Flight Delay Analysis and Predictions with Supervised Machine Learning

Group 21

Chen, Rachel, r52chen
Jajcanin, Igor, ijajcani
Selvadurai, Athithian, a6selvad
Zhang, Cong, c236zhang

Project Objective

The primary goal of this project is to analyze data gathered for flight delays in January of 2019 and 2020, to identify the airline, airport, day of the week and time with the best and worst delay record. Also we are getting into machine learning to predict possible future flight delays.

Data collection is based solemnly on departure time information as well as duration of the delay for participated airlines and airports. It does not collect other factors that could have been related to weather conditions, aviation system control or any other factors that could result in flight delay. Even though statistically the majority of departure delays are related to previous flight arrival delays, this information is not considered in this project.

The process includes data gathering, data cleaning, data analysis based on the airline, origin airport, as well as the departure time grouped by the day of the week. Analysis should also include departure times in four intervals, morning, daytime, evening and night. The final step of this project focuses on machine learning and flight delay prediction.

The main aspects of python library and techniques that we used are as follow:

Data cleaning/manipulation: pandas, numpy

visualization: matplotlib, seaborn

modelling: sklearn

machine learning algorithm: KNN

Introduction

Data Analysis

Data analysis is a process of capturing, cleaning, transforming and modeling data in order to gain insight on the available dataset. The focus here is to understand the data, find patterns, discover relationships within a dataset to answer some questions one might have. Data analysis has a longer history, traditionally data analysis was primarily reactive and focused on reporting, over time, it has evolved from dealing with smaller, structured dataset to steam and process more complex and unstructured data sources. More and more organizations are relying on data analysis to make informed business decisions. In this report, you will see us using the data analysis techniques to analyze the flight delay dataset to answer questions such as which airport or which air carrier has the worst flight delay records.

Machine Learning

For a dataset, good data analysis can set a solid foundation for machine learning. Machine learning uses algorithms to build models and learn from a dataset. Instead of traditionally human focus on programming, the machine is now learning on its own and keeps improving with each iteration. There are three types of machine learning: Supervised learning, Unsupervised Learning and Reinforcement Learning. For our project, we are using supervised learning. In supervised learning, the machine studies the training set data using assigned target and variables, the target will help the algorithm to correlate with the variables (features). The two popular supervised learning methods are classification and regression. Classification predicts discrete value while regression predicts the value of a continuous response variable. For our group project, we have decided to use supervised learning classification methods to predict whether a flight will be delayed or not.

Background

As flight delays are becoming increasingly common, they create major problems, negatively impact many stakeholders such as the aviation system, airlines, small businesses, and of course, the passengers. Taking passengers as an example, when flights are delayed, passengers face increased travel time and end up with increased expenses. Passengers will likely have to adjust their car rental, lodging plan etc. which also increase stresses. Our group thought it is important to analyze flight delay patterns, identify the airline, the airport, the day of the week and time that most flight delay occurs so passengers can be prepared and make alternative plans. By doing so, not only it will benefit the passengers, it will also help all parties involved.

Flight delay reasons are very complex, it could occur due to the problems at the original airport, the destination airport, weather, mechanical problems, any ground reasons, security problems, national aviation system decisions, air carries etc. Most of the time, flight delay happens due to multiple reasons. For the purpose of this project, we will not go into why and how the flight delay happens, but instead focus on the origin airport and the week of day and time that flight delay occurs.

We have also decided to utilize machine learning to predict the flight delay at the destination airport for the month of January in the upcoming years.

Data Analysis

Dataset Overview

There are two .csv files we used in our data analysis and machine learning exercises. They are downloaded from https://www.kaggle.com/divyansh22/flight-delay-prediction. This data is originally collected from the Bureau of Transportation Statistics, Govt. of the USA. These datasets contain all the flights in the month of January 2019 and January 2020.

As stated on the link above: “The two files contain all the flights in January. There are around 400,000 rows and 21 columns indicating the features of the flight including information about origin airport, destination airport, airplane information, departure time and arrival time.”

Here are the attributes information:

'DAY_OF_MONTH': Day of the month.

'DAY_OF_WEEK': Day of the week.

'OP_UNIQUE_CARRIER': Unique transport code.

'OP_CARRIER_AIRLINE_ID': Unique aviation operator code.

'OP_CARRIER': IATA code of the operator.

'TAIL_NUM': Tail number.

'OP_CARRIER_FL_NUM': Flight number.

'ORIGIN_AIRPORT_ID': Origin airport ID.

'ORIGIN_AIRPORT_SEQ_ID': Origin airport ID - SEQ.

'ORIGIN': Airport of Origin.

'DEST_AIRPORT_ID': ID of the destination airport.

'DEST_AIRPORT_SEQ_ID': Destination airport ID - SEQ.

'DEST': Destination airport.

'DEP_TIME': Flight departure time.

'DEP_DEL15': Departure delay indicator

'DEP_TIME_BLK': block of time (hour) where the match has been postponed.

'ARR_TIME': Flight arrival time.

'ARR_DEL15': Arrival delay indicator.

'CANCELLED': Flight cancellation indicator.

'DIVERTED': Indicator if the flight has been diverted.

'DISTANCE': Distance between airports.

Data Preparation

Since we have two separate sets of data with the same type of information, we decided to prepare the two datasets with different approaches but getting the same result. Once the two datasets are properly cleaned, modified and engineered, we merge them together to have one Panda Frame for all the data we want.

The first step of any good data clean up/preparation is deciding what library we will be using and import them.

1. 2019 Dataset Preparation

The files are in .csv format, thus we imported pandas and used pandas .read_csv() function to import the file into a pandas Dataframe. First we used pandas .head() function to get the first 5 rows of data, it is a quick way to check what data there is in a dataframe. To examine the dataset more closely, with the pandas.info() function, we are able to get a concise summary of the dataframe. This summary includes a list of all columns with their data types and the number of non-null values in each column. Here we can clearly see that the dataset is very big, over 500k rows, so we decided to get a random sample of rows with all the columns. Many datasets have more information than we need, the next step here is to look at all the features, focusing on our project objective, identifying the related feature from this dataset. Upon observation, we decided to drop the following columns: 'TAIL_NUM','OP_CARRIER_FL_NUM','ORIGIN_AIRPORT_ID','ORIGIN_AIRPORT_SEQ_ID','DEST_AIRPORT_ID','DEST_AIRPORT_SEQ_ID','CANCELLED','DIVERTED','ARR_TIME','Unnamed: 21'

These features are either duplicated information that we could get from other features or they simply just don't contribute to our final problem. Using ORGIN_AIRPORT_ID as an example, since we have the "ORIGIN' as the identifier for the origin airport already, there is no need for the ID column. The next thing we did was to check the missing value, using .isnull() and .any() combination, we found out there were only 1296 rows with NaN value, which is only 0.22% of the whole dataset. Since most machine learning algorithms do not support data with missing values, it is important to handle them before we go any further. In general, when there is a very small amount of data missing, we can drop them, that is what we did next.

Now we have cleaned the data, we will transform the data into a format that is more suitable to answer our questions and better for machine learning algorithms. This is when we started to use feature engineering. Since we want to calculate the flight delay amount in minutes, and there is no scheduled departure time in the original dataset, we decided to engineer the scheduled departure time using column 'DEP_TIME_BLK'. First we break the block into estimated departure start time and end time, then we use the following logic to estimate the scheduled departure time: If flight was not delayed, assume scheduled departure time is equal to actual departure time; otherwise, assume scheduled departure time is middle of the block. Another feather engineering, we did was binning. Since we are dealing with continuous time, we thought it would make more sense to bin the flight time into 4 sections: morning, afternoon, evening and night time flight. This will help us better explain the data. Using feather engineering, we added two columns into our dataset 'DEP_TIME_SCHEDULED' and 'TIME_GROUP'. The final step for preparing the 2019 dataset is to calculate the time delayed by subtracting scheduled departure time from actually flight take off time. Since this returns an integer, a function is applied to convert the integer into minutes. Thus, we got the final column we wanted to create 'DELAY'. Now we cleaned the data, added the variables we need by feather engineering, our dataset is ready to be analyzed.

2. 2020 Dataset Preparation

With 2020 data, we took a little bit different approach to be more "Python like" and we tried to use more Pandas available functionality.

Data is loaded in the same manner as 2019 by using .read_csv() function. We analyzed data frame and concluded that 2020 file contains same number as type of columns as 2019, and over 600k records. Column "Unnamed: 21" as all NaN and was obviously redundant and as it was of no significance the column was immediately dropped. Upon further the analysis, we dropped following columns

"OP_UNIQUE_CARRIER","OP_CARRIER_FL_NUM","OP_CARRIER_AIRLINE_ID","TAIL_NUM","ORIGIN_AIRPORT_SEQ_ID","DEST_AIRPORT_SEQ_ID","ARR_TIME","ARR_DEL15"

We did not drop the CANCELLED column as we wanted to make sure there is no DEP_TIME for any Cancelled flight and true there was not.

Next we wanted to make sure that 2020 data does not contain any null values and since there is only less than 1% of all data we concluded that dropping those records will not significantly affect our dataset. Therefore, we dropped those records and got data without any missing record.

Now that we cleaned the data we needed to transform the actual departure time, estimate scheduled departure time and find the delay. In order to do that we transformed the actual departure time from float to datetime format. The reason is that it would be easier to calculate deploy as difference between two datetime items. Additionally, in case when actual departure is on the next day, i.e. pasted midnight and scheduled time is before midnight, this way was easier to calculate delay. To find scheduled departure time the same as in 2019 we split DEP_TIME_BLK into two times. Those times were also converted from float to datetime data type. Since we already had function that was calculating scheduled time and to have the same algorithm we used the same function dep_time_scheduled().

The final step was to split the departure time into the same four groups as for 2019, MORNING, AFTERNOON, EVENING and NIGHT. Only this time we used four bins.

Note what we clean data for both 2019 and 2020 we analyzed what are common columns that are most relevant for data analysis. We concluded that these are columns what will give us desired analysis

```
'DAY_OF_MONTH','DAY_OF_WEEK','OP_CARRIER','ORIGIN','DEP_TIME','DEP_TIME_SCHEDULED','DEP_DEL15'
,'DISTANCE','DELAY','TIME_GROUP'
```

We excluded all canceled and diverted flights from the final DataFrame and concatenated both data frames.

Data Analysis

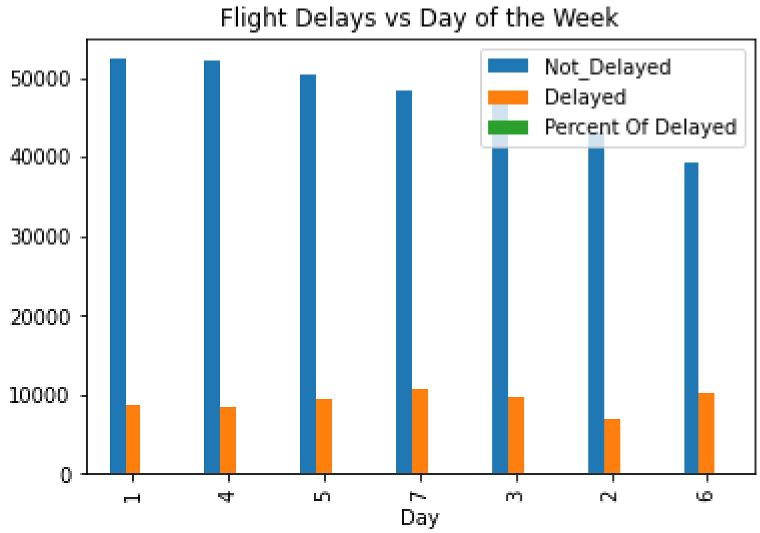

Fig 1. A breakdown of the flights per day of the week, 1 being Monday

As we can see in Fig 1. While the busiest days for flights seem to be Monday and Thursday they also seem to be the period of days with the least delays. In fact, the days with the highest amount of delayed flights are actually the weekend, Saturday and Sunday. Saturday flights are delayed the highest frequency with a 20.7% chance the flight will be delayed, and Sunday with a rate of 18.3% has the highest volume of delayed flights. Now if flight delays were correlated with flight volume i.e. a set percent that was constant we should see that the weekday business period, Monday-Thursday, which has the highest volume of flights should have the highest amount of delays but this isn't what is displayed. There could be a few reasons for this, perhaps airlines are more diligent during the work week and less so on the weekend when flights are likely leisure in nature, but it is clear there is an outside factor rather than just what day of the week it is.

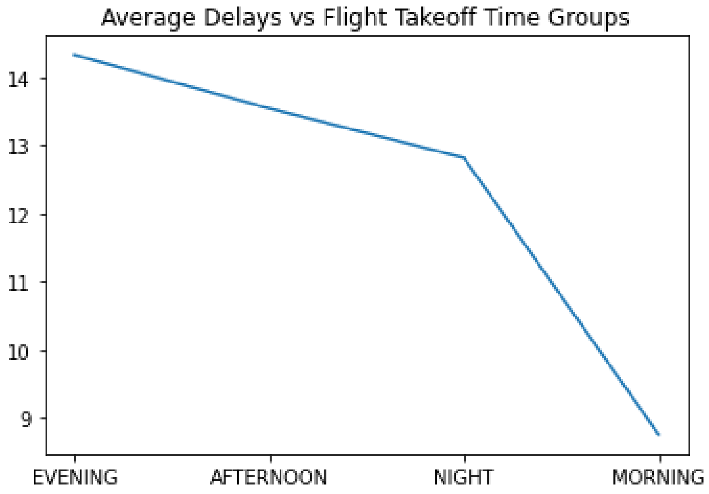

Fig 2. Average delay in flights based on their departure time, grouped into 4 groups

In Fig 2, we sorted our DEP_TIME_SCHEDULED data into 4 windows, 'NIGHT' or redeye flights, from 12 am to 6 am, 'MORNING' flights which were from 6 am to 12 pm, 'AFTERNOON' flights from 12 pm to 6 pm, and 'EVENING' flights which were from 6 pm to 12 am. What we see after plotting our data is that flights in the Evening and Afternoon see the largest delays on average, while redeye flights see smaller delays, and morning delays are significantly smaller. A potential causation here could be that at starting and after in the afternoon the airports have fewer employees staffed, increasing delays. It drops at night because there are fewer red eye flights than during the day so despite the lack of staff, the lack of flights decreases the delay values. Finally, in the morning, the airport is fully staffed, and everyone has had their coffee, or at least have more energy to work more diligently, decreasing delays.

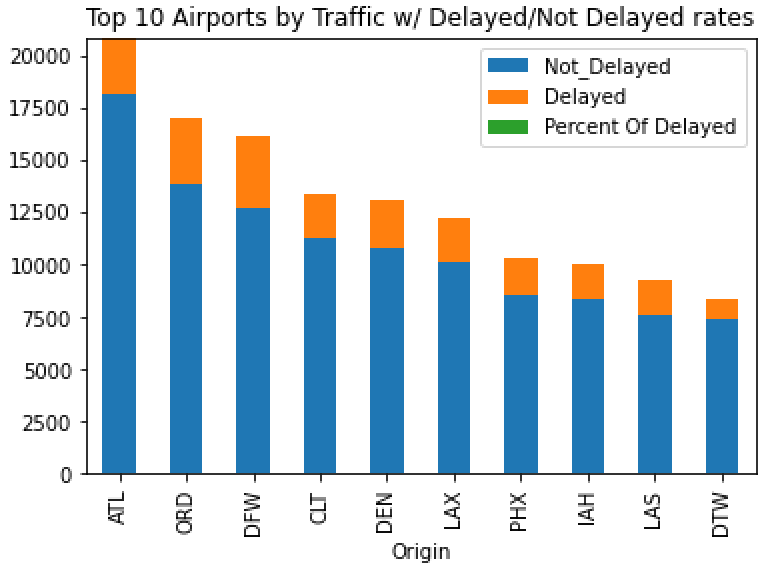

Fig 3. Breakdown of Flights, delayed or not, in the top 10 airports by flight traffic.

In fig 3 we see the top 10 most frequented airports, which unsurprisingly are all American Airports. We chose to show the top 10 most frequented airports, which are not necessarily the top 10 in delay rates, as we wanted to look at how the busiest airports handle delays, which is more indicative than a poorly run small airport with high delay rates. Our airports that see the fewest delays are on both ends of our graph, with Detroit Metropolitan Wayne County Airport having only a 11.2% rate of delay, and more shocking, Hartsfield-Jackson Atlanta International Airport having a delay rate of only 12.9%. This is surprising because ATL is by far the most trafficked airport in the world, but has one of the lowest delay rates, beating the next most efficient airport by almost 3%! This pattern is similar to our revelation from Day of the Week data, higher volume does not necessarily correlate to higher delay rates. Let's look at the highest delay rates, being Dallas Fort Worth International at a whopping 21.4% and Chicago's O'Hare International coming behind at 18.6%. Unfortunately, there does not seem to be a factor that is easily seen as these airports do share similar geography with at least one other airport that performs either better or worse, and are all international airports thus seeing the same types of flights leaving from the airports.

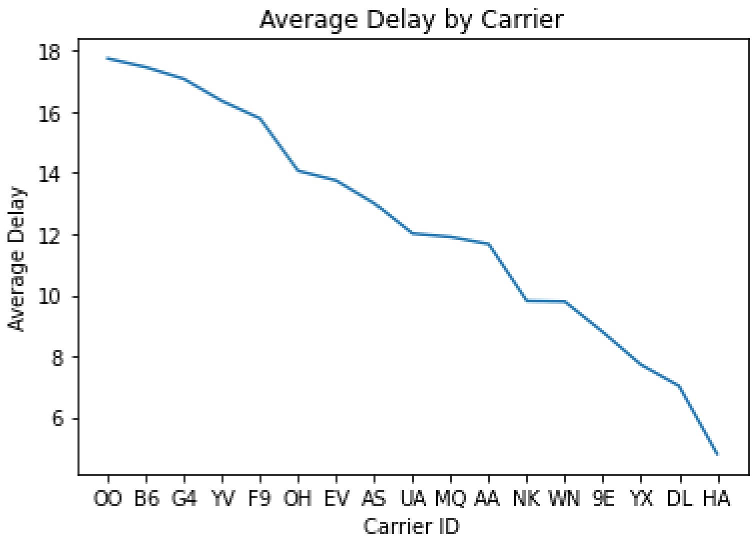

Fig 4. Average Delay by Airlines, via their unique carrier ID. Skywest, a regional airline has the highest delays, followed by JetBlue, a low-cost airline, and Allegiant Air, another low-cost airline. Compared to the lowest delay airlines, Delta and Hawaiian Airlines. Other than a useful aid at avoiding heavy delay airlines, it does seem the reasonable correlation shown here is that low-cost airlines offer service quality that matches their low price point, compared to Delta and Hawaiian, both top 10 airlines in the United States.

Machine Learning

KNN

We used K Nearest Neighbors ("KNN") to predict whether a flight is likely to be delayed. The features that we believed are relevant are as follows:

- Day of week
- Airline
- Origin airport
- Scheduled departure time
- Distance

The target variable that we are trying to predict is DEP_DEL15, which is an indicator of whether a flight is delayed (1 for yes and 0 for no). Of the features above, Airline and origin airport are categorical and in order to perform KNN, we first had to encode these into numerical labels.

The full dataframe has over a million rows and is computationally expensive to run KNN on since KNN checks the Euclidean distance between each training datapoint and all of its surrounding neighbors. For KNN, we randomly sampled 100,000 rows from the full dataframe and split 70/30 between the training/testing sets.

The KNN method produced a fairly high accuracy score of 0.845, meaning that it's accurate ~85% of the time in predicting whether a flight is likely to be delayed, based on the aforementioned features.

Linear Regression

We also used linear regression to predict the exact number of minutes of delay for a given flight. Linear regression does not work with categorical variables so we first converted each unique value of the categorical variables Distance Group and Departure Time Group to their own columns.

The regression had a very low accuracy score of about 3% but this makes sense since we are trying to predict the exact number of minutes of delay. However, when we calculate the average prediction error in minutes, it's only about 21 minutes which means on average, the delay time we predicted is about 21 minutes different than the actual delay time of each flight in the testing set.

Conclusion

As seen in our Linear Regression model, we have a fit score of 0.003 which shows extremely little correlation, practically none. However, this score does make sense because the delay of a flight is extremely hard to predict. While the dataset contains information like time of departure, origin and day of the week, none of these are obvious reasons why a flight might be delayed. There are a few reasons a flight might get delayed such as inclement weather which could be extrapolated by origin region and date but impossible to do with weather data. There could be a security or medical issue at the airport, or the plane might be having technical issues, which might be traceable via airline but even then, low cost and pricey airlines can use the same airplane models. If one flight gets delayed for whatever reason, then any connecting passengers shared between the two can cause a chain delay. All these potential reasons can cause flight delay but the dataset only gives us a record of past delays and where/when they've occurred. Really it's similar to predicting the weather, we know that there will be correlation but they don't necessarily imply causation. Ideally what we want to add to our dataset, and the model, are the dates of the delays so that they can be tracked to weather conditions in their respective, which is likely the most impactful, and well recorded, data we can use.

When we used KNN to predict the delay status of a flight, the accuracy score is quite high at about 85%, this gives us confidence that certain features such as day of the week, airline, scheduled departure time, departure airport, and distance are relevant factors impacting the delay status of a flight.