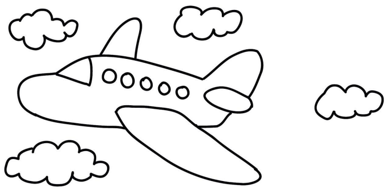


Flight Delays
And Cancellations

A Classification Analysis

By Adina Steinman

Business Problem

- Our team of strategy consultants has been hired by the FAA, Federal Aviation Administration, to optimize efficiency across US airports and minimize flight delays as much as possible
- Our goal is to create a model that is able to predict whether a flight will be delayed or not and see which features are most useful in determining our predictions
- We will build a model that optimizes **recall**. We want to ensure that all delayed flights are classified correctly and that none are mistakenly classified as non-delayed. It is important to ensure that the model classifies delayed flights as correctly as possible so the FAA can effectively strategize to minimize these delays

Datasets

- Kaggle's "2015 Flight Delays and Classifications" dataset consisting of 3 csv files; information on flights, airports, and airlines
- Data has 39 columns containing information on delay/arrival times, reason for delays, location of airport, airlines, etc.
- Data has 5,819,079 rows; a random sample of 10,000 was used for this analysis

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 10000 entries, 0 to 9999
Data columns (total 39 columns):
YEAR                        10000 non-null int64
MONTH                       10000 non-null int64
DAY                         10000 non-null int64
DAY_OF_WEEK                 10000 non-null int64
AIRLINE_CODE                10000 non-null object
FLIGHT_NUMBER               10000 non-null int64
TAIL_NUMBER                 9981 non-null object
ORIGIN_AIRPORT_CODE         10000 non-null object
DESTINATION_AIRPORT_CODE    10000 non-null object
SCHEDULED_DEPARTURE         10000 non-null int64
DEPARTURE_TIME              9862 non-null float64
DEPARTURE_DELAY             9862 non-null float64
TAXI_OUT                    9852 non-null float64
WHEELS_OFF                  9852 non-null float64
SCHEDULED_TIME              10000 non-null float64
ELAPSED_TIME                9828 non-null float64
AIR_TIME                    9828 non-null float64
DISTANCE                    10000 non-null int64
WHEELS_ON                   9847 non-null float64
TAXI_IN                     9847 non-null float64
SCHEDULED_ARRIVAL           10000 non-null int64
ARRIVAL_TIME                9847 non-null float64
ARRIVAL_DELAY               9828 non-null float64
DIVERTED                    10000 non-null int64
CANCELLED                   10000 non-null int64
CANCELLATION_REASON         148 non-null object
AIR_SYSTEM_DELAY            1890 non-null float64
SECURITY_DELAY              1890 non-null float64
AIRLINE_DELAY               1890 non-null float64
LATE_AIRCRAFT_DELAY         1890 non-null float64
WEATHER_DELAY               1890 non-null float64
AIRLINE                     10000 non-null object
AIRPORT_CODE                9184 non-null object
AIRPORT                     9184 non-null object
CITY                        9184 non-null object
STATE                       9184 non-null object
COUNTRY                     9184 non-null object
LATITUDE                    9173 non-null float64
LONGITUDE                   9173 non-null float64
dtypes: float64(18), int64(10), object(11)
memory usage: 3.0+ MB
```

Class Imbalance

- From our sampled dataset, 64% of data was classified as “delayed and 38% as non-delayed
- SMOTE technique was applied to handle this imbalance, and the synthetic dataset split was 50% delayed and 50% non-delayed. These resampled values were used to fit all our future models

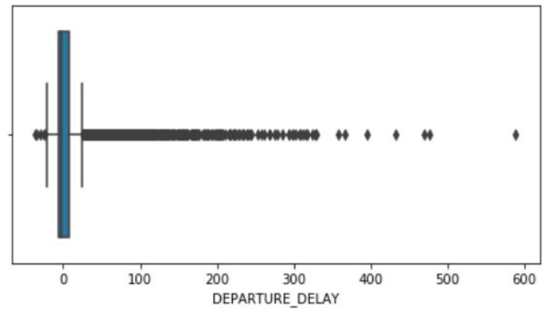
```
Original class distribution:

0    6384
1    3616
Name: DELAYED, dtype: int64
----------------------------------------
Synthetic sample class distribution:

1    4795
0    4795
Name: DELAYED, dtype: int64
```

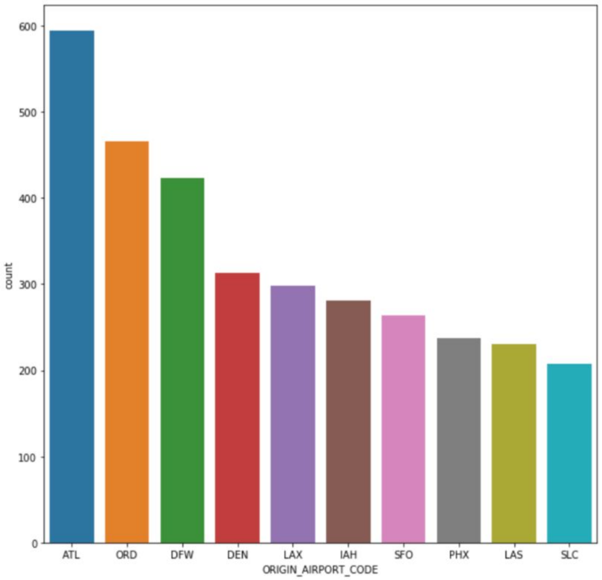
EDA

- Replaced NaNs with 0s or median values, then one-hot-encoding was applied to Months, Days of Week, Airlines and Airports (based on flight traffic

- Some extreme outliers in the target variable, departure delay, indicating that there could be some extreme cases/unique situations that caused certain flights to be delayed

- Feature engineering: created variable that binned airports based on flight traffic; heavy, medium, light and very light
-
- Certain airports are busier than others; particularly ATL, ORD, and DFW have the most air traffic, which makes sense as they are all hubs for Delta, United and American Airlines, respectively





Vanilla Model Results

model	train_accuracy	train_precision	train_recall	train_f1	test_accuracy	test_precision	test_recall	test_f1
knn	0.782	0.693	0.738	0.714	0.609	0.483	0.498	0.491
bayes	0.788	0.973	0.438	0.604	0.780	0.976	0.428	0.595
decisiontree	1.000	1.000	1.000	1.000	0.727	0.633	0.660	0.646
bagging	0.997	0.999	0.992	0.996	0.789	0.771	0.628	0.692
randomforest	1.000	1.000	1.000	1.000	0.799	0.796	0.629	0.703
Adaboost	0.813	0.780	0.689	0.731	0.798	0.757	0.685	0.719
XGBoost	0.833	0.829	0.692	0.754	0.811	0.797	0.671	0.729
SVM	0.859	0.884	0.711	0.788	0.788	0.791	0.598	0.681

Final Model

model	train_accuracy	train_precision	train_recall	train_f1	test_accuracy	test_precision	test_recall	test_f1
decisiontree	1.000	1.000	1.000	1.000	0.758	0.666	0.671	0.668
Decision Tree Tuned	0.792	0.687	0.779	0.730	0.774	0.672	0.739	0.704
XGBoost	0.835	0.819	0.696	0.752	0.821	0.804	0.672	0.732
XGBoost Tuned	1.000	1.000	1.000	1.000	0.785	0.718	0.676	0.696

- Decision Tree and XGBoost were tuned as their test_recall scores were the highest amongst the vanilla models
- Decision Tree final model included hyperparameters for criterion, max depth, minimum samples split and minimum split leaf. The result improved test_recall from 0.671 in vanilla model to 0.739 in final model
- XGBoost final model included hyperparameters for max depth, learning rate, subsample, minimum split loss, and number of estimators. The result improved test_recall from 0.672 to 0.676
- Resulting models will still misclassify some delayed flights as non-delayed (false negatives), however this is what the model aims to optimize

Conclusions and Recommendations

- While XGBoost increased our test recall score, it was only marginally, and the resulting model was overfit to our train data
- In future models, we should look to remove features that are not correlated with our target variable, as well as remove features that have low feature importances. Future work with more data cleaning could potentially lead us to stronger recall scores, without producing a model that is overfit
- FAA should allocate resources to improve the efficiencies of Late Aircraft Delay, Arrival Delay and Airline Delay, as these types of delays had the highest feature importances in our models

Thank You!

- Kaggle Dataset: https://www.kaggle.com/usdot/flight-delays
- GitHub: https://github.com/adinas94/phase3_project
- Contact Information
    - Email: adinasteinman@gmail.com
    - Linkedin: https://www.linkedin.com/in/adinasteinman/