



Prediction of US Airline Delays

Dennis Hofmann, Avantika Shrestha,
Isaac Zhao, Riddhi Thakkar, Edith Gomez

Introduction

- + Almost 3 million people fly in and out of U.S. airports a day ^[1]
- + In 2021, 16.85% (1.1 million) US flights were delayed ^[2]
- + Every minute of delay costs airlines about \$78 ^[3]



[1]<https://www.cnbc.com/2013/11/27/chart-of-the-day-airlines-losing-78minute-today.html>
[2]<https://www.transtats.bts.gov/HomeDrillChart.asp>
[3]https://www.faa.gov/air_traffic/by_the_numbers/

Objectives and Motivation

- + Goal: Predicting if a flight will be delayed will save airlines money and customer frustrations
- + Data can be used to optimize airline's planning



Dataset Description

- + Airline Reporting Carrier On-Time Performance Dataset ^[1]
 - + The Bureau of Transportation Statistics of the United States
 - + US domestic flights between 1987 and 2020
 - + Approximately 200 million domestic US flights
- + Daily Weather Data ^[2]
 - + Summary of the weather conditions member countries of the WMO
 - + Historical data from 1929 to the present
 - + Data from 1973 to the present being the most complete

[1]https://developer.ibm.com/exchanges/data/all/airline/?mhsrsrc=ibmsearch_a&mhq=+delay

[2]<https://www.ncei.noaa.gov/access/search/data-search/global-summary-of-the-day>

Airport Dataset

Feature	Description
Year	Year
Month	Month
DayofMonth	Day of Month
DayOfWeek	Day of Week (numeric)
FlightDate	Date of Flight
Flight_Number_Reporting_Airline	Flight Number
OriginCityName	Origin City Name
OriginState	Origin State
DestCityName	Destination City Name
DestState	Destination State
CRSArrTime	Computer Reservation System (scheduled) Arrival Time
CRSDepTime	Computer Reservation System (scheduled) Departure Time
ArrDelay	Arrival delay (minutes)
Diverted	1 = diverted
Distance	Distance between airports (miles)

Weather Dataset

Feature	Description
NAME	Airport name, state, and country
COORDINATES	Longitude and latitude
TEMP	Mean temperature (.1 Fahrenheit)
DEWP	Mean dew point (.1 Fahrenheit)
SLP	Mean sea level pressure (.1 mb)
STP	Mean station pressure (.1 mb)
VISIB	Mean visibility (.1 miles)
WDSP	Mean wind speed (.1 knots)
MXSPD	Maximum sustained wind speed (.1 knots)
GUST	Maximum wind gust (.1 knots)
MAX	Maximum temperature (.1 Fahrenheit)
MIN	Minimum temperature (.1 Fahrenheit)
PRCP	Precipitation amount (.01 inches)
SNDP	Snow depth (.1 inches)
FRSHTT	Indicator for occurrence of: Fog, Rain or Drizzle, Snow or Ice Pellets, Hail, Thunder, Tornado/Funnel Cloud

Questions the Dataset can answer

- + Can we predict if a flight will be delayed?
- + Which airport locations are prone to flight delays?
- + Does weather at different locations affect flight schedules differently?



Data Cleaning

- + Removed redundant variables
 - + Variables directly related to calculation of delay
 - + Removed variables with 40% missing values
 - + Target variable
- + Any flight with an arrival or departure delay greater than 15 minutes

FlightDate	Reporting_A	Origin	Dest	CRSArrTime	ArrTime	CarrierDelay	WeatherDelay	NASDelay	SecurityDelay	LateAircraftDelay	ArrDelay	ArrDelayMinutes
3/28/03	UA	LAX	JFK	615	617	NA	NA	NA	NA	NA	2	2
11/29/18	AS	LAX	JFK	1912	1851	NA	NA	NA	NA	NA	-21	0
8/28/15	UA	LAX	JFK	1634	1620	NA	NA	NA	NA	NA	-14	0
4/20/03	DL	LAX	JFK	619	616	NA	NA	NA	NA	NA	-3	0
11/30/05	UA	LAX	JFK	1653	1640	NA	NA	NA	NA	NA	-13	0
4/6/92	UA	LAX	JFK	2308	2248	NA	NA	NA	NA	NA	-20	0
12/3/12	VX	LAX	JFK	1900	1901	NA	NA	NA	NA	NA	1	1
12/1/04	HP	LAX	JFK	613	533	NA	NA	NA	NA	NA	-40	0
2/13/97	UA	LAX	JFK	1615	1640	NA	NA	NA	NA	NA	25	25
3/10/16	DL	LAX	JFK	712	719	NA	NA	NA	NA	NA	7	7
10/14/99	AA	LAX	JFK	717	722	NA	NA	NA	NA	NA	5	5
8/3/18	B6	LAX	JFK	839	913	11	0	23	0	0	34	34
6/1/06	AA	LAX	JFK	2332	2353	0	0	21	0	0	21	21

Joining the Datasets: Challenge

Airline Dataset

Feature	Description
Year	Year
Month	Month
DayOfMonth	Day of Month
DayOfWeek	Day of Week (numeric)
FlightDate	Date of Flight
Flight_Number_Reporting_Airline	Flight Number
OriginCityName	Origin City Name
OriginState	Origin State
DestCityName	Destination City Name
DestState	Destination State
CRSArrTime	Computer Reservation System (scheduled) Arrival Time
CRSDepTime	Computer Reservation System (scheduled) Departure Time
ArrDelay	Arrival delay (minutes)
Diverted	1 = diverted
Distance	Distance between airports (miles)

Weather Dataset

Feature	Description
NAME	Airport name, state, and country
COORDINATES	Longitude and latitude
TEMP	Mean temperature (.1 Fahrenheit)
DEWP	Mean dew point (.1 Fahrenheit)
SLP	Mean sea level pressure (.1 mb)
STP	Mean station pressure (.1 mb)
VISIB	Mean visibility (.1 miles)
WDSP	Mean wind speed (.1 knots)
MXSPD	Maximum sustained wind speed (.1 knots)
GUST	Maximum wind gust (.1 knots)
MAX	Maximum temperature (.1 Fahrenheit)
MIN	Minimum temperature (.1 Fahrenheit)
PRCP	Precipitation amount (.01 inches)
SNDP	Snow depth (.1 inches)
FRSHTT	Indicator for occurrence of: Fog, Rain or Drizzle, Snow or Ice Pellets, Hail, Thunder, Tornado/Funnel Cloud

Joining Datasets: Solution

Weather Dataset

Airport Name	Latitude	Longitude
ALTOONA BLAIR CO AIRPORT, PA US	40.29639	78.32028
ALTURAS MUNICIPAL AIRPORT, CA US	41.49139	120.56444
ALTUS AFB, OK US	34.65000	99.26667
ALVA REGIONAL AIRPORT, OK US	36.77306	98.66972
AMARILLO AIRPORT, TX US	35.22950	101.70420
AMBLER AIRPORT, AK US	67.10000	157.85000
AMELIA LAKE PALOURD, LA US	29.70000	91.10000
AMERADA PASS, LA US	29.45000	91.33333
AMES MUNICIPAL AIRPORT, IA US	41.99056	93.61889
ANAKTUVUK AUTO, AK US	68.16667	151.76667
ANCHORAGE ELMENDORF AFB, AK US	61.25306	149.79361

- + Found a third dataset with Airport Name and City to join on
- + Airport Names are different
- + Latitude and Longitude are different
- + Solution: Match records by nearest neighbor with Euclidean distance

Airline Dataset with Airport City Information

Airport Name	Latitude	Longitude	State	City
Altoona Blair County Airport	40.29640	78.32000	PA	Altoona
Alturas Municipal Airport	41.48300	120.56500	CA	Alturas
Altus Air Force Base	34.66710	99.26670	OK	Altus
Alva Regional Airport	36.77320	98.66990	OK	Alva
Rick Husband Amarillo International Airport	35.21940	101.70600	TX	Amarillo
Ambler Airport	67.10630	157.85699	AK	Ambler
Lake Palourde Base Heliport	29.69330	91.09870	LA	Amelia
Berwick Shore Base Heliport	29.66330	91.24070	LA	Berwick
Ames Municipal Airport	41.99200	93.62180	IA	Ames
Anaktuvuk Pass Airport	68.13360	151.74300	AK	Anaktuvuk Pass
Elmendorf Air Force Base	61.25100	149.80701	AK	Anchorage

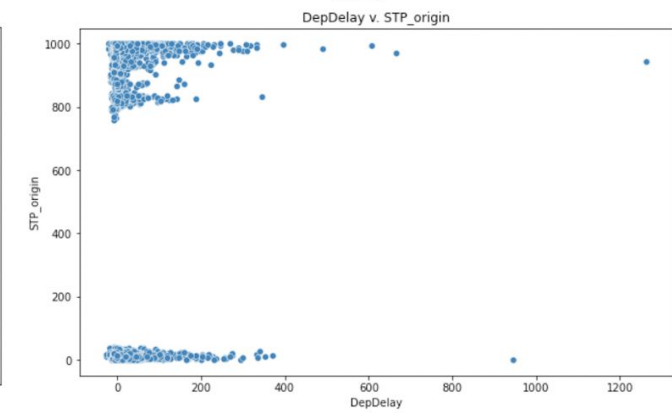
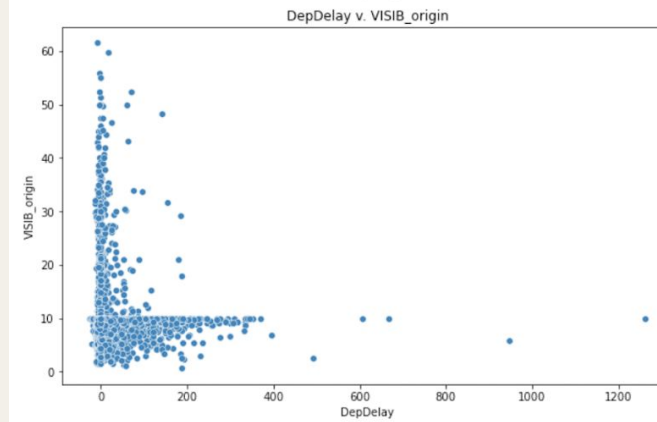
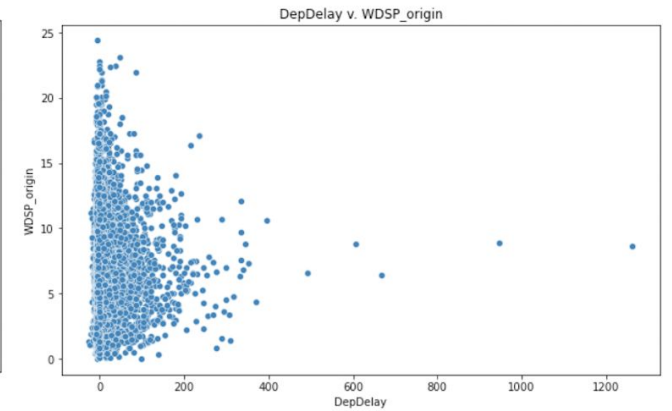
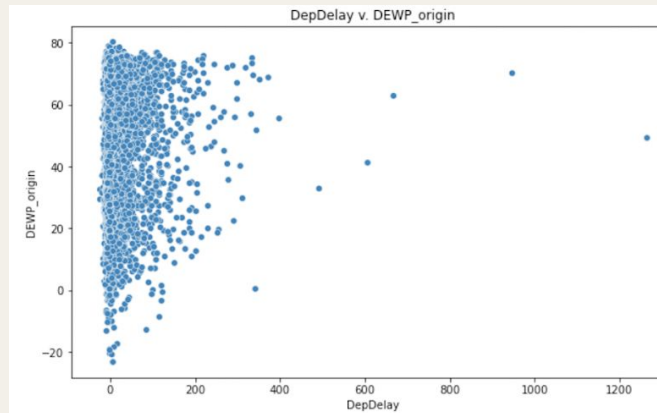
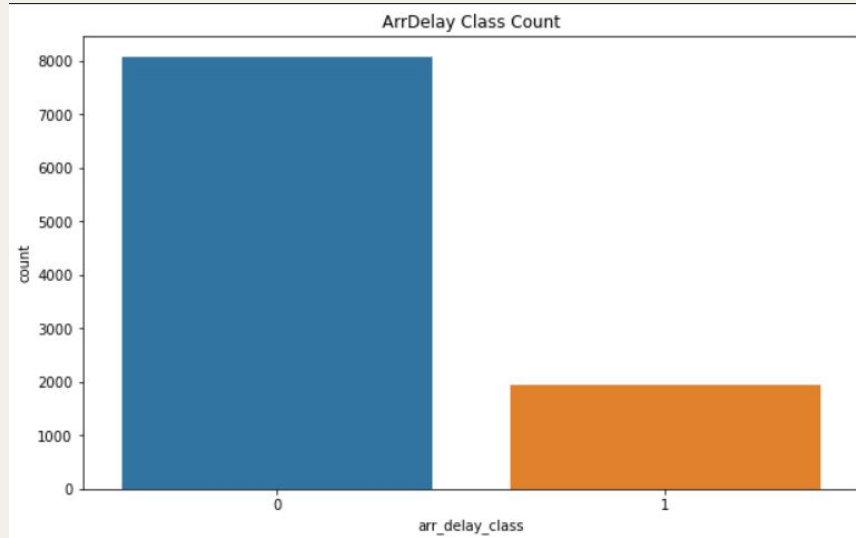
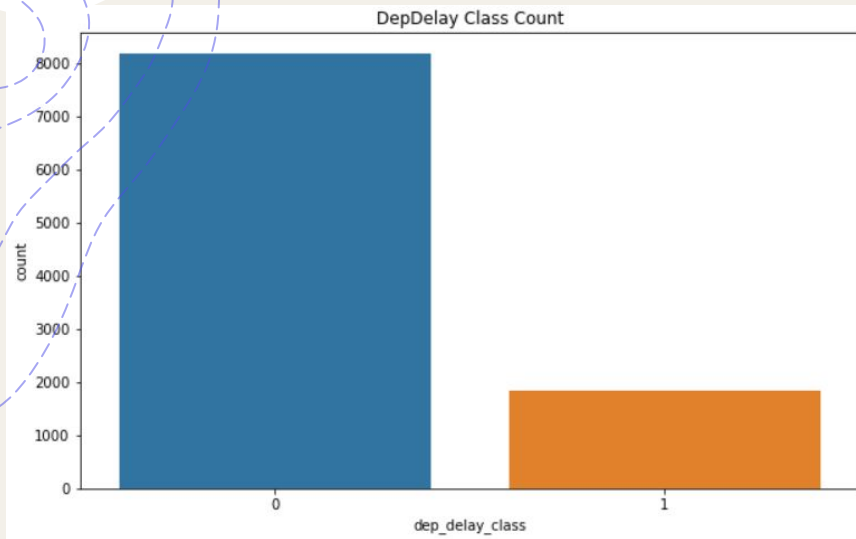
Reading In the Dataset

- + Memory issue
- + Solution:
 - + Using parameter `col_select` to only read in specific variables
 - + Using the parameter `chunksize` in the `read_csv` function

```
chunksize = 100000

tfr = pd.read_csv('full_df_2.csv', chunksize=chunksize, iterator=True, encoding='latin-1')
df = pd.concat(tfr, ignore_index=True)
```

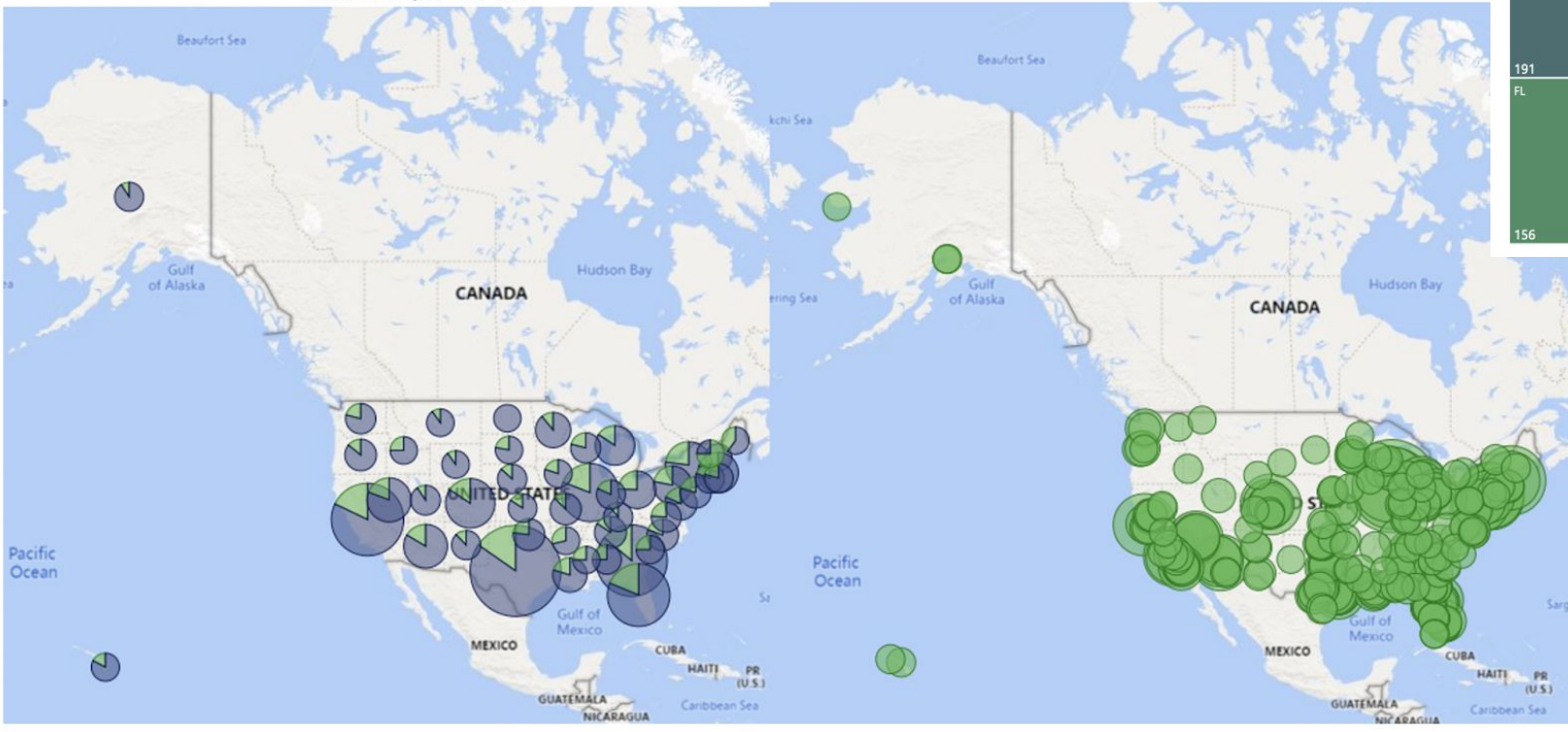
EDA Results



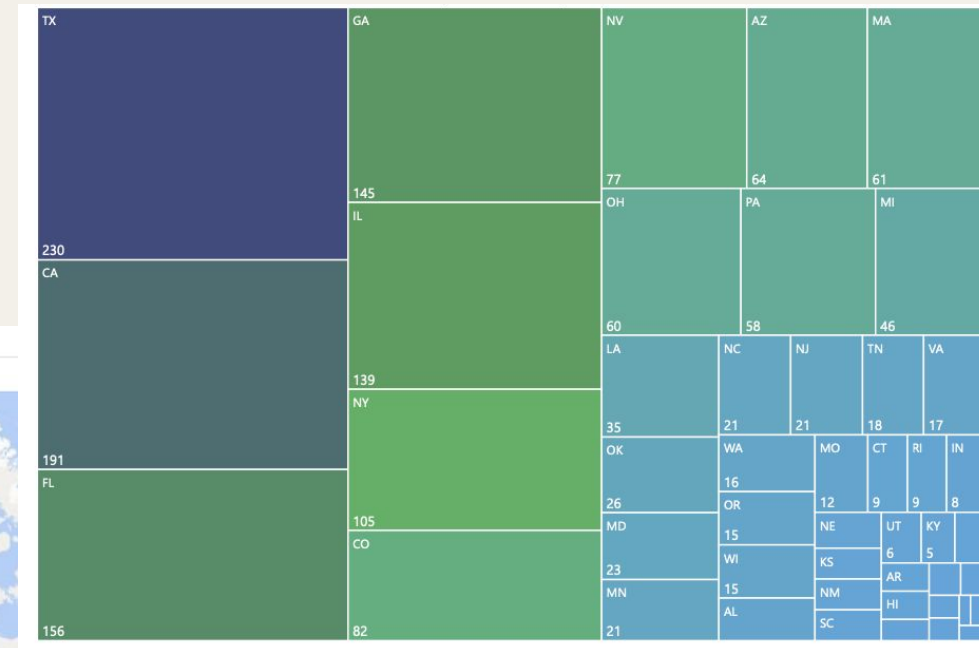
EDA Results

Flights by Destination Location

Status ● Delayed ● On Time



Number of Delayed Flights by destination Location

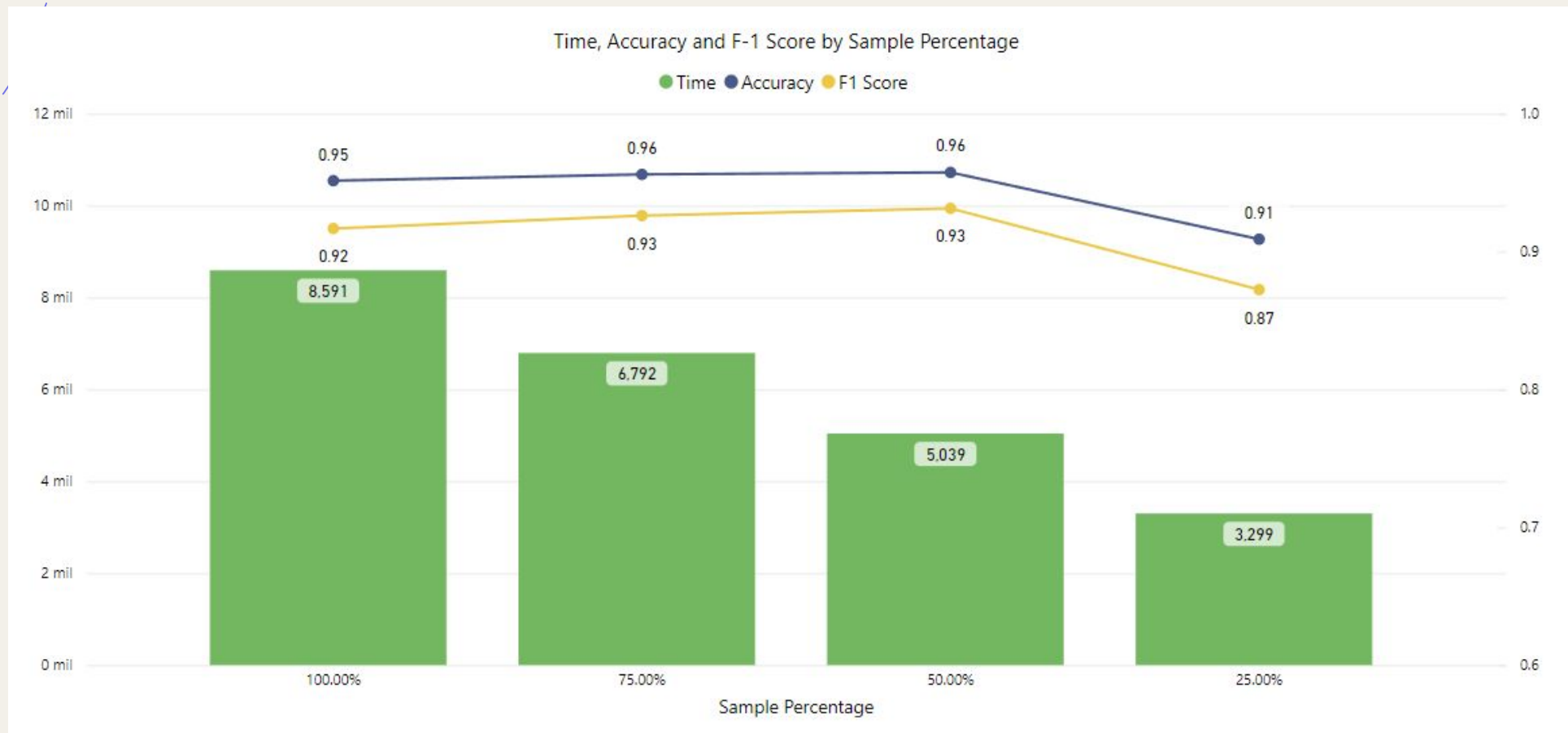




Machine Learning Method


- + Using Random Forest
 - + Faster, performs better, more interpretable

Class Imbalance

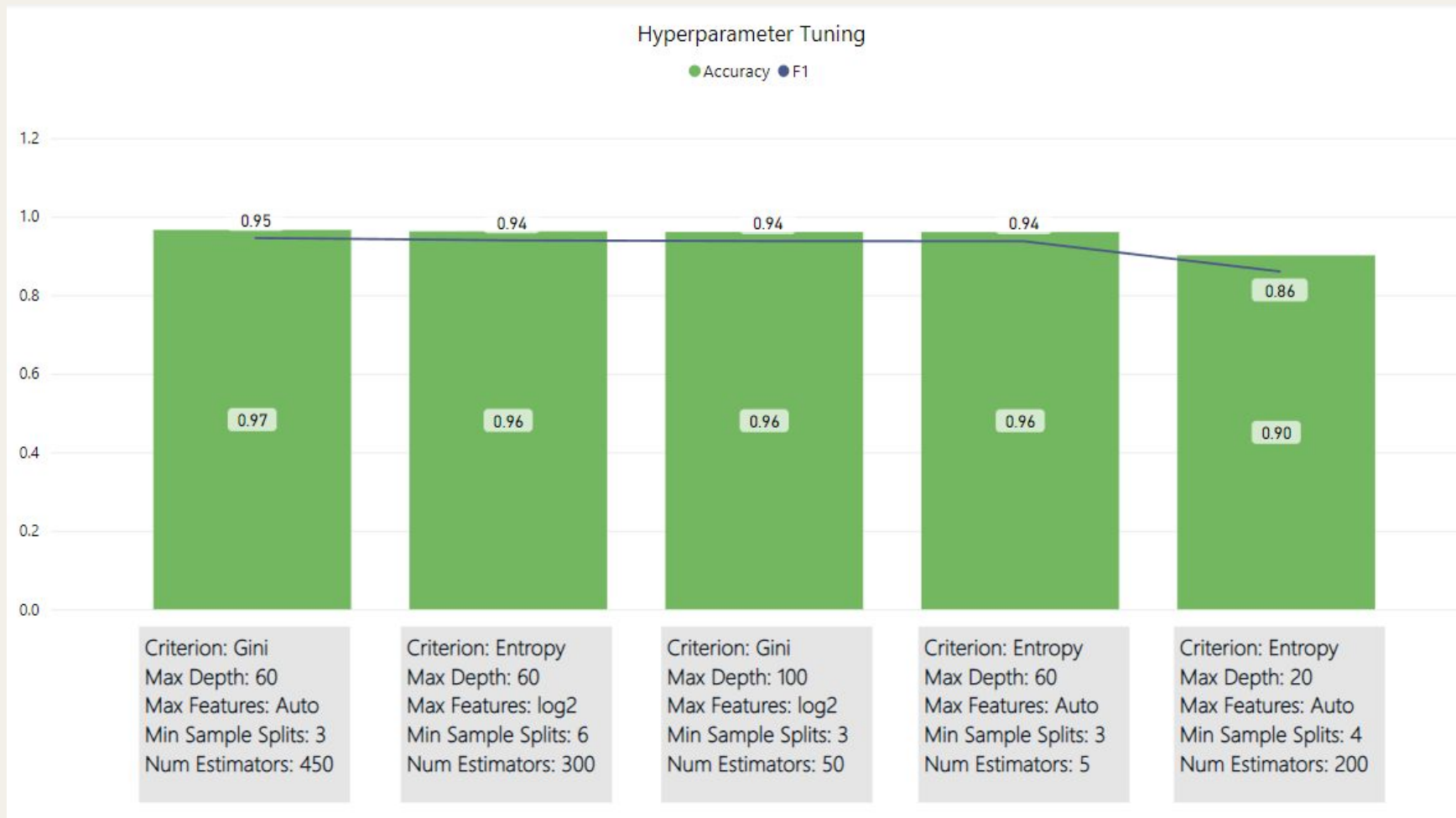


Random Forest Hyperparameter Tuning

- + Randomized Search CV
- + 15 runs

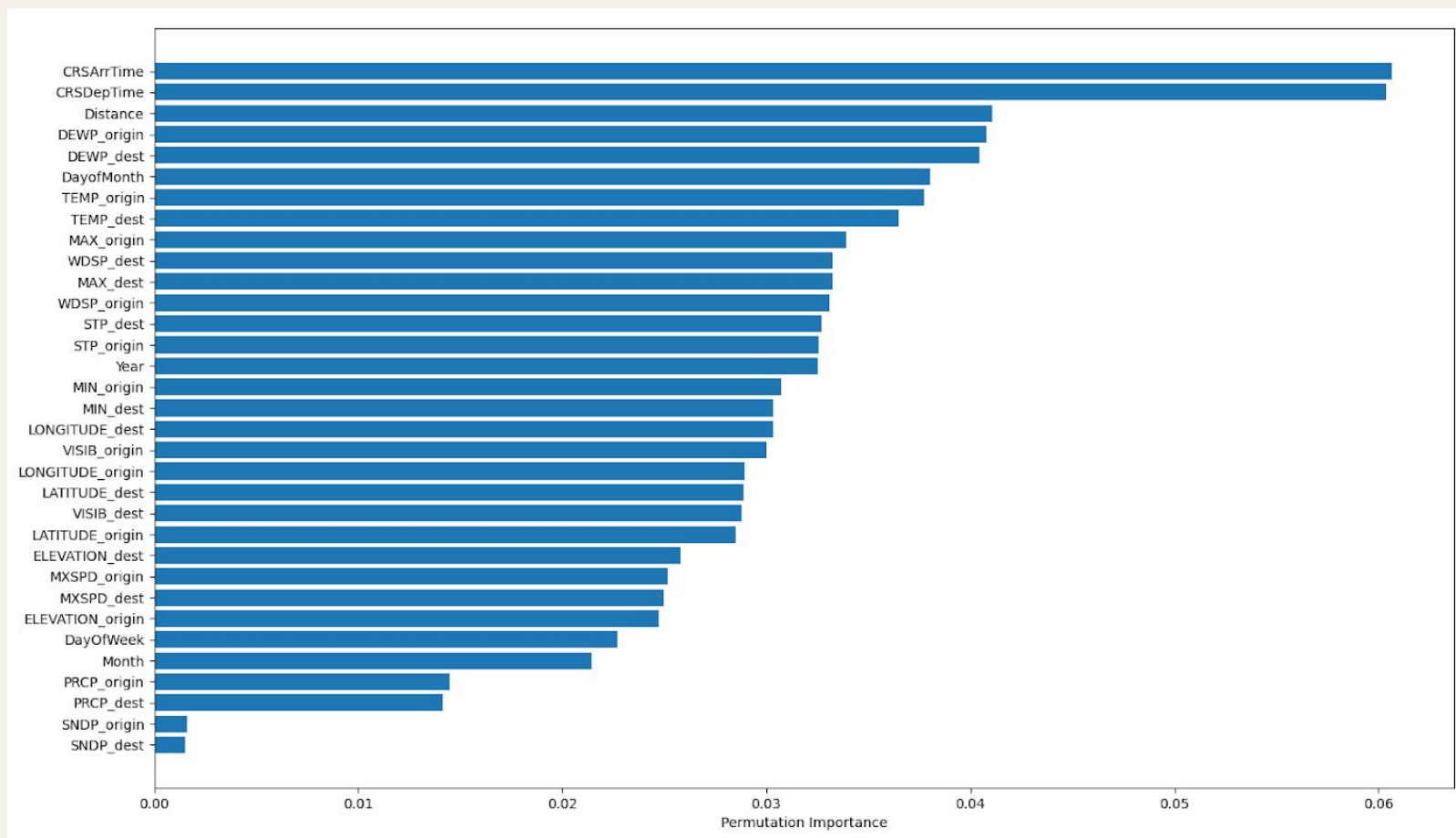
Hyperparamiter	Meaning	Values 
n_estimators	Number of trees	50, 100, 150, 200, 250, 300, 350, 400, 450, 500
max_features	number of features to consider when looking for the best split	'auto', 'sqrt', 'log2'
max_depth	The maximum depth of the tree	20, 40, 60, 80, 90, 100, None
min_samples_split	The minimum number of samples required to split an internal node	2, 3, 4, 5, 6
criterion	Function to measure the quality of the split	'gini', 'entropy'

Hyperparameter Tuning Results



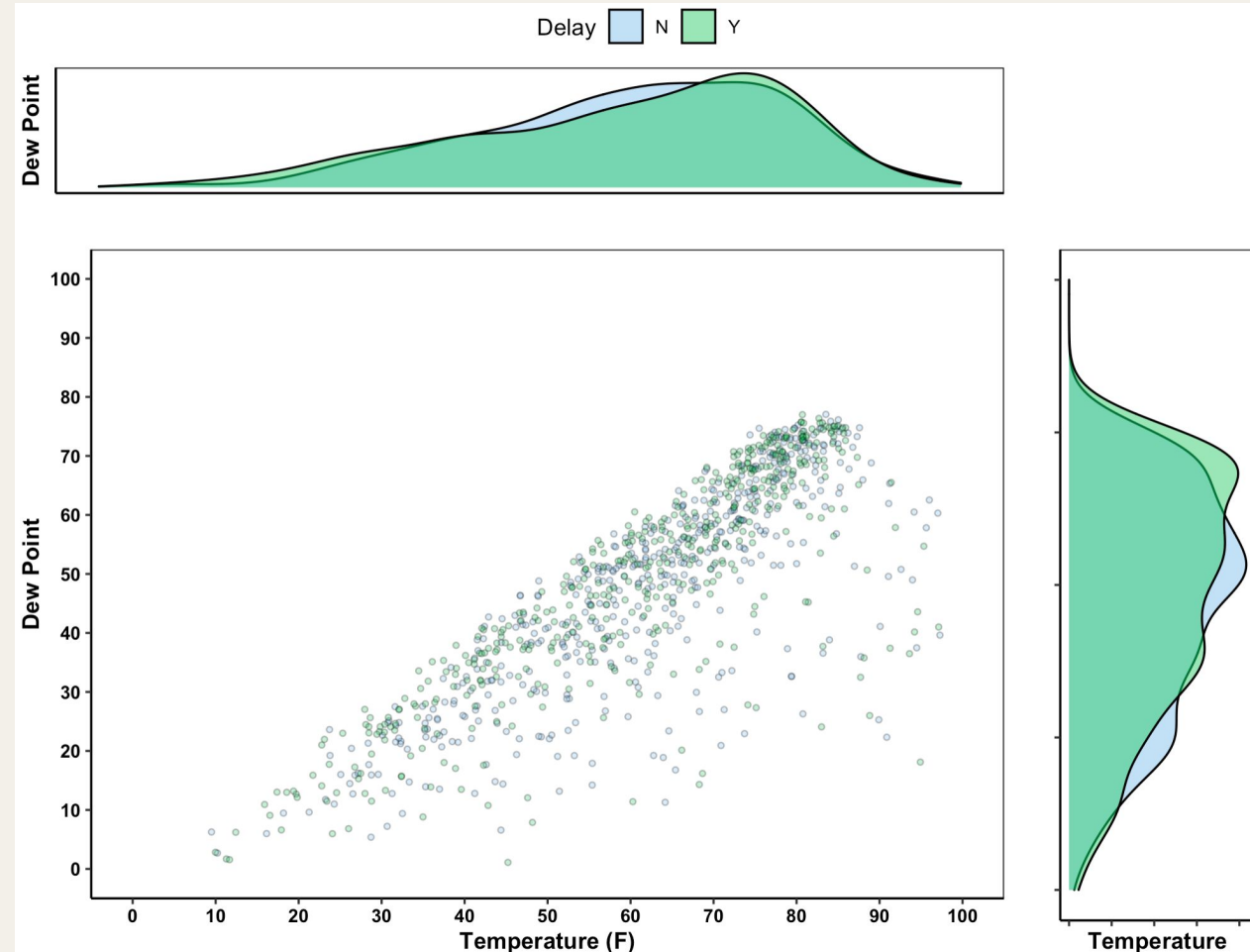
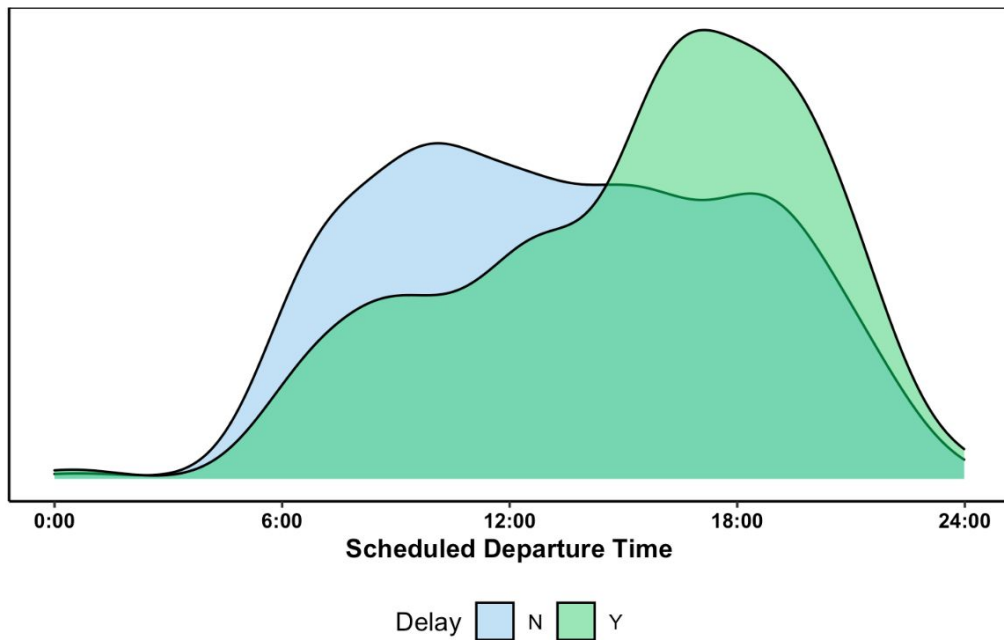
Results

- + Best: n_estimators: 450, min_samples_split: 3, max_features: auto, max_depth: 60, and criterion: gini
- + Accuracy: 0.97
- + F1: 0.95



EDA on Variable Importance

- Evening + Night departures more likely to have delays
- High dew point + high temperature = more clouds, fronts and other lifting forces, unstable air masses
- High dew points correlate with higher probability and severity of thunderstorms



Future Work

- + Test other models such as SVM, logistic regression, deep learning
- + Model for regression instead of classification to predict exact delay time
- + Delay times range from 15 min up to at most 10 hours
- + Web Scrape in real time all flights scheduled up to a week in the future and weather forecasts for those dates



Thank You

