BM Mini Project 2 Executive Summary

Objectives

For this project, our purpose was to find the model that has the best predictive power on whether or not a flight will be delayed by more than 15 minutes given new data similar to the "Delays" data set.

Methodology

In order to be able to achieve this goal of finding the best model, we first had to determine what kinds of models could achieve our goal of classifying whether or not a flight would be delayed. We decided to focus on logistic regression, random forest, decision tree, mlp, keras, and gradient boosting models to start with because these are the models we were most familiar with and because we know their predictive power is pretty strong while still having simple code.

After cleaning and training the data, we were able to fit the data using each model, and after this trial and error, we concluded that the random forest had the best predictive power of the three we originally limited ourselves to testing. This was because it had the highest accuracy of predicting of the three. This matched what we originally thought which was that the random forest model would likely have the best predictive power, but it was nice to see it confirmed compared to decision trees and logistic regression.

Results

As mentioned above, our final model that we decided upon was the random forest model. The hyperparameters we used for this model were 100 for the number of estimators and a max depth of 10. We decided these by using the grid search function in python to optimize our hyperparameters.

As for the model's performance metrics, it had a training accuracy of 0.818121 and a test accuracy of 0.817437. Considering the model, this is a pretty solid prediction accuracy overall. From there we can look at the variables that have the most impact on the model or are the most important. For our model, the top three most impactful variables were precipitation, the segment number a plane was on that day, and wind. An honorable mention goes to whether or not a plane departed from the Chicago O'Hare airport for being in the top 20 most impactful variables when no other airport was.

When you think about it logically, these variables impacting delays makes sense. If there is too much snow or precipitation in general, it is more likely for flights to be delayed. If a plane is making too many stops, by the end of the day, it might be delayed. Chicago usually has weather problems, so flights are likely to be delayed. All three of our top variables have the correlation of if they are high, the probability of flights being delayed is high.

Conclusion

In conclusion, we were able to successfully choose a model that can accurately predict 82% of the time whether or not a flight will be delayed by more than fifteen minutes if given similar data. This model could probably be improved overall, and would likely need closer to a 90-95% accuracy to be useful in our opinion, but we do believe we achieved our goal.