Elements that Affect Precipitation

Junghoon Kang

Precipitation is any product of water that falls under gravitational pull from clouds. Our goal in this project is to find out which factors determine the probability of rain. This prediction might help to figure out which factors have most relate to precipitation.

National Oceanic and Atmospheric Administration(NOAA) provide climate information for all States and we can find source in website NOAA:www.ncei.noaa.gov /access/search/data-search/global-summary-of-the-day. I used the last ten years (2010-2019) weather data observed at Madison Dane county airport. The dataset analyzed 3652 daily cases, which include two days of leap day (2012,2016). For our project, I used seven variables (1) Average Dew Point, (2) Maximum Temperature, (3) Minimum Temperature, (4) Precipitation, (5) Average Station Pressure, (6) Average Temperature,(7) Average Wind Speed. To easily determine whether it rained that day or not, I set "Rain" as True if precipitation over 0 and False if precipitation is 0.

First, in **Figure1**, we can see Madison's monthly amount of precipitation and the number of precipitations per month. From this table, we can notice that rainfall is relatively high in April-October, while the number of precipitations falls about 12 times per month on average regardless of the season. Therefore, it is considered that the indicators that are not affected by the season affect probability of precipitation, and the indicators affected by the season have the potential to determine the amount of precipitation.

For this project, I used principal component analysis(PCA) with the seven variables (exclude Precipitation) to find the dimensionality of the data. **Figure2** shows the plots that cumulative explained variance ratio by different of the number of components. Before scaling the data (blue line), the first component explained variance ratio was 80% which is high. So for this project, I used Standard Scaler, and after scaling the data(orange line), the first component variance ratio was 50%. However, from the plot, four components explained variance ratio was over 99%, so we can reduce dimension to 4 without losing information.

I had split my data into 75% train data and 25% test data. In the sklearn pipeline, I used (1)StandardScaler and (2)LogisticRegression. Through this model, it achieves about 74% of accuracy score, recall is about 56%, and precision is about 71%. Overall, it can't say it is a perfect model but still can be classified as a meaningful model.

Figure 3 shows the plot of the coefficient weight of our model. From the plot, we can conclude Average Temperature(TEMP) and Average Dew Point(DEWP) are the most important factor that determines precipitation and Minimum Temperature(MIN) and Average Wind Speed(WDSP) were relatively importance than other than left three variables. However, we can improve our model by excluding variable weight close to 0 or find another weather index that might can improve our model.

Based on statistical analysis, we identified factors that affect precipitation. We can find more theoretically accurate correlations through meteorology, but with these statistical analyses and observational data, without specific scientific knowledge, we can find meaningful correlations. We have identified that temperature, dew point, and wind speed affect precipitation. Using weather information and three elements will help us intuitively determine whether rain or not.

Figure 1a: Monthly: Amount of precipitations

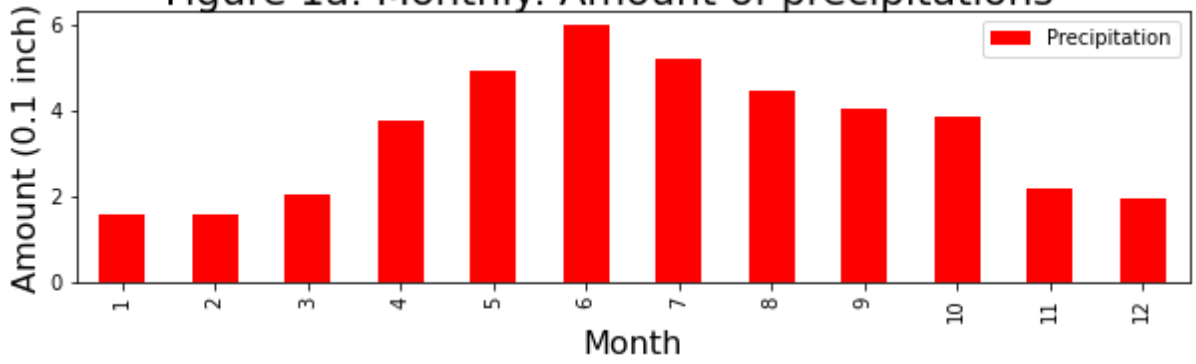

Figure 1b: Monthly: Number of times precipitations

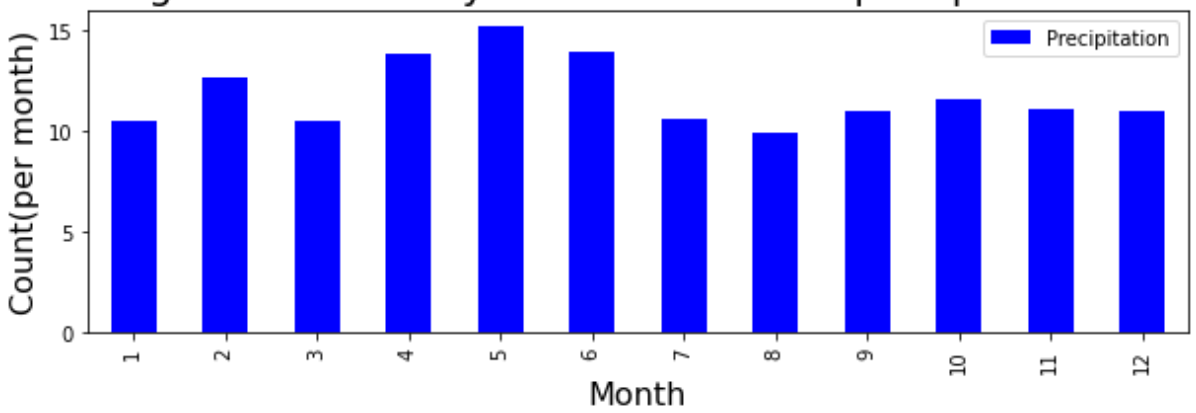

Figure 2: Principal Components of Breaks

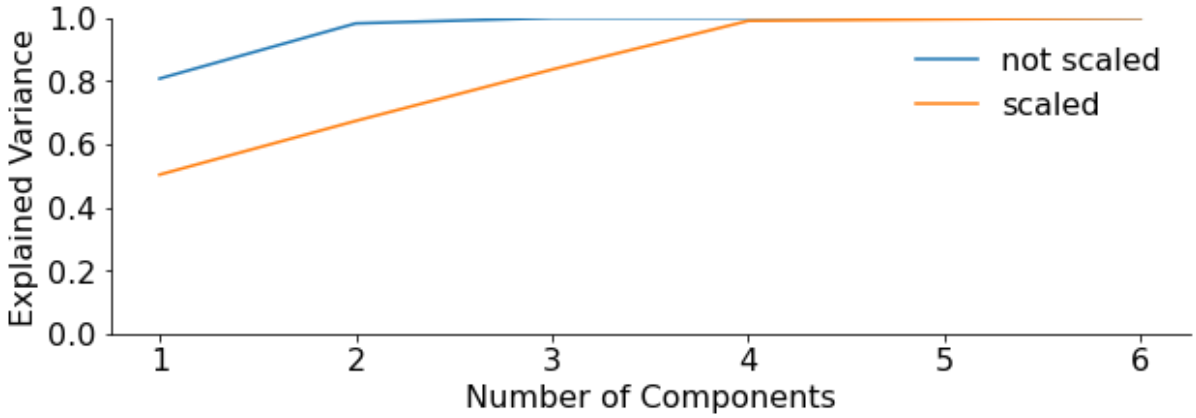

Figure 3: Logistic Regression Coefficients