Big Data Project

Insight in to Delays in the Airlines Industry

Prepared by Jason Nam, Shantanu Bisht, Muhammad Shahzad, Jiale Guan, Benjamin Rowley

Table of Contents

Problem Description

Problem Statement

For passengers, flight delays can cause inconvenience, missed connections, and cancellations, which can disrupt travel plans and cause financial losses. For airlines, flight delays can result in increased costs due to additional crew and fuel expenses, lower customer satisfaction, and potential legal liabilities. For the aviation industry as a whole, flight delays can lead to decreased efficiency and increased environmental impact due to longer flight times and increased emissions.

Therefore, predicting flight delays using current parameters can help airlines and other stakeholders take proactive measures to mitigate the impact of delays and minimize their occurrence. This can include scheduling alternative routes, adjusting flight times, and informing passengers in advance to avoid unnecessary inconvenience.

“Can we predict future flight delays using the current parameters of flight details gathered from flights in the US?”

Our goal is to create a flight delay machine learning model that can reliably predict a delay from existing parameters in the input dataset.

Background of the Situation

The airline business is a significant component of the transportation sector and has a significant impact on the world economy. Flight delays, which may harm travelers, airlines, and the entire supply chain, are among the biggest difficulties the business must overcome. Weather, technical problems, air traffic congestion, crew scheduling, and other reasons can all contribute to flight delays. These delays may lead to missed connections, aggravation for passengers, higher operating costs, and decreased revenue for airlines.

Current Attempts at Problem Solution

Flight delays are a problem that airlines and other industry stakeholders are working to solve using a variety of tactics. Some airlines have made investments in cutting-edge technology systems that can predict weather patterns and foresee flight delays. Airlines assess their flight itineraries on a frequent basis to optimise routing and lessen the effects of delays. Additionally, some airports have implemented new infrastructure and processes to reduce congestion and improve efficiency.

Current Resources Available

To alleviate flight delays, a variety of tools are available, including advanced data analytics, predictive models, and machine learning algorithms. These techniques can assist airports and airlines in spotting patterns and anticipating possible disruptions, allowing them to be proactive in reducing delays.

Reason for New Solution

Flight delays continue to plague the airline industry despite known solutions. To address the underlying causes of delays and guarantee a more seamless and effective experience for passengers, a new approach is required. Additionally, advancements in technology and data analytics have created new opportunities for developing more effective and scalable solutions.

Implications/Cost to Industry if Problem Not Resolved

Flight delays can have a big impact on the airline business and the whole economy if they are not resolved. Delays can result in decreased consumer satisfaction, reduced revenue for airlines, and harm to a brand's reputation. Moreover, delays can impact other sectors of the economy that rely on air transportation, such as tourism and business travel. Flight delays may cost airlines and the overall economy a lot of money if they go unchecked.

Cost Component	**Cost (in billions)**
Costs to Airlines	$8.3
Costs to Passengers	$16.7
Costs from Lost Demand	$3.9
Total Direct Cost	$28.9
Impact on GDP	$4.0
Total Cost	**$32.9**

Solution Summary

Conceptual Overview of the Solution

Predicting flight delays: Classification algorithms will be used to predict whether a flight is likely to be delayed or not, based on various features such as the airline, origin and destination airports, departure time.

Identifying delay causes: Classification algorithms can be used to identify the most common causes of flight delays. This information can be used to prioritize resources and make operational improvements.

Categorizing delays: Classification algorithms can be used to categorize delays based on Airline companies. This information can be used to better understand the impact of delays on passengers and to improve the accuracy of delay reporting.

Analyzing delay trends: Classification algorithms can be used to identify trends in delay patterns over time, such as which days of the week or months of the year have the highest delays, or which airlines or airports have the highest delay rates. This information can be used to make data-driven decisions about operational improvements or resource allocation.

Major Results (Output of the Solution)

The results of the delay prediction model was disappointing. With the best accuracy score from the different models tested being just above 64%, we concluded that with the given parameters, it was very difficult to predict a flight delay through predictive analysis.

Therefore, thought was given to analyse the relationship of each parameters to the delay classification labels. This was done to understand each parameter in depth to recommend what parameters are best and worst for flight delay classification.

Major Results (Output of the Solution)

```
Accuracy score of the Decision Tree model is 0.6342737076290826
Accuracy score of the Random Forest Tree model is 0.6258628680901029
Accuracy score of the K-Nearest Neighbors model is 0.6334023421808855
Accuracy score of the Logistic Regression model is 0.6418873404814139
```

Potential/Projected Savings or Improvements

Reduce airline delays: if our model is implemented, we could potentially reduce certain airline delays by accurately predicting most delays ahead of time based on a variety of factors within our model.

Save extra cost generated by airline delays: if our model can successfully predict potential delays, airlines as well as customers can save on additional costs that would have otherwise occurred due to delays

Improve customer satisfaction: reducing airline delays or accurately predicting potential airline delays can result in increasingly satisfied customers as they would then be able to plan their schedules around anticipated delays. One of the most significant frustrations for air travelers is the uncertainty that comes with flight delays. Passengers may need to rearrange their travel plans or may miss important meetings or events due to unexpected flight delays. However, by accurately predicting potential delays, airlines can provide passengers with real-time updates, enabling them to make informed decisions about their travel plans.

Enhance airport management ability: Predicting airline delays can result in enhanced airport management ability by providing airport operators with the necessary information to manage airport resources efficiently. If an airline is predicted to experience a delay, airport operators can adjust resources such as ground handling equipment, boarding gates, and staffing levels to better accommodate the anticipated delay

Tools Used/Required to Design, Test, and Implement

Python: The code is written in the Python programming language, which is a popular language for data science and machine learning.

NumPy: NumPy is a Python library used for numerical computing, which is used in this code to handle numerical data.

pandas: pandas is a Python library used for data manipulation and analysis, which is used in this code to load and manipulate data.

scikit-learn: scikit-learn is a Python library for machine learning, which is used in this code to define and train the machine learning models, as well as to preprocess the data.

Apache Spark: Apache Spark is a distributed computing framework used for big data processing, which is used in this code to read data from a CSV file.

Jupyter Notebook: Jupyter Notebook is an interactive computing environment that allows code to be executed in cells, which is used in this code to write and execute the machine learning code.

Excel: Excel is a spreadsheet can be used to analyze data. It is a very popular tool and can be used to create budgets, charts, and tables.

Projected Timeline of Implementation

Predicting delays: Will take at least a month to analyze airline delays based on time, destination, and departure.

Identifying delay causes: Will take days to find out delay causes.

Categorizing delays: Will also take days to categorizing these datasets.

Analyzing delay trends: It will take few hours as all delays variable are identified and categorized.

Solution Details

Assumptions Used or Made

The data is accurate: Assumptions are made that the data being used is accurate and reliable. If the data is incorrect or incomplete, it can negatively impact the accuracy of the model.

The data is representative: Assumptions are also made that the data being used is representative of the larger population or dataset. If the data is biased or unrepresentative, the model may not generalize well to new data.

The features are relevant: Assumptions are made that the features being used in the model are relevant to predicting flight delays. If irrelevant or redundant features are included, they can add noise to the model and reduce its accuracy.

The data is normally distributed: Many machine learning algorithms assume that the data is normally distributed. If the data is not normally distributed, it may be necessary to transform the data or use a different algorithm.

The data is independent: Assumptions are also made that the data points are independent of each other. If there is dependence between the data points, it can bias the model and reduce its accuracy.

The data is consistent over time: Assumptions are made that the patterns in the data are consistent over time. If the patterns change over time, the model may need to be updated or retrained to account for these changes.

Data Gathering and Import (Specific File and Format)

The code and data was processed in Databricks.

The code reads data from a specific file in CSV format using the spark.read.option method to specify the file path, header, and schema inference options. Here's the code used to read the data.

The file is located in the /FileStore/tables directory, and the file name is airlines_delay.csv. The file is assumed to have a header row, and the inferSchema option is set to true, indicating that the data types of the columns should be inferred from the data.

The toPandas() method is used to convert the Spark DataFrame to a Pandas DataFrame.

Data Cleanup and Normalization

Most of the numerical data pertaining to time used total minutes as the measurable unit. For the sake of machine learning process, we decided to keep the minute as the measurable unit. However, for visualization purposes, we have converted the measurable unit to hours.

For numerical data, we used K-Nearest Neighbor Imputer to impute missing numerical values and cleanup data. We also used Standard Scalar to normalize all numerical values.

For categorical data, we used Simple Imputer to implement a imputation strategy of using the most frequent values in the dataframe.

Data Transformation and Conversion

The data included categorical parameters.

- Airline Abbreviations (eg. DL, B6, US…)
- Airport From and Airport To Abbreviations (eg. ATL, HOU, PHX, …)

For these categorical data, we used One Hot Encoder to transform and convert the categorical data to numerical data.

- One Hot Encoder:
  eg. [0, 1, 0, … , 0, 0, 0]

Machine Learning Setup

The machine learning setup is a classification problem, where the goal is to predict whether a flight will be delayed or not based on various features such as the airline, airports, time of the flight, length of the flight, and day of the week.

The code reads in a dataset of airline delays, splits the data into training and test sets, and then defines and fits four different machine learning models:

- Decision Tree: A decision tree classifier is used with a maximum depth of 4.
- Random Forest: A random forest classifier is used with a maximum depth of 4.
- K-Nearest Neighbors: A K-Nearest Neighbors classifier is used with k=5.
- Logistic Regression: A logistic regression classifier is used.

Each of the machine learning models is set up using a pipeline that includes several steps for data pre-processing. The pre-processing steps involve handling missing values, scaling numeric features, and encoding categorical features.

The pre-processing steps are combined using a ColumnTransformer, which applies different preprocessing steps to different subsets of the input data based on the data types of the features.

Finally, the accuracy score of each model is calculated using the test set, allowing for a comparison of their performances.

Machine Learning Algorithms

Decision Tree: This is a type of classification algorithm that uses a tree-like model to make decisions. The model splits the dataset into smaller subsets based on the most significant feature that differentiates the classes. The splitting process is repeated until a certain stopping criterion is reached, such as a maximum tree depth or a minimum number of samples per leaf. The model then uses the resulting tree to make predictions on new data.

Random Forest: This is an ensemble learning algorithm that builds multiple decision trees and combines their predictions to make a final prediction. Each decision tree is built using a random subset of the training data and a random subset of the features. This helps to reduce overfitting and improve the generalization performance of the model. The final prediction is usually based on the mode (for classification) or the mean (for regression) of the predictions of the individual trees.

K-Nearest Neighbors: This is a type of instance-based learning algorithm that stores all available cases and classifies new cases based on their similarity to the stored cases. The algorithm uses a distance metric to calculate the similarity between the new case and the stored cases, and selects the K closest neighbors. The final prediction is usually based on the majority class of the K nearest neighbors.

Logistic Regression: This is a type of classification algorithm that models the probability of a binary outcome (e.g., yes or no) based on one or more predictor variables. The model uses a logistic function (sigmoid) to map the input variables to a probability score between 0 and 1. The decision boundary is usually set to 0.5, so any probability score above 0.5 is classified as the positive class, and any score below 0.5 is classified as the negative class. The model is trained using maximum likelihood estimation, and the coefficients of the logistic function are estimated using numerical optimization methods.

Additional Models Used

Additionally, a k-means clustering unsupervised model was used to try to discover hidden structure or groups within the data and visualize them.

Unfortunately, the results of the unsupervised learning failed to find patterns or relationships in the input data without any prior knowledge of what the output should be.

Reporting, Visualization, and Aggregation of Tools

Some reporting, visualization, and aggregation of tools include:

Jupyter Notebook: A web-based interactive computing environment that allows users to create and share documents that contain live code, equations, visualizations, and narrative text.

Matplotlib: A plotting library for the Python programming language and its numerical mathematics extension NumPy. It provides an object-oriented API for embedding plots into applications using general-purpose GUI toolkits like Tkinter, wxPython, Qt, or GTK.

Seaborn: A Python data visualization library based on Matplotlib that provides a high-level interface for drawing attractive and informative statistical graphics.

Sampled Input Data and Predicted Data

The input data consists of the features used for classification, which are extracted from the original dataset and split into training and test sets using the train_test_split method. Specifically, the input data consists of the following features:

- **Time:** Time of the flight
- **Length:** Length of the flight
- **DayOfWeek:** Day of the week of the flight
- **Airline:** Airline of the flight
- **AirportFrom:** Origin airport of the flight
- **AirportTo:** Destination airport of the flight

The target variable (predicted data) is the binary outcome of whether a flight is delayed or not. The target variable is extracted from the original dataset and split into training and test sets using the same train_test_split method. The target variable is stored in the y_train and y_test variables.

Each of the four machine learning models defined in the code is trained on the training set and used to predict the target variable for the test set. The accuracy of each model's predictions is then evaluated using the accuracy_score method from scikit-learn.

Sampled Input Data and Predicted Data

	Flight	Time	Length	Airline	AirportFrom	AirportTo	DayOfWeek	Class
0	2313	1296	141	DL	ATL	HOU	1	0
1	6948	360	146	OO	COS	ORD	4	0
2	1247	1170	143	B6	BOS	CLT	3	0
3	31	1410	344	US	OGG	PHX	6	0
4	563	692	98	FL	BMI	ATL	4	0

Python and any other Code (References to Libraries)

```python
import numpy as np
import pandas as pd
from sklearn.model_selection import train_test_split
from sklearn.metrics import accuracy_score
from sklearn.pipeline import Pipeline
from sklearn.impute import KNNImputer
from sklearn.preprocessing import StandardScaler
from sklearn.impute import SimpleImputer
from sklearn.preprocessing import OneHotEncoder
from sklearn.compose import ColumnTransformer
from sklearn.tree import DecisionTreeClassifier
from sklearn.ensemble import RandomForestClassifier
from sklearn.neighbors import KNeighborsClassifier
from sklearn.linear_model import LogisticRegression
```

```python
y = df.iloc[:,7]
x = df.iloc[:,1:7]
x_train, x_test, y_train, y_test = train_test_split(x, y, test_size = 0.3, random_state = 42)
```

```python
# Pre-process data: Numerical Xs

step_1a = Pipeline(
    [
        ("1A_i", KNNImputer(n_neighbors = 6)),
        ("1A_ii", StandardScaler())
    ])
```

```python
# Pre-process data: Categorical Xs

step_1b = Pipeline(
    [
        ("1B_i", SimpleImputer(strategy = 'most_frequent')),
        ("1B_ii", OneHotEncoder())
    ])
```

```python
# Pre-process data: Column transforms

num_x = ['Time', 'Length', 'DayOfWeek']

cat_x = ['Airline', 'AirportFrom', 'AirportTo']

step_1c = ColumnTransformer(
    [
        ('1C_i', step_1a, num_x),
        ('1C_ii', step_1b, cat_x)
    ])
```

```python
# Decision Tree

max_depth = 4

model_dt = Pipeline(
    [
        ('1C', step_1c),
        ('2_RF', DecisionTreeClassifier(max_depth = max_depth, criterion = "entropy"))
    ])

model_dt.fit(x_train, y_train)
```

```python
# Random Forest Tree

max_depth = 4

model_rf = Pipeline(
    [
        ('1C', step_1c),
        ('2_RF', RandomForestClassifier(max_depth = max_depth))
    ])

model_rf.fit(x_train, y_train)
```

Python and any other Code (References to Libraries)

```python
# K-Nearest Neighbors

n_neighbors = 5

model_knn = Pipeline(
    [
        ('1C', step_1c),
        ('2_RF', KNeighborsClassifier(n_neighbors = n_neighbors))
    ])

model_knn.fit(x_train, y_train)
```

```python
# Logistic Regression

n_neighbors = 5

model_lr = Pipeline(
    [
        ('1C', step_1c),
        ('2_RF', LogisticRegression())
    ])

model_lr.fit(x_train, y_train)
```

```python
print("Accuracy score of the Decision Tree model is " + str(model_dt.score(x_test, y_test)))
print("Accuracy score of the Random Forest Tree model is " + str(model_rf.score(x_test, y_test)))
print("Accuracy score of the K-Nearest Neighbors model is " + str(model_knn.score(x_test, y_test)))
print("Accuracy score of the Logistic Regression model is " + str(model_lr.score(x_test, y_test)))
```

Integration with External Applications

Databricks is a cloud-based platform that provides an integrated workspace for data engineering, data science, and machine learning. It is built on top of Apache Spark, a powerful open-source distributed computing framework that enables fast processing of large datasets.

The utilization of Databricks services allowed fast processing of our large data for machine learning purposes.

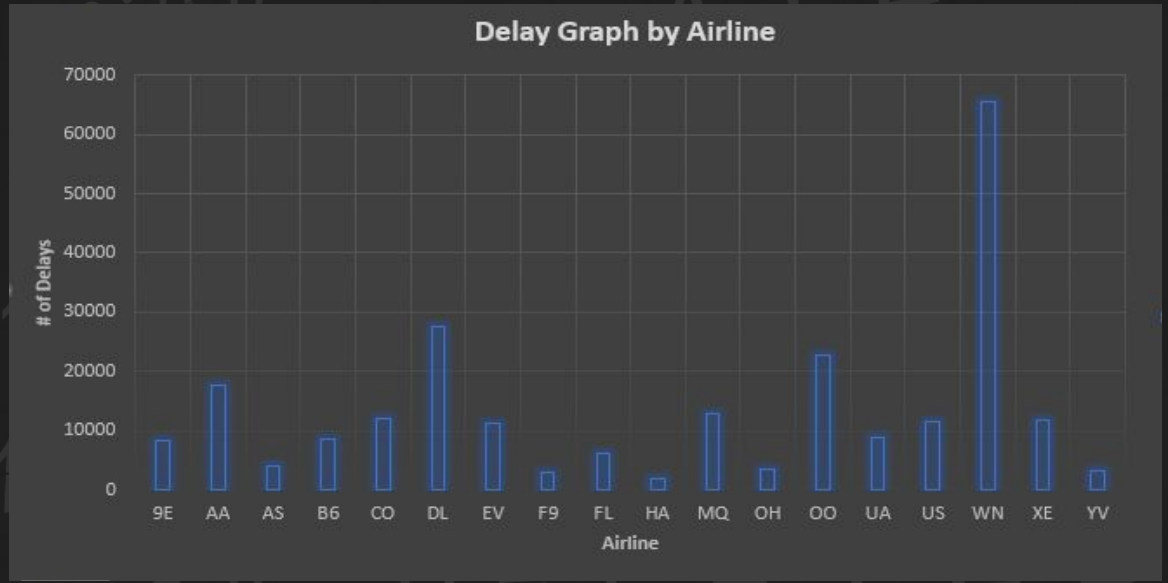
Results and Recommendation

Illustrations of Business Insights

The graph can be used to compare the delay performance of different airlines. It can help identify which airlines have the highest and lowest delays, and how they compare to each other.

Small airline operators like Hawaiian Airlines produce low delay numbers, while a big operator like Southwest Airlines produced the most number of delays during the given period.

This goes to show, we must provide a metric that can be used to group airlines based on their size and operational capacity.

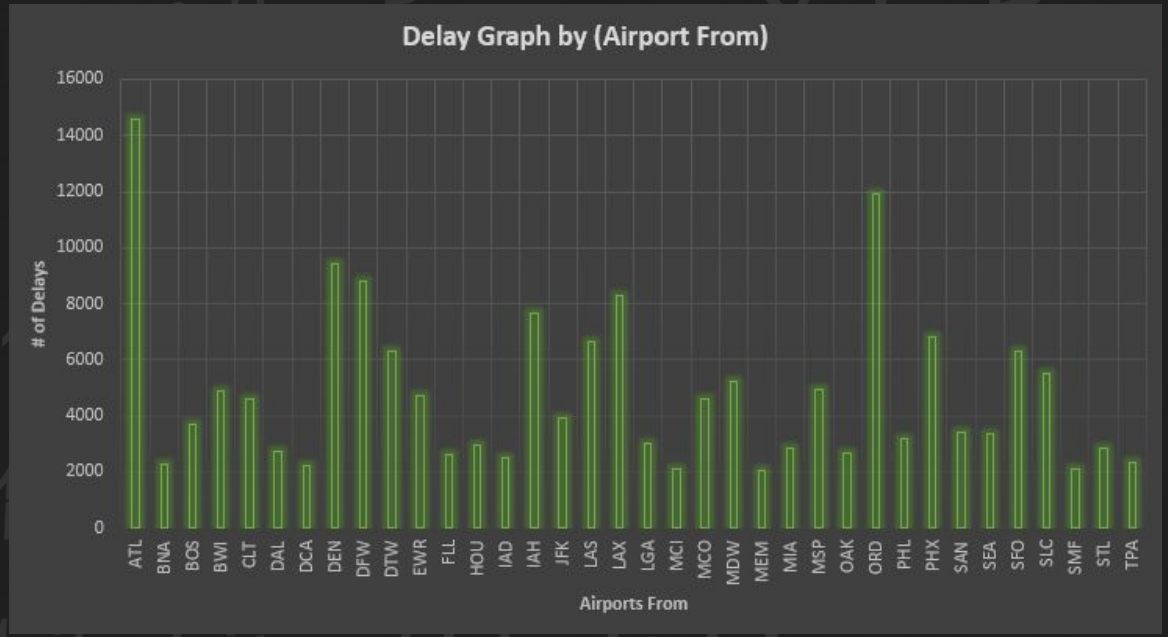


Illustrations of Business Insights

The graph can be used to compare the worst performing airports that flights are departing from.

Airports with small foot traffic, like BNA (Nashville International Airport), will most likely produce lower delays overall than airports with large foot traffic, like ATL (Atlanta International Airport).

The reasons may vary, from lack of flights departing from the airport, to high demand for flights that can cause overbooking and more complications.



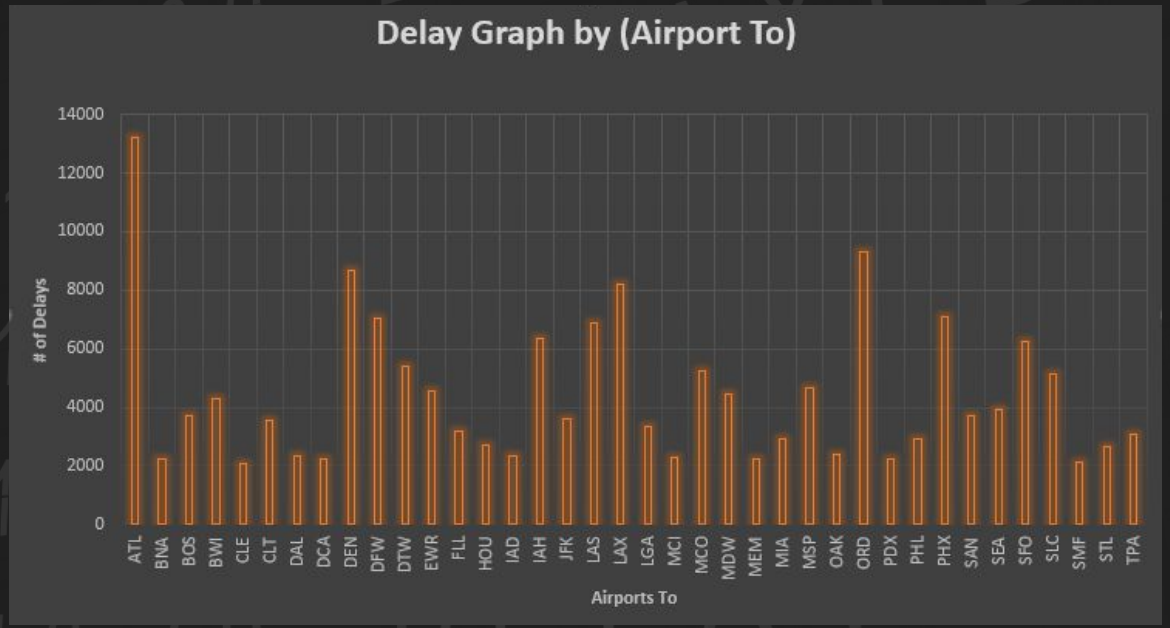
Illustrations of Business Insights

The graph looks almost identical to the previous Delay by Airport From graph.

This corroborates our previous point, that we should include metrics that can identify the operational size of different airports.

More flights arriving will also depend on the capacity of the airport to receive flights.

Demand and foot traffic can also ascertain the size of airports overall.



Recommendations on How to Improve Current Business Process for Big Data Use

- **Weather severity metric:** Weather can impact flight operations, and a severity metric can help in quantifying the impact.
- **Delay Analytics:** Complement real-time analytics with historical “flight delay” data
- **Airline Revenues:** Revenues of different airlines can help us categorize different airlines into small, medium, and large.
- **Airport Foot Traffic:** High foot traffic can lead to congestion, which can affect airport operations, such as delays.

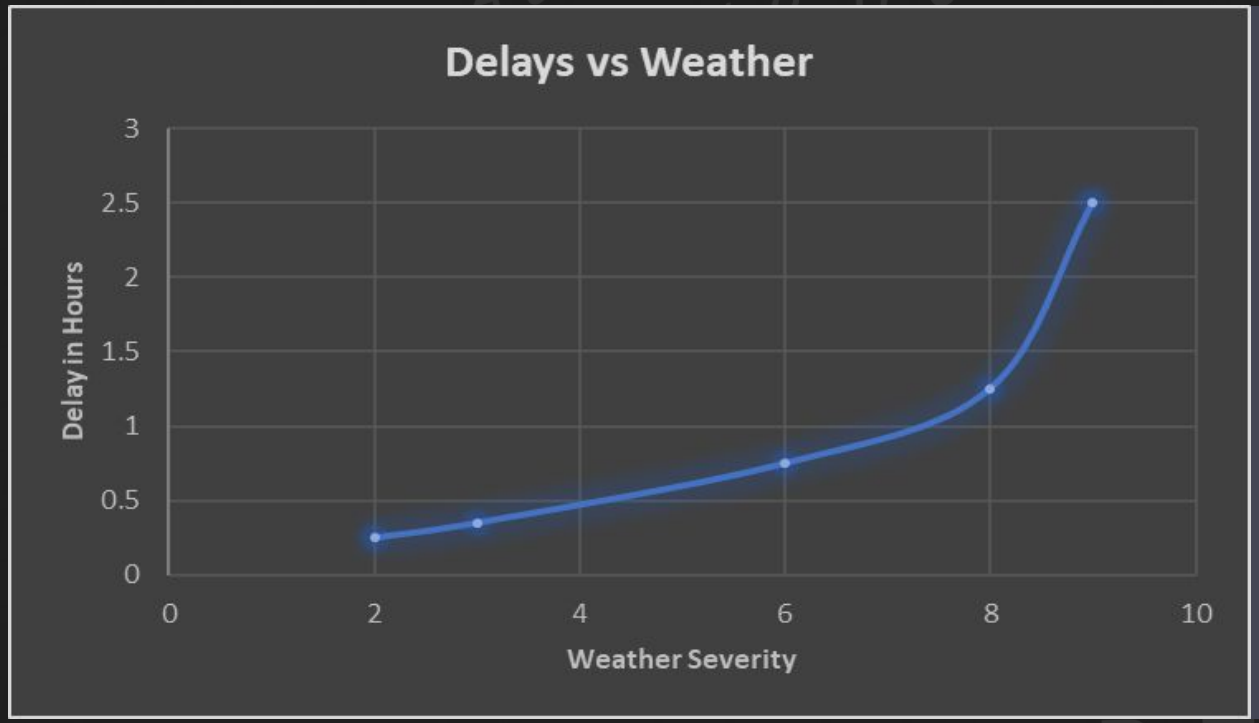
Keep record of every significant data that can be analysed and then utilized to be used in predictive analysis.

New Illustrations Provided to the Industry

- The data includes several relevant factors such as departure time, delay time, weather severity, weather conditions, airline revenue, and airport foot traffic.
- These factors directly relate to potential delays and can help the model account for the “complexity” of flight delays.
- By including these parameters, this data can provide more accurate predictions, resulting in an overall better predictive model.

Flight	Departure Time (24-Hours)	Delay Time (Hour)	Weather Severity	Weather	Airline	Revenues (Millions USD)	Airport
1	9:00	0.25	2	Cloudy	A	50	JFK
2	10:30	0.35	3	Sunny	B	150	LAX
3	12:30	0.75	6	Stormy	C	200	ATL
4	14:00	1.25	8	Snowy	D	300	ORD
5	17:30	2.5	9	Rainy	E	400	DFW

New Illustrations Provided to the Industry



- This is an example graph based on the new parameters
- The new parameters better illustrate the flight delay data.
- Leading to a better predictive model

Recommendations on Next Steps Based on Results

Unfortunately, our predictive analysis failed to predict flight delays accurately.

However, with improved flight delay analytics and relevant parameters, we can theorize the potential outcomes, and produce recommendations on next steps:

- Use the report to remind customers when a flight will be delayed to avoid disruptions to their travel plans - customer loyalty.
- Deduce demand beforehand to reduce delays overall.

Potential/Projected Savings or Improvements

We will aim to provide data that is more relevant with our model. With access to more pertinent and up-to-date data, the prediction models can capture the unique variables that may cause delays, such as weather conditions, air traffic congestion, airport foot traffic, and more. This data can then be analyzed and utilized to create more accurate prediction models, which can provide valuable insights into potential delays that airlines and airports can use to plan and adjust their schedules accordingly.

Projected Timeline of Implementation of Recommendation

Request a variety of airline companies to gather passenger data, such as time of the day that has the most amount of passengers taking flights, overall foot traffic, weather conditions that affect flight timings and patterns, etc. We will request them to gather this data for about one year.

We will then implement this real time gathered data into our prediction models in order to provide more accurate predictions of flight delays for a variety of airline companies. We expect this process of implementation to take approximately two months. After these two months, we will most likely have updated models and be able to provide better flight delay predictions.

Summary of Learnings from Project

- Flight delays are incredibly hard to predict using regular supervised classification models.
- More extensive analysis of data is required to fully understand the relationship and correlation between respective parameters. The limited analysis of data we were able to perform failed to train accurate predictive model.
- The input parameters have to be relevant to what we want to predict. A lot of the input parameters we had to work with had deficiencies pertaining to relevance, representation, and consistency.

References

“Flight Delays and Cancellations: A Guide | Canadian Transportation Agency.” Canadian Transportation Agency, https://otc-cta.gc.ca/eng/publication/flight-delays-and-cancellations-a-guide. Accessed 11 Apr. 2023.

“Flight Delays Cost More than Just Time, Airlines’ Reputation at Stake - Aviation Metric.” Aviation Metric, 15 Feb. 2022, https://aviationmetric.com/flight-delays-cost-more-than-just-time-airlines-reputation-at-stake/.

Smale, Natalie Lisbona &. Will. “The Airport Tech Helping to Prevent Delayed Flights - BBC News.” BBC News, BBC News, 7 Feb. 2022, https://www.bbc.com/news/business-60228430.

“Flight Delays Cost $32.9 Billion, Passengers Foot Half the Bill | Berkeley News.” Berkeley News, 18 Oct. 2010, https://news.berkeley.edu/2010/10/18/flight_delays/.

“Databricks - Sign In.” Databricks - Sign In, https://community.cloud.databricks.com/. Accessed 11 Apr. 2023.

Kemmer, Ryan. “Clustering on Mixed Data Types in Python | by Ryan Kemmer | Analytics Vidhya | Medium.” Medium, Analytics Vidhya, 25 Jan. 2021, https://medium.com/analytics-vidhya/clustering-on-mixed-data-types-in-python-7c22b3898086.