Subarna Joshi
CS 334
HW5

I did not have any non-discriminating features while building my first model. So, I did not remove any feature from my training dataset. I was able to add some data which helped to bump the ratio of positive and negative training data to 3:4. I tried adding new feature to my training data i.e. Wind Gust which helped a bit to improve the accuracy of the model. I guess, it could have helped a bit more, but a lot of data was missing in the Wind Gust feature and I had to use inputing method to complete the data set. I assumed the missing Wind Gust data was similar to previous day and the next day and used those data to fill the missing data. I also rescaled my data. I tried normalizing the data using Min-max scalar at first because most of my feature data did not have normal distribution and I also tried using the standardize the data using Standard Scaler, but I got the better prediction accuracy using Quantile Transformer Scaler. I do not have any categorical feature in my data so, I did not use one-hot-encoding. I used Repeated Random Test-Train Splits to test my model. I was able to improve my model accuracy by 3.5%, recall by 9%, precision by 5%, and F1 score by 7%. Using other testing methods also have similar outcome.

The accuracy of my model was about 88% in my first homework and I was only able to increase the accuracy of my model to 90.58%. I added a new feature i.e. wind Gust to improve the model but it didn't improve the model accuracy significantly. It might be because about 1:6 of the data in this column was missing and was filled using assumption that the data might be similar to previous and next day data. One of the other reasons, can be imbalanced data because currently the data set that I currently using is in ratio of 3:4. I could not use other feature such as Wind Chill and Heat Index which may improve the model because majority of data in these feature are missing in my dataset and It is not ideal to use inputing method in such case. The dataset that I used is not big enough as it only contains about 1200 data. I added few data which were generally the data of January, November and December because Boise have more rainfall during these months. These data also helped to improve the imbalance data. I guess, the accuracy of this model could have been increased if I was able to use more training dataset but I was limited by the number of data that I could download per day from that website that I used to get these data.

My model was already about 88% accurate in my first homework which is not bad for weather prediction model and to increase its accuracy significantly is very difficult.