SunTzuLombardi / **FlightsClassification**

Classify flights as delayed or not

☆ **0** stars **0** forks

☆ Star | Unwatch ▾

<> Code | Issues | Pull requests | Actions | Projects | Wiki | Security | Insights | Settings

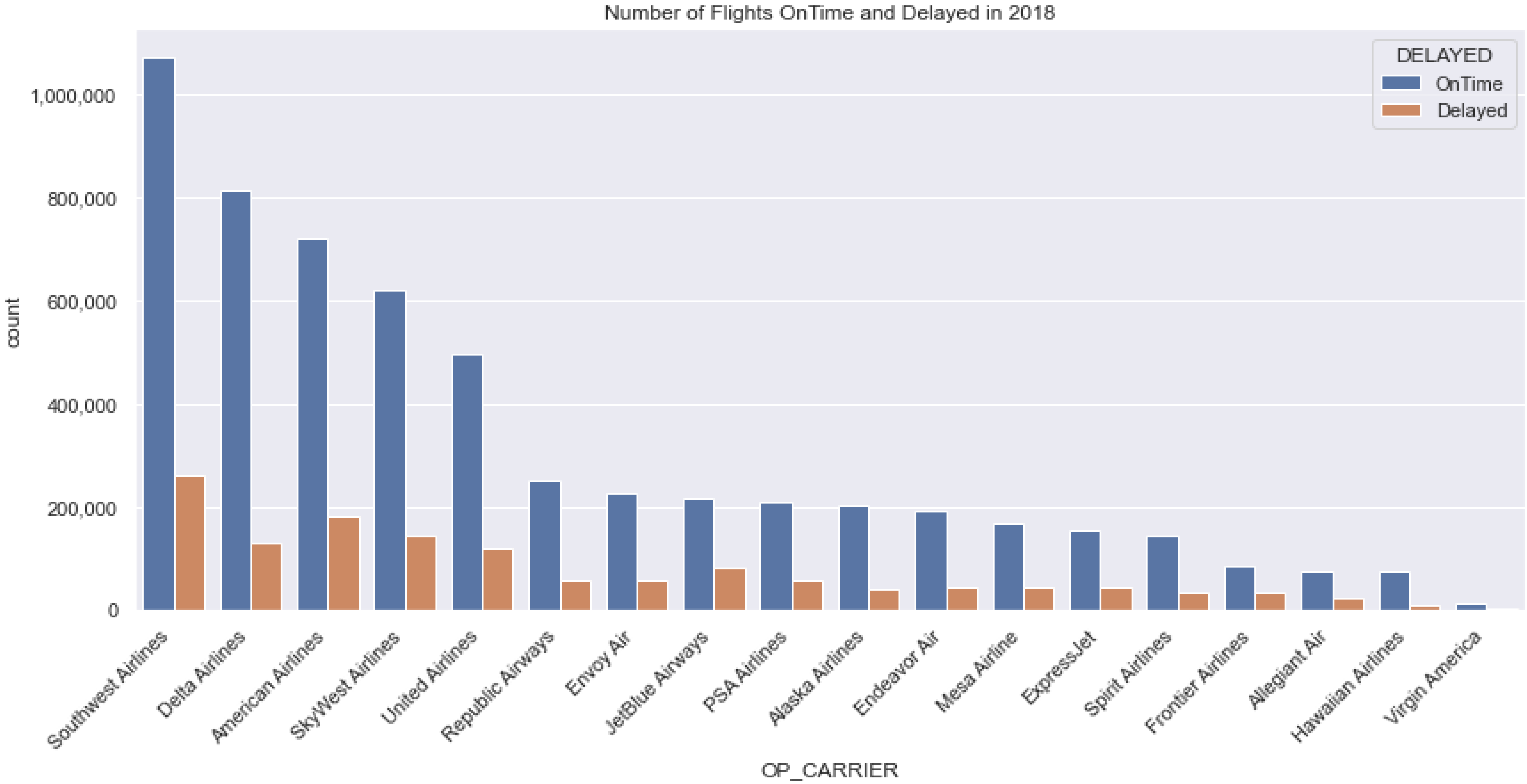
main ▾ ···

SunTzuLombardi Conclusion Edit ... 33 seconds ago 13

View code

README.md

Classification



Overview

This project encapsulates using Classification with Machine Learning for modeling 2018 Domestic Airline Flight Delays.

Business Problem

We are consulting with Southwest Airlines for Domestic Flights Analysis by looking at industry delays and routes for improvement opportunities. Which Airlines are usually late/early? Which routes are late/early?

Data

Airline and Cancellation Dataset on Kaggle by Yuanyu 'Wendy' Mu

All Records from United States Department of Transportation

2018 data containing 7.21M records 851MB Initial Features included:

FL_DATE - Date of Flight

OP_CARRIER - Flight Carrier

OP_CARRIER_FL_NUM - Flight Carrier Identifier

ORIGIN- Start Airport

DEST- Destination Airport

CRS_DEP_TIME - Computer Reservation System (CRS) Departure Time

DEP_TIME - Actual Departure Time

DEP_DELAY - Dep Time minus CRS Dep Time in Min

TAXI_OUT - Time To taxi

WHEELS_OFF - Time Wheels in Air

WHEELS_ON - Time Wheels on Ground

TAXI_IN - Time To taxi

CRS_ARR_TIME - Computer Reservation System (CRS) Arrival Time

ARR_TIME - Actual Arrival Time

ARR_DELAY - ARR_Time minus CRS_ARR_TIME in Min

CANCELLED - Flight Cancelled or not

CANCELLATION_CODE - Cancel Code

DIVERTED - Flight Was diverted or Not

CRS_ELAPSED_TIME -CRS scheduled Flight Time

ACTUAL_ELAPSED_TIME - Actual Flight Time

AIR_TIME - Time in the Air

DISTANCE - Distance of Flight

CARRIER_DELAY - Carrier Delay in Min

WEATHER_DELAY - Weather Delay in Min

CANCELLATION_CODE - Cancelled Code

NAS_DELAY - National Air Service Delay in Min

SECURITY_DELAY - Sec Delay in Min

LATE_AIRCRAFT_DELAY - Delay due to late Aircraft in Min

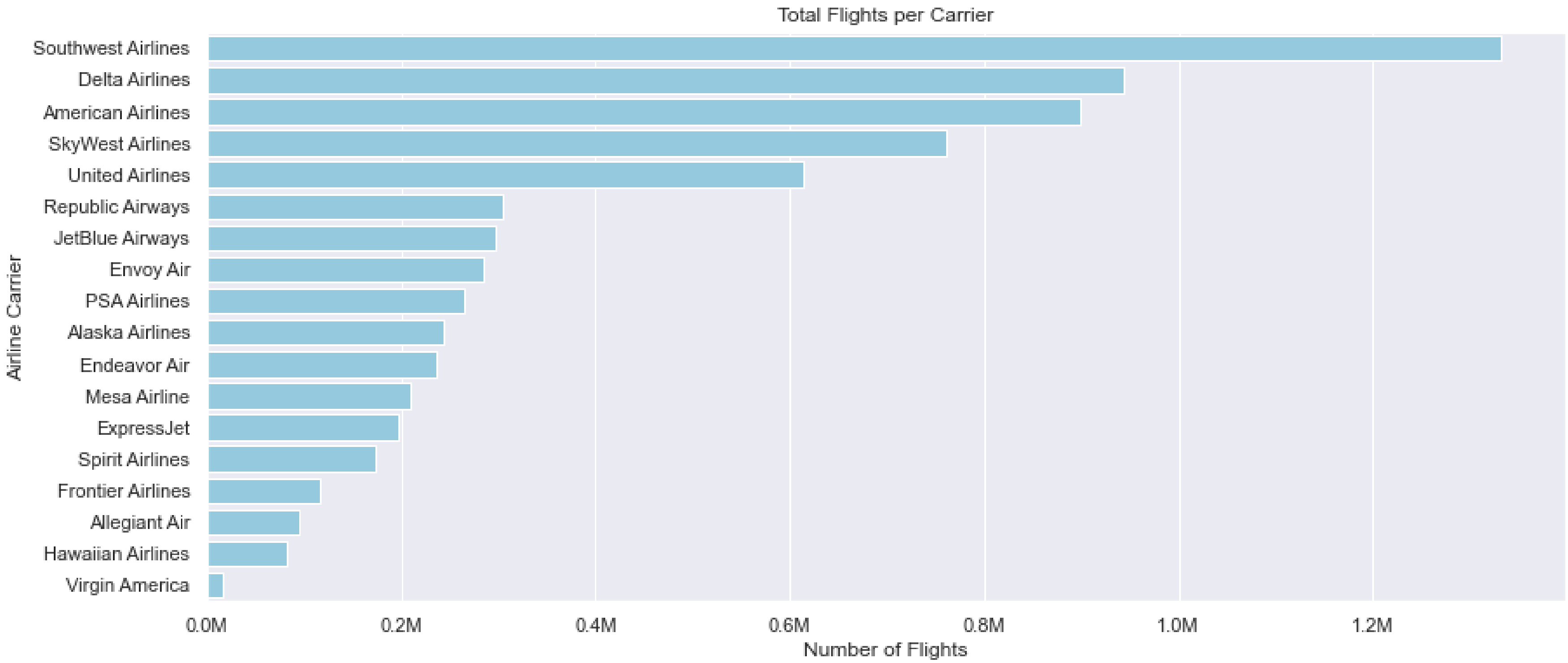
Methods

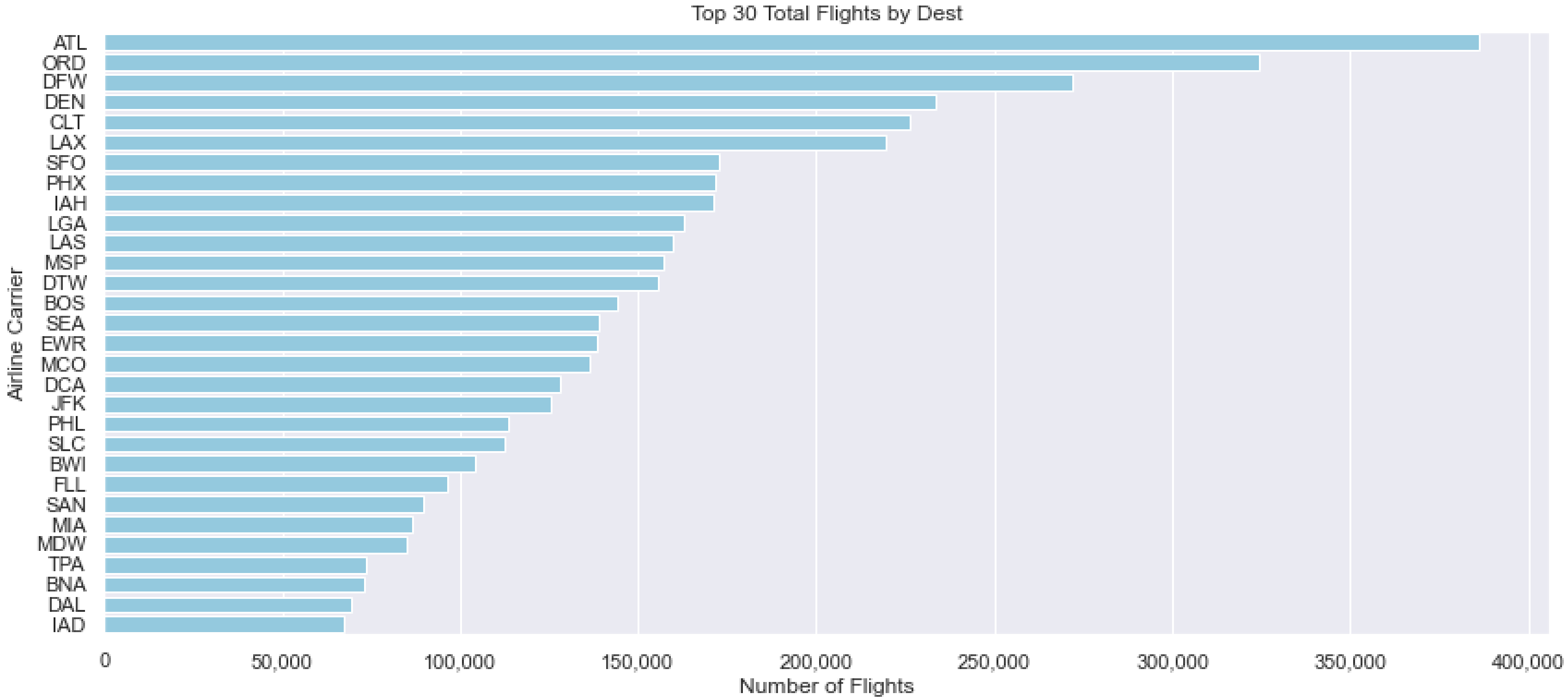
We performed Inferential Analysis of 7M+ recs looking at Airlines, Destinations of Flights, Delays, Times, We then reduced the Data set to just the Top 5 Airlines by numner of flights. We reduced the number of Origins and Destinations to the top 30 instead of the 358.

We also performed Classification Analysis with Machine Learning Algorithms Logistic Regression, Decision Trees, Random Forests, XGBoost

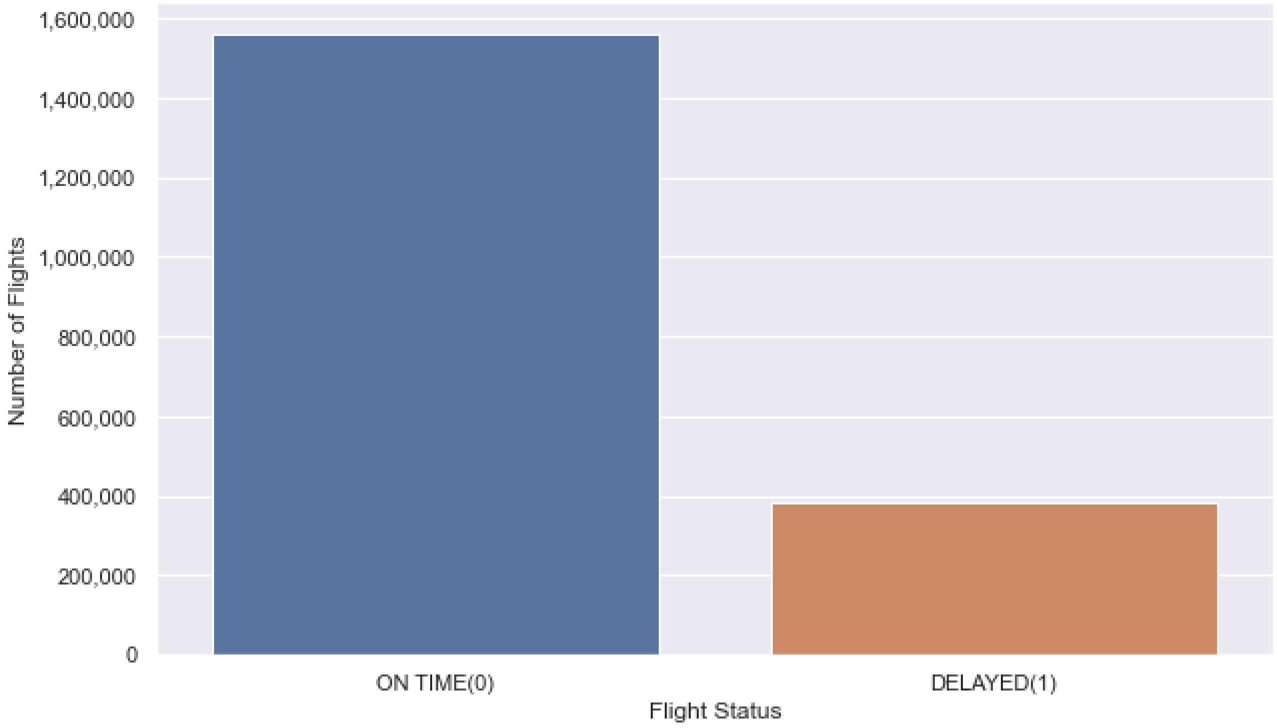
With GridSearch narrowing down the most optimal Hyperparameters to predict delayed flights and assess the strength and relationship and importance of the different features and their relation to delayed flight.

Results





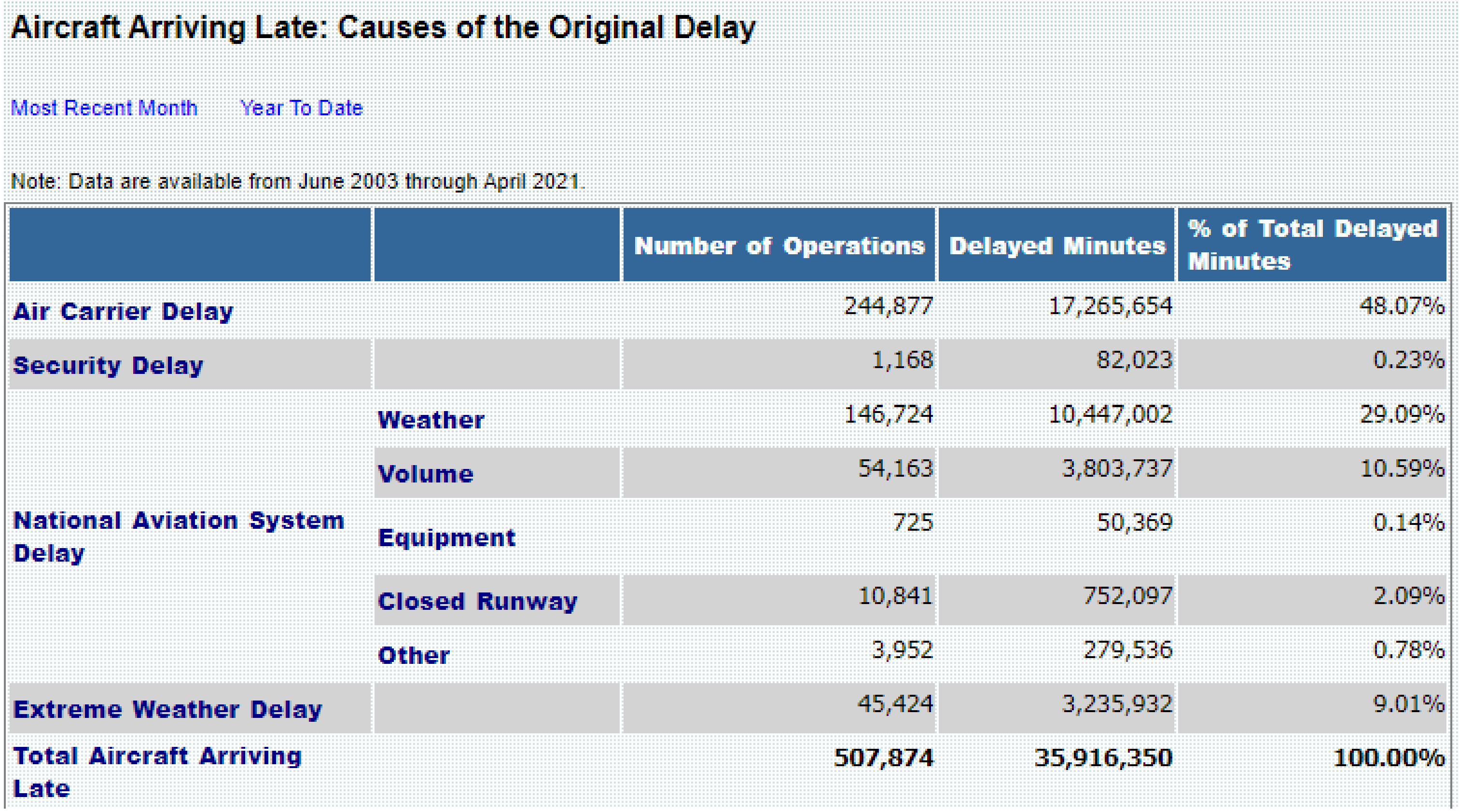
Modeling with continuous features Distance, Flight Time, Categoricals Weekdays, Months, Top 5 Airlines, Top 30 Origins and Destinations To Classify if Delayed or not. Delays are on Arrival Delays and >=15 mins



Best Predictive Results were found with the XGBoost algorithm With a Recall of 59%, Accuracy 66% , F1 value of .59

Model	Recall	Accuracy	F1
XGBoost	59%	66%	59%
Random Forest	59%	65%	57%
Decision Tree	39%	68%	55%

A flight is considered delayed when it arrived 15 or more minutes than the schedule (see definitions in Frequently Asked Questions). Delayed minutes are calculated for delayed flights only. When multiple causes are assigned to one delayed flight, each cause is prorated based on delayed minutes it is responsible for. The displayed numbers are rounded and may not add up to the total.



Aircraft Arriving Late: Causes of the Original Delay

Most Recent Month Year To Date

Note: Data are available from June 2003 through April 2021.

		Number of Operations	Delayed Minutes	% of Total Delayed Minutes
Air Carrier Delay		244,877	17,265,654	48.07%
Security Delay		1,168	82,023	0.23%
National Aviation System Delay	Weather	146,724	10,447,002	29.09%
	Volume	54,163	3,803,737	10.59%
	Equipment	725	50,369	0.14%
	Closed Runway	10,841	752,097	2.09%
	Other	3,952	279,536	0.78%
Extreme Weather Delay		45,424	3,235,932	9.01%
Total Aircraft Arriving Late		**507,874**	**35,916,350**	**100.00%**

Pulled from BTS Bureau Trans StatNote: Airlines report late-arriving aircraft as a category of the cause of delay when a previous flight with same aircraft arrived late, causing the present flight to depart late. Airlines do not report the cause of delay for the first late flight that caused the second delay. Using data reported by the airlines for other categories of delay causes, the page displays calculations of the causes of delay for the late arriving aircraft category. These calculations use the percentages of delay minutes reported by the airlines in the air carrier, national aviation system, security and weather categories and assign them proportionately to the late arriving aircraft category. The displayed numbers are rounded and may not add up to the total.

Conclusions

We predicted 59% of Delayed flights but Recall needs to be more accurate.
SouthWest Airlines should focus on Reducing Backup Delays as that is 50% of All Delays.

Challenges
The Large Dataset and finding an appropriate model for the complexity of the data was a challenge.

Next Steps

Try to model again with reduction of origins and destinations: Top 20, Top 10.
Another additional approach is to try PCA analysis for understanding the value of certain features.

For More Information

See the full analysis in the Jupyter Notebook or review this presentation

For additional info, contact Daniel M. Smith at danielmsmith1@gmail.com

Repository Structure

```
├── code
│   ├── init.py
│   ├── __.py
├── data
├── images
├── init.py
├── README.md
├── presentation.pdf
├── gitbhub.pdf
├── notebook_Classification.pdf
├── Classification.ipynb
```

Releases

No releases published
Create a new release

Packages

No packages published
Publish your first package

Languages

● Jupyter Notebook 100.0%