|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | | | | |
|  |  |  | |  |
| How to reliably predict airline delays | | |
| Shantel Culver | Tosin AdeboteAshley Drewry | Carolina PachecoMelissa Flores | Jiaxuan Ji | Sylvia Nanyangwe | | DS4A – Team 16 |

Table of Contents

[Meet Team 16 3](#_Toc84443614)

[Overview 4](#_Toc84443617)

[The problem we are looking to solve 4](#_Toc84443618)

[How to accomplish this with data 4](#_Toc84443619)

[Why does this matter? 4](#_Toc84443620)

[Potential Audience 5](#_Toc84443622)

[About the Data 6](#_Toc84443624)

[Sources 6](#_Toc84443623)

[Data Wrangling and Cleaning 7](#_Toc84443623)

[HEADING 8](#_Toc84443625)

**Team 16**

MEET the TEAM

**Tosin Adebote**

Data Analyst

**Shantel Culver**

Data Analyst

**Ashley Drewry**

Team Lead

**Melissa Flores**

Technical Lead

**Jiaxuan Ji**

Data Analyst

**Sylvia Nanyangwe**

Data Analyst & Visualization

**Carolina Pacheco**

Data Analyst & Visualization

|  |  |
| --- | --- |
|  |  |
| OVERVIEW The airline industry provides air transportation for passengers and cargo by using aircraft. Flight delays and cancellations can significantly affect customer experience and cause financial loss to the business. The total cost of delays from 2016-2019 was over US$23 billion. Define the specific problem that should be solved Assist customers in making better decisions when booking flights and help businesses mitigate financial loss due to delays.   * Predict flights that will be delayed allowing passengers to avoid those flights if they chose. * Allow businesses to do root-cause analysis on flights with habitual delays as well as apply mitigation strategies to avoid negative customer experiences related to delays.   Figure 1 | |
|  | |  |
| The table above shows cost estimates of delays reported from the Federal Aviation Administrations in 2019  <https://www.faa.gov/data_research/aviation_data_statistics/media/cost_delay_estimates.pdf> | |  |

**How to accomplish this with data**

* Analyze flight delay data: detecting seasonality, airport and connection flights patterns and any correlation.

**Why does this problem matter?**

Delays have costs airlines over US$23 billion. Additionally, negative customer experiences further erode a company’s reputation and future profit by loss of business.

|  |  |
| --- | --- |
|  |  |
| Potential Audience  * Anyone who is booking flight arrangements (for self or others) would be interested. This information would allow them to make informed decisions regarding flights and/or airports. * Airline executives could benefit by identifying frequently delayed flights and do deep dives into cause and effects.  |  |  |  | | --- | --- | --- | |  | Shape  Description automatically generated | Shape  Description automatically generated with low confidence | | Business  Airports, airlines, booking agents, any business profiting from the sale of air transportation. |  | Consumer  Business travelers, recreational travelers, frequent flyers, budget conscious individuals, etc. |   We believe our delay insights will be mutually  beneficial for the business as well as the consumer. | |

Flights

* November 25 has the MOST # of flights during 2018 (﻿22160)
* November 22 has the LEAST # of flights during 2018 (﻿12290)

Airlines

* Total airlines in 2018 = 18

Airports

* There are over 5000 airports in the USA
* The following are the busiest airports in the USA:

ATL

LAX

ORD

DFW

DEN

JFK

SFO

LAS

SEA

CLT

MCO

MIA

PHX

EWR

IAH

* Orlando International Airport (MCO) - 45M Passengers
* Miami International Airport (MIA) - 44M Passengers
* Phoenix Sky Harbor International Airport (PHX) – 44M Passengers
* Newark Liberty International Airport (EWR) – 43M Passengers
* George Bush Intercontinental Airport (IAH) – 41M Passengers

|  |  |
| --- | --- |
| ABOUT THE DATASOURCESAirline Delay and Cancellation Data, 2009-2018  * US flights * Jan 2009-Dec 2018 * 6.43m for 2009, 6.45m for 2010, 6.07m for 2011, 6.10m for 2012, 6.37m 2013, 5.82m for 2014, 5.82m for 2015, 5.62m for 2016, 5.67m for 2017, 7.21m for 2018 * Has column for departure delay, a column where e 2% of entries have missing data for this variable * Has unique carrier code and the flight number. * Captures arrival and departure delays (with how long the delay was rather than a flag). * Link: <https://www.kaggle.com/yuanyuwendymu/airline-delay-and-cancellation-data-2009-2018>  January Flight Delay Prediction  * US flights * Jan 2019- Jan 2020 * 587k observation for 2019 and 607k for 2020 * Data is very complete * There is a flag column indicating it was delayed by 15 min or more. * This dataset has the unique carrier code and the flight number. * Captures arrival and departure delays (with how long the delay was rather than a flag). * Link: <https://www.kaggle.com/divyansh22/flight-delay-prediction>  Feb 2020 US Flight Delay  * US flights * Feb 2020 * 574k observations * Has a flag for departures that were more than 15 minutes late, but not how long the departure was. * Departure time and the departure delay flag are missing data for 1% of entries * Has unique carrier code but not the actual flight number. * Has 9 columns when the previous two datasets had 28 and 21, respectively. * Captures only departure delays, with a flag rather and not the duration of the delay. * Link: <https://www.kaggle.com/rowhitswami/feb-2020-us-flight-delay>   \*\* The original source for all data sets from Kaggle is the [Bureau of Transportation Statistics](https://www.bts.gov/). EDADATA WRANGLING AND CLEANINGOPERATIONALIZING THE VARIABLES  1. **What is the definition of a delay?**  * Per the [Federal Aviation Administration](https://www.faa.gov/), a flight is considered delayed when it is 15 minutes later than the scheduled time.  1. **What fields are we looking at to determine the delay?**  * The dataset we are using has 28 columns present. We are focusing on the DEP\_DELAY column which represents the number of minutes the flight is delayed beyond the scheduled departure time.    prepping the data Using the Panda’s built-in libraries, we performed the following items on our working dataset.   * Removed duplicate rows * Removed unnecessary data * Cancelled flight rows as we are not analyzing this portion of the dataset * Flights with a negative DEP\_DELAY value * Converted date and time columns to date/time variables * Added geopoints for the airports to use for mapping purposes   **Missing Values**  There are 18 columns with missing values. The number of missing values in these columns are close; this could  indicate they are in the same rows.  A picture containing text, air conditioner  Description automatically generated  White lines indicate missing values. Most of the missing values in those 18 columns are in the same rows. The rows with missing values might belong to *cancelled* or *diverted* flights. When looking at “CANCELLED” and “DIVERTED” columns, the number of cancelled flights is approximately the same as the number of missing values in columns about flight information.  Text  Description automatically generated  Text  Description automatically generated  Text  Description automatically generated  Diverted flights also cause missing values. The missing values are because of *cancelled* or *diverted* flights. Therefore, we decided to drop these columns.  Chart, histogram  Description automatically generated  Chart, bar chart  Description automatically generated  **Frequency distribution of delays:**  Chart  Description automatically generated  Chart, bar chart, histogram  Description automatically generated  Chart, histogram  Description automatically generated  \*There are over 5000 airports in the USA  \*Check Top 20 busiest airports  **Diverted & Cancelled Flights:**  Chart, bar chart, histogram  Description automatically generated  Preliminary graph analysis might indicate ‘cancelled’ and ‘diverted’ flights are the same; however, there are a total of 17,859 and 116,584 ‘diverted’ and ‘cancelled’ flights respectively.  Chart, bar chart  Description automatically generated  Diverted  0.0 7195587  1.0 17859  Chart, bar chart  Description automatically generated  Cancelled  0.0 7096862  1.0 116584  Boxplot of delays by airlines and airports  Chart, box and whisker chart  Description automatically generated  A picture containing box and whisker chart  Description automatically generated | |
| Dashboard 1 | |
|  |  |
| Modeling **Overview**  Prior to our modeling implementation phase, we have looked at the underlying nature of our dataset, which provides high-level methodology to select statistical models which adhere the most to both our research goal and the data. Since our model’s response variable is binary aiming to categorize flights as delayed or not, we have reviewed several classification models such as KNN, Logistics, Random Forest, and Neural Network.  The aforementioned exploratory data analysis suggests that the frequency distribution of delayed flights is highly skewed. One potential issue stemming from this skewness is that our model is less likely to identify delayed flights given only a relatively small portion of input are labeled as “delayed.” The below graph shows the prediction result from one classification model, which from the first look, indicating the result leans towards flights which are not delayed.  Calendar  Description automatically generated with low confidence  Another potential problem follows is that the independent variables could be highly correlated. We have utilized correlation maps (see below) to verify the property of heterogeneity.      While the below correlation map does seem suffer co-variance problem, certain classification methods such as logistics regression is more likely to underperform when includes excessive parameters. Therefore, we decide to narrow down the number of independent variables to consider by Random Forest to measure the high-level dynamics across each variable, and then using Logistics Regression to delve into one or a couple of indicators which explain the most to the variance of our target variable, the delay-flag. We also consider logistics regression combined with Ridge and LASSO for the purpose of regularization and minimize the likelihood of overfitting. Finally, to better resolve the underlying intricacy and flaws of our data, we attempt more advanced deep learning model such as Neural Network.  The model was designed to predict the probability that a flight is delayed by 15 minutes or more. In other words, the following models are binary classifiers. The data from 2014 to 2017 was used to train the model. Data from 2018 was used to evaluate the performance of the models.  The models were run using Google Colab. Since the instance could be closed at any time and availability depends on the use of Google resources at that moment, this required output from tuning to be saved periodically on Google Drive.  **Random Forest & Multinomial Logistics Regression**    **Graphical user interface, text, application  Description automatically generated**  We have implemented both random forest and multinomial logistics regression to explore the dynamic across independent variables. For our random forest model, we have selected 12 variables, which accounts to explain 61% of variance in delayed flights. For our multinomial logistics regression, we have conclude all independent variables, and achieve an accurate rate of 0.915. While this predication rate seems impressive, we are concerned that this might attribute to the skewness of data, and hence, decide to focus more on single logistics regression in the following section.  **Single Logistics Regression**  **Chart  Description automatically generated**  **Text, letter  Description automatically generated** | |
| To select DISTANCE as an explanatory variable, our logistic model has achieved an accuracy rate at 0.83 for our test set samples. This seemingly suggest that that distance could serve as indicator to predict the likelihood of delay. | |

|  |
| --- |
|  |

|  |
| --- |
| Neural Network Configuration The use of a neural network was motivated by the fact that none of the variables were strongly correlated with a flight departure being delayed by 15 minutes or more. This means that there may be more complex relationships between the variables.  Since neural networks require numerical data, categorical variables such as month, day of the week, and carrier were converted into dummy variables. Since geopoints were added to the departure and destination airports, the names of the airports were dropped from the data and the latitudes and longitudes were ingested instead. The input data was normalized to increase training speed and potentially improve the performance of the model.  A random sample of 50% of the training data was used to train the model since there were problems with consuming too much RAM in Google Colab. It is interesting to note that the random sample did produce better results than using 2014 data for training. It could be that using more recent data produces better results.  The Sequential model was implemented in Keras and was tuned using Kera Tuner[[1]](#footnote-1). A binary cross entropy loss function, which is widely used for binary classification, was selected with an Adam optimizer, which is often recommended as the default optimizer. The model has two hidden layers, both which use ReLU activation functions, which are less computationally expensive than other activation functions[[2]](#footnote-2). Two hidden layers were selected because there were only 58 features fed into the model. Dropout layers were not considered since there were millions of observations in each year of training data and therefore the risk of overfitting was low. The output layer uses a sigmoid activation function to predict which class was more likely for a given observation.  One project looking at departure delays in San Francisco solved this problem with two layers[[3]](#footnote-3) and theoretical work suggests that two hidden layers can approximate any function[[4]](#footnote-4), so instead the focus was on adjusting the node size. The input layer has 58 nodes, one for each feature. Since the input data is normalized to be in the same range, bias nodes are not necessary. Since a sigmoid activation function was used, we would have chosen one node for the last layer.  Next, Keras Tuner was used to optimize the model. The optimizer identified that 480 nodes in the hidden layer and a learning rate of 0.001 would be optimal. We have more nodes in the hidden layer because there are patterns that we can identify by extending to more nodes.    After four epochs, the loss was not reduced further. The accuracy even began to decrease after six epochs, suggesting the model began to overfit. Neural Network Results   The figure above summarizes the performance of the optimized neural network. Since delays (labeled as 1) were rare, it was more difficult for the neural network to learn this class. In contrast, the neural network worked better for non-delayed flights (labeled as 0).  While the downside of deep learning models is the lack of explainability, tools have been developed to identify which features are important. Scott Slundberg, who works on Explainable AI at Microsoft Research, developed a Python tool called SHAP (SHapley Additive exPlanations[[5]](#footnote-5)) to identify which features are important and show the directionality.  A picture containing graphical user interface  Description automatically generated  The figure above suggests that Feature 27, the month of January, and Feature 57, the carrier, are the most influential in determining if a flight will have a delayed departure. However, this figure was created only using a small subset of the training data (1000 observations) due to computational constraints. The month of January has many delays due to weather and the holiday season which starts before Christmas and extends into after the New Year. While it is surprising not to see the month of December as a feature that influences the prediction, this may be due to delays in January being longer, and therefore more flights being 15 or more minutes late. |

|  |
| --- |
|  |
| Conclusion The goal of this exercise was to predict departure delays of at least 15 minutes. A model that predicts delays can help customers plan and airlines to intervene to reduce delays. The neural network’s prediction is better than guessing. It even reveals that the time of year and carrier are important in predicting whether a flight is on time. This suggests that carriers can play a large role in preventing some delays. However, airport operators will also have to make changes given that the departure airport was an important feature for predicting if a flight is delayed.  Given the impact of time of year on delays, future work can consider using weather data to predict delays. Since weather can be predicted up to 10 days in advance, it can provide value to both customers and airline or airport operators. Additionally, future neural network models can use data on previous delays that day to predict if there will be a future delay. This can be accomplished using a Long Short-Term Memory (LSTM) layer. This is well-suited for this problem because it can better leverage our time series data. The data is of every flight in the United States from 2014 to 2018. It includes departure and arrival times. This allows us to leverage the fact a delay in one location could produce a delay in the connecting location. Future work can consider using bagging techniques to leverage the skill of different types of models. | |

1. Keras Team. (2019, May 31). *Keras documentation: Getting started with KerasTuner*. Keras Documentation. Retrieved October 22, 2021, from https://keras.io/guides/keras\_tuner/getting\_started/ [↑](#footnote-ref-1)
2. Glorot, X., Bordes, A., & Bengio, Y. (2011). Deep Sparse Rectifier Neural Networks. *Proceedings of the 14th International Con- Ference on Artificial Intelligence and Statistics*, *15*. [↑](#footnote-ref-2)
3. Wang, Y. & Sun, Y. (2018). Predicting Delays in Flight Departure Time at SFO. Retrieved October 22, 2021, from http://cs230.stanford.edu/projects\_spring\_2019/reports/18680989.pdf [↑](#footnote-ref-3)
4. Goodfellow, I. (2015). *Deep Learning.* The MIT Press. [↑](#footnote-ref-4)
5. Slundberg, S. (2018). *Welcome to the SHAP documentation — SHAP latest documentation*. Welcome to the SHAP Documentation. Retrieved October 22, 2021, from https://shap.readthedocs.io/en/latest/index.html [↑](#footnote-ref-5)