



Airline Delays Analysis and Prediction

By Group: FlightDelay Inspector

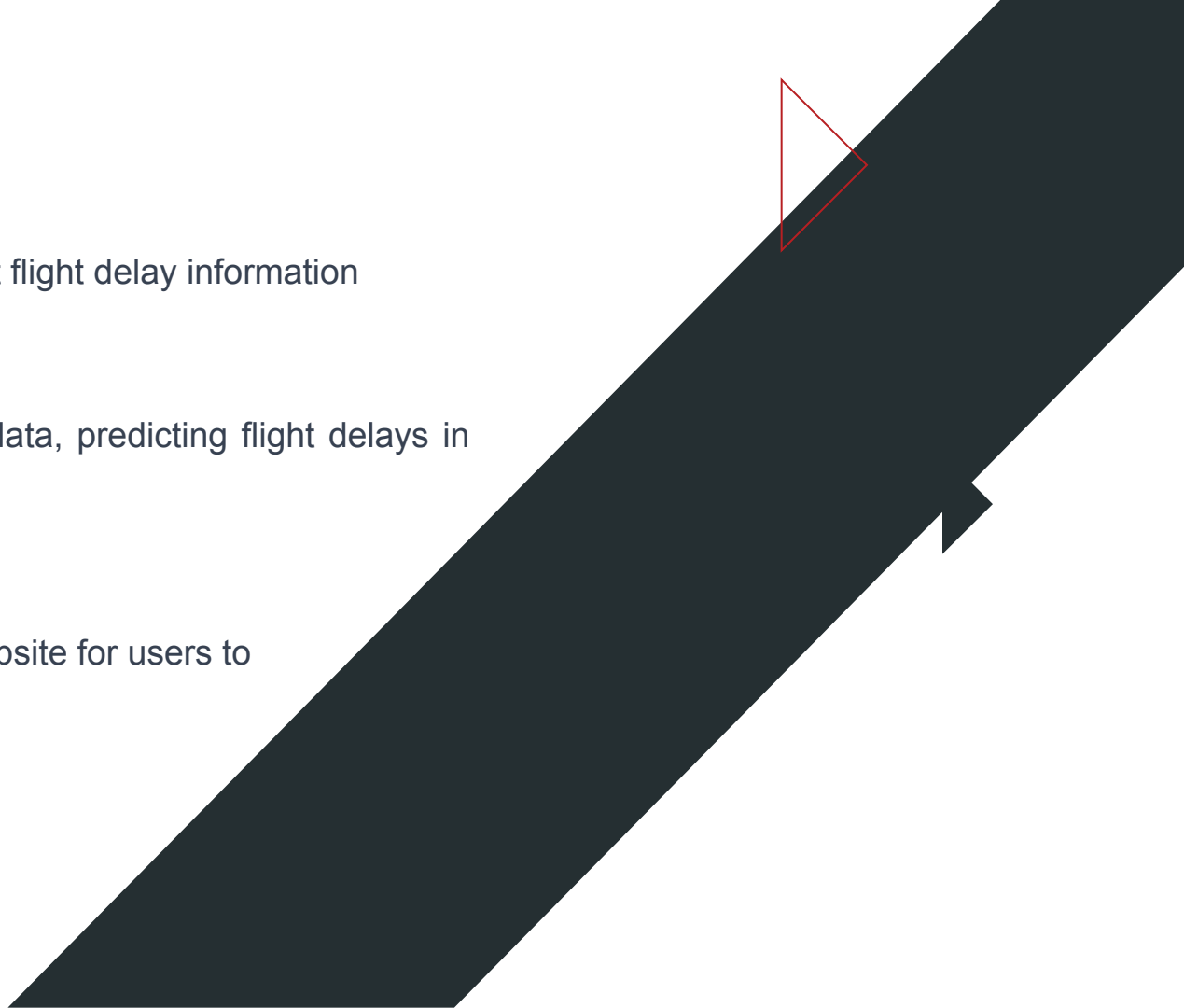
12/05/2023

Overview

Developed a system to predict flight delay information

Analyzed weather and flight data, predicting flight delays in real time.

Developed an AWS-based website for users to easily predict flight delays.





CONTENTS



Data Processing & Analyzing

Data Visualization

Data Prediction

Live Demo

Data Processing

Data Source

Bureau of Transportation Statistics and National Centers for Environmental Information

Data Clean

Select relevant attributes and deal with outliers

Data Join

Join the two sources of data

Database

Deploy Cassandra database on AWS

Data Analyzing

Identify 20 factors for analyzing. Categorized these into five dimensions.

PySpark job for Extract-Transform-Load. Creation of 31 CSV files for subsequent visualization.

Date & Time

"YEAR", "MONTH", "DAY_OF_WEEK", and "DEP_TIME"

Location

State, City, Top 10 States, Top 20 Cities, and Elevation

Continuous
Weather

Average temperature, dew point, visibility, wind speed, maximum sustained wind speed, precipitation, and snow depth

Categorical
Weather

Severe weather conditions like fog, snow, hail, thunderstorms, and tornadoes.

Data Analyzing - Date

Graph_1_Year

YEAR	num_delays	avg_delay_time
2020	820606	35.50
2021	1843107	38.46
2022	2397653	40.21

Graph_3_Weekday_2022

DAY_OF_WEEK	num_delays	avg_delay_time
Monday	357617	40.66
Tuesday	300125	36.60
Wednesday	302976	36.84
Thursday	352762	40.84
Friday	386698	42.52
Saturday	321613	41.90
Sunday	375862	40.95

Graph_2_Month_2022

MONTH	num_delays	avg_delay_time
January	165662	42.03
February	164400	39.57
March	205964	39.11
April	206645	40.41
May	214752	38.73
June	230234	42.70
July	227472	42.91
August	216696	42.53
September	172476	35.38
October	180211	34.10
November	184932	36.46
December	228209	45.65

Data Analyzing - Time

Graph_4_Hour_2020

HOURL	num_delays	avg_delay_time
0	3533	94.02
1	1613	93.61
2	475	134.64
3	225	108.03
4	120	122.71
5	4687	25.58
6	23951	21.88
7	32413	29.96
8	39572	31.46
9	44861	32.77
10	50035	31.78
11	53578	32.63
12	52362	33.36
13	54207	33.66
14	56663	32.25
15	59258	32.15
16	57478	33.99
17	59644	33.82
18	60463	33.90
19	52927	38.78
20	49905	39.43
21	32392	50.95
22	20395	56.57
23	9741	73.82

Graph_4_Hour_2021

HOURL	num_delays	avg_delay_time
0	14058	98.02
1	5351	118.92
2	1701	148.67
3	637	148.47
4	306	141.17
5	11977	18.95
6	58639	18.84
7	70822	30.12
8	87970	30.40
9	90088	33.67
10	102079	32.75
11	112361	32.74
12	109301	33.97
13	116833	33.17
14	121262	33.64
15	122411	35.28
16	125330	36.15
17	129310	37.50
18	136937	38.51
19	126818	42.74
20	111449	46.50
21	93520	51.23
22	61166	60.86
23	32356	77.19

Graph_4_Hour_2022

HOURL	num_delays	avg_delay_time
0	21154	101.42
1	7462	132.38
2	2051	169.44
3	784	141.64
4	393	157.88
5	26684	14.62
6	76360	19.92
7	89581	29.92
8	107125	32.98
9	111052	37.14
10	130355	36.07
11	134089	36.43
12	139315	35.91
13	149193	35.56
14	152233	36.42
15	159502	36.41
16	160702	38.06
17	165981	38.56
18	172118	39.24
19	159789	43.69
20	153446	45.88
21	124736	51.72
22	97623	56.57
23	55364	71.41

Data Analyzing - Location

Graph_5_State

ORIGIN_STATE_NM	num_delays	avg_delay_time
Texas	599781	39.39
California	550515	33.25
Florida	472194	40.95
Illinois	340586	38.48
Georgia	283414	31.61
North Carolina	224176	39.35
New York	224011	48.06
Nevada	188618	33.69
Virginia	178018	46.00
Washington	171244	28.65
Arizona	171055	33.85
New Jersey	122663	44.72
Michigan	113052	42.89
Tennessee	108909	37.72
Maryland	107728	30.33
Pennsylvania	102477	46.14
Hawaii	97936	25.45

Graph_9_Elevation

ELEVATION_CATEGORY	num_delays	avg_delay_time
100	2224810	39.48
200	985778	36.81
300	859888	41.15
400	527559	34.23
500	21215	57.27
600	7554	61.69
700	181061	33.97
800	26372	39.36
900	21559	43.70
1000	20070	51.76
1100	6304	52.58
1200	4521	58.23
1300	105631	34.50
1400	25270	42.70
1500	4997	67.79
1600	647	93.30
1700	19160	42.54
1800	2329	59.02
1900	5479	82.10
2000	1406	79.42
2100	2083	62.64
2200	1324	90.42
2300	300	90.54
2400	5320	81.52

Data Analyzing - Weather

Graph_16_WDSP

WDSP_CATEGORY	num_delays	avg_delay_time
5	1358701	37.34
10	2772172	38.57
15	791213	40.61
20	129558	47.00
25	9170	54.59
30	522	80.19
35	21	54.33
40	9	63.56

Graph_13_VISIB_Category

VISIB_CATEGORY	num_delays	avg_delay_time
High_VISIB	376	46.16
Low_VISIB	56077	54.35
Medium_VISIB	5004913	38.63

Graph_18_PRCP

PRCP_CATEGORY	num_delays	avg_delay_time
2	5026078	38.73
4	31063	48.97
6	2329	53.35
8	1703	72.01
10	190	83.05
14	3	32.67

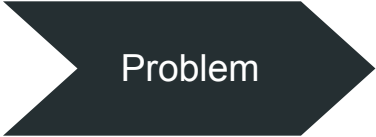
Data Analyzing



Practicability

Not just statistics

Useful for both the airline industry and passengers



Problem

Code efficiency:

- Cache
- Join
- Broadcast

Data Visualization



Techniques

Frontend - React and JavaScript

Chart Plot - Chart.js and D3

Server - Deployed on EC2

Demo - We will present live Demo at the end

Data Visualization

[Home Page](#)[Data Source](#)[Data Analysis](#)[Flight Delay Prediction](#)

Introduction

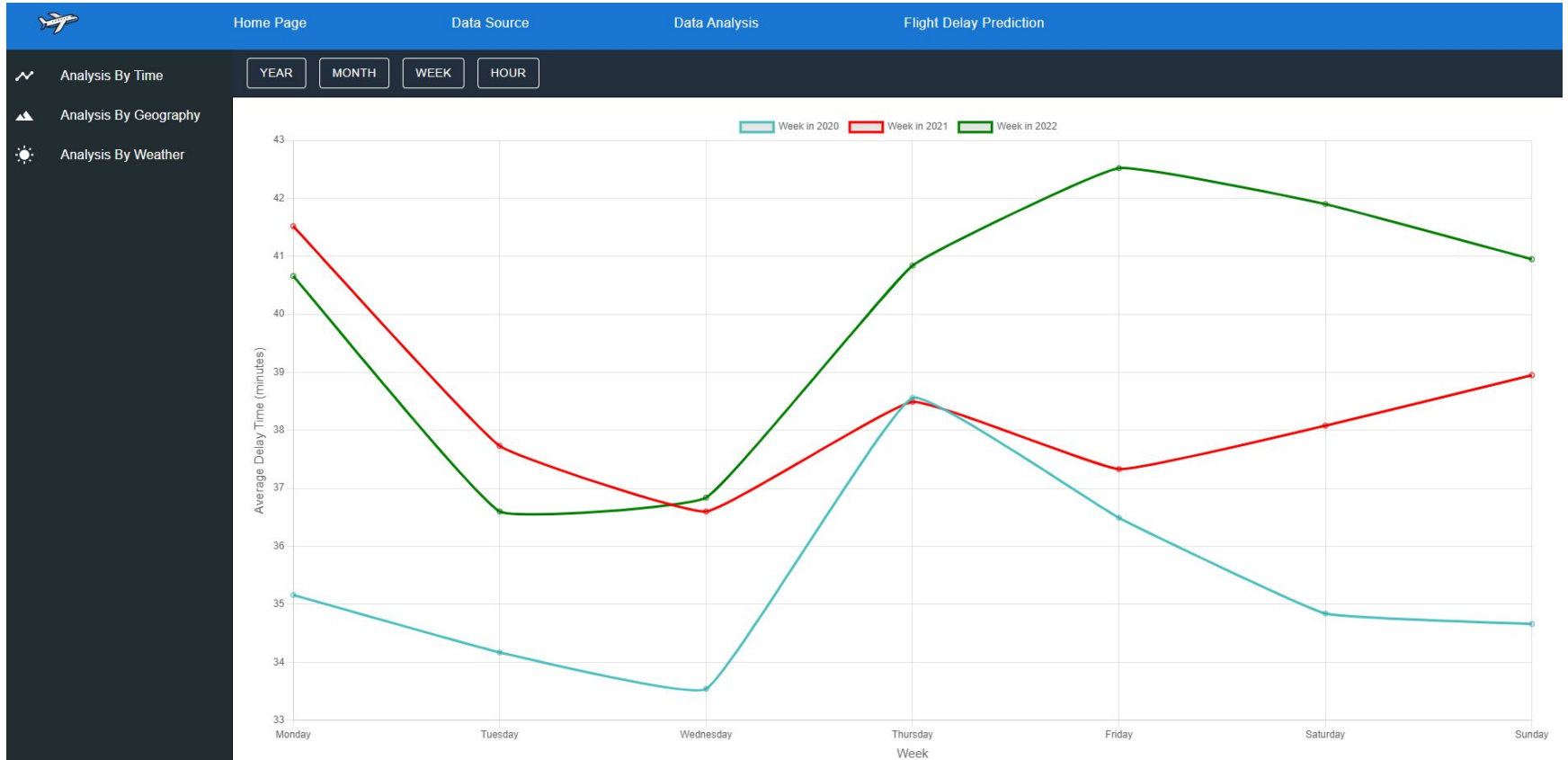
Inspired by the Kaggle dataset <https://www.kaggle.com/threnjen/2019-airline-delays-and-cancellations>, our group decided to further investigate the relationship between airport weather conditions and flight delays/cancellations. Noticed that this Kaggle dataset is made by joining two public datasets from where limited attribute are included and only data in 2019 was used. From the source datasets, we found that the source dataset contains a wealth of relevant attributes, which has rich potential that we can do more research by using more attributes and expanding the target year from a single year to several years.

Source Dataset 1

Attribute	Description
Station	Station number (WMO/DATSAV3 possibly combined w/WBAN number)
DATE	Given in mm/dd/yyyy format
LATITUDE	Given in decimated degrees (Southern Hemisphere values are negative)
LONGITUDE	Given in decimated degrees (Western Hemisphere values are negative)
ELEVATION	Given in meters
NAME	Name of station/airport/military base
TEMP	Mean temperature for the day in degrees Fahrenheit to tenths.
DEWP	Mean dew point for the day in degrees Fahrenheit to tenths.
VISIB	Mean visibility for the day in miles to tenths.

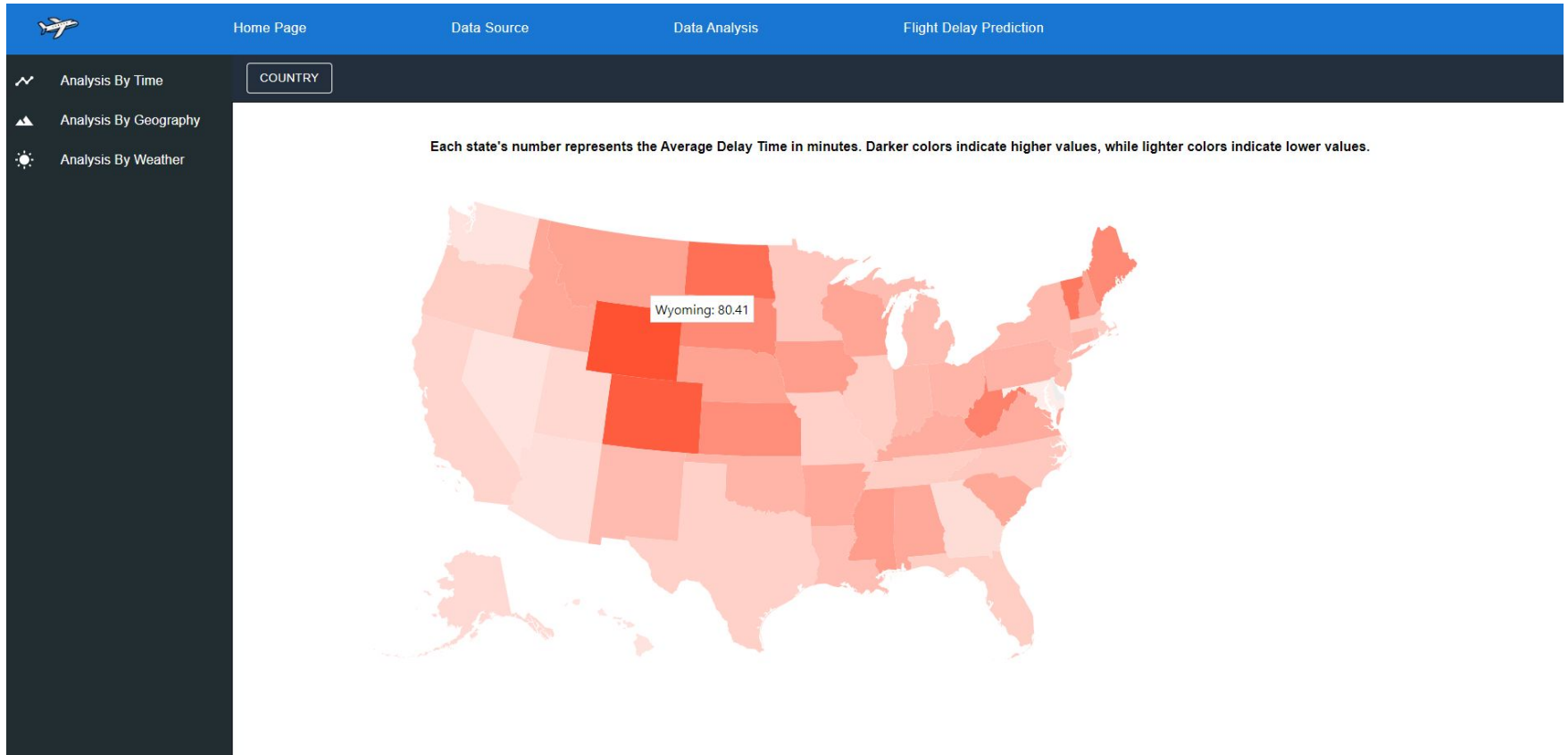
Data Visualization

Line Chart



Data Visualization

Geographical Chart



Data Visualization

Bar Chart



Data Visualization

Prediction Model (Will Explain Model Later)

Enter The Below Information To Predict Your Flight Delay

Date
12/14/2023

Current Hour
23

IATA Airport Code
LAX

Average Temperature
40

Max Temperature
60

Min Temperature
30

Visibility
High Visibility (>1 miles)

☐ Fog ☒ Rain or Drizzle ☐ Snow or Ice Pellets ☐ Hail

☐ Thunder ☒ Tornado or Funnel Cloud

START PREDICT

Enter The Below Information To Predict Your Flight Delay

Date
12/14/2023

Current Hour
23

IATA Airport Code
LAX

Average Temperature
40

Max Temperature
60

Min Temperature
30

Visibility
High Visibility (>1 miles)

☐ Fog ☒ Rain or Drizzle ☐ Snow or Ice Pellets ☐ Hail

☐ Thunder ☒ Tornado or Funnel Cloud

START PREDICT

Prediction Result:
The Expected Delay is 0-15 minutes. (Prediction confidence: 84.76%)

Data Prediction

Problem Definition

- Binary classification task:
 - Class 0 — Flights with normal departure (delay less than 15 mins)
 - Class 1 — Flights in an abnormal state (delay exceeding 15 mins / cancelled)
- Input: flight information features and weather conditions features
- Output: predicted label in $\{0, 1\}$, confidence level

Data Prediction

Data Preprocessing and Feature Engineering

- Handling missing features of cancelled flights with **sampling**
- **Normalizing** the numeric features
- Processing non-numeric features with **one-hot coding**
- Dealing with class imbalance by **oversampling**

Data Prediction

Machine Learning Model and Metrics

- Metric:
 - Balanced F-1 score
- Machine Learning Models:
 - Ridge Classifier
 - KNN
 - Decision Tree Classifier
 - Random Forest Classifier
 - MLP



Live Demo
