STAT3215: Final Project Proposal

Leon Nguyen ldn19001, Charitarth Chugh chc21001

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 variable selection and the multiple linear regression model and how both can be improved. The objective of this model is to predict the on-time performance of flights based on an optimal set of predictors. This paper also seeks to answer how long typical departure delay is, and what the major causes of departure delay are.

How to recreate dataset used in the paper

The dataset used in the original paper utilizes a subset of the attributes provided from the Bureau of Transportation Statistics (BTS). They have cleaned and pre-processed the dataset but have also imputed values which may contribute to bias and unintended consequences on regression. This was done used additive regression, bootstrapping, and predictive mean matching used the "Hmisc" package in R studio. The raw data set contained approximately one million observations. We can acquire the raw superset of data directly from the BTS. They provide files of domestic flights by month, so we will need to merge them into one dataset for the year of 2016. We may need to filter out non-domestic flights and any observations that have a lot of null values. If there are some or few null fields in a given observation, we will input zero because this indicates that the associated regressor is not a cause of delay, and thus should not be considered for said observation.

The original dataset which is publicly available from the Bureau of Transportation Statistics has over 23 attributes, but for the sake of simplicity the original authors have kept the eight regressors and the response variable of departure delay (`DepDelay`). The process by which they choose which regressors is not fully elaborated on in the original analysis, but upon analyzing the different attributes and selecting the one related to delay time, we chose predictors which agreed with those used in the analysis. We will validate that the variable selection used in the original analysis is the best to use, and use model validation to see if we can improve the MLR model by reducing the number of regressors and/or explaining more of the variation from the relationship between the departure delay and predictors.

Extension/enhancement to implement

We will start cleaning the dataset to remove miscellaneous information that has no impact on flight delay, such as flight duration and destinations.

First, we will address non-normality of the residual 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 tests would give us a better understanding of the extent of which regressors influence flight delay.

Once we transform the necessary regressors, we will examine methods to find a better model based on a set of potential predictors. We want to explain as much variation as we can while using an adequate number of regressors according to the Principle of Parsimony. Our goal is to either show that using the initial eight regressors is the best MLR model, or to find a simpler model with less regressors that explains the same or larger amount of variation.