Shirley Li
10/22/22

Final Report:
Predicting RDU Flight Delays and Cancellations

Problem statement

Flying has always been stressful, and even more so in our post-covid world. These days it seems like flying has gotten worse: more delayed/cancelled flights, more missed connections, more lost luggage, and far worse airline customer service than previously. And flights are more expensive than ever!

I hope to alleviate at least one piece of the puzzle: delays and cancellations. I created a model that predicts whether flights leaving Raleigh-Durham airport in North Carolina will be on time, delayed by less than 1 hour, less than 2 hours, more than 2 hours, or cancelled (a multiclass classification model). My random forest classifier model had the following metrics on the test set:

- Accuracy: 0.89
- Precision: 0.89
- Recall: 0.89
- F1-Score: 0.89
- ROC-AUC score: 0.96

The Random Forest Classifier had the following hyperparameters found through a RandomizedSearchCV: {'n_estimators': 425, 'max_features': 'auto', 'criterion': 'gini'}.

Travellers in the RDU area will be able to use the tool to decide whether or not to book certain flights and airlines on certain days of the week, month, and year to selected destinations. Travel agencies booking from the RDU airport will also be able to take this into consideration. Hopefully having this tool will ameliorate some of the unavoidable stress of flying!

With more widespread use airlines and the RDU airport might be held accountable for their delays/cancellations.

Data Wrangling

I began with a dataset from Kaggle: Airline Delay and Cancellation Data, 2009 - 2018. I found this unsatisfactory, however, as so much has changed in the world and airline industry since 2018. The data may as well be irrelevant for today!

The Kaggle data was sourced from the United States Department of Transportation Bureau of Transportation Statistics. I decided to download my own dataset: flights all over the US over the time period July 2021–June 2022 (the latest available at the time). I would have liked to do a longer period of time, say 2018-2022. But it was already a humongous amount of data and I nearly didn't have enough space to save it on my computer!

For each month from July 2021–June 2022 I downloaded the following columns from the website:

Quarter = QUARTER
Month = MONTH
DayofMonth = DAY_OF_MONTH
DayOfWeek = DAY_OF_WEEK
FlightDate = FL_DATE
Operating_Airline = OP_UNIQUE_CARRIER
Flight_Number_Operating_Airline = OP_CARRIER_FL_NUM
Origin = ORIGIN
OriginCityName = ORIGIN_CITY_NAME
OriginStateName = ORIGIN_STATE_NM
Dest = DEST
DestCityName = DEST_CITY_NAME
DestStateName = DEST_STATE_NM
CRSDepTime = CRS_DEP_TIME
DepTime = DEP_TIME
DepDelayMinutes = DEP_DELAY_NEW
DepDel15 = DEP_DEL15
DepartureDelayGroups = DEP_DELAY_GROUP
CRSArrTime = CRS_ARR_TIME
ArrTime = ARR_TIME
ArrDelayMinutes = ARR_DELAY_NEW
ArrDel15 = ARR_DEL15
ArrivalDelayGroups = ARR_DELAY_GROUP
Cancelled = CANCELLED
CRSElapsedTime = CRS_ELAPSED_TIME
ActualElapsedTime = ACTUAL_ELAPSED_TIME
AirTime = AIR_TIME
Distance = DISTANCE
CarrierDelay = CARRIER_DELAY
WeatherDelay = WEATHER_DELAY
NASDelay = NAS_DELAY
SecurityDelay = SECURITY_DELAY
LateAircraftDelay = LATE_AIRCRAFT_DELAY

The concatenated dataset was 605186 rows × 33 columns.

To wrangle the data, I first dealt with miscategorized data types and changed them to the correct ones: categories, objects, date_time, etc. I also wanted to make the data more readable so I renamed the Carriers column. I wasn't quite ready to decide what columns to drop just yet.

Instead, I decided to subset the dataset so I didn't have as many values overall. First I tried subsetting by NC flights, but this was still too big. Eventually I decided to only work with departing flights from RDU. This final dataset was 52645 rows x 33 columns.

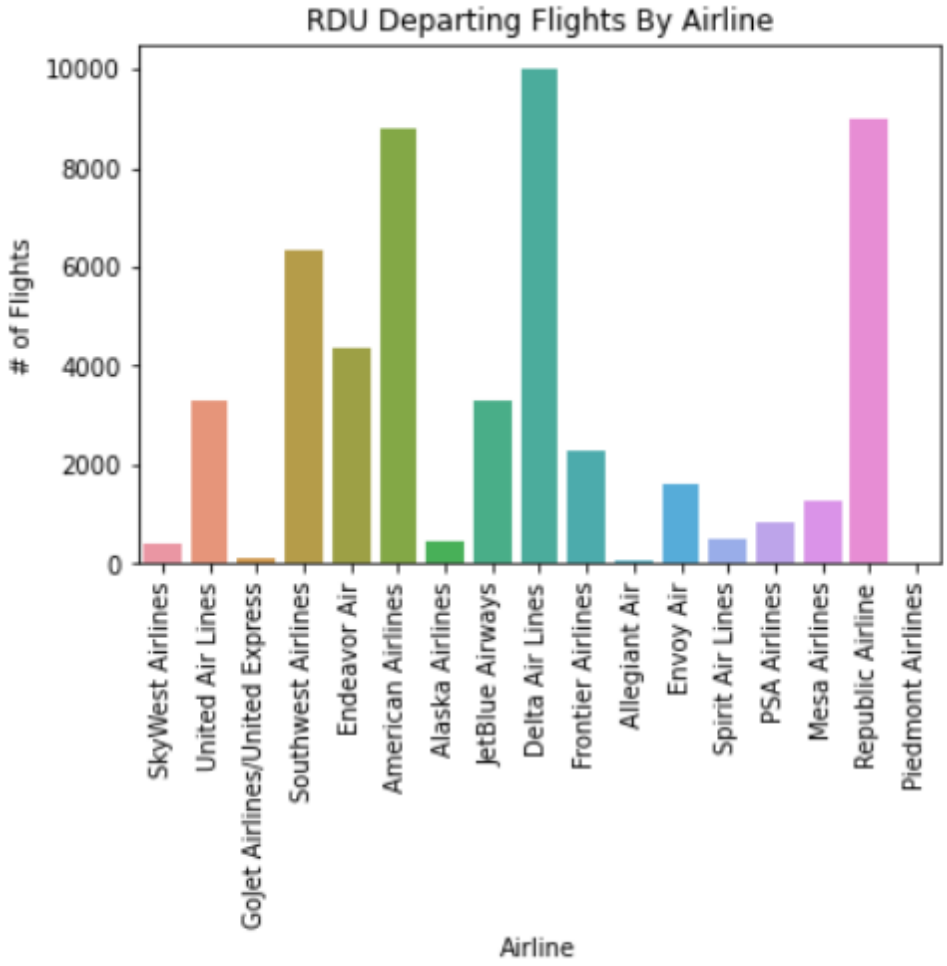
Exploratory Data Analysis

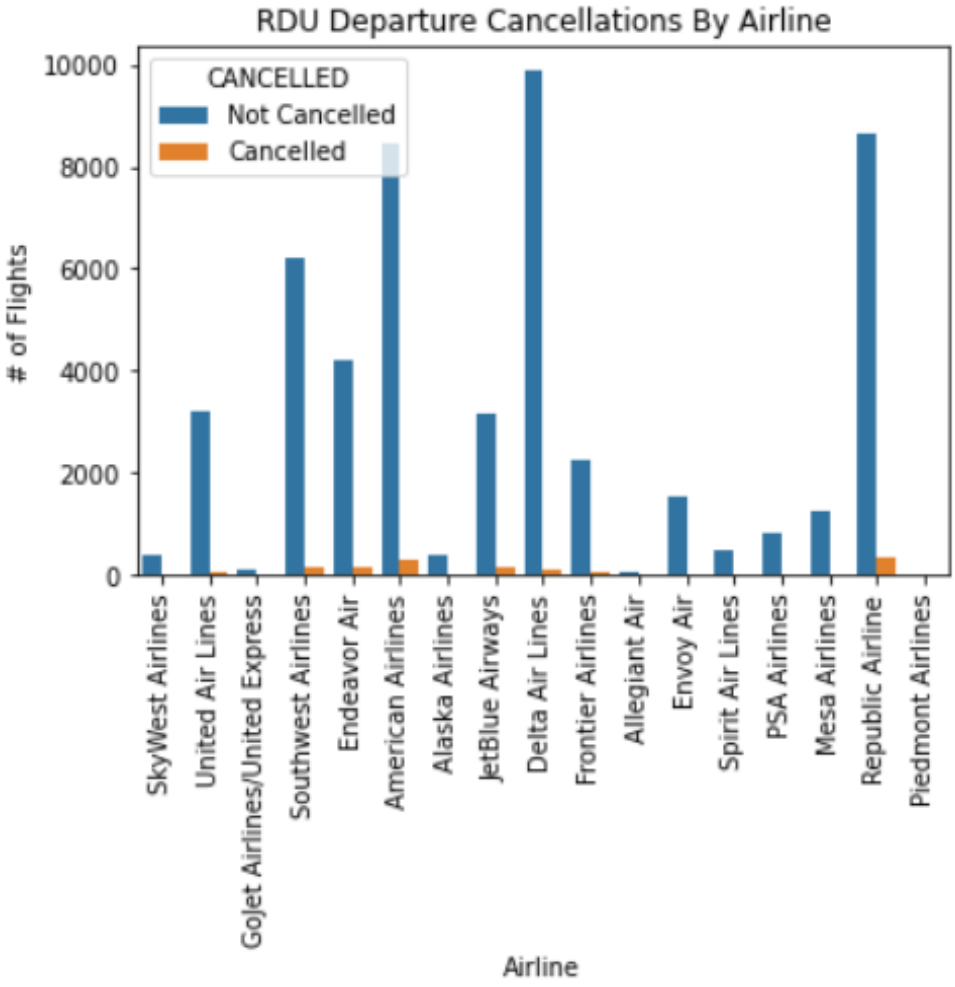
For my exploratory data analysis, I decided to subset my RDU dataset into two simple binary classification problems: flight delays (DEP_DEL15, more than 15 min delay) and flight cancellations (CANCELLED). Later on I decided to use a more specific result: on time, delayed by less than 1 hour, less than 2 hours, more than 2 hours, or cancelled (a multiclass classification model).

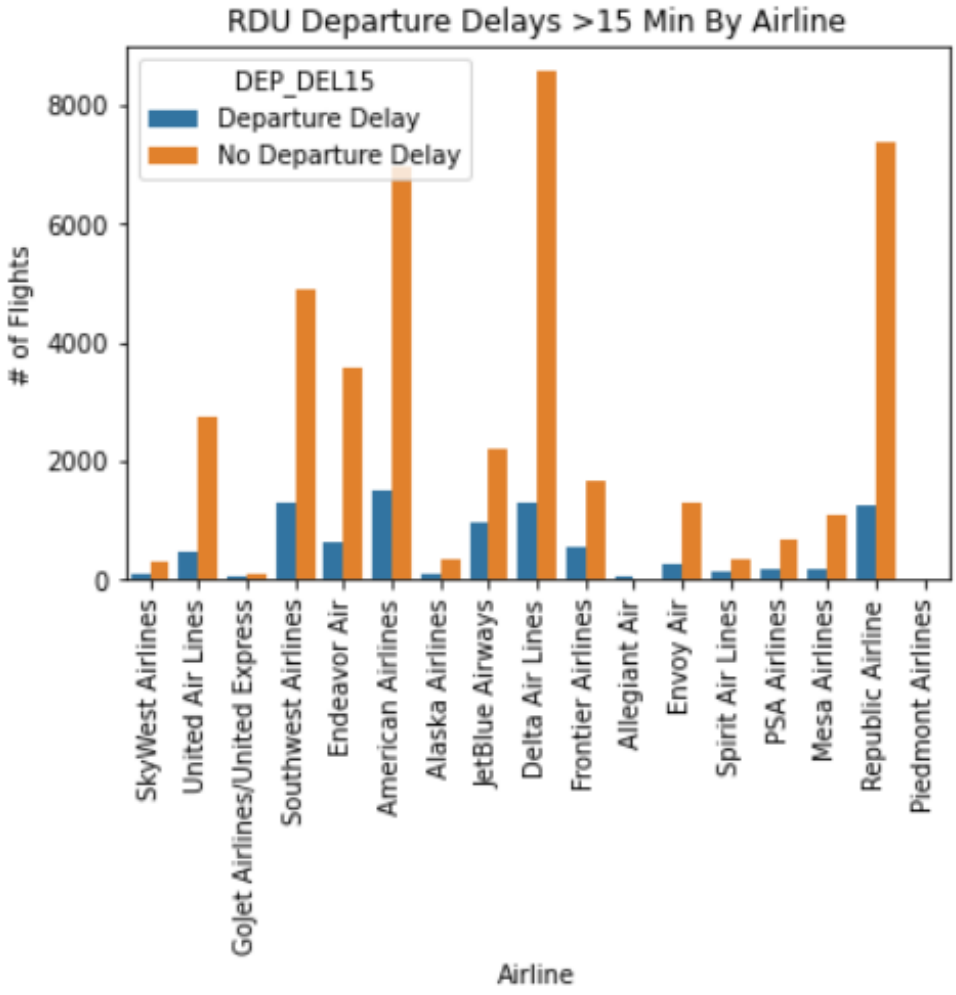
I wanted to see if there were particular features that affected delays and cancellations.

1. Airline

Some airlines are more popular at RDU. Are there worse offenders for flight delays and cancellations? I wanted to capture the fact that some airlines have more departing flights overall so shouldn't be penalized on that account.


RDU Departing Flights By Airline
# of Flights
10000
8000
6000
4000
2000
0
SkyWest Airlines
United Air Lines
Goljet Airlines/United Express
Southwest Airlines
Endeavor Air
American Airlines
Alaska Airlines
JetBlue Airways
Delta Air Lines
Frontier Airlines
Allegiant Air
Envoy Air
Spirit Air Lines
PSA Airlines
Mesa Airlines
Republic Airline
Piedmont Airlines
Airline

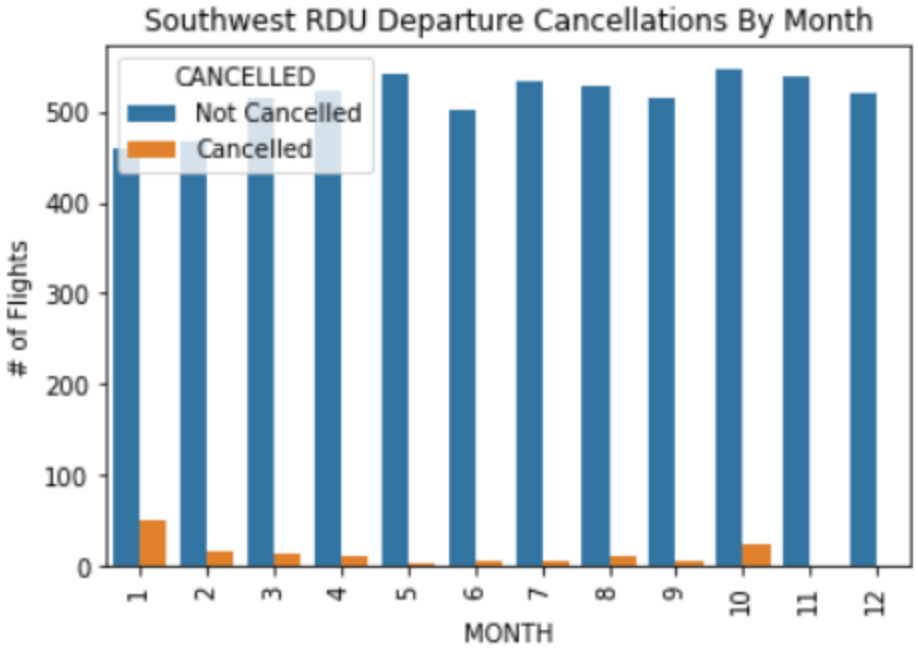

RDU Departure Cancellations By Airline
CANCELLED
Not Cancelled
Cancelled
# of Flights
10000
8000
6000
4000
2000
0
SkyWest Airlines
United Air Lines
Goljet Airlines/United Express
Southwest Airlines
Endeavor Air
American Airlines
Alaska Airlines
JetBlue Airways
Delta Air Lines
Frontier Airlines
Allegiant Air
Envoy Air
Spirit Air Lines
PSA Airlines
Mesa Airlines
Republic Airline
Piedmont Airlines
Airline

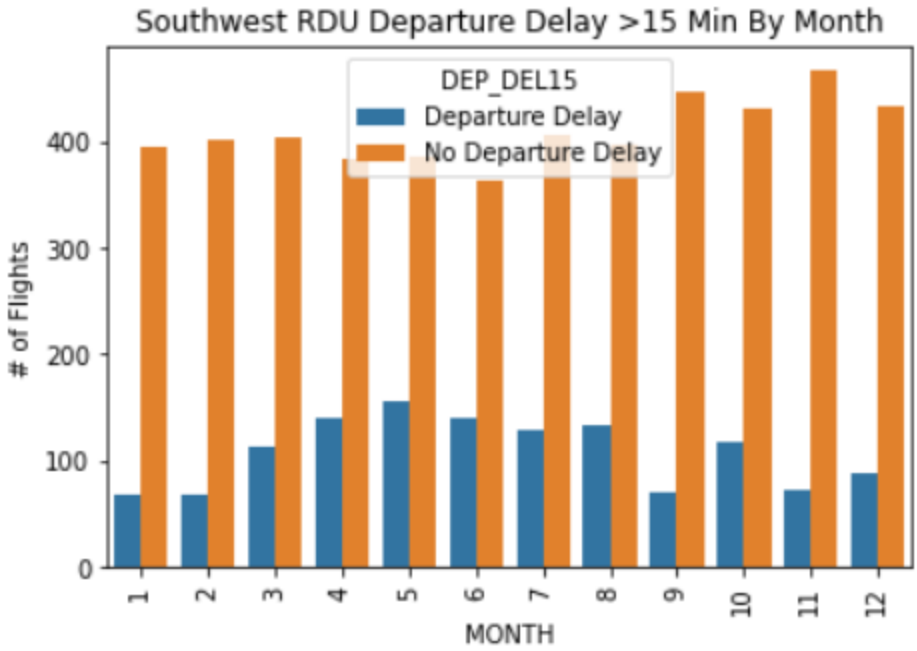


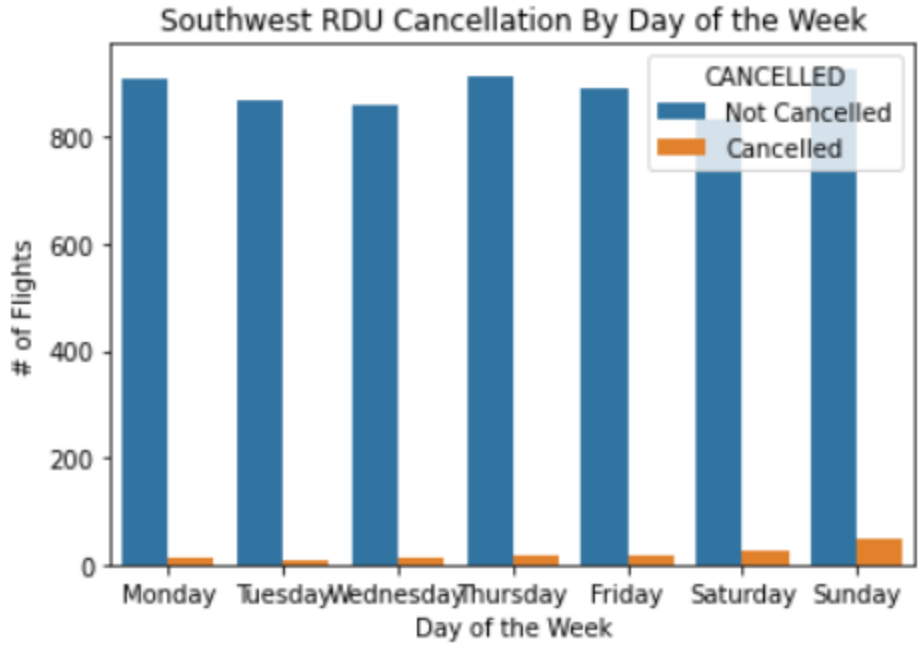
As far as flight cancellations, Delta is an extremely popular carrier, but its cancellation rate is lower than American, Republic, Endeavor, Southwest, and Jetblue. As far as flight delays, Jetblue looks like a fairly bad offender: nearly half of its flights were delayed by at least 15 minutes!

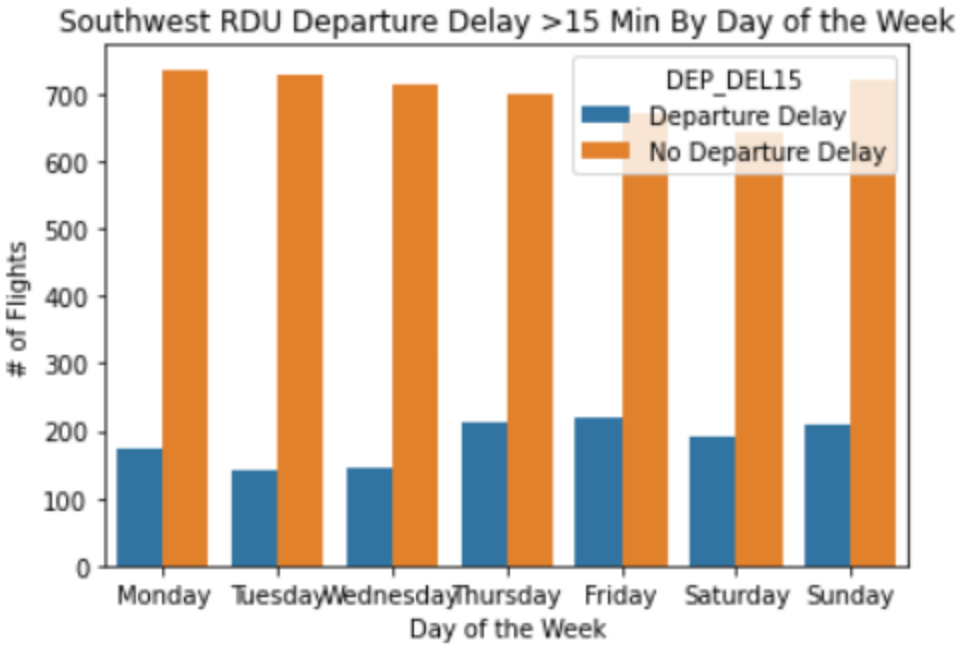
2. Month and Day of the Week

I was also curious whether specific dates affected delays and cancellations. I arbitrarily chose to subset on Southwest (who has their own terminal at RDU–the airport has two terminals total). I looked at different months and different days of the week.


Southwest RDU Departure Cancellations By Month
CANCELLED
Not Cancelled
Cancelled
# of Flights
0
100
200
300
400
500
1
2
3
4
5
6
7
8
9
10
11
12
MONTH


Southwest RDU Departure Delay >15 Min By Month
DEP_DEL15
Departure Delay
No Departure Delay
# of Flights
0
100
200
300
400
1
2
3
4
5
6
7
8
9
10
11
12
MONTH

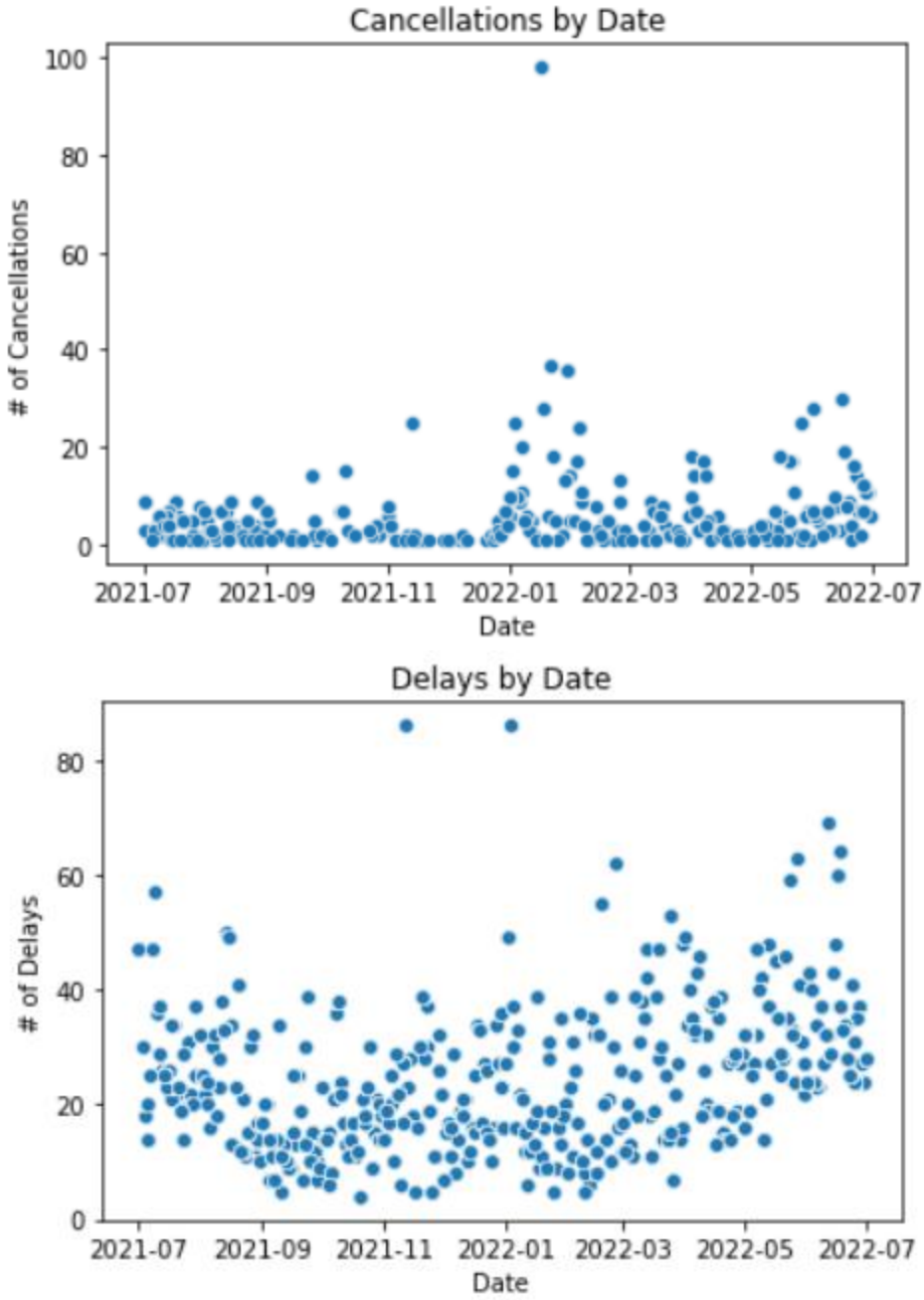




Southwest had the most cancelled flights in January and October, and on Sundays. It had the most flight delays of >15 min in May and Thursday, Friday, and Sunday.

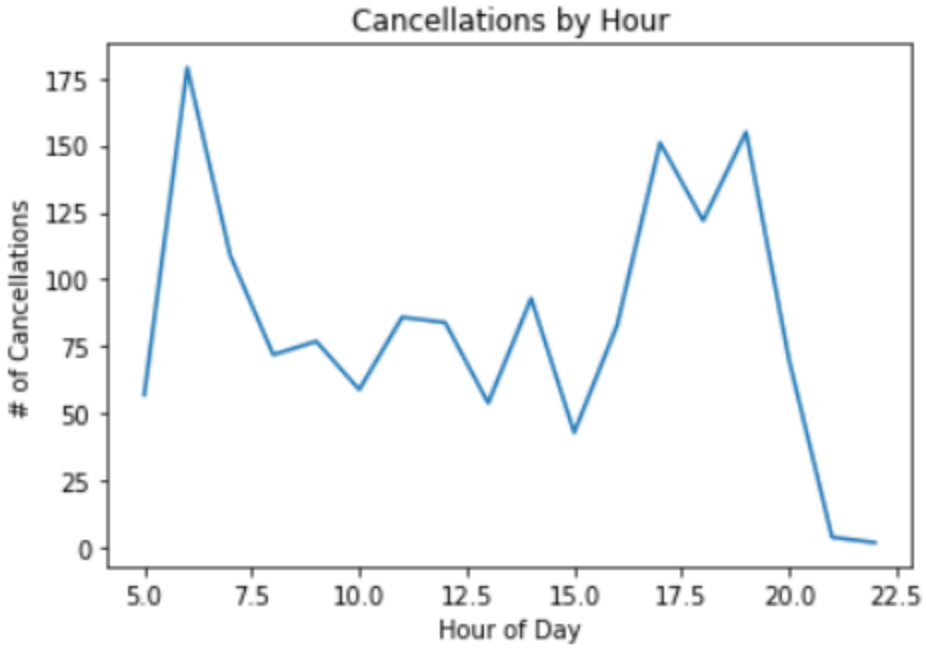
3. Dates

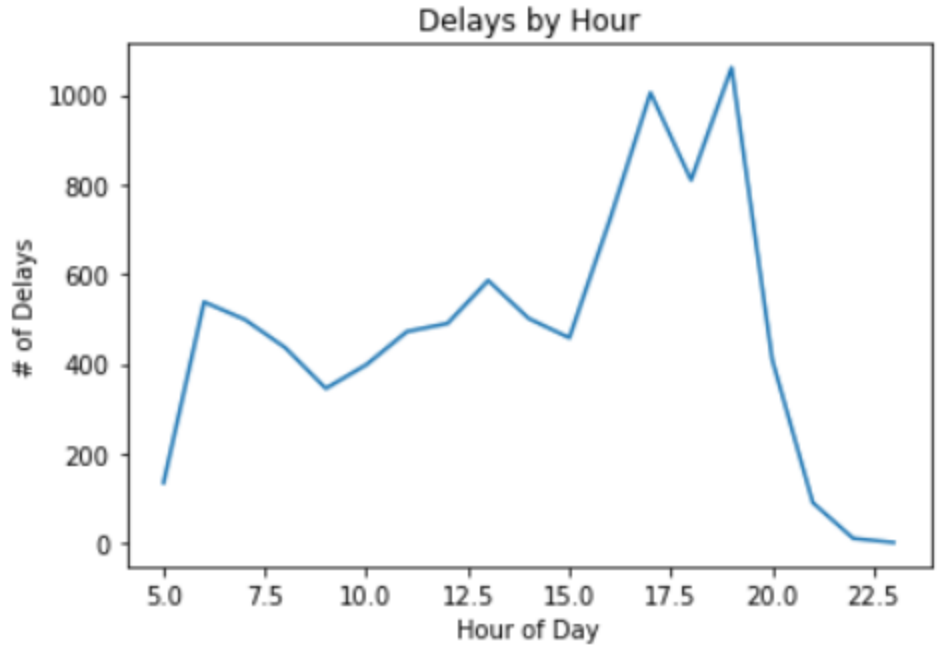
What about dates overall?



Cancellations seem worst in January. Delays are fairly evenly spread but slightly better in the fall from Sept-Nov.

4. Time of Day





It looks like cancellations are worse early in the day around 6am and later in the day around 5-8pm. Delays are worst between 5-7pm.

If nothing else, we've demonstrated that these features are somewhat important to our question!

Data Preprocessing

Based on what I learned through my Exploratory Data Analysis, I could make some sensible decisions about how to move forward.

Setting a Multiclass Classification Problem

The first thing that I did was decide on the problem specifically–which could take many forms. We have delays of varying time frames and we have cancellations. We solve a binary classification problem (on time or delayed/cancelled), or a multiclass classification problem (on time, delayed by 15 min, 30 min, 45 min, etc. all the way up to 180 min, or cancelled).

I decided to create a simpler multiclass classification problem. I wanted something that would be specific enough to be useful to travellers, but not so specific that it would be difficult to train. I created a new result column:

0 = no delay
1 = delay of 1 hour or less
2 = delay of 2 hours or less
3 = delay of more than 2 hours
4 = cancelled

This is what my model will predict.

Feature Selection

Next I pared down my features. I didn't want collinearity so I got rid of redundant data (a few columns measured similar information in slightly different ways). I decided that I only cared about departure delay and not about arrival delay. My new dataset had 52645 rows x 12 columns.

#	Column	Non-Null Count	Dtype
0	QUARTER	52645 non-null	category
1	MONTH	52645 non-null	category
2	DAY_OF_MONTH	52645 non-null	category
3	DAY_OF_WEEK	52645 non-null	category
4	CARRIER	52645 non-null	object
5	FL_NUM	52645 non-null	object
6	DEST	52645 non-null	object
7	CRS_DEP_TIME	52645 non-null	object
8	CRS_ARR_TIME	52626 non-null	object
9	CRS_ELAPSED_TIME	52645 non-null	float64
10	DISTANCE	52645 non-null	float64
11	RESULT	52645 non-null	category

Other Feature Engineering/Preprocessing

Finally I performed other engineering/preprocessing in order to get the data ready for modelling.

1. Binning

I divided departure and arrival time into four bins: Morning (3:00am to 9:00am), Midday (9:00am to 3:00pm), Afternoon_Evening (3:00pm to 9:00pm), and Late_Night (9:00pm to 3:00am). These bins were more useful than the exact time of day.

2. Resampling

I resampled the data to correct our data imbalance between “on time” and the other possibilities. An on time departure was roughly four times as likely as all the other possibilities combined, and posed problems for modelling.

3. Dummy variables

I created dummy variables for our categorical and object types: 'QUARTER', 'MONTH', 'DAY_OF_MONTH', 'DAY_OF_WEEK', 'CARRIER', 'DEST', 'DEP_TIME_BINS', and 'ARR_TIME_BINS.'

4. Train/test sets

I split the data into X and y and divided it into training and test sets.

5. Scaling

Once I had the train/test sets, I was able to scale/standardize the numeric features 'CRS_ELAPSED_TIME' and 'DISTANCE' using a logistic scaler. I fit on the training set and transformed the training and test sets.

Modelling

Model Selection

I wanted to test a variety of models:

0. Dummy Classifier
1. KNN Classifier
2. Logistic Regression
3. Random Forest Classifier
4. Gradient Boosting Classifier
5. (SVM Classifier)

Challenges with modelling

Imbalanced Dataset

Initially I did not resample my dataset. As we recall we had a highly skewed/imbalanced dataset with far more on time departures than delays and cancellations. As a result my models had good ROC-AUC scores but their precision, recall, and F1 scores did not outperform the dummy classifier.

Once I resampled my dataset, this problem disappeared.

Long run times

A few of my models were apparently very computationally expensive and would not run even after waiting hours! The Support Vector Machine Classifier never ran so I removed it from my final documentation. I also could not run a robust hyperparameter search with RandomizedSearchCV on Logistic Regression or the Gradient Boosting Classifier. They simply took too long.

The best model

We achieved the best result with Random Forest. Both the KNN Classifier (very close) and the Random Forest Classifier far outperformed the Dummy Classifier, Logistic Regression, and the Gradient Boosting Classifier.

This is how the different models performed on the test set:

	Accuracy	**Precision**	**Recall**	**F1 Score**	**ROC-AUC**
Dummy Classifier	0.5	0.25	0.5	0.33	0.5
KNN	0.85	0.86	0.85	0.85	0.96
Logistic Regression	0.52	0.39	0.52	0.43	0.62
Random Forest	0.89	0.89	0.89	0.89	0.96
Gradient Boosting	0.55	0.56	0.55	0.48	0.72

Since we have a multiclass classification problem, we can't graph the ROC-AUC curve.

The Random Forest Classifier had the following hyperparameters found through a RandomizedSearchCV: {'n_estimators': 425, 'max_features': 'auto', 'criterion': 'gini'}.

Takeaways

1. For Travellers

Given this model, travellers hoping to avoid delays and cancellations can predict what will happen with their flights. The model has a reasonable amount of specificity. It predicts whether

a flight will be on time, delayed by less than 1 hour, less than 2 hours, more than 2 hours, or cancelled. If a traveller doesn't mind a delay of less than 1 hour, they can book any flight that falls into either category: on time or delayed by less than 1 hour.

2. For Travel Agencies

Travel agencies booking flights out of RDU will be able to determine whether or not given flights meet their criteria.

3. For Airlines and RDU Airport

Finally if travellers/travel agencies start using this tool, airlines will be held accountable for their record of delays/cancellations. Those with a worse record will not receive as many bookings, and this will affect their bottom line.

Future Research

1. Thresholding

At present we haven't set the threshold for positive classifications for each category. To do so we'd need to create a different process for the model: start with a binary classification model and build a multiclass model on top of that.

However thresholding could have a few major benefits. First an optimal threshold value could improve our model significantly. Second it could more clearly quantify uncertainty for travellers and travel agencies.

2. More Dates/Airports

Due to data limitation constraints, I only used flights departing from the RDU airport from July 2021–June 2022. Given more CPU and computational power, I would definitely use a longer period of historical flight data, say from 2018-2022. In addition, we could expand this model to flights departing from and arriving at RDU, flights departing from and arriving in NC, and even flights from/to the most popular airports in the US.

3. Adding Features Like Weather Forecast

It would also be worthwhile to add commonsense features like weather forecast. Everyone knows that extreme weather like rain or snow, icy conditions, and other events affects flights! However, I couldn't find an easy API that could pull historical weather data. The Wunderground API was taken offline in 2018 and the other alternatives are complicated and/or costly.