**DELIVERABLE 3: README.md *(Draft 2)***

**Group 4: Final Project**

**January 2020 Flight Delay Analysis:**

**Flight Delays, Best and Worst Airports and Airlines**

1. **PRESENTATION TOPIC**

**January 2020 Flight Delay Analysis: Flight Delays, Best and Worst Airlines and Airports**

1. **Project Overview**
2. Predictive Problem: Predict the arrival of flight delays with real world data
3. Determine the Best and Worst Airlines for flight delay arrivals and departures
4. Determine the Best and Worst Airports for flight delay arrivals and departures
5. **Why we chose this topic?**
6. We thought the topic would be interesting.
7. Flight delays are prevalent and a hot topic in all news cycles.
8. The purpose of this project is to use exploratory data analysis and to build machine learning models to predict flight delay arrivals for airlines and airports in the United States.
9. A flight is considered delayed if it arrived at (or departed) the gate 15 minutes or more after the scheduled arrival (or departure) time as reflected in the airline carriers’ Computerized Reservation System.
10. We wanted to share our results and provide top 10 lists of the Best and Worst Airlines and Airports for flight delay arrivals and departures.
11. **Data Exploration Phase of Project**

**\*Purpose:** To build machine learning models to predict flight delay arrivals for airlines and airports in the United States.

**Step 1:** **Dataset**

* Found a dataset on Kaggle.com titled, “January Flight Delay Prediction.” There were two sets of data for January 2019 and January 2020 flight delays and we used the January 2020 dataset.

**Step 2: Database**

* A database was setup using AWS
* ER Diagrams were created using Quick DBD *(picture below)*
* PgAdmin – Postgress SQL Database Management tool *(screen shot below)*
* Created Flight Delay Database Query Tables *(listed below)*

**ER Diagrams**

*Graphical user interface, text, application, chat or text message

Description automatically generated*

**PgAdmin**

*A screenshot of a computer

Description automatically generated with medium confidence*

**Flight Delay Database Query Tables**

Table 1:

CREATE TABLE Flights (  
    DAY\_OF\_MONTH int,  
    DAY\_OF\_WEEK int,  
    OP\_UNIQUE\_CARRIER varchar,  
    OP\_CARRIER\_AIRLINE\_ID int,  
    OP\_CARRIER varchar,  
    TAIL\_NUM varchar,  
    OP\_CARRIER\_FL\_NUM int,  
    ORIGIN\_AIRPORT\_ID int,  
    ORIGIN\_AIRPORT\_SEQ\_ID int,  
    ORIGIN varchar,  
    DEST\_AIRPORT\_ID int,  
    DEST\_AIRPORT\_SEQ\_ID int,  
    DEST varchar,  
    DEP\_TIME float,  
    DEP\_DEL15 float,  
    DEP\_TIME\_BLK varchar,  
    ARR\_TIME float,  
    ARR\_DEL15 float,  
    CANCELLED float,  
    DIVERTED float,  
    DISTANCE float  
);

**Table 2:**

CREATE TABLE Arrival\_Airports (  
    airport varchar PRIMARY KEY,  
    airport\_name varchar  
);

**Table 3:**

CREATE TABLE Departure\_Airports (  
    airport varchar PRIMARY KEY,  
    airport\_name varchar  
);

**Table 4:**

CREATE TABLE Airlines (  
    carrier varchar PRIMARY KEY,  
    carrier\_name varchar  
);

**The Foreign Key Constraints:**  
ALTER TABLE Flights  
ADD FOREIGN KEY (DEST) REFERENCES Arrival\_Airports(airport);ALTER TABLE Flights  
ADD FOREIGN KEY (ORIGIN) REFERENCES Departure\_Airports(airport);ALTER TABLE Flights  
ADD FOREIGN KEY (OP\_UNIQUE\_CARRIER) REFERENCES Airlines(carrier);

**Step 3: Machine Learning Models**

* Classification Model testing

1. Logistic Regression Model:
   * Training Accuracy: 93.21%
   * Testing Accuracy: 93.14%
   * Precision (weighted average): 93%
   * Recall (weighted average): 93%
   * F1-Score (weighted average): 93%
2. Random Forest Classifier:
   * Training Accuracy: 90.78%
   * Testing Accuracy: 89.69%
   * Precision (weighted average): 90%
   * Recall (weighted average): 90%
   * F1-Score (weighted average): 87%
3. Gradient Boosting Classifier:
   * Training Accuracy: 93.21%
   * Testing Accuracy: 93.21%
   * Precision (weighted average): 93%
   * Recall (weighted average): 93%
   * F1-Score (weighted average): 93%

Best Model for our Project: Gradient Boosting Classifier

It has the highest testing accuracy (93.21%) among the three models, which shows it works well with unseen data. It has a strong balance between precision, recall, and F1-score, indicating that it performs well across different metrics.

The model is not only good at identifying true positives but also at minimizing false positives and false negatives, making it the most reliable choice. We believe this will be the best model for our final project.

Calendar

Description automatically generated with medium confidence

Chart, line chart

Description automatically generated

Chart, line chart

Description automatically generated

Chart, treemap chart

Description automatically generated

1. **Technologies, languages, tools and algorithms used throughout the project**

* Python technologies used within the machine learning model:
  + Gradient Boosting Classifier
  + Numpy
  + Panda
  + Skylearn
  + Sqlalchemy
* Jupyter Notebook
* Tableau
* CSV Editor
* PgAdmin
* Quick DBD
* Canva
* PowerPoint
* Github
* AWS

1. **Results of Analysis**

**Analysis of Flight Delays**

**Time of Day**

- The hour of departure plays a role in the on time arrival of flights.

- Early AM flights from 4 AM to 9 AM show the best on-time performance. This may be due to less traffic and the fact that severe weather is less likely.

- Late night flights have the highest delay rates. This could be caused by less resources during this time of day and the impact of delays accumulated during the day.

**Airlines**

- The performance of airlines varies greatly.

- Southwest Airlines seems to manage their schedule the best, with the lowest percentage of delayed flights.

- Airlines like PSA Airlines and Alaska Airlines have a high percentage of delays, potential issues with the way they manage operations, scheduling, or their routes

**Airports**

- The efficiency of airports also contribute to delayed flights.

- Airports, such as Owensboro Daviess County Regional and Quincy Regional-Baldwin Field, effectively manage departures, resulting in no delayed flights. These are small airports, which have less risk of high air traffic volume and less operational challenges.

- Airports like Adak Island, Pago Pago International, and Ogden-Hinckley have very high rate of delayed departures. These airports could be facing challenges such as weather, resources, or logistical problems.

Predictive Modeling

- The Gradient Boosting Classifier resulted in strong performance for predicting delays. Precision, recall, and accuracy are high, which shows that the model is reliable and could be useful when used against real world datasets.

- The model is more accurate in predicting on-time flights, it provides reasonable predictions for delayed flights. Further tuning could improve this.

- With the impact of flight delays on passengers and airlines, a model like this could be valuable for improved scheduling, customer satisfaction, and reduced cost by improving delay rates.

**Please find below additional visuals of our analysis results:**

A picture containing text, screenshot, number, font

Description automatically generated

A picture containing text, screenshot, number, font

Description automatically generated

A picture containing text, screenshot, number, font

Description automatically generated

A picture containing text, screenshot, number, font

Description automatically generated

**BEST AIRLINES**

A picture containing text, screenshot, number, font

Description automatically generated

**WORST AIRLINES**

A picture containing text, screenshot, number, font

Description automatically generated

1. **Visualization Dashboard**

* Tableau
* Working Tableau dashboard link:

<https://public.tableau.com/app/profile/kennan.foust/viz/ProjectTableau_16832615644700/Dashboard1?publish=yes>

A screenshot of a graph

Description automatically generated with low confidence

1. **Future Analysis Recommendations**
2. Impact of Weather: If weather data is available, investigate the impact of weather conditions on flight delays. This would reveal how different weather conditions affect flight punctuality.
3. Predicting Flight Cancellations: Build a model to predict flight cancellations. This could help passengers and airlines better anticipate and manage cancellations.
4. Route Analysis: Analyze specific routes to identify which ones are more prone to delays or cancellations.
5. Other machine learning models worth considering:

- Support Vector Machines (SVM)

- XGBoost – Could be more efficient and perform better than the Gradient Booster Classifier

1. **What we would have done differently if we had more time?**
2. Present additional data on the best and worst hours for on-time arrivals.
3. Screen shots of data below

A picture containing text, screenshot, number, font

Description automatically generated

A screenshot of a table

Description automatically generated with low confidence

1. **PowerPoint Presentation**

**Group 4 FINAL Presentation**

1. Link to PowerPoint presentation in main branch.

File title is “Group 4 Final Project Presentation”