U. S. Flight Delays and Cancellation Analysis

INFO 5304 - Data Science in the Wild

Final Project Report

Hyein Baek (hb437)

Haomiao Han (hh696)

John Lin (hl2357)

Jianang Wang (jw2594)

Abstract

In this paper, we propose machine learning-based algorithms that predict the length of delay (if any) of flights and whether a flight will be cancelled or not. Using a dataset containing domestic flight delays and cancellations from 2016 to 2018 provided by the U. S. Department of Transportation as well as a dataset containing U.S. weather events during the same time period, we performed data cleaning and feature engineering on the datasets and experimented with various machine learning models. We were able to achieve a result of 82.3% accuracy when predicting flight cancellations and 70.7% accuracy when predicting flight delays.

Background

We have all experienced flight delays and cancellations ourselves - the delays and cancellations are frustrating and stressful to deal with, and could ruin an otherwise perfect trip. In fact, flight delays are not only frustrating but also costly: to passengers, airlines, airport operators as well as the U. S. economy as a whole. A 2010 study conducted by UC Berkeley showed that domestic flight delays could cost $32.9 billion per year to the American economy. It would be great if we could reduce the economical, social and environmental cost created by flight delays and cancellations.

Therefore, our team would like to approach this problem from a data science perspective, and attempt to address the problem by analyzing historical data on flight delays and cancellations and creating a tool to predict delays/cancellations on future flights. If we could successfully create such a tool, we believe that it will be quite useful to travellers and airlines, and greatly improve our travel experience.

Dataset

The flight dataset we have used is the 2008-2019 Flight Delays and Cancellations dataset. This dataset is compiled by a Kaggle user and the original data is provided by the U.S. Department of Transportation. We only used the portion containing flight information from 2016 to 2018.

- The file size of the dataset is 2.2 GB. It contains 28 columns and around 18.5 million rows.
- Features included are:
    - Flight Information: Date of the flight, airline, flight number, origin and destination airports, departure time (both scheduled and actual), arrival time (both scheduled and actual), air time (both scheduled and actual), distance travelled;
    - Cancellation/Delay Information: Cancellation and delay reasons by category and delay time (if applicable - a lot of the flights are not delayed/cancelled).
- One bias we found in this dataset is that there are only 265,138 flight cancellations recorded. Compared to the total number of 18,505,725 flights, only about 1.4% of flights are cancelled. Therefore, we will need to adjust our machine learning algorithms to eliminate this bias.

During the second half of the semester, we decided to augment the original dataset with a weather dataset, and selected the US Weather Events (2016 - 2019) dataset compiled by Moosavi et al.

- The file size of the dataset is 585 MB. It contains 13 columns and around 5 million rows.
- Features included are:
    - Event ID: a unique ID for each event;
    - Date and Time: the start and end time for a given weather event;
    - Type of the event: can be one of the following: Severe Cold, Fog, Hail, Rain, Snow, Storm and Other Precipitation;
    - Severity of the event;
    - Location Information: The airport code of the location where the event is recorded, as well as its zip code, city, county, state, time zone, and latitude/longitude.

Preprocessing and Feature Engineering

After loading our flight dataset using `pandas`, we examined the dataset and determined that several columns of the dataset is not necessarily for our machine learning model. For example, a lot of the columns are related to the actual departure/arrival/taxi time of a flight, which we cannot use since if we were to use the model to predict a flight in the future, we would not have these data beforehand. For the flight cancellation model, we would also need to drop the `DEP_DELAY` column as cancelled flights do not have a departure delay (since they never departed).

After dropping these columns, we checked if our dataset contains empty cells or cells with null values. We found that there are 23 missing values in the `CRS_ELAPSED_TIME` column and 4,744 missing values in the `DEP_DELAY` column. Given that we have more than 18 million rows, dropping these rows would not make a large difference to our model. We then decided to drop these rows.

We did some preprocessing on the weather dataset before combining with the flight dataset. We first transformed the airport code format in the weather dataset so that it aligns with the airport code used in the flight dataset. After that, we used `get_dummies` from `pandas` on both `Type` and `Severity` producing a cross one hot vector with length of 13 representing the type of weather and the severity of that weather. Then, we used the `groupby` function on `date` and `Airport` together and aggregated the sum. With this step, we produced a dataset table that shows the number of counts of weather and the severity reports at each airport on each day. If there are no reports of weather on a certain day, we assume it is sunny/clear, and none of the one hot vector variables are active.

We then proceeded to join the weather dataset with our flight dataset. Because our assumption is that the weather at the origin and the destination of each flight would both have some effect on the cancellation and delay time. We used `left merge` (with respect to the flights dataset) from `pandas` on both `ORIGIN` and `date`, which outputs a dataset table containing the weather at the origin airport on the date of each flight. Then, we merged again using the same strategy for the `DEST` and `date` to add the weather at the destination airport for each flight as well. This process is equivalent to doing a Left Join twice in SQL.

We then used the `LabelEncoder` feature from `scikit-learn` to encode all categorical variables into numerical values so that we can use the dataset to build machine learning models. In another version, we used `get_dummies` from `pandas` to encode all categorical variables into one hot vector, because many variables do not have any ranking meaning behind them.

Dates in both datasets are originally represented in a year-month-day format in a single cell, which we think was not desirable. Thus, we splitted the date into three columns, with one column for month, one for day, and one for day of the week.

To deal with data imbalance, we randomly sampled 265,000 flights that are not cancelled. We then combined these randomly sampled flights with all of the cancelled flights (265,138 flights at this point), and used that to train our model for predicting flight cancellations. As there is no data imbalance for the delayed duration of the flights, we used all available data from the dataset to build our model for predicting flight delays.

Developing and Testing Machine Learning Algorithms

We used 80% of the data as the training set and 20% of the data as the validation set.

For our flight cancellation prediction model, we experimented with two algorithms - Logistic Regression and Random Forest - to create the model.

For our flight delays prediction model, we first treated it as a regression problem and experimented with Linear Regression and Random Forest algorithms. However, our initial results were not very positive as the predictions are not very accurate. We then decided to treat it as a classification problem and attempted two models:

- Under the first model, we treat the problem as a binary classification problem where the target variable is either 0 (a flight departed on time or earlier than schedule) or 1 (a flight departed later than schedule);
- Under the second model, we treat the problem as a multiclass classification problem where the target variable is either 0 (a flight departed on time or earlier than schedule), 1 (0 < departure delay <= 30), or 2 (departure delay > 30).
- We initially wanted to create a 4-class model as well (with class 1 referring to 0 < departure delay <= 15 and class 2 referring to 15 < departure delay <= 30); however, the model performance was less than desirable as its accuracy was consistently under 50%. Thus, we decided not to use this model.

We then used Random Forest and Neural Network algorithms to create the models.

To test if augmenting the original dataset with the weather dataset has improved the performance, we run each model twice, once with the weather dataset and once without the weather dataset.

Results

Below are our results for predicting flight cancellations:

ML Algorithm	**Performance without weather dataset (Accuracy)**	**Performance with weather dataset (Accuracy)**
Logistic Regression	0.651	0.671
Random Forest	0.823	0.821

Below are our results for predicting flight delays:

Regression	**ML Algorithm**	**Performance (R^2 Score)**
	Linear Regression	0.0129
	Random Forest	0.0674

Binary Classification	**ML Algorithm**	**Performance without weather dataset (Accuracy)**	**Performance with weather dataset (Accuracy)**
	Random Forest	0.680	0.706
	Neural Network	0.626	0.636

Multiclass Classification	**ML Algorithm**	**Performance without weather dataset (Accuracy)**	**Performance with weather dataset (Accuracy)**
	Random Forest	0.512	0.552
	Neural Network	N/A	0.570

Discussion and Future Work

We have concluded a few takeaways given our results:

1. Augmenting the original flights dataset with the weather dataset only marginally improved the performance of the model for predicting flight delays by around 2 to 3 percentage points, whereas it hasn't improved the performance for the flight cancellation prediction model at all. This is a little surprising given that our data visualization shows

that more than 50% of flight cancellations happened due to weather. In the future, we may need to take other weather aspects into account, such as temperature or wind speed. From our real life experiences, we also know that flight delays and cancellations can also happen because of adverse weather conditions on the flight route (that don't necessarily affect either the origin or the destination city). Since we did not know the route each flight took, we can only use the weather data at the origin and destination airport; ideally, our model should take the weather on the flight route into consideration too.

2. The recent outbreak of coronavirus made us realize that there could be more potential causes of flight delay and cancellation such as disease outbreak, major sports and political events, and holidays. Our current model mostly focuses on the weather impact, which we found to be the top reason for flight cancellations during 2016 to 2018. In the future, we are looking to expand the feature vector of our datasets to improve the diversity and complexity of these models. If we plan to use our dataset to predict flights in the future, we would also need to include some more recent datasets due to the dramatic changes that were made to flight scheduling recently.

References

Please refer to the hyperlink(s) on each page for all references.