ISYE 6420: Project
Spring 2023
Deborah Chang

Airline Delays with Logistic Regression: Bayes or Classic?

Abstract

In this paper, we compare the classical logistic regression model with the Bayesian logistic regression approach using flight delay data. Applying noninformative priors and various assumptions, we come to the conclusion that in this particular case, both models performed very closely or about the same. In the real world, however, we might expect various results that lead to the conclusion that one approach is significantly better than the other.

Introduction

According to Travel and Leisure, 22% of flights in 2022 were delayed or canceled, with Allegiant Air having the lowest performance and Hawaiian Airlines with the highest on-time metric at 85.43% (Fox). As much as airlines try to avoid flight delays, it's not guaranteed. Factors such as weather (which was especially severe in winter 2022 during the holidays), air traffic, system outages, and fuel are beyond their control. Even so, delays may lead to customer complaints about the length of delay, bad reviews, subsequent delays, and potential impact on other flights traveling in or out. Real Time flight status trackers such as FlightStats, FlightAware, Google itself, or enabling the iPhone Messages app to track for you have given passengers more visibility and guidance with travel planning. The US Department of Transportation provides information on what is potentially compensated should there be a delay or even cancellation. However, there is no law requiring airlines to provide compensation, but they are required to provide any changes to the flight status within a week prior and 30 min after they are aware of them. They do provide some refund or voucher should any trips be canceled.

Many AI and machine learning methods have been applied to increase efficiency of airline operations and reduce the probability of a delay. For example, insights on staff capacity, average check-in timing, or volume of flight bookings can be leveraged to optimize departure preparedness for airlines. A specific example is from Altexsoft where airlines are able to have real time data on the aircraft's technical state to optimize the flight route.

Given data on a set of flights, including arrival and departure airport, day of the week, time, and whether they were delayed, we will compare the classical and bayesian approach for logistic regression. The model can help predict delays for future flights and enable visibility for potential customers on what flights are better to book as well as for airlines to evaluate what factors are causing delays.

American	278	1:35P	54	Cancelled
jetBlue	1884	2:19P	24	Cancelled
DELTA	409	2:47P	77	On Time
jetBlue	284	3:53P	9	Cancelled
DELTA	2649	1:59P	77	Cancelled
DELTA	1839	3:53P	76	On Time
UNITED	1601	1:55P		Cancelled
jetBlue	328	1:59P	23	Cancelled
spirit	140	2:20P		Cancelled
UNITED	1764	3:00P		Cancelled
jetBlue	228	3:50P	1	Cancelled

Figure 1. Real Time Flight Board. Source: Fox, Alison. "The Airlines with the Most Delays This Year, According to the Bureau of Transportation Statistics." Travel + Leisure, Travel + Leisure, 27 Oct. 2022.

Data

<u>Features</u>

The data (Chacko) is from Kaggle and is called: "Airlines Dataset to predict a delay." There are over 530K records and 8 different fields:

- ID: Indicator
- Airline: the type of airline the flight is on
- Aircraft: type of plane/flight number
- AirportFrom: the airport the flight is coming from
- AirportTo: the airport the flight is heading to
- DayOfWeek: the day of the week (from 1 to 7)
- Time: the time of the flight departure in minutes from midnight (referring to the comments)
- Length: flight length in minutes
- Delay: indicates whether the flight was delayed or not (binary)

There are some assumptions applied to this data:

- All flights are domestic in the United States
- The time of the flight departure are all standardized regardless of which location/time zone the flight leaving from

Here are some observations about the dataset and features:

The average length of the duration of these flights are around 132.2 minutes, or about 2.2 hours. 55% of the flights in this dataset did not experience a delay. The average flight departure time is around 802.73 min from midnight, or around 1:30PM.

The most common departure day is the middle of the week, or around Day 3-4 and followed by weekends [see Figure 2 below]. This could make sense since it's generally better to fly on weekdays with lower costs and potentially less people and crowds.

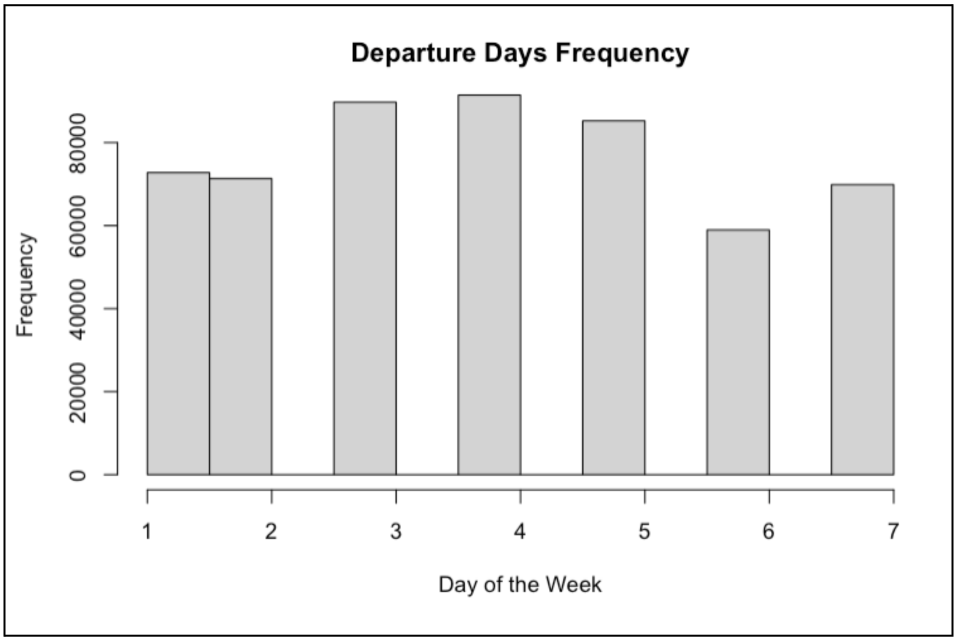
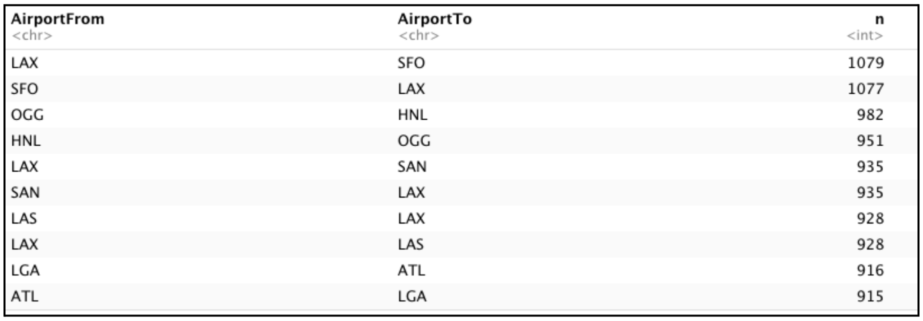


Figure 2. Histogram of Departure Days of Flights

In terms of the airports that the flights departed from and arrived at, most arrivals are from Hartsfield-Jackson, O'Hare, Dallas/Fort Worth, Denver, and Los Angeles. These same 5 airports above are also common destinations. The most common flight paths are LAX to SFO, SFO to LAX, OGG to HNL, HNL to OGG, and LAX to SAN.



AirportFrom <chr>	AirportTo <chr>	n <int>
LAX	SFO	1079
SFO	LAX	1077
OGG	HNL	982
HNL	OGG	951
LAX	SAN	935
SAN	LAX	935
LAS	LAX	928
LAX	LAS	928
LGA	ATL	916
ATL	LGA	915

Figure 3. Top Flight Paths

The airlines with the most flights are Southwest, Delta, Skywest, American, and Envoy Air.

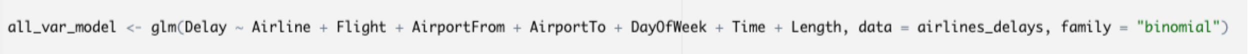
Southwest has the most delays at 27% of all delays. Almost 50% of delays are for SFO - LAX and back; however when layering on the airline, Dallas to Houston flights for Southwest cover almost 40% of delays. 34% of all delays occur during the middle of the week, but this could be more of a scale factor since there are more flights on the third or fourth days of the week. The delays based on time of departure are more evenly distributed.

Models and Results

First we need to convert these categorical variables to numeric: Airline, Flight, AirportFrom, and AirportTo so that we can apply the logistic regression models using R.

Classical Approach

Next, I apply the frequentist model using this equation:

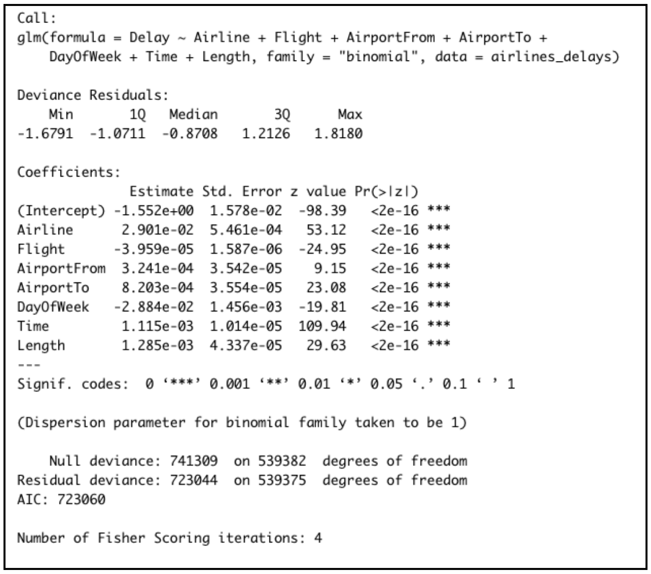


```
all_var_model <- glm(Delay ~ Airline + Flight + AirportFrom + AirportTo + DayOfWeek + Time + Length, data = airlines_delays, family = "binomial")
```

Figure 4. Frequentist Logistic Regression Model

The glm model applies the logit link and is applied to the binary response variable and covariates. It's based on the log odds ratio, and the prediction would then be taken based on

Here were the results:



```
Call:
glm(formula = Delay ~ Airline + Flight + AirportFrom + AirportTo +
    DayOfWeek + Time + Length, family = "binomial", data = airlines_delays)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.6791  -1.0711  -0.8708   1.2126   1.8180

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.552e+00  1.578e-02  -98.39   <2e-16 ***
Airline      2.901e-02  5.461e-04   53.12   <2e-16 ***
Flight      -3.959e-05  1.587e-06  -24.95   <2e-16 ***
AirportFrom  3.241e-04  3.542e-05    9.15   <2e-16 ***
AirportTo    8.203e-04  3.554e-05   23.08   <2e-16 ***
DayOfWeek   -2.884e-02  1.456e-03  -19.81   <2e-16 ***
Time         1.115e-03  1.014e-05  109.94   <2e-16 ***
Length       1.285e-03  4.337e-05   29.63   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 741309  on 539382  degrees of freedom
Residual deviance: 723044  on 539375  degrees of freedom
AIC: 723060

Number of Fisher Scoring iterations: 4
```

Figure 5. Frequentist Logistic Regression Model Results

The Pearson residual plot displays the difference between the predicted and the actual response data.

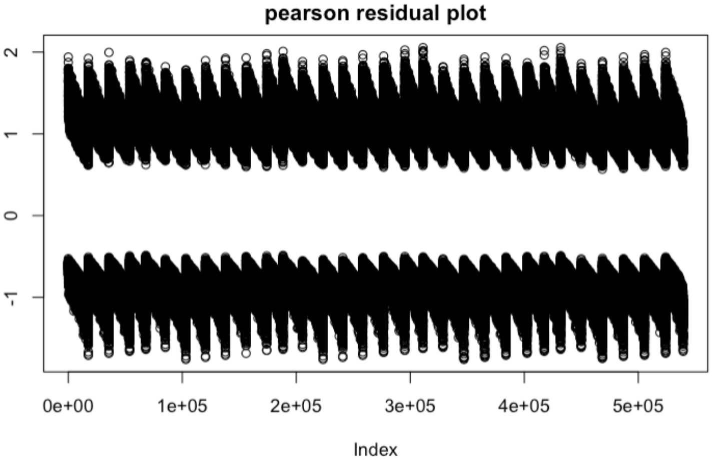


Figure 6. Frequentist Logistic Regression Model Residual Plot

We then select the best model to analyze with variable selection - using the Akaike information criterion (AIC) the full model is still preferred.

Now we will split 75% of the data into training data and the remaining to test the model on. We train the model and then make predictions based on the test data and get the accuracy metrics. After training the model and making predictions on the test data, the accuracy comes back to around 58.41%. Considering this is a realistic dataset, it's fairly decent. Here is the ROC curve:

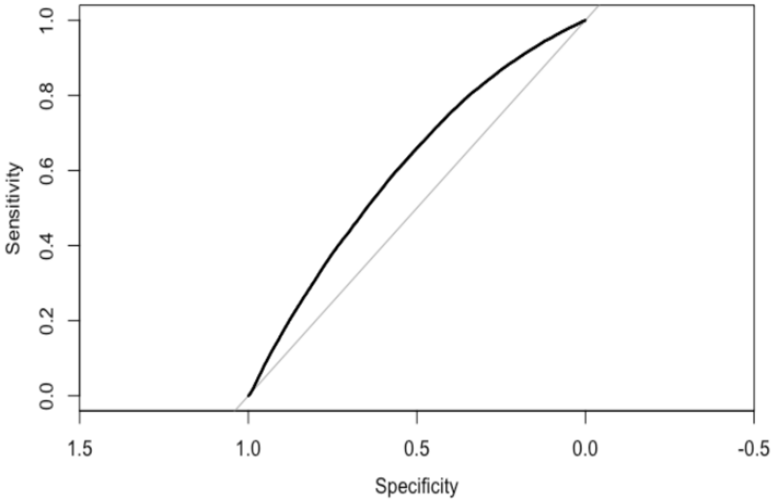


Figure 7. Frequentist Logistic Regression Model ROC Curve

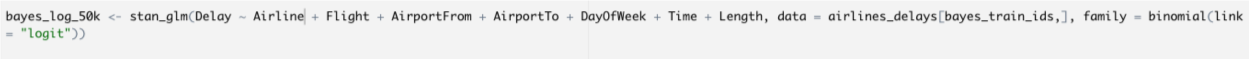
The AUC is around 0.6, which is somewhat poor.

Bayesian Approach

Now, we investigate the Bayesian method and apply logistic regression here. We'll use rstanarm to apply the Bayesian method.

Because stan_glm runs into error or runs slower when the data is larger than a certain volume, I sample 50,000 records from the data to build the model on. I then run the model on these records with default priors. Then another 50,000 records from the data are sampled as test data to be predicted on.

Here is the formula that was used:



```
bayes_log_50k <- stan_glm(Delay ~ Airline + Flight + AirportFrom + AirportTo + DayOfWeek + Time + Length, data = airlines_delays[bayes_train_ids,], family = binomial(link
= "logit"))
```

Figure 8. Bayesian Logistic Regression Model

The graph below shows which variables seem significant:

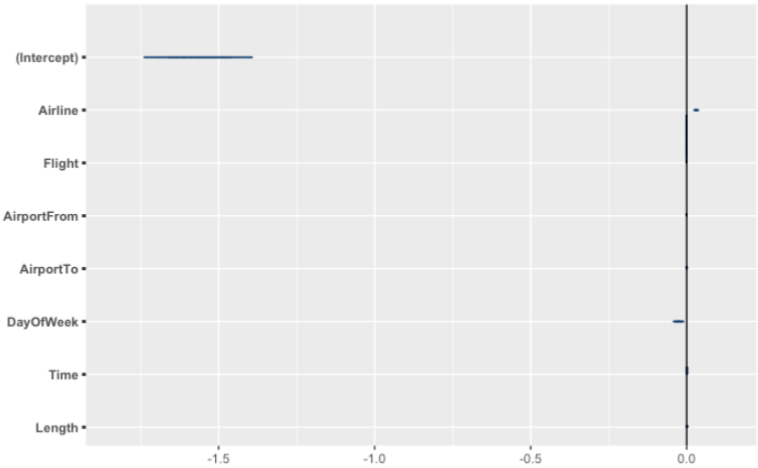
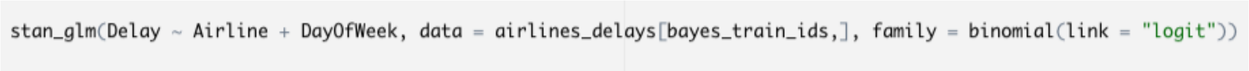


Figure 9. Bayesian Logistic Regression Model Variable Significance Plot

The accuracy is around 52% after applying the training model on the test data and comparing it with the observations. The significant variables seem to be just the airlines and day of the week. The 95% credible set for the predictor airlines is $[2.6 * 10^{-2}, 3.3 * 10^{-2}]$ and for day of the week it's $[-4.1 * 10^{-2}, -2.3 * 10^{-2}]$.

We re-train the model on just these 2 variables:



```
stan_glm(Delay ~ Airline + DayOfWeek, data = airlines_delays[bayes_train_ids,], family = binomial(link = "logit"))
```

Figure 10. Bayesian Logistic Regression Model with 2 Covariates

Going through the same method with prediction, the accuracy is slightly better at 55%.

Analysis

Comparing these two models, the model for both the frequentist and Bayesian approach seem to be decent. However, the dataset itself may be more challenging to be explained by logistic regression. Both models seem to be decent in prediction but the frequentist method does seem to perform slightly better. However, from a Bayesian standpoint the advantages of applying knowledge from a strong prior does help with increasing the knowledge of whether a delay will occur or not.

Conclusion

This project was to compare the difference between classical logistic regression and Bayesian logistic regression. Future work could include analyzing certain packages or software that would allow for more iterations for the Bayesian approach and allow the model to run on a larger dataset. Also, more work can be done to specify more informative priors and run more training models on each method and compare them. More fields pulled on the dataset could help with making better predictions, given that the categorical variables needed to be updated to numerical for the purposes of the modeling steps. Overall, the logistic regression modeling can be viewed from both a frequentist and Bayesian standpoint.

References

Chacko, Jims. “Airlines Dataset to Predict a Delay.” *Kaggle*, 21 June 2022, https://www.kaggle.com/datasets/jimschacko/airlines-dataset-to-predict-a-delay?resource=download.

“Flight Delays & Cancellations.” *US Department of Transportation*, United States Department of Transportation, 10 Oct. 2017, https://www.transportation.gov/individuals/aviation-consumer-protection/flight-delays-cancellations.

Fox, Alison. “The Airlines with the Most Delays This Year, According to the Bureau of Transportation Statistics.” *Travel + Leisure*, Travel + Leisure, 27 Oct. 2022, https://www.travelandleisure.com/most-delayed-airlines-2021-2022-6814429.

Team, Content. “An Airline Leverages AI to Reduce Operations Costs by More than 3% by Predicting Flight Delays More Accurately.” *CrowdANALYTIX*, 18 June 2020, https://www.crowdanalytix.com/an-airline-leverages-ai-to-reduce-operations-costs-by-more-than-3-by-predicting-flight-delays-more-accurately/.