Predicting Flights' Arrival Delay Time with Big Data

Washington University in St. Louis

Fall 2023

Professor Hossein Amini

Group 62

Luvia Wang (523180)

Kuan Lin (520389)

Yuanrong Xia (517386)

Lilo Hsiung (519636)

2023.12.7

Executive Summary

Problem Statement

Our study aims to identify factors affecting flight delays and proposes insights to enhance on-time rates.

Description of Data

Utilizing the Airline Flight Delay and Cancellation Data (January 2023 - August 2023) from the US Department of Transportation, our dataset comprises 784MB of information, with 4,545,422 rows, including 33 variables. The process requires big data tools.

Data Analysis Process

1. Data Cleaning: Removing observations with missing values in key predictor columns
2. Data Preprocessing: LabelEncoder transforms text labels into numerical values. StandardScaler normalizes numerical data.
3. Variable Selection: Calculate correlation and select predictors that are statistically significant: Dep_delay, Dep_time, Taxi_out, wheels_off
4. Model and Result: A multi-linear regression model is employed. The adjusted R-square is 0.668. Test MSE is 237.33.

Conclusion

1. Customers can import departure time in the model to predict delay times and make decisions
2. As the time goes by from morning, afternoon, evening, to midnight, the predicted delay time for the flights decreases
3. The model has limitations regarding a modest Adjusted R-square, the removal of missing data, and a dataset mainly focused on operational predictors.

1. Problem Statement

According to the Transportation Security Administration (TSA), Americans do not appear to be spending less on travel during the 2023 Memorial Day holiday, despite the ongoing presence of Qualcomm, with air travel in the United States exceeding 2019 pre-pandemic levels.

TSA said 9.79 million people took to the skies this holiday, more than during the long weekend in 2019. Among them, more than 2.7 million passengers traveled by air on Friday, the highest number since the outbreak.

It can be seen that flight is still an important choice for Americans to travel. However, according to the data from Department a (Exhibit 1), we found that from 2018 to 2023 till now, the punctuality rate of flights has basically remained at about 77%, although in 2020-2021, the punctuality rate has exceeded 80%, but it has dropped to the original level since 2022.

Overall, an on-time rate of less than 80 percent and a delay rate of more than 20 percent are disadvantages for passengers when choosing an airplane as a means of transportation. For airlines, this is an obstacle to increasing revenue and a breakthrough in finding new profit points.

Therefore, through the collected flight delay data, our team studied the specific factors that affect flight delay, and proposed possible plans to improve the flight on-time rate.

2. Description of Data

The data source of our dataset is Airline Flight Delay and Cancellation Data, January 2023 - August 2023 and collected from US Department of Transportation, Bureau of Transportation Statistics. And its size is 784MB. The data set consists of 33 variables including flight date, flight number, actual departure time and so on. To process regression analysis, we set

ARR_DELAY which means the difference in minutes between scheduled and actual arrival time and early arrivals show negative numbers, as our response variable. Data types are also included int, object and float.

Exhibit 2 shows specific data names, types, and descriptions.

3. The Reason of Big Data

Generally speaking, we define big data as the data that we cannot analyze in time using traditional processes or tools. In our data set, we totally need to process over 700mb data which include over 4,545,422 rows and 33 columns. Using traditional tools is time-consuming and the process of data processing is more tedious. So we decided to use the big data tool pySpark to analyze the data set.

4. Data analysis

4.1 Data Cleaning

To simplify the process of the data cleaning, we delete all the observations that have missing values in these five predictor columns [DEP_TIME, DEP_DELAY, TAXI_OUT, WHEELS_OFF, DISTANCE] and our target column [ARR_DELAY], making sure that we won't run into error in the following process. Also, we want to make sure we don't use the same data to train our model. So we delete all duplicate data before really organizing the dataset in order to prevent overfitting in the future model prediction.

4.2 Data Preprocessing

Before doing any analysis or prediction, we have to organize the data into the format we can address in the following problems. First, we observed that both text data and numerical data exist in the Flight_2023 dataset. Therefore, we utilize the method called LabelEncoder to encode the text labels with values between 0 and n_classes-1, transforming all text attributes into categorical values. After that, we have the dataset with all numerical values to standardize the data in each column and can calculate the correlation between the target variable and the other predictor variables.

Now with all continuous data, we normalize all of them to ensure that no single attribute disproportionately influences our future model, and to mitigate the effects arising from differing units of measurement, thereby preventing any attribute from carrying excessive weight in the model's calculations. Using StandardScaler, we can calculate each value by subtracting the mean and dividing by their standard deviation to get the normalization value ((x-mean)/Standard deviation).

After preprocessing the data, we can select the important attributes in the model to predict our target variable by looking at the correlation between the target variable and the other predictors.

4.3 Variable selection

To select appropriate variables for prediction, we look into correlation matrix (Exhibit 3) among all the variables. Dep_delay variable is highly correlated with our predicted variable Arr_delay, the correlation coefficient is 0.98.

Besides, we choose some other variables that make sense in practical situations. They are dep_time, taxi_out, wheels_off and distance.

Intuitively, high taxi out and wheels off time can contribute to congestion at the departure airport, affecting subsequent departures and arrivals at the destination airport. Also, those ground operations can impact the overall turnaround time between flights, influencing scheduling and potential delays. As for the actual departure time, different times of the day may influence the takeoff and landing of fights. Additionally, although the correlation coefficient between distance and arr_delay is smaller, even less than 0.01, we are curious whether long-distance flights are more likely to delay.

4.4 Model and Result

Our method for prediction is multi-linear regression and the model looks like the following:

$$arr\_delay = \beta_0 + \beta_1\ dep\_time + \beta_2\ dep\_delay + \beta_3\ taxi\_out + \beta_4\ wheels\_off + \beta_5\ distance + \epsilon$$

At first, we randomly split the whole dataset into training and test sets in an eight-to-two ratio. Then we fit the model with a training sample and get the statistics and estimated coefficient for different variables.

After running the regression, we get the result.

	coef	std err	t	P>|t|	[0.025	0.975]
DEP_TIME	-1.6026	0.211	-7.598	0.000	-2.016	-1.189
DEP_DELAY	102.8798	0.083	1241.519	0.000	102.717	103.042
TAXI_OUT	12.6719	0.084	151.260	0.000	12.508	12.836
WHEELS_OFF	0.5316	0.211	2.524	0.012	0.119	0.944
DISTANCE	-0.0650	0.083	-0.784	0.433	-0.228	0.098

The adjusted R-square of the model is 0.668 and predictors are statistically significant except distance. In terms of interpretation, holding all else fixed, when departure time increases by 1 standardized unit, the estimated arrival delay would decrease by 1.60 units, meaning that flights would have smaller delay time or even arrive earlier in the evening compared with morning. For other variables, dep_delay, taxi_out and wheels_off, when one of them increases one scandalized unit holding all else fixed, arr_delay is expected to increase 102.88, 12.67 and 0.53 units respectively.

Then we use test set to evaluate our model and the test MSE is 237.33.

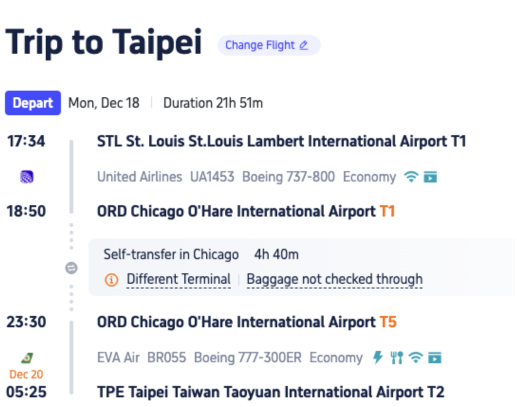
5. Conclution

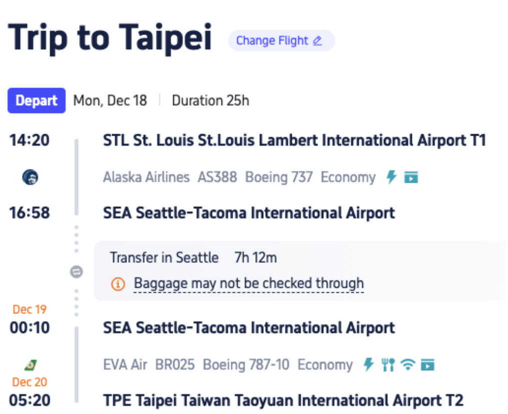
5.1 Application

After training and testing the model, we used it in two scenarios to see if we could get useful information or predictions. The two scenarios are 1. predicting as a customer and 2. analyzing as an airline.

Predicting as a Customer

As consumers, we often struggle to decide which flight to book. Other than considering factors such as price and airline preference, it's also crucial to consider potential delays. Our model aims to provide insights into predicted delay times, offering valuable information for informed decision-making. To illustrate the model's efficacy, we conducted a comparative analysis of two trips from STL to TPE, denoted as options 1 and 2. Option 1 involves a layover in Chicago (ORD) with departure times of 17:34 and 18:50, while option 2 features a layover in Seattle (SEA) with departure times of 14:20 and 00:10. The specific trip details are shown in the accompanying visuals.





To accurately anticipate potential delay times, we adhere to a structured set of steps:

1. Calculate the median of factors unrelated to departure time, as these are details unavailable during the ticket booking process.
2. Standardize the input data, mirroring the preprocessing steps employed during model training.
3. Input the data into the model, encompassing departure times for all four flights and the calculated median for the remaining factors.
4. Consolidate the flights within the same option.
5. Generate and display the conclusive result.

Result:

```
STL_ORD_TPE: 79.09704896963453
STL_SEA_TPE: 87.6501081418329
```

The outcome, as displayed, shows that the model predicts a delay time of 79.1 minutes for Option 1 and 87.7 minutes for Option 2. Consequently, based on our model's analysis, opting for Option 1 is advisable, as it suggests a potentially shorter delay time compared to Option 2.

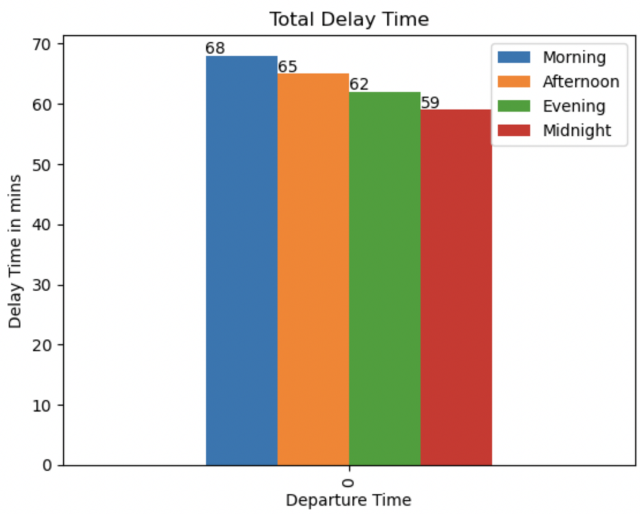
Analyzing as an Airline

Imagine an airline expressing interest in utilizing this model for strategic adjustments to reduce delays. Within the model's framework, four variables—Dep_time, Dep_delay, Taxi_out, and Wheels_off—are considered, with the latter three tied to operational aspects. These operational factors can only be improved by digging into the operation process and figuring out what causes the delay. Conversely, Dep_time, linked to forecasting, allows us to take action when arranging the flight schedule.

To analyze what time in a day might cause a longer delay time, here are our steps:

1. Divide the day into four distinct sections: [Morning: 0600-1159 [Afternoon: 1200-1759] [Evening: 1800-2359] [Midnight: 2400-0559]
2. Standardized the input data, applying the same methodology as the model training process.
3. Inputting the data and computing delay times within the assigned four sections.
4. Averaging the delay results across all four sections.
5. Presenting the outcome and visually depicting the results in a plot

Result:



As shown in the plot, the morning section exhibits the highest delay time at 68 minutes, succeeded by 65 minutes in the afternoon, 62 minutes in the evening, and 59 minutes at midnight. This indicates a gradual decrease in predicted delay times as the day progresses from morning to midnight. This insight can prove valuable for airlines in optimizing flight schedules, guiding them to strategically arrange flights as the day unfolds, potentially mitigating delay-related challenges.

5.2 Limitation

After building the model and apply the model to real-world scenarios, we found 3 major limitations:

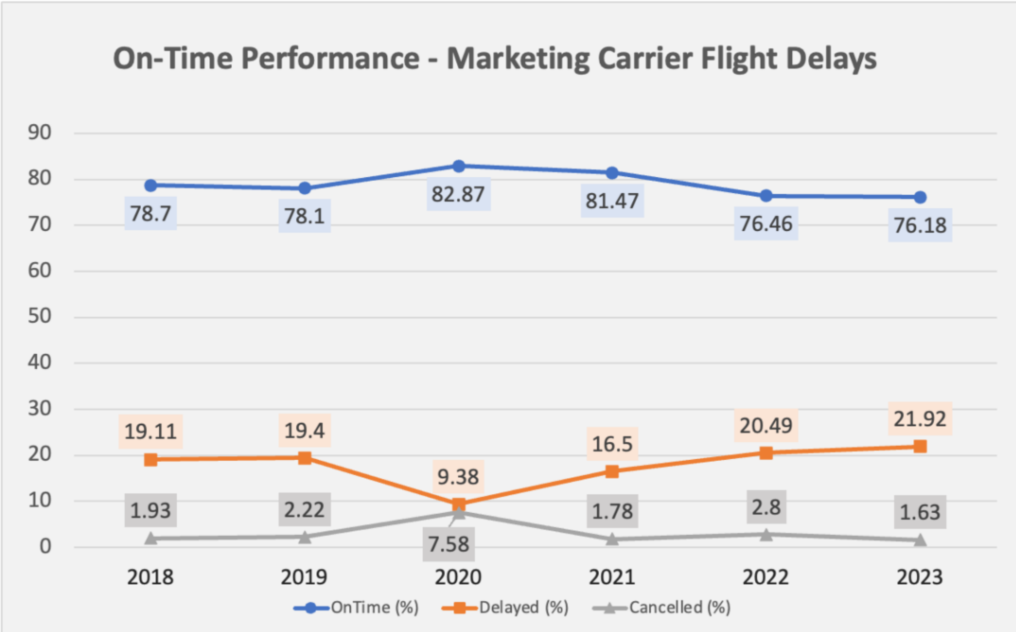
1. The Adjusted R-square, currently registering at a modest 0.668, encapsulates the model's current explanatory power, serving as a baseline from which we can strategically navigate for improvement. With this model, we have the opportunity for further analysis and methodological adjustments, creating a pathway to a model with higher performance.

2. While cleaning the data, we removed more than half of the missing values. This deletion, although necessary for data integrity, introduces a challenge as it may diminish the dataset's completeness. Hence, the model could perform better if we explore alternative approaches for handling missing data that strike a balance between preserving data quality and optimizing predictive performance.

3. Most of the predictors in our dataset pertain to operational aspects, posing a challenge in our forecasting endeavors. To improve the model, we can explore ways to broaden the dataset's predictive scope by incorporating a more diverse set of predictors that capture a comprehensive range of influential factors.

Appendix:

Exhibit 1



(source:https://www.transtats.bts.gov/Marketing_Annual.aspx?heY_fryrp6lrn4=FDFG&heY_fryrp6Z106u=M&heY_gvzr=E&heY_fryrp6v10=E).

Exhibit 2: Variable Description

Updated Header	Data Type	Description
FL_DATE	object	Flight Date (yyyymmdd)
AIRLINE_CODE	object	Unique Carrier Code. When the same code has been used by multiple carriers, a numeric suffix is used for earlier users, for example, PA, PA(1), PA(2). Use this field for analysis across a range of years.
DOT_CODE	int64	An identification number assigned by US DOT to identify a unique airline (carrier). A unique airline (carrier) is defined as one holding and reporting under the same DOT certificate regardless of its Code, Name, or holding company/corporation.
FL_NUMBER	int64	Flight Number

ORIGIN	object	Origin Airport
ORIGIN_CITY	object	Origin Airport, City Name
DEST	object	Destination Airport
DEST_CITY	object	Destination Airport, City Name
CRS_DEP_TIME	int64	CRS Departure Time (local time: hhmm)
DEP_TIME	float64	Actual Departure Time (local time: hhmm)
DEP_DELAY	float64	Difference in minutes between scheduled and actual departure time. Early departures show negative numbers.
TAXI_OUT	float64	Taxi Out Time, in Minutes
WHEELS_OFF	float64	Wheels Off Time (local time: hhmm)
WHEELS_ON	float64	Wheels On Time (local time: hhmm)
TAXI_IN	float64	Taxi In Time, in Minutes
CRS_ARR_TIME	int64	CRS Arrival Time (local time: hhmm)
ARR_TIME	float64	Actual Arrival Time (local time: hhmm)
ARR_DELAY	float64	Difference in minutes between scheduled and actual arrival time. Early arrivals show negative numbers.
CANCELLED	float64	Cancelled Flight Indicator (1=Yes)
CANCELLATION_CODE	object	Specifies The Reason For Cancellation
DIVERTED	float64	Diverted Flight Indicator (1=Yes)
CRS_ELAPSED_TIME	float64	CRS Elapsed Time of Flight, in Minutes
ELAPSED_TIME	float64	Elapsed Time of Flight, in Minutes
AIR_TIME	float64	Flight Time, in Minutes
DISTANCE	float64	Distance between airports (miles)
DELAY_DUE_CARRIER	float64	Carrier Delay, in Minutes

Exhibit 3: Correlation Matrix for Selected Variables and Response Variable

	Dep_time	Dep_delay	Taxi_out	Wheels_off	Distance	Arr_delay
Dep_time	1.0000					
Dep_delay	0.0087	1.0000				
Taxi_out	-0.0767	-0.1048	1.0000			
Wheels_off	0.9191	-0.0103	-0.0343	1.0000		
Distance	-0.0782	0.0050	0.0173	-0.0856	1.0000	
Arr_delay	-0.0117	0.9811	0.0197	-0.0236	0.0068	1.0000

Data Source:
https://www.transtats.bts.gov/Marketing_Annual.aspx?heY_fryrp6lrn4=FDFG&heY_fryrp6Z106u=M&heY_gvzr=E&heY_fryrp6v10=E

Code:
https://github.com/johnny880624/Big_Data_Final.git