Machine Learning Flooding Prediction Project

Joey Binz
Clemson University
CPSC 6300 Applied Data Science
Spring 2022
School of Computing
jbinz@clemson.edu
April 24, 2022

1.0 INTRODUCTION

This project sought to determine if data driven machine learning could predict flooding in certain areas of the world given the correct data. In an age of the world where sea level rise and global warming are becoming more and more prevalent the ability to predict flooding could prove invaluable. Knowing flooding in advance allows people to safely evacuate, businesses to move merchandise and inventory away from danger, and property owners to prepare their properties for flooding ahead of time. This project serves as a proof of concept for machine learning based flood prediction.

There are two primary data sources for this project. The flood (gauge height) data was recorded daily by a U.S. Geological Survey station in Charleston SC based off of the Cooper River [3]. Precipitation data was recorded by NOAA and was recorded as daily totals from Charleston International Airport [2]. In total there was 2.22 MB of numerical data (about 35 years worth of data) in the dataset used.

2.0 SUMMARY OF EXPLORATORY DATA ANALYSIS

The data set (as stated in the introduction) comes from two primary sources. The combined data set contains water gauge height (ft) and precipitation totals (in) recorded daily dating back to October of 1986 and ending in March of 2022. The completed dataset has 12949 observations (in both columns) with 382 unique gauge height observations and 296 unique precipitation observations.

Gauge height (the output variable) predictors for this project were hypothesized to be precipitation and the previous week's recordings for precipitation and gauge height. **Figure 1** shows the precipitation plotted with gauge height across the data set. **Figures 2 & 3** show the seasonal decomposition of both gauge height and precipitation to identify trends over time.

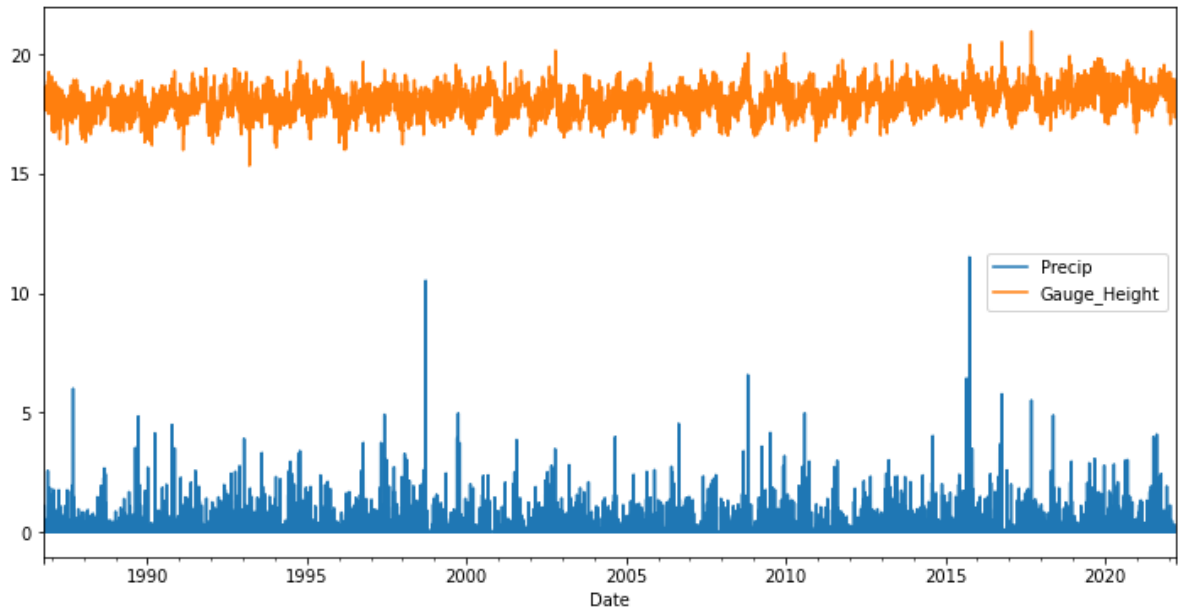

Figure 1. Precipitation & gauge height data plotted against time

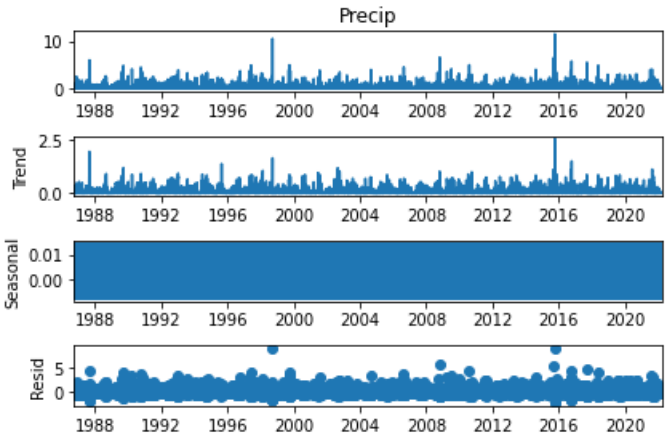

Figure 2. Precipitation Seasonal Decomposition

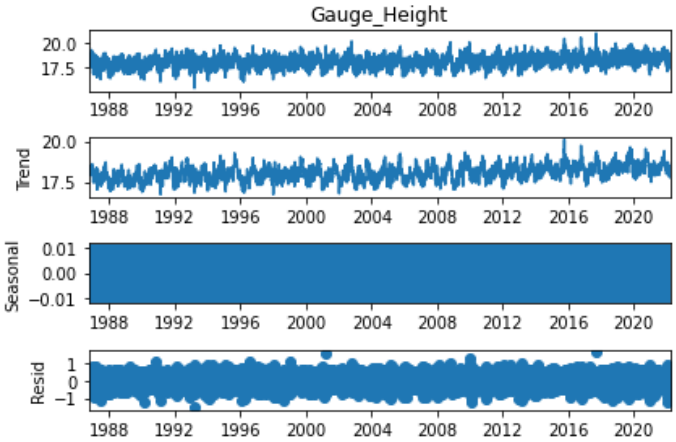

Figure 3. Gauge Height Seasonal Decomposition

Cleaning the data involved removing a small amount of null gauge height values (surprisingly there were no null values with the precipitation data), getting rid of a few garbage values, combining the two data sets into one on a datetime index, scaling the data with a MinMaxScaler, and taking a few extra steps to prepare the data for an LSTM.

In order to fully take advantage of an LSTM model (the details of which are outlined in section 3 of this report) the data was modified to contain three columns: two columns containing the precipitation and gauge height for the previous day (t-1) and one column for the gauge height of

the current day (t). This allowed the model to train on the previous day's gauge height and precipitation data as input and output the next day's gauge height [1].

	var1(t-1)	var2(t-1)	var2(t)
1	0.000000	0.440497	0.433393
2	0.000000	0.433393	0.447602
3	0.000000	0.447602	0.440497
4	0.000000	0.440497	0.422735
5	0.000000	0.422735	0.495560
...	...	...	...

Figure 5. Formatted LSTM data

It is worth noting that in the early stages of this project data obtained from USGS stations near Pickens, SC was used to implement the first model of this project before realizing the data was unusable. The implementation of the LSTM model was done only with the new data obtained from Charleston SC which spans a much longer time and has far more entries.

3.0 SUMMARY OF MACHINE LEARNING MODELS

This project implemented and evaluated two machine learning models, a basic linear regression model and a Time Series Long Short Term Memory Recurrent Neural Network model (LSTM). The linear regression model (as mentioned in section 2) was implemented with a set of data containing precipitation and gauge height from Pickens, SC while the LSTM was implemented with the data from Charleston, SC.

3.1 LINEAR REGRESSION MODEL

The linear regression model was chosen as an entry point for the project largely because of its simplicity and the hypothesis that rainfall data would be highly correlated to flood data. The specific linear regression model used was OLS from the statsmodels.api module. The Least Squares method was used with the OLS model. Rainfall (inches) was used as the input for the model, and river discharge maximum (cubic feet per second) was used as the output for the model. The model's error rates are reported below:

Test Accuracy: 28.96%

R-Squared: 0.272
F-Statistic: 937.1

There seems to be some underlying issues with the model chosen and the data causing poor accuracy. The linear regression model likely fails to account for the normal streamflow fluctuations on days where there is no rainfall. Additionally the sample size of data was only about 2500 observations due to corrupted data. For future implementations of this project more observations would be ideal.

3.2 TIME SERIES LSTM MODEL

A time series LSTM (Long short term memory) Recurrent Neural Network was chosen for the second iteration of the project. The model takes patterns over time in the data into consideration when making predictions. When making predictions the model takes in a previous week of flood data/gauge height (ft) & rainfall data (inches) and outputs predictions for future gauge height levels. The model's error rate is reported below:

RMSE: 0.304

The time series LSTM model appears to fit the data quite well. A RMSE of 0.304 suggests that predictions made are within 0.304 ft of the test data.

3.3 MODEL COMPARISON

The LSTM model is a much better fit for this data than the previously tried linear regression model. The LSTM is able to pick up on both long & short term patterns present in the dataset & apply them to make much more accurate predictions. This logically makes sense, since there are many other factors that come into play to determine flooding (such as tides, previous rainfall, rainfall of surrounding areas, or weather phenomenon) than just precipitation.

3.4 PREDICTIONS

A few useful/impactful predictions were run with the model to show how it could be useful.

Scenario 1:

A tropical storm has just hit Charleston causing increased flooding from the storm surge & increased precipitation on the day of the storm. Given the previous week's rain & flood data predict the coming week's river levels after the storm has passed to see how long flooding will last.

Input:

Precip	Gauge_Height
0.14	18.61
0.41	18.39
0.00	18.50
0.00	18.72
0.00	19.26
0.02	19.63
5.51	20.96

Output (Gauge Height ft):

```
[[18.60398 ]
 [18.401196]
 [18.513252]
 [18.706606]
 [19.190554]
 [19.528522]
 [20.528883]]
```

Scenario 2:

Even in a normal week with little to no rainfall it would be useful to ship captains & deckhands to know the predicted height levels of the river for navigational purposes. Given the previous week, predict the upcoming week's water levels.

Input:

Precip	Gauge_Height
0.00	17.38
0.00	17.28
0.00	17.39
0.00	16.94
0.01	17.83
0.93	16.85
0.00	15.33

Output (Gauge Height ft):

```
[[17.565973]
 [17.484573]
 [17.574142]
 [17.211918]
 [17.938492]
 [17.109343]
 [16.011488]]
```

Scenario 3:

Heavy rainfall has hit the Charleston area in the past few days. Predictions on the river's water levels could greatly help riverside businesses know weather to begin to sandbag doors & prepare for the worst in the coming week. Given the previous week's data, predict the next week's flooding.

Input:

Precip	Gauge_Height
0.00	19.55
0.00	19.51
0.00	19.35
0.41	19.08
1.37	19.14
1.61	19.55
11.50	20.20

Output: (Gauge Height ft)

```
[[19.455662]
 [19.418886]
 [19.272451]
 [19.010565]
 [19.024414]
 [19.385881]
 [19.567827]]
```

4.0 SUMMARY AND CONCLUSION

This project set out to determine if there were ways to predict flooding using data driven algorithms. This goal was achieved through the implementation of a time series LSTM and with flood and rain data from Charleston, SC. Given the results of this project data driven models appear to be effective at predicting flooding and water levels given the proper data to train with.

The results of this project should serve as a proof of concept to any experts in the field of or adjacent to hydrology that data driven time series machine learning is a viable tool to help influence impactful decisions.

If more time and resources were available this project could be scaled up by utilizing data from more locations. The original aspirations for this project were to have it use global flood data, however due to issues in obtaining the data this had to be scaled back. Implementing a graph based LSTM network was another potential improvement to give the model a higher performance. Additionally accounting for more factors that can impact flooding such as winds, flooding of surrounding areas, or the future weather forecast could have been added to create a more complex (and hopefully more accurate) model.

5.0 REFERENCES

[1] Brownlee, J. (2020, October 20). Multivariate time series forecasting with lstms in Keras. Machine Learning Mastery. Retrieved April 24, 2022, from https://machinelearningmastery.com/multivariate-time-series-forecasting-lstms-keras/

[2] National Centers for Environmental Information (NCEI). (n.d.). Daily summaries station details. Daily Summaries Station Details: CHARLESTON INTL. AIRPORT, SC US, GHCND:USW00013880 | Climate Data Online (CDO) | National Climatic Data Center (NCDC). Retrieved April 24, 2022, from https://www.ncdc.noaa.gov/cdo-web/datasets/GHCND/stations/GHCND:USW00013880/detail

[3] United States Geographical Survey. (n.d.). USGS surface-water daily data for the nation. Retrieved April 24, 2022, from https://waterdata.usgs.gov/nwis/dv?referred_module=sw&search_criteria=site_tp_cd&submitted_form=introduction