Machine learning models are difficult to improve. It's an art involving cutting noise, transforming data, and deriving new data from existing data or using new data. As machine learning models become better they become exponentially more difficult to improve. I could not improve my model by 5% as I preprocessed the data and cleaned it up for homework 4, have identified likely the near max number of features from my data that could improve the model and identified all the features that didn't work from these attempts to perform feature engineering.

For preprocessing in homework 4 I immediately realized that there were many features that were going to be useless. This included features such as sunrise and sunset that had no recurring patterns that were going to be useful for predicting the rain, however I did try this and it seemed to make the model worse rather than better. This also included removing data such as conditions, description, and icon that either gave away the answer for when it's going to rain or just didn't relate to when it was going to rain at all such as the station. Furthermore, I planned ahead to remove data with mostly missing features such as severe risk. I believe this is part of the reason that I was not able to increase my model by 5%. I already put thought into cleaning up and processing the data on the last assignment.

The other reason why I believe that I was not able to increase accuracy by 5% is because I already used most of the useful data for homework 4 so it was already highly accurate. Therefore, it became exponentially harder to increase the accuracy. My homework 4 accuracy was already 91.9% and I was only able to get it up to 95.1%. It was already hard enough from homework 4 to raise the accuracy but it became even more difficult as I was able to successfully raise it even more. First I was able to raise the data from 91.9% to 93.5% by normalizing the data so that sea level pressure was not too valued in the overall model. From here adding seasons to the features from the datetime data and removing noise in the dataset I was able to raise the accuracy from 93.5% to 95.1%. The noise I removed included the 'feelslike' data and repetitive features such as 'cloud cover' when 'visibility' already existed and 'uv index'/'solar radiation' when 'solar energy' already existed and essentially measured the same thing. I also removed 'dew' as this was related to humidity and 'moonphase' as this has no effect on the rain which was proven when removing it did not affect the accuracy.

Finally, I reached the point in which there wasn't really much else I could do to raise the accuracy. The noise was already removed from the data by thinking through what was repetitive and what was useless for predicting rain. I already normalized the data and created as many new useful features as I could. I even tried adding features such as the hour that the sunset and sunrise occurred or how many times it has already rained in the week, month, etc. However, this was all proven useless and none of this worked or made the model worse. I could have added more data about the weather in visiting areas but in reality I do not believe this would have helped. This is mostly because I proved that I had all the features I needed to predict rain by methodically removing features and monitoring changes in accuracy. Removing all those features that either improved the accuracy or did not. The biggest reason I believe this wouldn't have helped though is because my area under the curve metric is at 96.9%. This is essentially the highest my area under the curve can be without overfitting or being too good of a model to realistically work on new data. Yes the model likely could have been improved a little more but since the area under the curve was already so high this proved that any other features would have had to be incredibly creative and useful as it becomes exponentially harder to improve a model the better it is.