Predicting the Length of Flight Delays

Andrea Cao, Anlin Li, Zheyi Li, Manuel Rodriguez Ladron De Guevara, Mingyang Wang, Shuhan Yang.

Introduction

Flights delays are a common problem that affects millions of people a year, causing adverse economic effects. We present an analysis using different regression models to predict flight delay times. We train on a dataset containing flights during December 2016. Our goal is to firstly, learn associations and potential correlations among the different factors of variations, such as departure and arrival airports, airline, or day of the month. Secondly, we evaluate different models such as linear regression, decision trees or random forest to find the best performing model.

Data

Our dataset contains 34,314 US national flights from 2016 that departed specifically from ORD (Illinois) or DFW (Texas). There is a total of 26 variables, including our response variable, the arrival delay. This is the difference in minutes between scheduled and actual arrival time. We further clean the rest of the 25 predictor variables, as there are some variables that are deterministically related to the response variable. After some data preprocessing, including outlier removal and the logarithmic transformation of some skewed variables, we retain the most important variables such as *day of the month, airline (carrier), origin, destination, departure time, taxi in, taxi out,* among others. In total, we retain 12 predictor variables.

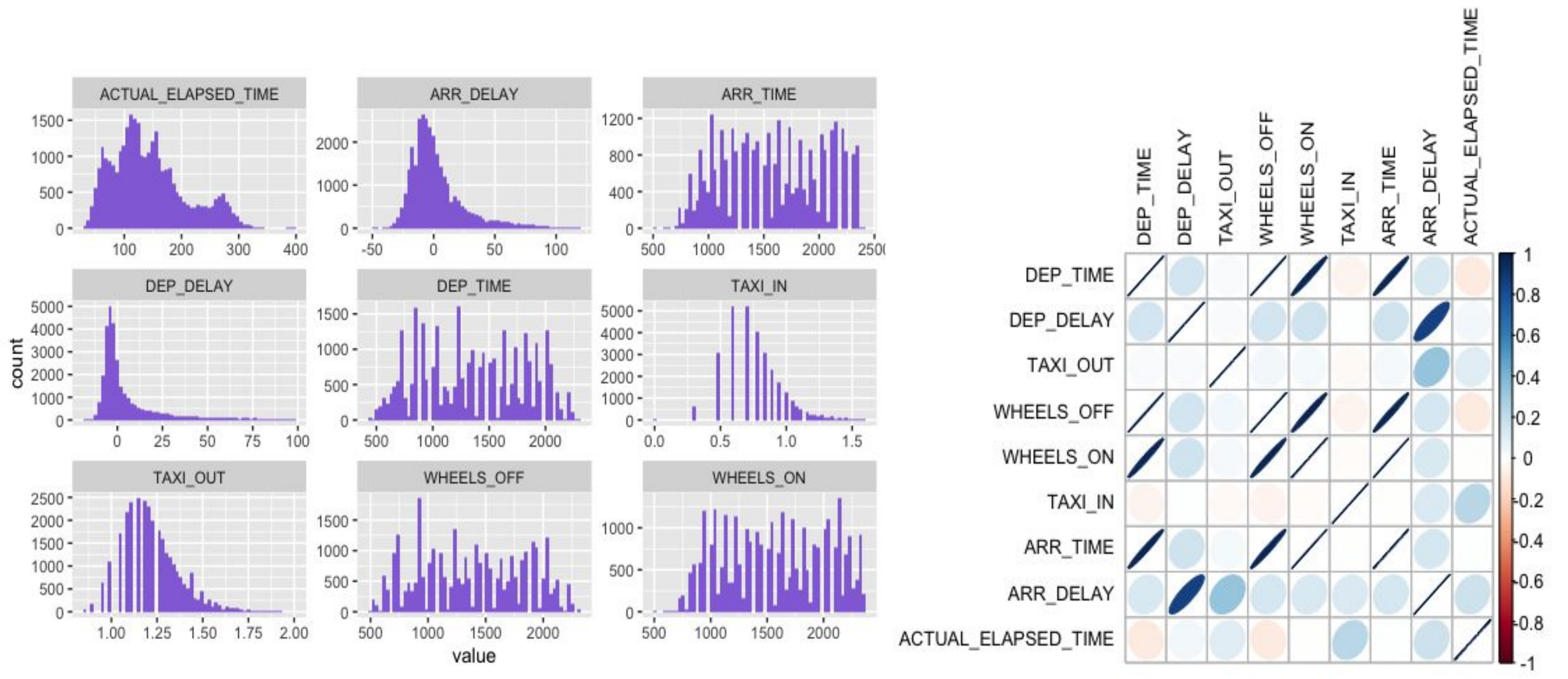

Figure 1.. Histograms of important numeric variables (left) and correlation plot (right).

The figure above shows histograms of the most important numeric variables, including the response variable *arr_delay* (left). A correlation plot (right) shows strong correlations between some variables. For example, expectedly *arr_delay* is strongly correlated to *dep_delay*, as well as *wheels_on* and *wheels_off*. *arr_time* is also correlated to both wheels modes. There is no readily apparent visual association between the response variable and any of the predictor variables.

Methods

After data pre-processing, our final dataset contains 30,609 data points. We split the data into a training set and a test set, where 70% is used for model training and 30% for model testing. We use Mean-Squared Error (MSE) to select the best model among the following models: Linear regression, Decision Tree, Random Forest, Extreme Gradient Boosting (XGB), and K-nearest Neighbor (KNN). We use Linear regression to identify important predictor variables.

Analysis

We present a summary of the performance of the models using MSE in Table 1. Linear regression and Random Forest are our best performing models, achieving the lowest MSE score. We also show a diagnostic plot for the linear regression model in Figure 2, which indicates the relation between our model and the data. Figure 2 shows that the regression line is a good fit to the data.

- Linear Regression achieves an MSE of 57.650, and an Adjusted R-squared of 0.889, which indicates that the model is informative. Adjusted R-squared indicates the goodness of fit, and a value of 1 indicates that the model perfectly fits the data.
- K-Nearest-Neighbor: Since KNN algorithm only use numeric factor, the *CARRIER, ORIGIN, DAY_OF_MONTH* and *DEST* columns were dropped in this classification process. We hypothesize that removing these many factors affects the performance of the model, achieving an MSE of 128.47.
- Random Forest: Since the random forest algorithm limits the factors in one column to be 53, the *DEST* column was dropped. Even though it loses some information, it gives a fairly good MSE. For fairness, we perform linear regression using the same dataset that Random Forest uses, and we obtain an MSE of 73.123.
- Extreme Gradient Boost reaches an MSE of 70.398, with *DEP_DELAY* being the most important factor followed by *TAXI_OUT*.
- Decision tree model tends to overfit the data, which leads to a poor performance on the test set. It reaches an MSE of 103.458.

Methods	MSE
Linear Regression	57.650
K-Nearest-Neighbor	128.4709
Random Forest	**55.661***
Extreme Gradient Boost	70.398
Decision Tree	103.458

Table 1. Comparison between different models. *Random Forest was performed using one fewer variable (DEST).

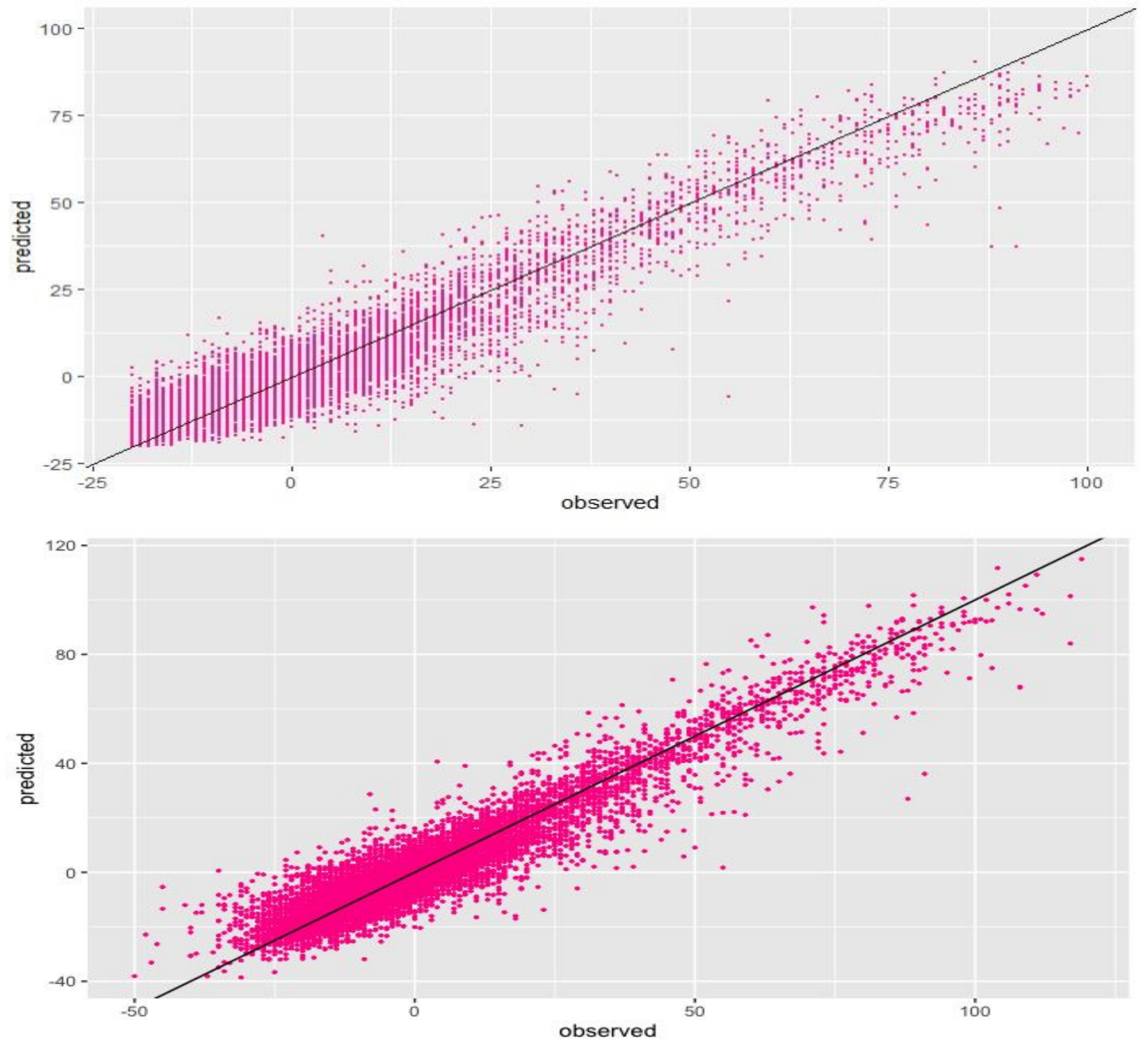

Figure 2. Diagnostic plots for Random Forest (above) and Linear Regression (below).

Conclusion

We present an analysis to select the best model to predict the response variable, *arrival delay.* We find that the most important predictor variables are *CARRIER, DEP_DELAY,* and *TAXI_OUT.* The best models are linear regression and random forest. Due to limitations on the number of variables that the random forest algorithm implementation in R can take, we find that dropping the *DEST* variable gives the best performance using this model. For comparison fairness, we also compute linear regression without *DEST* variable, achieving a higher MSE of 73.123. This, however, imposes a limitation on the model's prediction. We conclude that both linear regression and random forest are the best models, achieving a similar MSE score when the former uses the *DEST* variable. In random forest, the most important variables are *DEP_DELAY, TAXI_OUT*.

References: Alex Reinhart (http://rosmarus.refsmmat.com/datasets/datasets/flight-delays/)