STAT3215: Final Project Proposal

Leon Nguyen 2880311, Charitarth Chugh

2023-10-27

The analysis to be replicated The analysis that this paper covers assesses the on-time performance of domestic flights based on observations from the year 2016. We will recreate the multiple linear regression model. The paper also discusses the usage of a random tree and random forest model, but because this falls outside the scope of the course, we will only focus on the MLR model and how it can be improved. The objective of this model is to predict the on-time performance of flights based on various factors.

How to recreate dataset used in the paper We will focus on only the variables used in the original analysis. The data which is publicly available on the [________] has [] attributes, but for the sake of simplicity we have kept the eight regressors and the response variable of departure delay (`DepDelay`)

Extension/enhancement to implement We will examine methods to address non-normality of the errors. We will delve into applying transformations (potentially log) onto regressors and the response variable because of their right-skewed distribution. Once a normal distribution of errors has been achieved, we can draw inferences and perform tests that we would not have been able to do otherwise. No inferences or formal hypothesis testing was done to test the significance of the regressors. Since the variables are not correlated with each other, we can assume they are independent and do ANOVA testing without worrying about interacting terms. This would test the impact of each regressor on the model and analyze the amount of variation explained by the relationship with the regressor and response. The conclusions from these test would give us a better understanding of the extent of which regressors influence flight delay.