030001	-0.064059	-0.185663	-0.049752	0.088532	-0.056618	-0.055894	0.126252	0.057611	0.211284	0.096073	0.021216	0.040860	-0.014314	0.064752	-0.002276	0.017255	-0.030854	0.052290	1.000000
year	-0.000000	0.008689	-0.001730	0.030846	0.004122	0.020940	0.021078	-0.002222	0.009161	-0.030598	0.003190	-0.003879	-0.001718	-0.005572	0.001549	-0.003286	-0.001018	-0.004119	0.002684	-0.033382
month	-0.000000	0.010845	0.003252	0.147109	0.191573	0.166770	0.166709	0.187757	0.206410	-0.003875	0.186045	-0.100490	-0.109028	-0.054855	-0.109076	-0.092587	-0.107580	-0.051076	-0.091830	0.030991
week	0.000000	0.011884	0.003082	0.147178	0.190377	0.166927	0.166853	0.186281	0.205537	-0.005825	0.184823	-0.098968	-0.108374	-0.051919	-0.110232	-0.090822	-0.106810	-0.047854	-0.093966	0.032850
day	0.000000	0.008743	-0.003097	0.007402	0.007226	0.008889	0.008850	0.005171	0.008619	-0.007857	0.007054	0.002162	0.002462	0.000618	0.002895	0.001942	0.003098	0.001532	0.002586	0.003335

Figure 6 Correlation metrics for each variable with every other variable.

### 3.4 HANDLING DATA IMBALANCING BY CONVERTING TO A BINARY

#### CLASSIFICATION PROBLEM

In order to handle data imbalance and make it easier to work on model building for multivariate time series data we converted the multi-class classification problem to a binary classification problem. The categories D0, D1, D2, D3, D4 consisted of drought conditions representing abnormally dry, moderate, severe, extreme, and exceptional drought and none class indicating no drought. This dataset had a huge data imbalance in the target class. Thus, we combined the drought categories into one class and no drought categories into another class. Thus, making the problem a binary classification problem. After combining the drought classes, we observe that there are 1,652,230 and 1,104,566 values for no drought conditions in the target variable.

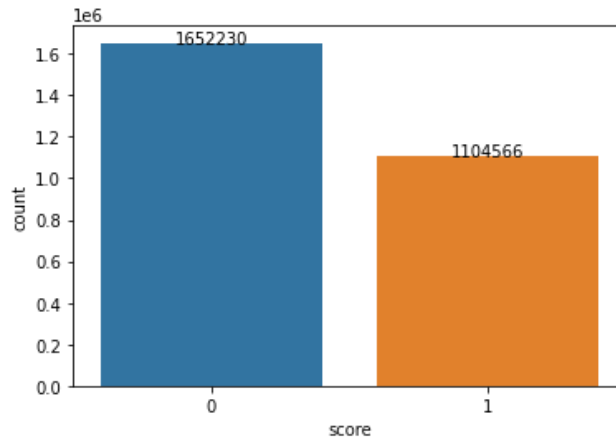


Figure 7 Data Distribution after converting to binary classification

### 3.5 TIME SERIES DATA PRE-PROCESSING

Given a sequence of numbers for a time series dataset, we can restructure the data to look like a supervised learning problem. We can do this by using previous time steps as input variables and use the next time step as the output variable. The Idea is shown by example in Figure 4.1. Here, we use the time step before 180 days of the current time step as an input feature.

date	X1	X2	Y		date	p_X1	p_X2	p_Y	X1	X2	Y
1	10	4	0	➔	1	NaN	NaN	NaN	10	4	0
2	12	5	1		2	NaN	NaN	NaN	12	5	1
...	...	...	...		...	...	...	...	...	...	...
180	15	8	3		180	10	4	0	15	8	3
181	20	9	4		181	12	5	1	20	9	4

Figure 8 preparing data to include past knowledge.

## Chapter 4. EXPERIMENTS AND RESULTS

This section explains the details and results of the models built to predict drought conditions using the weather and soil data, the experiment setup, models descriptions and results are shown below.

### 4.1 EXPERIMENTAL SETUP

Before we started model building, we cleaned the dataset and dropped all the null values in the target variable and stored it in a csv file that consisted of 2M rows. We used the Azure Machine Learning Cloud platform for implementing the Machine learning model. The platform provided collaborative notebooks and good scaling capabilities to run the ML models on huge data set of 2M rows. However, a lot of errors specific to Azure were encountered while running tensor flow libraries. Since we didn't have much time in hand, we switched to working on google collab which could efficiently run the models.

### 4.2 ARIMA MODEL – BASELINE MODEL

A time series is a sequence where a metrics are recorded over regular time intervals. Here the weather and soil parameters needed to predict drought are recorded on weekly basis. Thus, we have prepared our time series data to be monthly for ARIMA model as a week is a small duration of time for drought conditions to change while a year is extremely a huge duration of time. It must be noted that drought conditions can also occur during some seasons of the year.

ARIMA is short for 'Auto Regressive Integrated Moving Average' is actually a class of models that 'explains' a given time series based on its own past values, that is, its own lags and the lagged forecast errors, so that equation can be used to forecast future values. An ARIMA model is characterized by 3 terms/parameters: p, d, q. 'p' is the order of the 'Auto Regressive' (AR) term. It refers to the number of lags to be used as predictors. And 'q' is the order of the 'Moving Average' (MA) term. It refers to the number of lagged forecast errors that should go into the ARIMA Model. d is the number of differencing required to make the time series stationary

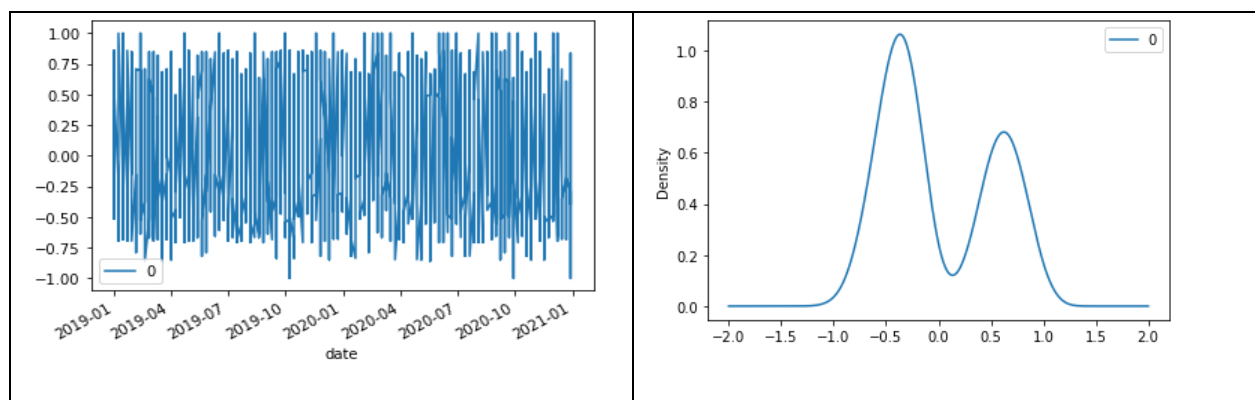


Figure 9 transformed data in left and residual graph on right

#### 4.2.1 Model Details:

```
count 198.000000
mean -0.004243
std 0.529266
min -0.999762
25% -0.432490
50% -0.151576
75% 0.493182
max 1.000819
```

##### ARIMA Model Results

Dep. Variable:	D.score	No.	Observations:	1722
Model:	ARIMA(5, 1, 0)		Log Likelihood	-1346.886
Method:	css-mle		S.D. of innovations	0.529
Date:	Sat, 18 Dec 2021		AIC	2707.773
Time:	03:24:04		BIC	2745.932
Sample:	1		HQIC	2721.890

	coef	std err	z	P> z	[0.025	0.975]
const	0.0002	0.004	0.046	0.964	-0.007	0.007
ar.L1.D.score	-0.8188	0.024	-34.747	0.000	-0.865	-0.773
ar.L2.D.score	-0.6785	0.029	-23.082	0.000	-0.736	-0.621
ar.L3.D.score	-0.5267	0.031	-16.907	0.000	-0.588	-0.466
ar.L4.D.score	-0.3742	0.029	-12.713	0.000	-0.432	-0.317
ar.L5.D.score	-0.2110	0.024	-8.940	0.000	-0.257	-0.165

Roots	Real	Imaginary	Modulus	Frequency
AR.1	0.5628	-1.1881j	1.3147	-0.1796
AR.2	0.5628	+1.1881j	1.3147	0.1796
AR.3	-1.4205	-0.0000j	1.4205	-0.5000
AR.4	-0.7395	-1.1764j	1.3895	-0.3393
AR.5	-0.7395	+1.1764j	1.3895	0.3393

Figure 10: Arima model details

#### 4.2.2 Results:

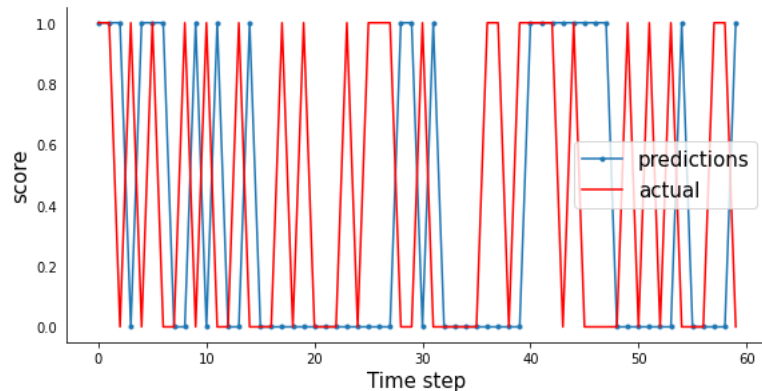


Figure 11: Arima model results

#### Performance metrics

RMSE value: 0.708

MAE value : 0.50

Accuracy: 0.49

Recall: 0.49

Precision: 0.47

F1 score: 0.48

### 4.3 XGBOOST MODEL

XGBoost stands for “Extreme Gradient Boosting”, where the term “Gradient Boosting” originates from the paper Greedy Function Approximation: A Gradient Boosting Machine, by Friedman. XGBoost is used for supervised learning problems, where we use the training data (with multiple features) to predict a target variable.

XGBoost can also be used for time series forecasting, although it requires that the time series dataset be transformed into a supervised learning problem first. It also requires the use of a specialized technique for evaluating the model called walk-forward validation, as evaluating the model using k-fold cross validation would result in optimistically biased results.

#### 4.3.1 Implementation details

##### 4.3.1.1 XGBoost Classifier with Multiclass Classification

The model has been trained using 10,000 randomly picked instances, with `n_estimator` value of 100, booster value is selected as ‘gbtree’, 0.3 learning rate, and ‘reg\_lambda’ which is L2 regularization as lambda value.

#### 4.3.1.2 XGBoost Classifier with Binary classification

Converting the problem into binary classification helped to improve the results, moreover, XGBoost classifier with Binary classification provided the best accuracy score among all the models. The recall, precision and F1 Score is also better than other classification models.

#### 4.3.1.3 XGBoost Regression

To convert the project into regression, we built a model by taking continuous values in the class variable as in the original data, the multi-label target variable can be used by taking average value of the labels available for a particular row. We then forecast the future value for drought predictions and took the round value of predicted result to convert the problem in classification and evaluated the model.

### 4.3.2 Results

The model is tested on 1000 instances, with and without inducing past knowledge. The model without past knowledge gives a performance accuracy of 29%. After inducing the past knowledge, the model performs slightly better with the accuracy of 56%. XGboost Classifier with binary classification shows the best performance. The table below shows the comparison of binary classification and multi class classification with XGBoost classifier and XGBoost Regression model results.

	MultiClass Classification	Binary Classification	Regression Problem
Accuracy	0.5672	0.664	0.552
Precision (with weighted average)	0.5672	0.664	0.552
Recall (with weighted average)	0.5289	0.6384	0.5317
F1-Score (with weighted average)	0.5260	0.5937	0.5334
MAE	0.5894	0.3669	0.3579
RMSE	1.1698	1.0996	0.9953

Figure 12: XGBoost model results and comparison of Multiclass classification, binary classification and XGBoost regression

The prediction results are shown in the figure below.

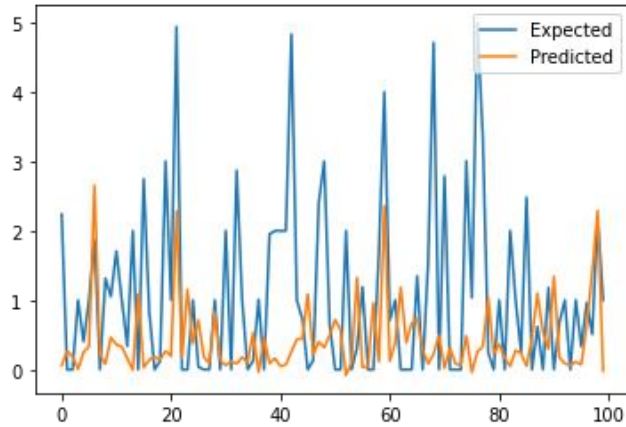


Figure 13: Prediction results of XGBoost models

## 4.4 LSTM

For drought prediction using multivariate time series forecasting, we have used the popular Long Short-Term Memory (LSTM) which is a type of Recurrent Neural Network(RNN) for model building. We have used the Keras API integrated with TensorFlow deep learning library.

*Reformatting/Reshaping the data:* Here the data is split and reformatted to provide time series dependency between predictor and target attributes i.e., each drought score depends on previous 14 values of each predictor attribute.

*Scaling:* Data is scaled so that all attributes have data in the same ranges. With a generalized dataset, each attribute has proportional influence on the target variable.

### 4.4.1 Model Details

We have used a 'sequential' model. It involves defining and adding layers to the model one by one in a linear manner, from input to output. Defining the layers of the model and configuring each layer with a number of nodes, activation function, and connecting the layers together into a cohesive model. In the first layer that input shape is defined by passing the argument. In the case of this LSTM model, we have used 14 X 21 as we use 14 previous instances, each having 21 predictor attributes.

- *Dropout:* It is a regularization technique to reduce overfitting of the training dataset making the model robust. This is done by dropping out some number of layer outputs. Ex: adding an argument 0.3 in Dropout() function implies 30% of inputs will be dropped in each update.



- *Dense layer:* It is nothing but a regular fully-connected Neural Network layer, which is used for bringing down the output dimensionality. Here adding a Dense Layer with one neuron in the output produces a single value that corresponds to drought score.

Model: "sequential"		
Layer (type)	Output Shape	Param #
lstm (LSTM)	(None, 14, 128)	76288
leaky_re_lu (LeakyReLU)	(None, 14, 128)	0
lstm_1 (LSTM)	(None, 14, 128)	131584
leaky_re_lu_1 (LeakyReLU)	(None, 14, 128)	0
dropout (Dropout)	(None, 14, 128)	0
lstm_2 (LSTM)	(None, 64)	49408
dropout_1 (Dropout)	(None, 64)	0
dense (Dense)	(None, 1)	65
Total params: 257,345		
Trainable params: 257,345		
Non-trainable params: 0		

Figure 14: LSTM model summery

*Compiling the model:* The model is compiled, to optimize loss function using various algorithmic optimization procedures. Also choosing performance metrics which are tracked during the training process. In this model we have used optimization procedure stochastic gradient descent(SGD), to optimize loss 'binary\_crossentropy'. We have chosen performance metrics Root mean squared error (RMSE) which is tracked during training the model

- *Class BinaryCrossentropy:* Computes the cross-entropy loss between true labels and predicted labels.
- *Class SGD:* Gradient descent (with momentum) optimizer.

*Early stopping:* Model is underfit if its trained very little, also the model may overfit due to too much training. Both these cases result in poor model. Hence to avoid both these cases we have used early stopping. This monitors the loss on the training dataset and stops the training process as soon as there are signs of overfitting.



*Fitting the model:* Fitting the model with training configurations like epochs - number of loops required to train the dataset and batch size - number of samples passed in each loop. In our model we have used epochs = 50 and the batch size = 32

```
Epoch 1/50
2757/2757 [=====] - 1404s 507ms/step - loss: 0.4500 - root_mean_squared_error: 0.2492 - accuracy: 0.5991 - val_loss: 0.3553 - val_root_mean_squared_error: 0.1972 - val_accuracy: 0.6822
Epoch 2/50
2757/2757 [=====] - 1806s 655ms/step - loss: 0.4479 - root_mean_squared_error: 0.2482 - accuracy: 0.5993 - val_loss: 0.3549 - val_root_mean_squared_error: 0.1969 - val_accuracy: 0.6822
Epoch 3/50
2757/2757 [=====] - 10591s 4s/step - loss: 0.4476 - root_mean_squared_error: 0.2480 - accuracy: 0.5993 - val_loss: 0.3532 - val_root_mean_squared_error: 0.1957 - val_accuracy: 0.6822
Epoch 4/50
2757/2757 [=====] - 9355s 3s/step - loss: 0.4474 - root_mean_squared_error: 0.2479 - accuracy: 0.5993 - val_loss: 0.3537 - val_root_mean_squared_error: 0.1960 - val_accuracy: 0.6822
Epoch 5/50
2757/2757 [=====] - 3314s 1s/step - loss: 0.4473 - root_mean_squared_error: 0.2478 - accuracy: 0.5993 - val_loss: 0.3542 - val_root_mean_squared_error: 0.1964 - val_accuracy: 0.6822
```

Figure 15: Training of LSTM model

#### 4.4.2 Results

Model is evaluated on the validation dataset. Here we have used data from year 2010 to 2011 for evaluating the model performance

*Using the model for Predictions:* The trained model is now used for making predictions of drought score. Here we have used data from 2012 to 2020 to make predictions.

```
accuracy 0.6322511843984236
f1 score 0.6335786859549086
precision score 0.6349330596059473
recall score 0.6322511843984236
AUC&ROC 0.5009367352214011
MAE 0.36774881560157635
RMSE 0.6064229675742636
```

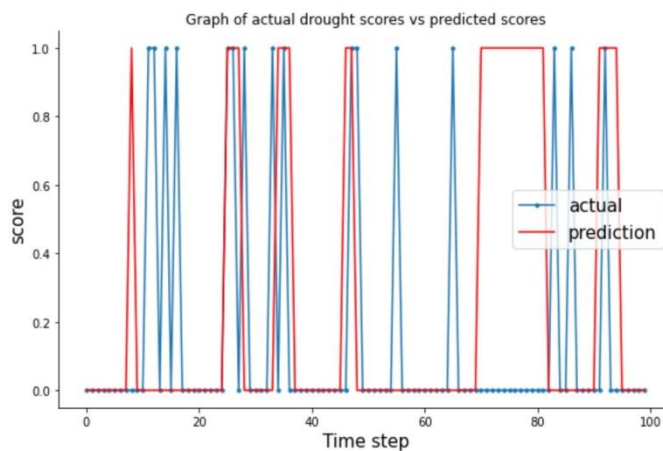


Figure 16: Prediction and actual results mapping of LSTM model

## 4.5 BIDIRECTIONAL LSTM

Bidirectional LSTMs are an extension of traditional LSTMs that can improve model performance on sequence classification problems. In problems where all timesteps of the input sequence are available, Bidirectional LSTMs train two instead of one LSTMs on the input sequence. The first on the input sequence as-is and the second on a reversed copy of the input sequence. This can provide additional context to the network and result in faster and even fuller learning on the problem.

### 4.5.1 *Implementation details*

Our implementation of Bidirectional LSTM is same as LSTM with the bidirectional layers. It can be referred from the section 4.3.1. We implemented bidirectional LSTM for multiclass as well as binary class classification.

### 4.5.2 *Results*

The bidirectional LSTM so far provided the best RMSE and MAE score among all the models, also the prediction is better than any other model if the problem is treated as a regression problem.

The table below shows the results of Bidirectional LSTM model.

<b>Results</b>	<b>MultiClass classification</b>	<b>Binary Classification</b>
Accuracy	0.5421	0.542
Precision (with weighted average)	0.321	0.395
Recall (with weighted average)	0.417	<b>0.86</b>
F1-Score (with weighted average)	0.253	0.426
MAE	<b>0.1936</b>	0.574
RMSE	<b>0.2449</b>	0.757

Figure 17: bidirectional LSTM binary classification and multi-classification results

## 4.6 COMPARISONS

The comparison of the results obtained using all the models is shown below.

MODELS	ARIM A	XGBOOST MULTI- CLASS	XGBOOST BINARY	XGBOOST REGR.	LSTM CLASSI.	LSTM REGR.	BI- LSTM MULTI CLASSI.	BI-LSTM BINARY CLASSI.
<b>ACCURACY</b>	0.49	0.5672	<b>0.664</b>	0.552	0.552	0.552	0.5421	0.542
<b>PRECISION</b>	0.47	0.5672	<b>0.664</b>	0.552	0.520	0.552	0.321	0.395
<b>RECALL</b>	0.49	0.5289	0.6384	0.5317	0.55	0.5317	0.417	<b>0.86</b>
<b>F1-SCORE</b>	0.48	0.5260	0.5937	0.5334	0.522	0.5334	0.253	0.426
<b>MAE</b>	0.50	0.5894	0.336	0.6579	0.67	0.6579	<b>0.1936</b>	0.574
<b>RMSE</b>	0.708	1.1698	1.0947	0.9986	0.6695	0.9953	<b>0.2449</b>	0.757

Figure 18: Comparison between results of all the models implemented

The table above provides comparison between all the models implemented. Almost all models give the similar accuracy between 49% to 56%. XGBoost Classifier with binary classification provided the best accuracy of 66%. The bidirectional LSTM provided best recall of 86% for binary classification however the accuracy is still low for the model. Overall Bi-LSTM works best if project is treated as a regression model. Moreover, for the classification problem, XGBoost provides best result. The baseline model ARIMA gives comparable results to LSTM that shows the robustness of the model without deep learning network.

## Chapter 5. CHALLENGES AND FUTURE WORK

From understanding the problem to running a machine learning model, the process required to overcome a plethora of challenges. This section states the challenges faced during different phases of the project. And the further work that can be done to improve the project quality.

### 5.1 CHALLENGES FACED

#### *More Data doesn't help:*

The problem provided some of the insights on machine learning models and the dataset being used by the models. It was observed that even though training a model on just 10% of the data and training it on entire dataset does not improve the accuracy.

#### *Domain knowledge*

It is important that one understands a problem thoroughly before building a machine learning problem. Having domain knowledge is important as it helps in deciding which features are important, which features might not create an impact. Is the model performing as it should perform based on human common sense. All these tasks requires to have a deep understanding of each and every feature of the model and knowing about the field of the problem.

#### *Dealing with the temporal data*

Making prediction on time series data is not as straightforward as working on normal data. In the normal machine learning classification problem only the features contribute to the prediction results, in temporal data, the time or a season and the previous data also needs to be included as a feature in order to have accurate predictions. Some of the ways to overcome this challenge is discussed in section 8.1

#### *Imbalanced data set*

The combination of weather and soil data used here is completely imbalanced and the consequences of that can be seen in section 8.2.2. The XGBoost Model is struggling to make predictions of the classes for which the dataset does not have enough instances. One of the ways to handle imbalanced data is to use under-sampling. For this problem we have selected randomly picked 0.6M instances of class 0 out of 1.2M instances.

#### *Huge Dataset*

The training dataset has over 19,000,000 instances. And it includes 21 features. To process this huge amount of data, high RAM capacity is required. One solution that can be used is to process the data in smaller chunks. The pseudocode used to process the data in patches of 2M rows is shown in Figure. We also used under-sampling methods and removed the instances which contained null values in the target variable to reduce the data size.

#### *Handling missing data:*

As described in the section 3 we have used approaches like copying same values for 1 week, eliminating the rows without target value, averaging 7 rows to 1 row and KNN imputation for missing values in the target variable.

## 5.2 FURTHER WORK

### *More domain knowledge Required:*

More research can be carried out to understand how the different soil and weather parameters irrespective of a geographical region affect drought conditions.

### *Data from Various locations other than US:*

Though generic weather and soil parameters not specific to a region is used to build the model but the model is built on data is collected from regions all over US. Thus it must be tested against the data from geographical areas other than US.

### *Consideration of Climate Change:*

The challenge of concept drift must be addressed as the drought conditions can be impacted by climatic changes.

### *More Advanced Models:*

The dataset can be trained using more advanced methods and deep learning models to get better values of performance.

## Chapter 6. CONCLUSION

Drought is one of the natural disasters that impacts human lives in the long term, it impacts not only human lives but also the entire life cycle of that area. It impacts agriculture sector, climate condition, wildlife, nature, and many other factors, which at the end results in government spending lots of money to maintain the quality of life in the area. Predicting drought before hand can help everyone save lives and cost. If the drought condition is predicted in the earlier stages, it helps to take steps to reverse the drought condition and in the worse case maintain or prepare for the drought.

It helped to learn that how the different soil and weather parameters irrespective of a geographical region affect drought conditions. The data is collected from regions all over US. These models can be extended to other geographical areas. The limitation of this approach is concept drift that is not considered due to climatic changes. Another limitation is that more soil data can be included in the data to build the prediction model.

This report shows the prediction of drought condition using machine learning and deep learning techniques with the help of weather and soil data. It shows the use of timeseries data with imputing past knowledge can help the models predict drought in a better accuracy. Though the prediction is still not as good as it should be to be used in a real world. The XGBoost model with binary classification and Bidirectional LSTM with binary classification provides good score among the other models such as ARIMA, and LSTM. Almost all models gives the similar accuracy between 49% to 56%. XGBoost Classifier with binary classification provided the best accuracy of 66%. Th bidirectional LSTM provided best recall of 86% for binary classification however the accuracy is still low for the model. Overall Bi-LSTM works best if project is treated as a regression model. Moreover, for the classification problem, XGBoost provides best result. The baseline model ARIMA gives comparable results to LSTM that shows the robustness of the model without deep learning network.

It can be said that if the problem is treated as a regression problem, LSTM based models provides good rmse score and if treated as a classification problem, XGBoost provided better results.

## Chapter 7. BIBLIOGRAPHY

- [1] N. D. M. Center, "predicting Drought," 2021. [Online]. Available: <https://drought.unl.edu/Education/DroughtIn-depth/Predicting.aspx>.
- [2] Z. A. A. N. N. e. a. Hao, "Global integrated drought monitoring and prediction system.," <https://doi.org/10.1038/sdata.2014.1>, 2014.
- [3] H. E.-G. A. A. Morteza Lotfirad, "Drought monitoring and prediction using SPI, SPEI, and random forest model in various climates of Iran," *Journal of Water and Climate Change* ;, 2021.
- [4] R. L. Elizaveta Felsche, *Applying machine learning for drought prediction in a perfect model framework using data from a large ensemble of climate simulations*, vol. 21, Department of Geography, Ludwig Maximilian University of Munich, Munich, Germany, 2021.
- [5] C. Minixhofer, "Predict Droughts using Weather & Soil Data," 2021. [Online]. Available: <https://www.kaggle.com/cdminix/us-drought-meteorological-data/code>.