Predicting Flight Delays

By: Sunny & Wes

Table of Contents

Introduction

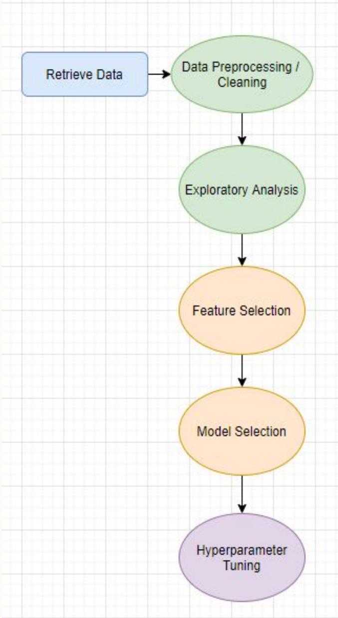
Company: Sunley Airlines

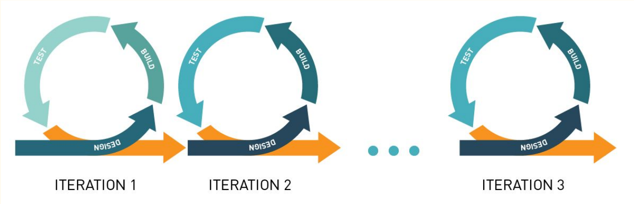
Problem: Customer satisfaction is down due to an increase in flight delays.

Solution: Machine Learning modelling to retrieve insights into the cause of these delays.

Workflow

Model Process

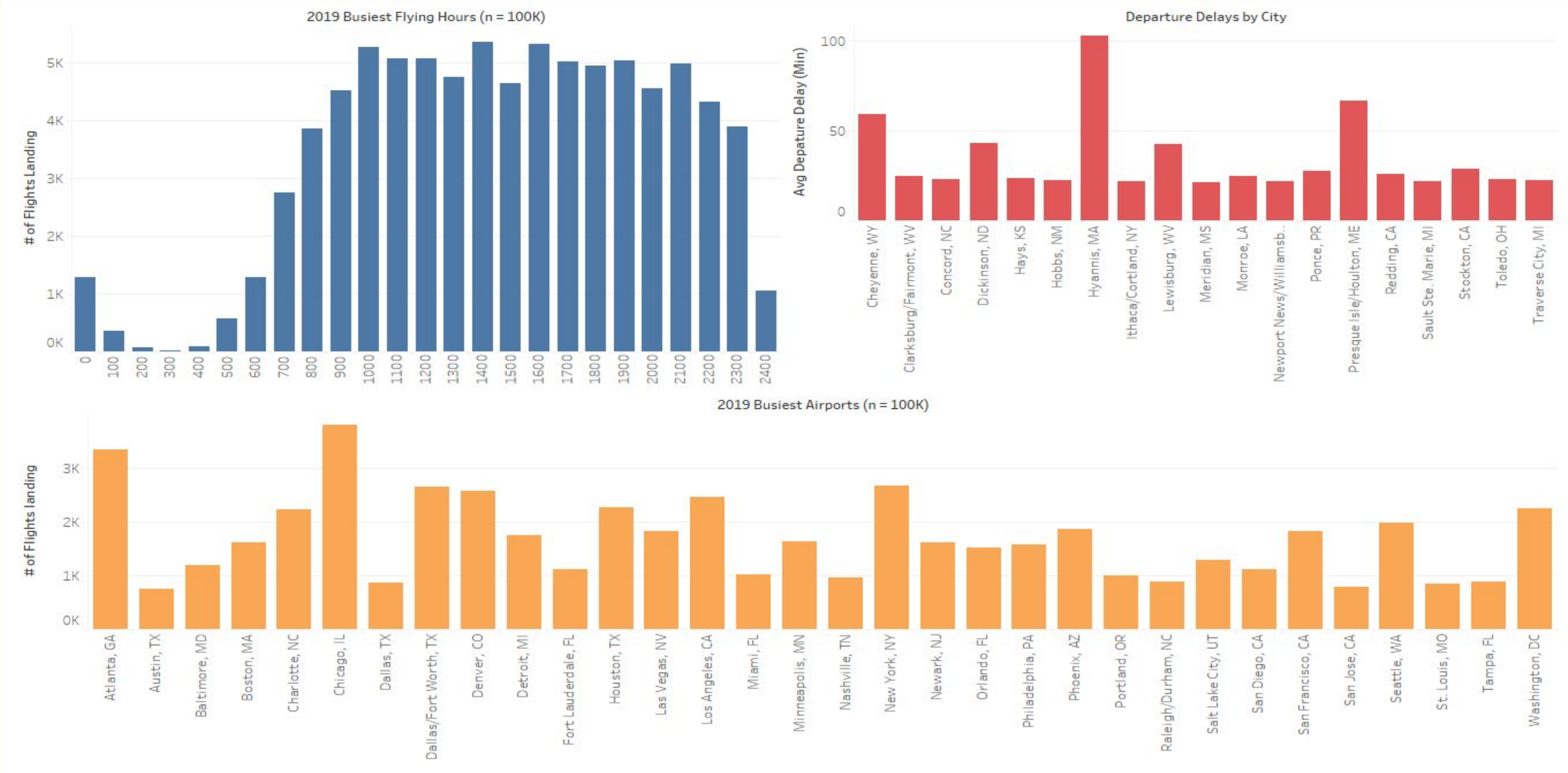




= Final Model

- Adjust features
- Add weather data (type and severity)
- Merge with Fuel and Passenger tables

Potential Causes of Arrival Delays



Imbalanced Dataset / Outliers

Criteria: An arrival delay is defined as a delay greater than 15 minutes. (https://www.fly.faa.gov/flyfaa/usmap.jsp)

Imbalance: 18.6% of our sample data is considered delayed.

Solution: Take a subsample of the data to perform our analysis.

Handling Outliers: Arrival delays that were greater than 180 minutes were removed.

Model 1 (Linear Regression)

OLS Regression Results

Dep. Variable:	arr_delay	R-squared:	0.014
Model:	OLS	Adj. R-squared:	0.014
Method:	Least Squares	F-statistic:	28.93
Date:	Thu, 29 Jul 2021	Prob (F-statistic):	1.54e-50
Time:	12:37:40	Log-Likelihood:	-25866.
No. Observations:	18320	AIC:	5.175e+04
Df Residuals:	18310	BIC:	5.183e+04
Df Model:	9		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	1.561e-16	0.007	2.13e-14	1.000	-0.014	0.014
crs_dep_time	-0.0113	0.010	-1.179	0.238	-0.030	0.007
arrival_hour	0.0650	0.011	6.118	0.000	0.044	0.086
arr_time	-0.0924	0.008	-10.947	0.000	-0.109	-0.076
actual_elapsed_time	-0.0648	0.020	-3.256	0.001	-0.104	-0.026
air_time	-0.0796	0.021	-3.711	0.000	-0.122	-0.038
distance	0.0986	0.048	2.075	0.038	0.005	0.192
miles_per_min	0.0126	0.016	0.767	0.443	-0.020	0.045
total_taxi_time	0.0452	0.009	4.828	0.000	0.027	0.063
avg_dest_taxi_time	0.0685	0.008	8.845	0.000	0.053	0.084
dest_traffic	0.0211	0.008	2.796	0.005	0.006	0.036

Omnibus:	20711.105	Durbin-Watson:	1.959
Prob(Omnibus):	0.000	Jarque-Bera (JB):	2366131.014
Skew:	5.845	Prob(JB):	0.00
Kurtosis:	57.434	Cond. No.	4.44e+15

Model 2 (Random Forest Regressor)

Feature Engineering

- Expected Miles per Minute
- Total Taxi Time
- Destination Traffic (Number of flights landing at each airport)
- Average Origin Taxi Time

Results

- MAE = 33.7845
- R^2 = 0.5008

Model 3 (Random Forest Regressor)

Feature Engineering (Exclusions)

- Total Taxi Time
- Destination Traffic

Feature Engineering (Additions)

- Average Destination Traffic Time
- Average Origin Departure Delay

Results

- MAE = 12.5642
- $R^2 = 0.5905$

Results

Feature Significance

- Average Origin Delay significantly improved the model.

Potential Scheduling Adjustments

- Prior to finalizing flight details, add additional time when departing from cities with a large departure delay.
- Plan flights from 12am-7am for the cities with the largest departure delays.
- Preemptively warn customers of potential delay to increase customer satisfaction

Challenges

Data Cleaning/Exploration

- Duplicate Data
- Missing Values
- Joining tables

Feature Selection/Engineering

- Over 80 different features
- Finding correlated features

References

- https://github.com/lighthouse-labs/mid-term-project-I
- https://www.fly.faa.gov/flyfaa/usmap.jsp
- https://en.wikipedia.org/wiki/Flight_length
- Bureau of Transportation Statistics