Flight Data Analytics

Noor Kaiser, Olivia Lee & Zach Moring

Boston University MET Computer Science Department

CS 777 Big Data Analytics

Term Project

Spring 2023

INTRODUCTION

Logistics is a field which resists easy solutions. The problem of getting goods and people from one place to another is often intractable. Logistics systems rely on the coordination of dozens, hundreds, or thousands of individual vehicles, goods, people, and worksites. These facts make the field of logistics an ample source of both large data sets and interesting problems. This project attempts to analyze the relationship between individual factors of the airline arrival/departure record and the delays which result.

THE DATASET

Domain-Specific Knowledge

All flights in North American airspace are registered ahead-of-time with the government – the FAA in the United States, etc. The system used to register and record flights is called the Computer Registration System, or CRS. Flight registrations and planning are carried out well in advance of their occurrence, without specific knowledge of the weather conditions, ground conditions, etc. that may affect the flights. Thus, it is very common for planes to arrive early or late, depart early or late, and so on. For our dataset, it is useful to distinguish between *planned* arrival/departure and *actual* arrival/departure. In our dataset, *planned* arrival/departure is encoded into the columns prefixed with CRS, while the *actual* arrival and departure are unmarked.

The dataset also indicates the cause of a given delay. Causes may be one of the following options: Carrier, weather, NAS, security, or late aircraft. Late aircraft delays and weather delays are self-explanatory. A delay is marked as a carrier delay when it is delayed due to a process operated by the carrier or a private contractor, such as aircraft cleaning, delayed meal service delivery, baggage complications, and so on. Security delays result from security incidents or delays related to processes operated by TSA, the NSA, or other security interests. Lastly, NAS delays are those delays which are caused by the airport, including overcrowded skies, mistakes from air traffic control, etc.

Columns

The dataset used for this analysis carries information on airline delay and cancellations for the years 2009 through 2018. The dataset includes 28 columns and 62 million rows over the 10 years. Each row represents one flight, and the attributes describe each record. In our models, we will perform a regression analysis that predicts the variable “ARR_DELAY”, representing the total delay on arrival in minutes, and a classification that will label whether a flight is predicted to be delayed or on time. The list below outlines all 28 columns and their descriptions:

- *FL_DATE*: Date of the flight, yy/mm/dd
- *OP_CARRIER*: Airline Identifier
- *OP_CARRIER_FL_NUM*: Flight Number
- *ORIGIN*: Starting Airport Code
- *DEST*: Destination Airport Code
- *CRS_DEP_TIME*: Planned Departure Time

- *DEP_TIME*: Actual Departure Time
- *DEP_DELAY*: Total Delay on Departure in minutes
- *TAXI_OUT*: The time duration elapsed between departure from the origin airport gate and wheels off
- *WHEELS_OFF*: The time point that the aircraft's wheels leave the ground
- *WHEELS_ON*: The time point that the aircraft's wheels touch on the ground
- *TAXI_IN*: The time duration elapsed between wheels-on and gate arrival at the destination airport
- *CRS_ARR_TIME*: Planned arrival time
- *ARR_TIME*: Actual Arrival Time
- *ARR_DELAY*: Total Delay on Arrival in minutes
- *CANCELLED*: Flight Cancelled (1 = cancelled)
- *CANCELLATION_CODE*: Reason for Cancellation of flight: A - Airline/Carrier; B - Weather; C - National Air System; D - Security
- *DIVERTED*: Aircraft landed on airport that out of schedule
- *CRS_ELAPSED_TIME*: Planned time amount needed for the flight trip
- *ACTUAL_ELAPSED_TIME*: ARR_TIME+TAXI_IN+TAXI_OUT
- *AIR_TIME*: The time duration between wheels_off and wheels_on time
- *DISTANCE*: Distance between two airports
- *CARRIER_DELAY*: Delay caused by the airline in minutes
- *WEATHER_DELAY*: Delay caused by weather
- *NAS_DELAY*: Delay caused by air system

- *SECURITY_DELAY*: Delay caused by security
- *LATE_AIRCRAFT_DELAY*: Delay caused by late aircraft

DATA PREPARATION

Data Cleaning

Our dataset consisted of 10 separate CSV files (one for each year of flight data) with 28 features. However, we saw several data inconsistencies, redundancies, and missing values in our dataset. The following was done to eliminate these issues:

- Drop column 28, i.e., 'Unnamed: 27'
- Replace NAs with 0s for all delay type columns, i.e., 'CARRIER_DELAY', 'WEATHER_DELAY', 'NAS_DELAY', 'SECURITY_DELAY', 'LATE_AIRCRAFT_DELAY'
- Replace NAs in column 'CANCELLATION_CODE' with the letter 'E'
- Drop all remaining NAs from the dataset
- Create binary variables for delay type columns

Additionally, we realized that each of the CSV files had a different set of carriers for whom the data was collected for that specific year. To maintain consistency for the data over the course of the 10 years, we found the intersection of the carriers that were present in the dataset in each year. Therefore, we only retained the data for the following carriers:

- carriers = ['DL', 'B6', 'WN', 'AS', 'F9', 'EV', 'HA', 'AA', 'UA', 'OO'].

Since we intended on studying the impact of various carriers on the delay in flights, we used one hot encoding for the carrier label to end up with the following:

- column ['OP_CARRIER'] -> columns ['OP_CARRIER_AA', 'OP_CARRIER_AS', 'OP_CARRIER_B6', 'OP_CARRIER_DL', 'OP_CARRIER_EV', 'OP_CARRIER_F9', 'OP_CARRIER_HA', 'OP_CARRIER_OO', 'OP_CARRIER_UA', 'OP_CARRIER_WN']

Finally, we created two new variables to replace the ‘ORIGIN’ and ‘DEST’ columns. The original columns contained the IATA codes for the origin and destination airports for each instance in the dataset – this information did not seem to be particularly useful to train our models. Instead, we created new columns for level of traffic at origin and destination airports as follows:

- If number of flights per year >= 100,000 flights, ‘ORIGIN_TRAFF’ and ‘DEST_TRAFF’ == high traffic
- If number of flights per year < 100,000 flights and > 10,000 flights, ‘ORIGIN_TRAFF’ and ‘DEST_TRAFF’ == medium traffic
- If number of flights per year <= 10,000 flights, ‘ORIGIN_TRAFF’ and ‘DEST_TRAFF’ == low traffic

We intended on using these two variables to study whether the busyness of the origin or destination airports have an impact on the delay in flights.

Final Dataset

After processing the data, we obtained 38 features and approximately 56 million rows in our final dataset. This dataset will be used to build, train, evaluate and compare our classification and regression models.

DIMENSIONALITY REDUCTION

Principal Component Analysis (PCA) was used for dimensionality reduction of our dataset. Two separate implementations of PCA were used for the regression and classification models, while retaining the same methodology for both models.

To optimize the number of components, k, in PCA, the PCA was calculated for a range of values of *k*. In this case, PCA was calculated with the *k* = 1, 2, 3, 4, and 5. For each value of *k*, the explained variance ratio (EVR) was calculated and plotted against the *k*-values to generate a scree plot, showcasing the proportion of the explained variance as a function of the number of principal components. The best value of *k* is the value that provided a sufficient amount of explained variance while keeping it as low as possible. Keeping with the common rule of thumb, the optimal number of components chosen was the smallest *k* value that explained at least 80% of the total variance.

As seen in the scree plot below, the optimal number of components is 2 for our dataset.

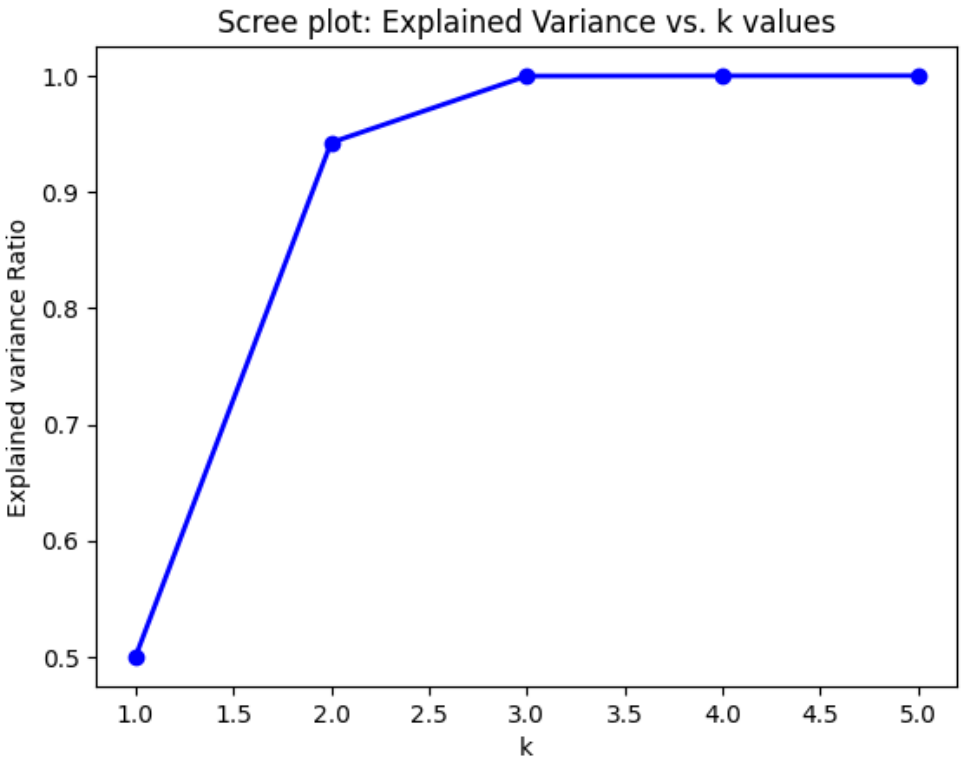

Figure 1: Scree plot for optimal k in PCA

MODELS

Linear Regression: Predicting arrival delay time

This model will perform a multiple linear regression to predict the amount of time a flight is delayed for (if at all) in terms of when it arrives at its destination airport ('ARR_DELAY'). Canceled and diverted flights are excluded from this dataset.

The regression models built were as follows:

- Using all the remaining variables in the dataset

- Using dimensionality reduction (PCA) and regularization (Lasso Regularization, Ridge Regularization)

Performance metrics such as accuracy, MSE, RMSE, R-squared and adjusted R-squared will be used to evaluate the models built using the training dataset, on the testing dataset.

Logistic Regression: Predicting whether a flight is on time or delayed

This model will perform a binary classification using logistic regression to predict whether a flight is delayed or on time. Canceled and diverted flights are not included in this data set. The columns chosen for this classification, before dimensionality reduction, exclude the information regarding actual flight times and types of delay to assume that the flights have not taken off to fairly predict delay. A new column "DELAY" is made and is the class label in this classification where 1 represents a delay and 0 for no delay. Flights that are considered delayed are flights with an arrival delay of more than five minutes.

For hyperparameter tuning, grid search was performed for the regularization parameter, elastic net parameter, and maximum number of iterations. The regularization parameter searches through values 0.01, 0.1, and 1, elastic net with 0.0, 0.5, and 0.1 and maximum number of iterations at 10 and 100. With a train, test, and validation split at 0.7, 0.2, and 0.1 respectively, the optimal hyperparameters chosen by the validation set were as follows:

Logistic regression with PCA:

- Regularization Parameter: 1.0
- Elastic Net: 0.0

- Maximum Number of Iterations: 10

Logistic regression without PCA:

- Regularization Parameter: 0.01
- Elastic Net: 0.0
- Maximum Number of Iterations: 100

PERFORMANCE

Linear Regression

The linear regression model using all the variables from the dataset obtained an accuracy of 94% on the training set, which we realized might have been the case because the model was overfitting to the training data. We incorporated PCA and Regularisation techniques in the regression models in order to reduce the chances of overfitting. When we eliminated certain variables in the dataset based on these techniques as well as intuition, we obtained a shockingly low accuracy of 13%.

Logistic Regression

Our logistic regression model after hyperparameter tuning and dimensionality reduction has an ROC AUC of 0.582 for classifying whether a flight is delayed or on time. Interestingly, the model performed better without dimensionality reduction, with an ROC AUC of 0.715.

CONCLUSION

According to the performance of these regression and classification models, it appears that the models perform better without dimensionality reduction. This could suggest that the features in the dataset are all important and should not be reduced. Additionally, this analysis may require a larger and more complex dataset, in order to improve the model's performance while maintaining a certain level of simplicity and generalisability.

Another possible explanation can be found in the nature of the data itself. Given that the data we used chronicled *only* the outcomes, it is possible that our data does not contain enough causative data to form meaningful conclusions. Future work could augment this dataset by adding features including weather, news events (such as protests, closures, and union actions), and other “causative” features. Accounting for these, we would expect to see robust results more in-line with intuition.