Project Summary

Batch details	Capstone_GRP3_GGN_JAN22
Team members	Ankit Chauhan Ankush Dujania Anukriti Singh Chauhan Aryan Singh Bhadauria Shivansh Verma Srikar Yalamarthy
Domain of Project	Predictive Analysis
Proposed project title	Flight Delay Prediction
Group Number	3
Team Leader	Anukriti Singh Chauhan
Mentor Name	Mr. Muppidi Srikar

Dataset name – Flight Data

Introduction to the problem/domain/background details:

Commercial airlines are a backbone of the worldwide transportation system, bringing a good-sized socio-economic utility via means of permitting cheaper and less complex long-distance travel. After more than half a century of mainstream adoption (specially in the US), airline operations have seen primary optimizations and these days function with terrific reliability even withinside the face of exhausting engineering challenges. Still, the present-day passenger is once in a while inconvenienced through aircraft delays, disrupting an otherwise exacting system and causing tremendous inefficiencies

at scale in 2007, 23% of US flights had been more than 15 minutes past due to depart (federal definition of flight delay), levying an aggregate price of $32.9bn on the US economy. Suboptimal weather situations had been the direct reason for ~17% of those delays, suggesting better expertise of aircraft unfriendly weather should enhance airline scheduling and notably reduce delays.

Problem Statement:

How can we predict the future delays of departure of flights based on the previous delay data?

Business problem/ Impact in business of your problem/Need for this study/Abstract:

- Airlines are required to provide passengers with information concerning modifications to the standing of the flight if the flight is scheduled to depart within seven days. Airlines are needed to grant these status updates half-hour (or sooner) once the airline becomes aware of a standing change. The flight status information must, at a minimum, be provided on the airline' website and via the airline's telephone reservation system.
- Also, once a flight is delayed for a half-hour or more, the airline should update all flight standing displays and different sources of flight data at U.S. airports that are under the airline's management within 30 minutes once the airline becomes aware of the problem.

- Our predictions will fetch the pattern of delays in flights and optimize the traffic pattern. By optimizing this, variable businesses can generate better profits and control their variable costs such as fuel consumption and per minute losses incurred due to delay in take-off. We presume that time above 15 minutes in takeoff is considered a negative income for the business. Using our ML model we can detect the peak hours of delay, we can predict the delay based on climate conditions.

Variable identification:

Independent variables:

- AIRPLANE_ID
- YEAR
- MONTH
- DAY_OF_WEEK
- DEP_TIME_BLK
- DISTANCE_GROUP
- SEGMENT_NUMBER
- CONCURRENT_FLIGHTS
- NUMBER_OF_SEATS
- CARRIER_NAME

- AIRPORT_FLIGHTS_MONTH
- AIRLINE_FLIGHTS_MONTH
- AIRLINE_AIRPORT_FLIGHTS_MONTH
- AVG_MONTHLY_PASS_AIRPORT
- AVG_MONTHLY_PASS_AIRLINE
- FLT_ATTENDANTS_PER_PASS
- GROUND_SERV_PER_PASS
- PLANE_AGE
- DEPARTING_AIRPORT
- LATITUDE
- LONGITUDE
- PREVIOUS_AIRPORT
- PRCP
- SNOW
- SNWD
- TMAX
- AWND

Target Variable: DEP_DEL15

Variable information/Data description

- MONTH: Month
- YEAR: Year
- DAY_OF_WEEK: Day of Week
- DEP_DEL15: TARGET Binary of a departure delay over 15 minutes (1 is yes)
- DISTANCE_GROUP: Distance group to be flown by departing aircraft
- DEP_BLOCK: Departure block
- SEGMENT_NUMBER: The segment that this tail number is on for the day
- CONCURRENT_FLIGHTS: Concurrent flights leaving from the airport in the same departure block
- NUMBER_OF_SEATS: Number of seats on the aircraft
- CARRIER_NAME: Carrier
- AIRPORT_FLIGHTS_MONTH: Avg Airport Flights per Month
- AIRLINE_FLIGHTS_MONTH: Avg Airline Flights per Month
- AIRLINE_AIRPORT_FLIGHTS_MONTH: Avg Flights per month for Airline AND Airport
- AVG_MONTHLY_PASS_AIRPORT: Avg Passengers for the departing airport for the month
- AVG_MONTHLY_PASS_AIRLINE: Avg Passengers for the airline for the month
- FLT_ATTENDANTS_PER_PASS: Flight attendants per passenger for airline
- GROUND_SERV_PER_PASS: Ground service employees (service desk) per passenger for airline
- PLANE_AGE: Age of departing aircraft
- DEPARTING_AIRPORT: Departing Airport
- LATITUDE: Latitude of departing airport
- LONGITUDE: Longitude of departing airport

- PREVIOUS_AIRPORT: Previous airport that aircraft departed from
- PRCP: Inches of precipitation for the day
- SNOW: Inches of snowfall for the day
- SNWD: Inches of snow on the ground for the day
- TMAX: Max temperature for the day
- AWND: Max wind speed for the day
- AIRPLANE_ID: Airplane ID

Future Work/Methodology:

We will analyze and create features:

'DATE': create the date from the year, month and day of the week

'LOW': The lower value of DEP_TIME_BLK

'HIGH': A higher value of DEP_TIME_BLK

'TIMESTAMP': create a timestamp with the date and lower value of DEP_TIME_BLK

'WIND_CHILL': the perceived temperature due to the cooling effect of wind blowing

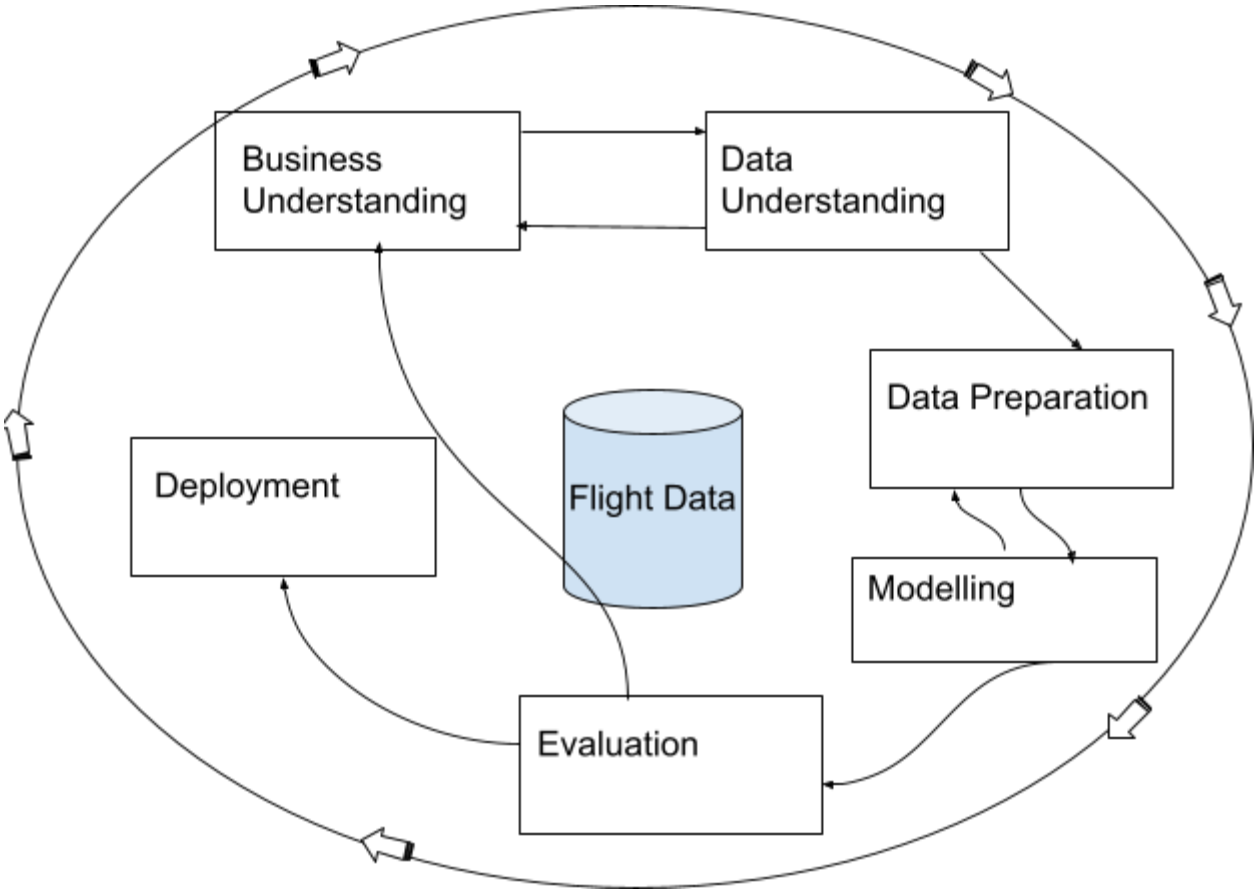
'PRCP_SNOW_RATIO': ratio of precipitation and snow

'PLANE_AGE_AIRLINE_AIRPORT_FLIGHTS_MONTH_RATIO': ratio of plane age and airline and airport flight months.

'SEAT_DISTRIBUTION': Ratio of seats and in concurrent flight CONCURRENT_FLIGHTS

'SEAT_DISTRIBUTION_NORMALISED': normalised values of the ratio of seats and in concurrent flight

We will then start analyzing what variables affect the delay in departure and drop the rest of the variables, followed by training the model and then finally predict if the departure delay is more than 15 minutes.



1) BUSINESS UNDERSTANDING

- Accessing our present situation
    - We know that 23% of US flights had been more than 15 minutes past due to departing, levying an aggregate price of $32.9bn on the US economy. The delay in the departure of flight is usually affected due to weather conditions, technical difficulty, etc.
- Inventory of resources
- Necessities, assumptions, and problems
- Risks and possibilities
- Terminology
- Costs and benefits
- Determining our goals
    - Criteria for business success
    - Criteria for success
- Developing our project plan
    - Project plan
    - Initial assessment of techniques and tools

2) DATA UNDERSTANDING

- Exploring data
    - Key Attribute distribution

        - Relationships between the small number or pairs of attributes
        - Attributes of important sub-populations
        - Simple aggregation results
        - Simple analysis of statistics
- Report for data Exploration
- Verifying quality of data
    - Whether the information is complete, covering all the required cases.
    - Whether the information is correct, and in case it contains errors, how often do they occur?
    - If the information contains missing values, and if so, where they occur, how they are represented and how common they are.
- Report Quality

3) DATA PREPARATION

- Cleaning your data
    - Report of the data cleaning
    - Constructing required data
    - Derived attributes
    - Generated records
    - Integrated data
    - Merged data
    - Aggregations

4) MODELING

- Generating test design
- Build the model
    - Describing characters of the current model that we might find useful in future.
    - Adjusting the parameter setting that is used in producing the model.
    - Listing the rule procedures for rule-bases models, along with an assessment of the overall accuracy of the model and its coverage.
    - In case the model is opaque, we need to create a list of technical information on it, like neutral network topology, sensitivity, accuracy and other behavioral descriptions that have been generated during the modeling process.
    - Describing the interpretation and behavior of the model.
    - Stating conclusions about data patterns, if any.
- Assessing the model
    - Executing the validation tests and evaluating results, as per the evaluation criteria.
    - Make a comparison between the results of the comparison and interpretation.
    - Creating ranking for the results, with regard to the evaluation criteria and success.
    - Interpreting the results in terms of business in this stage, as far as possible.
    - Checking whether the models are readable.
    - Checking the credibility of the obtained results.
    - Analyzing the potential for the incorporation of each result.

- Evaluating the specific features of each of the modeling techniques and finding out the reason behind why certain parameter settings and modeling techniques lead to impressive or poor results.

5) EVALUATION

- Evaluating the results
  - Assessing of the Data Modeling results
  - Approved final models
- Reviewing the process
- Determining the subsequent steps
  - List of actions that we might possibly take
  - Decisions

6) DEPLOYMENT

- Monitoring the plan and its maintenance
- Plan for monitoring and maintenance
- Production of the final report
  - Final report
  - Final presentation
- Reviewing the project
  - Experience documentation

Timeline Chart (Weekly plan):

Tabular timeline of the work to be done: -

Week 1-2	16-05-2022	Performing EDA & various feature engineering techniques on the data
Week 3-4	30-05-2022	Interim Report submission on 02-06-2022; Start with model building
Week 5-6	13-06-2022	Submission of Final Report on 27-06-2022
Week 7	27-06-2022	Submission of Final report and present the final model with presentation on 29-06-2020

References (Data set source/Journals/articles)

https://machinehack.com/hackathons/data_engineering_championship/overview

Extra Information

The above is the link for an ongoing competition organized by MachineHack. The last date for submitting the solution for this is May 30, 2022, as per the website. The information given in the above sub-titles is subjected to the information already mentioned on the website.

Declaration: This is to declare that the dataset that we are using for our capstone project does not have any relevant legality associated with it and can be used to showcase the work we do on it as a presentation in Great Learning.