Flight Analysis

●●●

Predicting Flight Delays

Josh Friedman

Problem

- 51% of flights between Boston and Chicago were delayed in 2015.

- Delays can cause airline customers to miss important family events, interviews and meetings and can require a customer to spend more money.

- Overall, delays increase stress and airline customers would like to avoid them.

Solution

- Understand the most important factors that cause a delay and build a predictive model

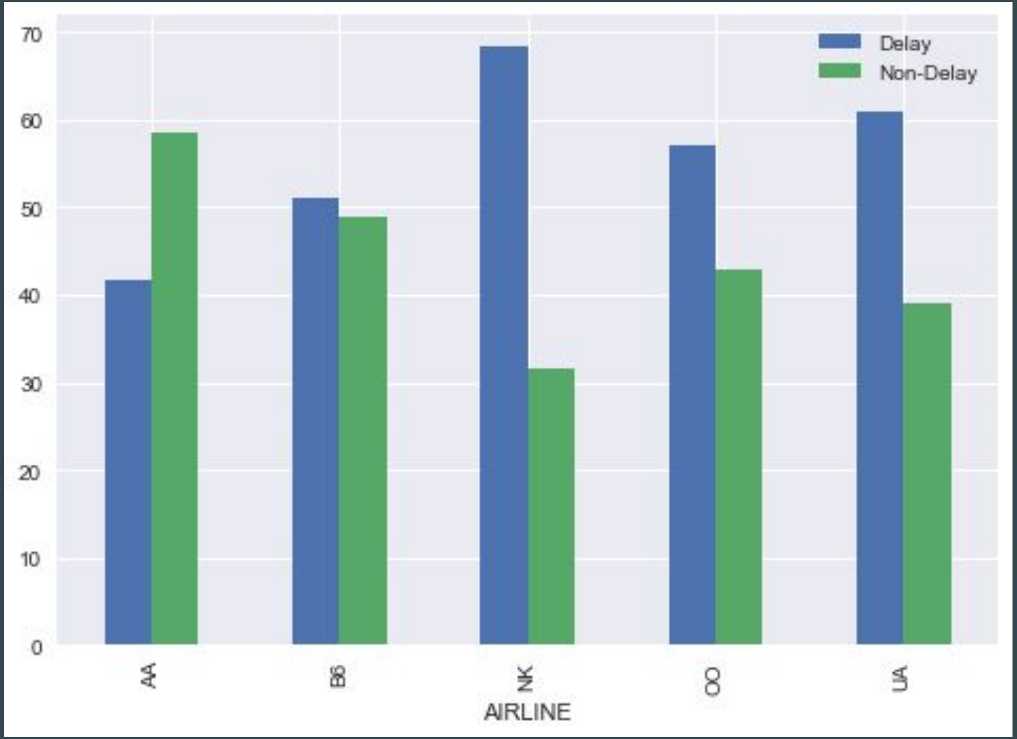
- Customers can then choose their travel options more intelligently

Steps

1. Gather Data
    a. Flight delay data from the Department of Transportation
    b. Weather data from NOAA's National Centers for Environmental Information (NCEI)
2. EDA and Data Storytelling
    a. Compare delay rates among airlines, days of the week and seasons
    b. Run hypothesis tests to determine if there are differences in delay rates
3. Build Predictive Model
    a. Compare Logistic Regression, Random Forest, and K-NN

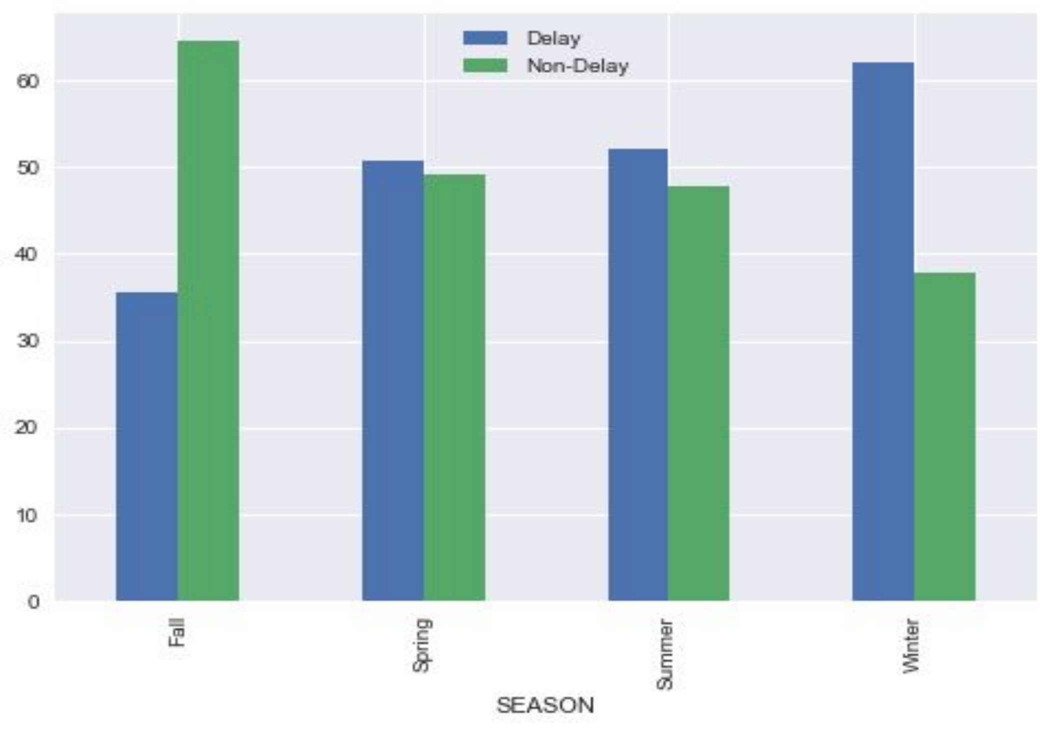
Data Storytelling

Which airlines have the highest and lowest percentage of delays?



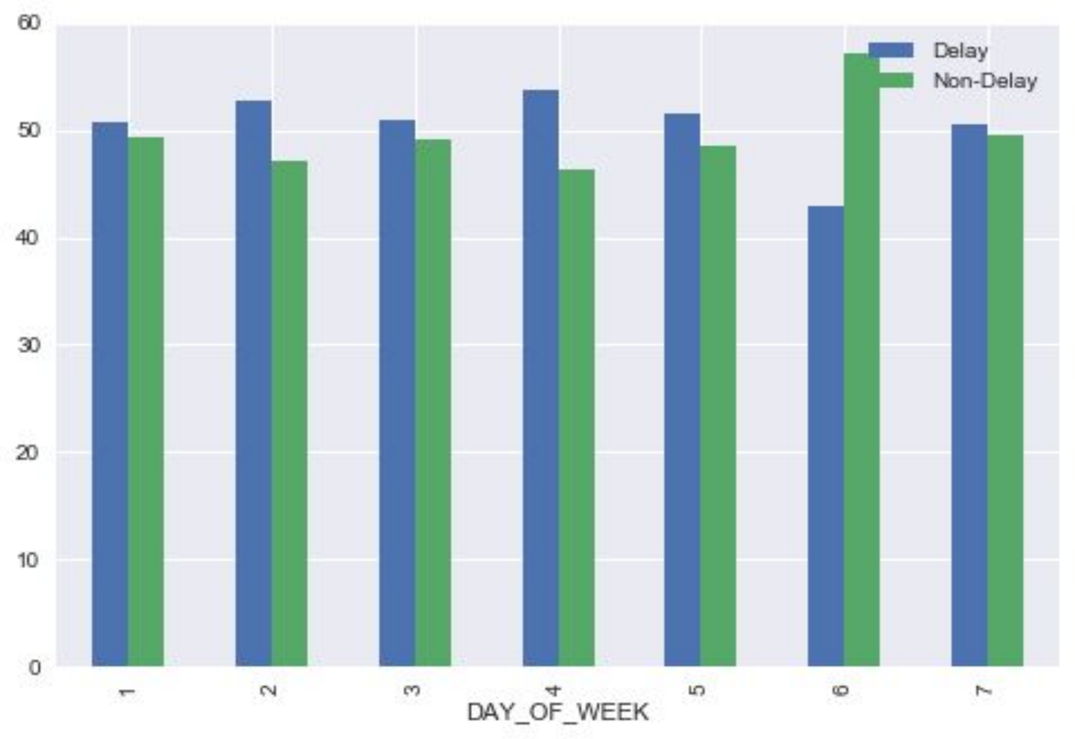
Data Storytelling

Which seasons have the highest and lowest percentage of delays?



Data Storytelling

Which days of the week have the highest and lowest percentage of delays?



Hypothesis Tests

- Conducted 2 proportions ztests to test the differences in proportions of delays

- 5 tests were conducted:
  - Airline - American vs. United
  - Season - Winter vs. Spring/Summer/Fall
  - Day of the week - Friday vs. Saturday

Results

- Each test provided a p-value that was less than .05 - able to reject null hypothesis that the proportions in delays were the same
- Conclusions
    - American delay rate < United
        - 95% CI: (-.232 and -.155)
    - Winter delay rate > spring
        - 95% CI: (.06 and .16)
    - Winter delay rate > fall
        - 95% CI: (.21 and .31)
    - Winter delay rate > summer
        - 95% CI: (.052 and .148)
    - Friday delay rate < saturday
        - 95% CI: (-.15 and -.006)

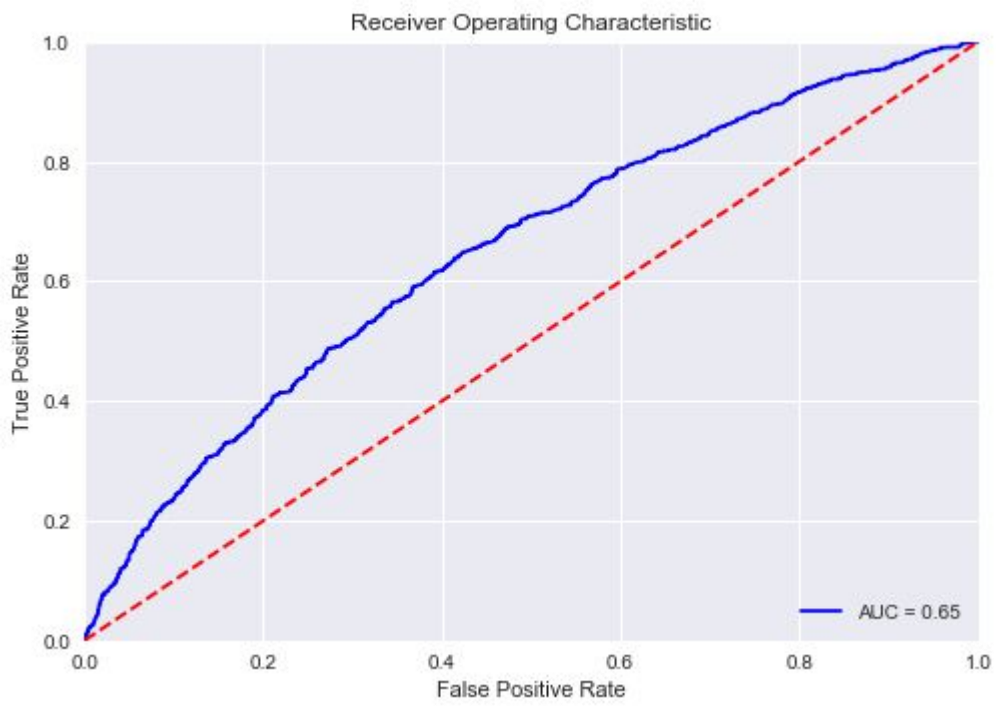
Predictive Models

- Data Prep
  - Normalized PRCP, SNOW, and AWND
  - Created dummy variables for DAY_OF_WEEK, SEASON, AIRLINE, ORIGIN_AIRPORT, and DESTINATION_AIRPORT.
- Model Fitting and Evaluation
  - GridSearchCV to determine the optimal parameters
  - Felt accuracy score was appropriate because the success rate in the dataset (delay rate) was around 50%
  - Used a confusion matrix to determine the type 1 and type 2 errors that were made with each model
  - Built ROC curves and used AUC score as my primary metric for comparison. The model with the highest AUC score was selected.

Logistic Regression

- Accuracy Score: 61%
- Confusion Matrix Results:
  - True Positive: 1010 were delays and predicted correctly
  - True Negative: 938 were not delays and predicted correctly
  - False Positive: 634 were not delays and predicted incorrectly
  - False Negative: 603 were delays and predicted incorrectly

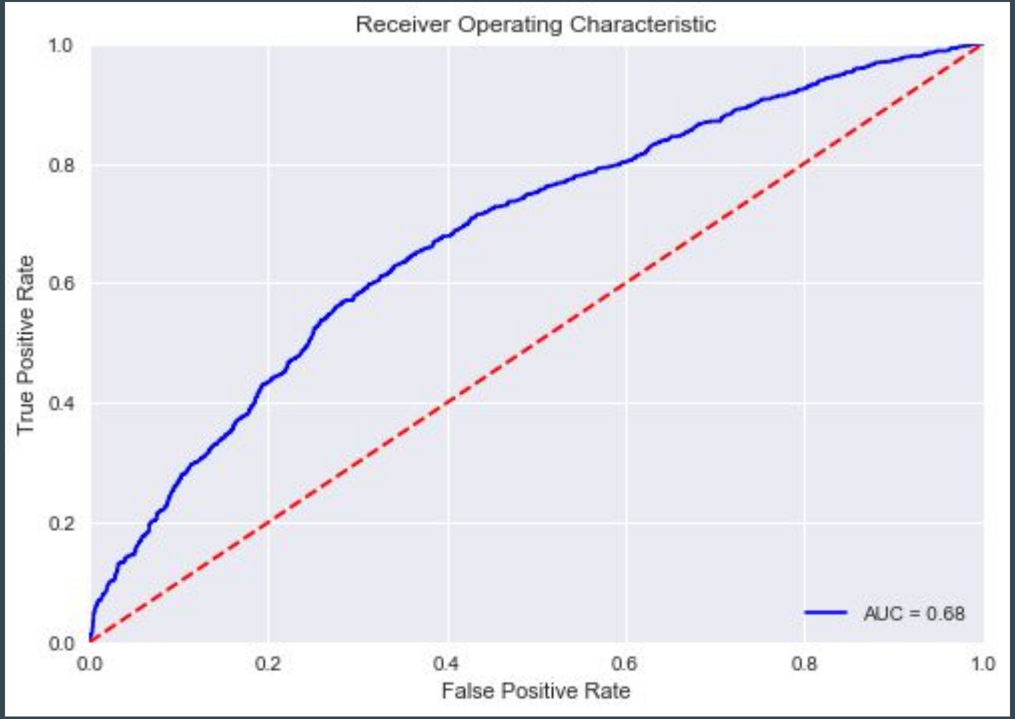
AUC: .65



Random Forest

- Accuracy Score: 64%
- Confusion Matrix Results:
  - True Positive: 1085 were delays and predicted correctly
  - True Negative: 942 were not delays and predicted correctly
  - False Positive: 599 were not delays and predicted incorrectly
  - False Negative: 559 were delays and predicted incorrectly

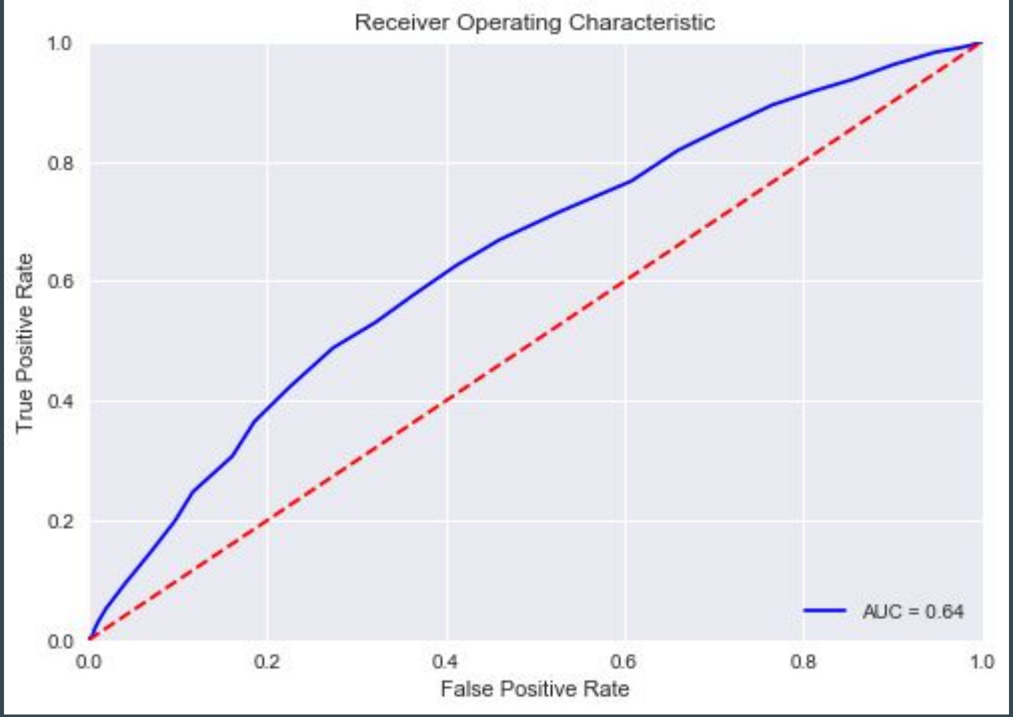
AUC: .68



K-NN

- Accuracy Score: 61%
- Confusion Matrix Results:
    - True Positive: 950 were delays and predicted correctly
    - True Negative: 979 were not delays and predicted correctly
    - False Positive: 562 were not delays and predicted incorrectly
    - False Negative: 694 were delays and predicted incorrectly

AUC: .64



Insight and Recommendations

- Selected Random Forest - highest AUC score of .68
- Precipitation and wind are the most important features in determining a delay
- Winter has the highest proportion of delays
    - Weather factors are often more severe in the winter time and thus it would make sense that winter caused the most delays of any season.
    - Customers should keep this in mind and should give themselves more time to travel in the winter
- American Airlines caused the least delays and it seems that if you are on a time crunch traveling between Chicago and Boston, it would be a smarter choice than other airlines.
- If leaving on a weekend, Friday would be a smarter choice than Saturday.

Next Steps

- Utilize the rest of the dataset to get a deeper understanding of the best airlines and airports when under a time crunch.
  - For example, if a customer had an interview in New York and was flying from the Cleveland area, should they fly United or American out of Akron/Canton or Cleveland? Should they fly to JFK or Laguardia?
- Analyze time of day to get a deeper understanding of the impact of timing on delays. Should my flight be at 6am or 1pm?
- Feed customers upcoming flight information to the model and use the results to inform travel decisions.