Predicting Significant Flight Delays using Supervised Learning

NATE NETZNIK

Objective

Identify which flights will likely be delayed or cancelled to help consulting firms make informed travel decisions

- Many consultants travel frequently over long distances for business purposes
- Managers would be interested in how to minimize risk of flight delay or cancellation

Data

Dataset obtained from Kaggle

Took random sample of 100,000 flights from data spanning between August 2021-July 2022

Target variable to be predicted indicates whether a flight was delayed

- For our purposes, “delayed” flights includes flights that were cancelled

Predictors include

- Quarter
- Month
- Day of Week
- Airline
- Origin
- Destination
- Scheduled Departure Time
- Scheduled Arrival Time
- Distance

Methods

Constructed two models

- Random Forest: generates several decision trees – sequences of tests that lead to predictions – then averages predictions from decision trees
- Logistic Regression: generates a probability that a given flight will be delayed or cancelled

Results

	Random Forest	Logistic Regression
Overall Accuracy	**61.3%**	60.4%
% Predicted delays that were actually delayed	**30.8%**	30.5%
% Predicted non-delays that were actually not delayed	84.5%	**84.8%**

Conclusions

The **Random Forest model** is the better performing model overall.

The model does **not yield reliable delay predictions** (only ~30% are correct)

- However, most predicted non-delays are correct (~85%)

Future Work

Continue fine-tuning model parameters to further improve performance

Consider other classification models such as a naïve Bayes probabilistic model

Consider using a new set of predictors