An Intelligent Approach to Flight Delay Prediction & Analysis

Guided By:

Mr. Tushar Kute

Mrs. Trupti Joshi

Submitted By:

Chaitanya Larokar (220943025010)

Vaibhav Gavli (220943025052)

Pushpak Paunikar (220943025029)

Akash Kunjir (220943025016)

Introduction

- In the last few years, air travel has become a common, easy and affordable mode of transport.
- Over 100,000 commercial flights operate on a daily basis, around 200,000 flights arrive delayed each year which causes huge losses for airlines as well as their customers.
- Flights can be delayed due to various reasons, some of them being, extreme weather conditions, heavy air traffic, due to air carrier, etc.
- Predictive machine learning model of a classifier & regression using flight delay information.

Objective of Project

- The objective is to examine the impact of various factors on the arrival delay for domestic flights in the United States.
- It uses a machine learning model to classify and predict the arrival delays of various flights in different airports & predict the delay time of flight using regression, during the years 2019 - 2021.
- The machine learning classification & regression algorithms are then evaluated with standard metrics and hence compared.

Overview of Tools Used

- Python
- Spark
- Machine Learning
- Tableau

Spark MLlib

tableau SOFTWARE

Project Summary

- Understanding of business need
- Collection of data
- Understanding of flight data
- Preparing data for modeling
- Creating predictive models

Dataset

- Data source :
  Bureau of Transportation Statistic (BTS) (https://www.bts.gov/)

- Data content:
  Flight delay features of different airports for year 2019 and 2021

- Data understanding :
  Available variables
  Relationships

- Data Preparation :
  Missing values
  Redundant variables

Data Cleaning Operation

- Define the problem
- Collect the data
- Check for duplicates
- Handle missing vales
- Standardize the data
- Validate the data

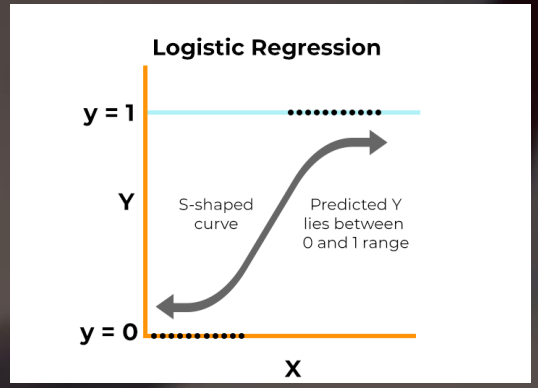
Models

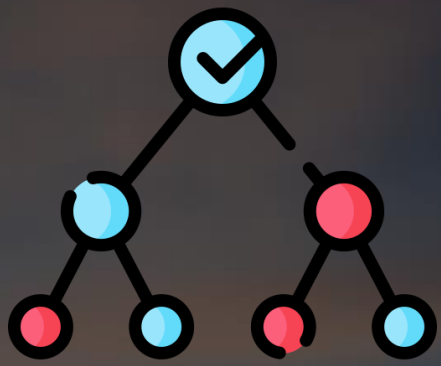
Classification:

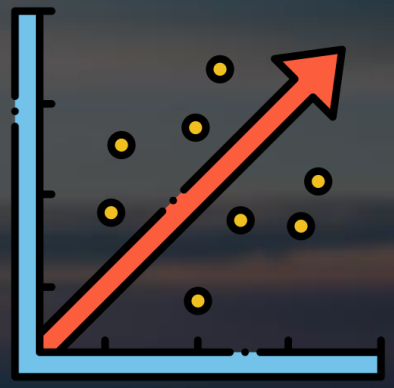
- Logistic Regression
- Decision Tree Classifier
- Random Forest Classifier

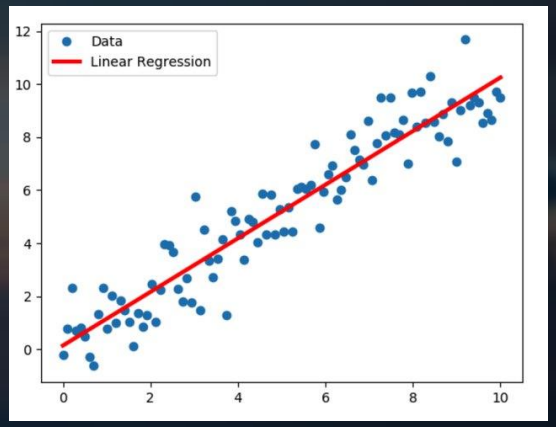
Regression:

- Linear Regression
- Decision Tree Regressor
- Random Forest Regressor









Classification

Algorithm	Train Accuracy	Test Accuracy
Logistic Regression	1.0	1.0
Decision Tree Classifier	1.0	1.0
Random Forest Classifier	1.0	1.0

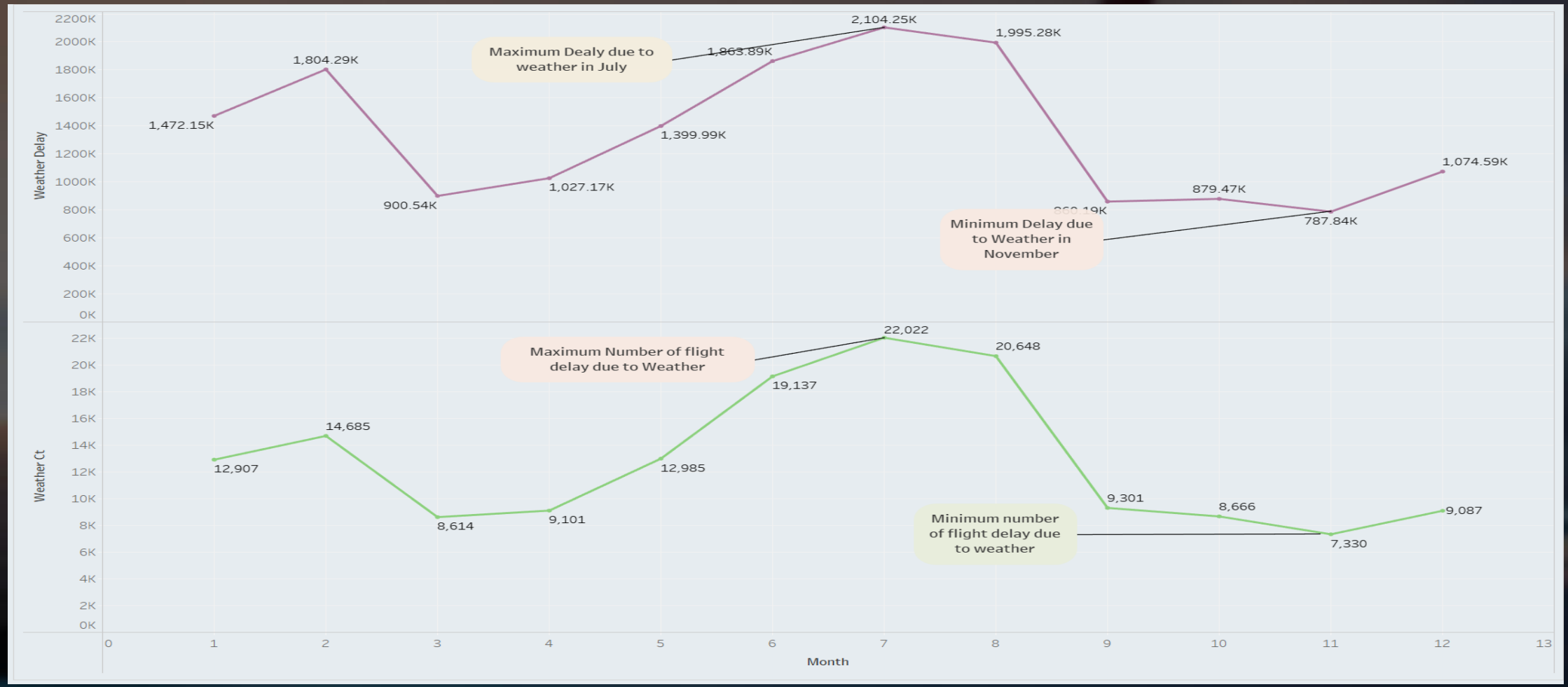
Regression

Algorithm	Train R Squared	Test R Squared	MSE	MAE
Linear Regression	0.99	0.99	1.08e-07	2.39e-06
Decision Tree Regressor	0.92	0.92	0.07096	0.07411
Random Forest Regressor	0.93	0.90	0.083091	0.04552

Analysis

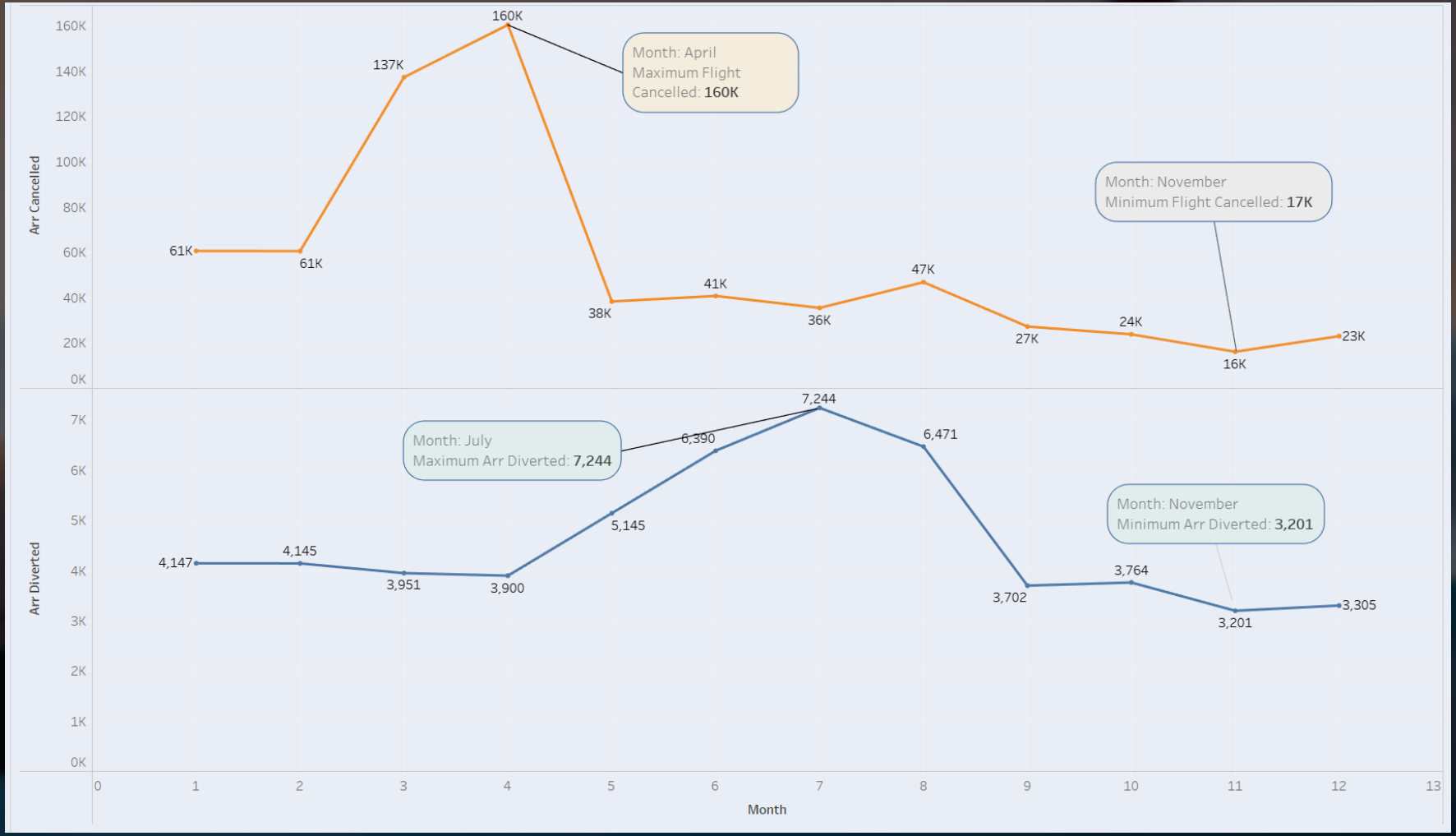
Analysis using Tableau

Preliminary Findings - Year Wise Weather Delay & Weather Ct Vs Month



- It is observed from this chart that maximum delay due to weather is in July and minimum delay is in November.

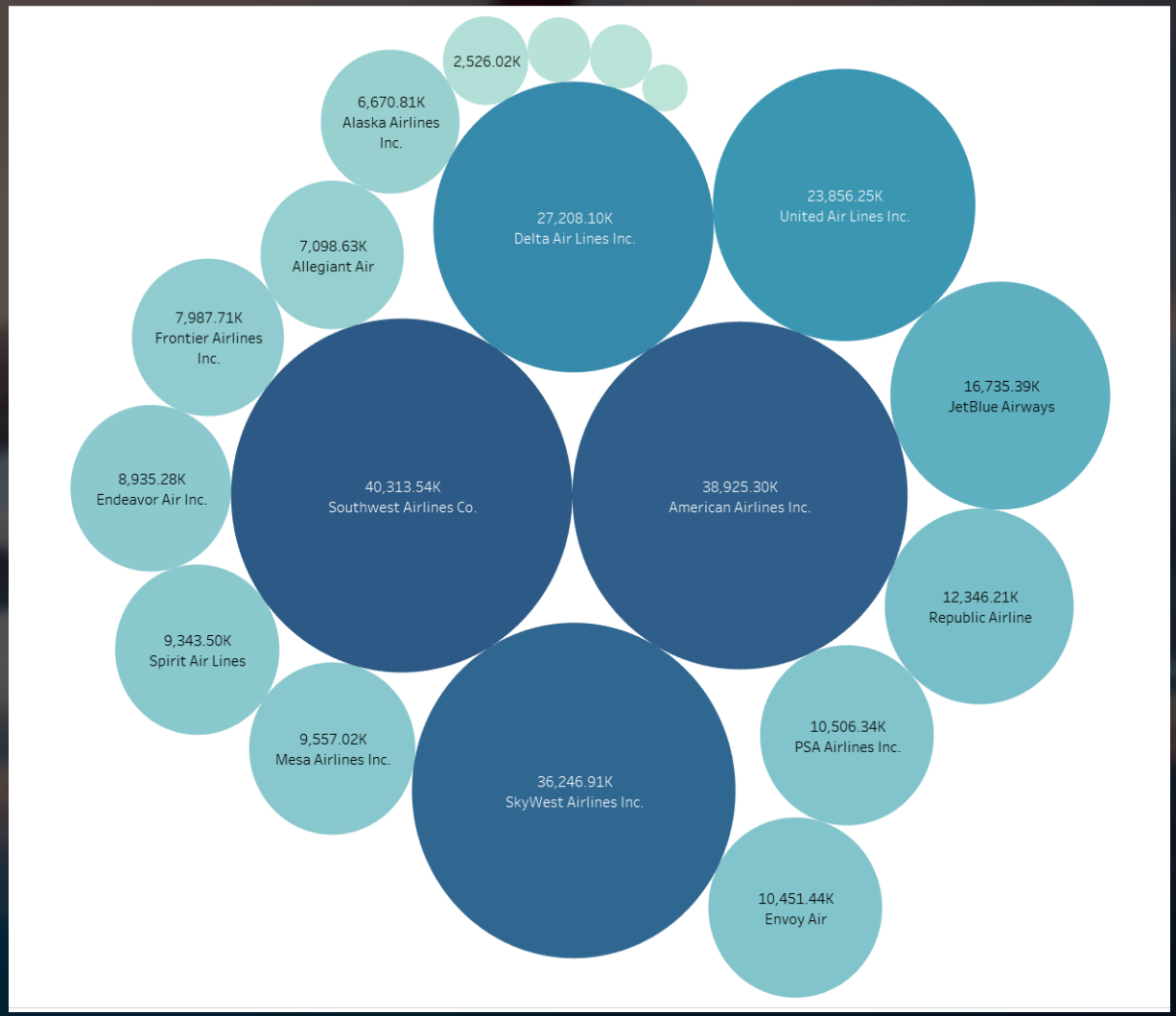
Month Wise Cancelled Flight



- From the above chart it is observed that maximum number of cancelled flight is in April and minimum number of cancelled flight is in November.

Carrier Performance

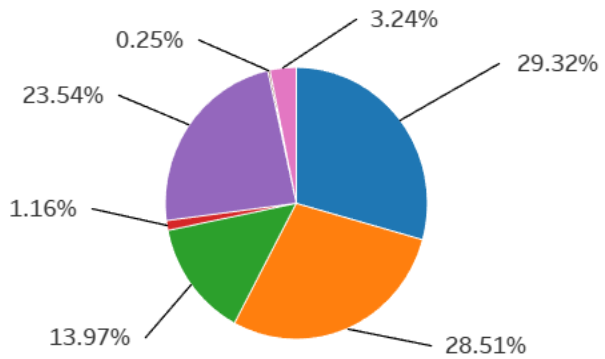
- Bubble chart shows the air carrier wise performance.
- Arr delay (delayed flight in minutes) with respect to Air carrier
- From above chart it is observed that ExpressJet Airline has minimum delay hence it has good performance.
- On other hand Southwest Airlines has maximum delay hence it has poor performance.



Dashboard 1

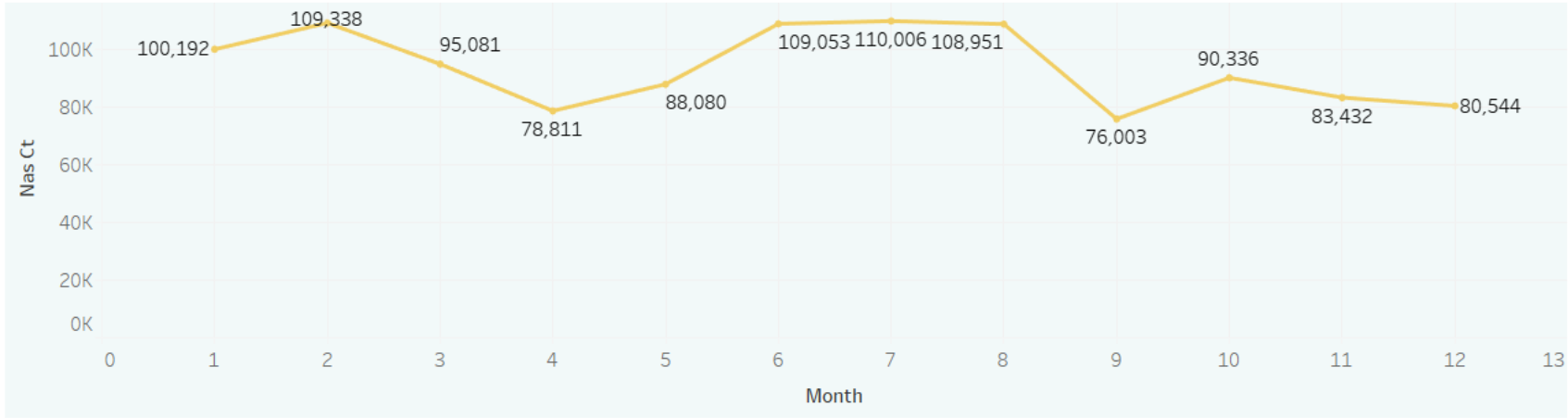
Year
2019 2022

% Flight Delay

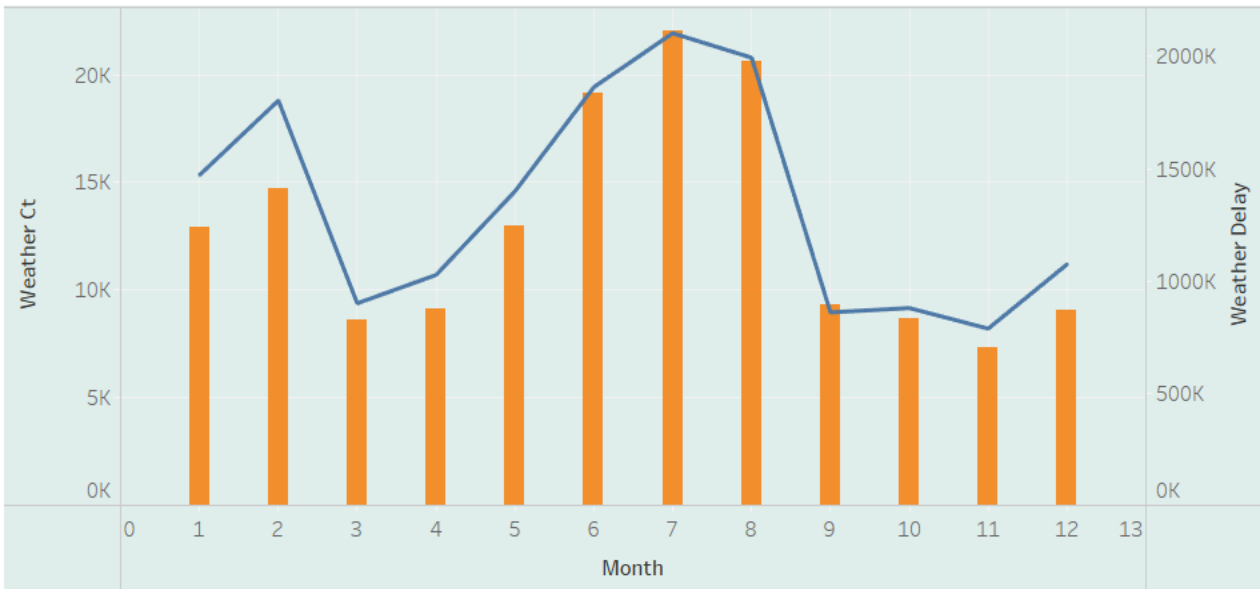


Factors
Air Carrier .. Cancelled National Avi.. Weather De..
Aircraft Arri.. Diverted Security Del..

Year Wise Nas Dealy & Nas Ct Vs Month

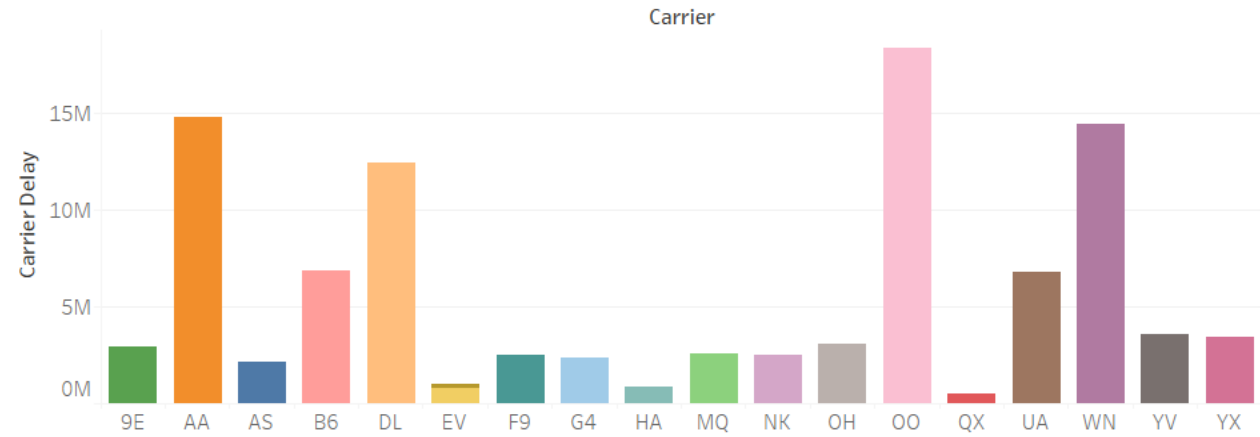


Number of Flight delay due to weather and delay (minute) by month



Measure.. Weather Ct Weather De..

Performance of Carrier



Carrier Name
Alaska Airli.. Allegiant Air American Ai.. Delta Air Lin.. Endeavor Ai.. Envoy Air ExpressJet .. ExpressJet

Analysis

- Among all the factors affecting flight delay, carrier delay, delay due to National Aviation System, Aircraft Arriving late are the major factors.
- In month of February maximum number of flights were delayed due to heavy air traffic which is controlled by NAS.
- The numbers of flights delayed and the total time of delay due to Weather is maximum from June to august.
- SkyWest Airline has maximum numbers of flights delay due to Air Carrier (performance of crew members).

Conclusion

Overall, the flight delay prediction and analysis project has provided valuable insights into the potential of data analytics and machine learning in the aviation industry. By developing accurate and efficient prediction models, this technology has the potential to transform the way airlines and airport authorities manage air travel disruptions, thereby improving the overall customer experience

Hence two stage, classification and regression, machine learning engine was designed and built to classify whether if a flight will arrive late or on time and predict the number of minutes by which a flight arrive late.


Thank You