**HW5: Improving your Model: CS 334 Algorithms of Machine Learning:**

* We initially run the logistic regression by holding back 25% of data set for test set. This gives an ***AUC of 54.4% and 89.13% accuracy***
* Next we tried to convert all the metrics to log scale for temperature and wind speed and we tried to bucket the wind speed and max temperature into low, medium and high categories
  + Adding multiple metrics like snow fall did not change the accuracy or AUC

Doing this also didn’t improve the AUC and accuracy

* For the same dataset we tried to update the classification of rain definition by reducing it to >0.01 precipitation
  + Doing above doesn’t improve the AUC (***53.7%),*** but it reduced the accuracy down to **70%**
* For the above dataset, now we added rainy days in Nampa and removed the snow days in bogus basin, as it was not giving us better accuracy
  + Doing above improved the AUC to ***67%*** and brought back the accuracy close to ***80%***
  + This has also improved our precision value to 10/25=***40%,*** which was previously ***12.5%***
* Using different validation methodologies, below were the accuracy scores:
  + K fold validation: 70.73%
  + Stratified k fold validation: 75.62%
  + LOCOV: 78.63%
  + Repeated random test-train splits: 79.27%
* **Key observations:**
  + Relaxing the precipitation criteria in combination with adding Nampa rainfall as a factor is biggest factor in improving the AUC score
  + Bucketing the numerical categories into low, medium and high categories didn’t change the AUC and accuracy by >3%. But it started populating data into the True negative category and hence I continued to use this option
  + Also transforming the temperature data into log-scale didn’t help much in the data set I have chosen. But can definitely be helpful if the dataset changes. So I continued to use this method for validation.
* **Conclusion:**
  + The model began with AUC of 50%, accuracy of 89% and Specificity of 0%. By adding multiple factors, I was able to change the AUC to 67%(13% improvement), accuracy got reduced to 80%(9% loss), while the Specificity improved from 0% to 40%
  + Repeated random test\_train splits provides best accuracy of 79.27%

