Flight Delays: A Bayesian Logistic Regression Analysis

Bayesian Machine Leaning Final Paper

Ryan Lipps (rhl8pk@virginia.edu), Sydney Mathiason (qex8sd@virginia.edu)

2023-12-11

Problem Description

As the holidays approach, we are interested in informing travelers on the risk of flight delays that would add to an already stressful season. By analyzing data on airline delays, we hope to gain insight on factors contributing to delays and guide travelers on how best to anticipate them with the support of measures of uncertainty.

Data Overview

The data analyzed for this project is from the U.S. Department of Transportation's Bureau of Transportation Statistics and consists of information on domestic, commercial flights from January 1, 2015 through December 31, 2019. The dataset includes several columns such as date of flight, scheduled departure and arrival times, actual departure and arrival times, airline (carrier), origin and destination airport, cancellations, and reason for delay. A full data dictionary is provided in Appendix 1.

For the purposes of our analysis, we reduced the features to the list below. Links to relevant exploratory data analysis plots and notes are provided for each feature.

- Hour - hour of scheduled departure time; extracted from scheduled departure time (see Appendix 2a)
- Day of week - day of week for scheduled departure time; extracted from date (see Appendix 2b)
- Day of year - day of year for scheduled departure time (leap year 2016 included); extracted from date (see Appendix 2c)
- Month - month for scheduled departure time; extracted from date (see Appendix 2d)
- Destination - destination airport 3-digit code (see Appendix 2e)
- Origin - origin airport 3-digit code (see Appendix 2f)
- Carrier - airline 2-digit code (see Appendix 2g)
- Delayed - binary outcome with 1 for delayed, 0 for not delayed; imputed from arrival delay being greater than 20 minutes

To give a holistic interpretation of flight delays, we based the time components in our analysis around departure times as opposed to arrival times, while interpreting delays as being based on arrival time only. Our intention behind this interpretation of delay is that generally travelers are more concerned with whether or not they are delayed in arriving to their destination, regardless of if they were delayed in departure. For instance, one might miss a connecting flight if their arrival to a stopover airport is significantly delayed, whereas if a flight is delayed departure, time may be made up during the flight and upon arrival.

It is important to note that the raw dataset for all of 2019 is missing explicit scheduled departure time (for which we based the model's time components). To correct for this, we imputed the scheduled departure time by subtracting the departure delay values from the actual departure time values. To ensure this was a viable imputation we checked the calculations against scheduled departures for 2015 and 2017 and achieved complete agreement.

Due to computational limitations, we took a stratified sample of flights by date, reducing to 90,280 observations from the original 34 million observations. We performed exploratory data analysis on both the stratified set and the original set and did not find a significant difference between the two. Thus we believe our modeling and analysis is representative of the full dataset.

We used one-hot encoding to include the relevant categorical predictors of carrier, origin, and destination in the models. As this significantly increases the memory and computational load in the modeling process, we chose to limit our analysis to the airlines with the highest number of flights, being Southwest (WN), Delta (DL), United Airlines (UA), American Airlines (AA), and Skywest (OO). Additionally, we limited origin and destination airports to the top 25% of airports by incoming and outgoing flight volume, respectively. This simultaneously reduces computational load and limits our analysis to the highest volume airports and airlines.

Lastly, we excluded the month of December 2019 from model training. This functions as a test set for evaluating model predictive performance.

Probability Modeling

We have opted to use Bayesian logistic regression for the model basis, as our objective is to predict the probability of a flight being delayed. Bayesian logistic regression provides interpretable coefficients that represent the impact of each predictor on the log-odds of a flight being delayed. This interpretability is crucial when trying to understand the influence of predictors on the probability of flight delays. Furthermore, the output of Bayesian logistic regression provides probability distribution estimates, aiding in informed decision making and providing easily interpretable measures of uncertainty.

Approach

To consider the impact of increased model complexity due to the large number of categorical predictors in our data, we evaluated three different Bayesian logistic regression models: one model consisting of all predictors outlined in the data overview section (see Appendix 3a); a second model with the seasonal time components, origin, and destination (see Appendix 3b); and a final model with the seasonal time components and carrier (see Appendix 3c).

In all of the model specifications, we used Gaussian priors centered on zero with a standard deviation of one. For the categorical predictors we assumed no covariance. For the time components, we specifically used the ZeroSumNormal function used by Benjamin T. Vincent, who employs this function to “transform the normally distributed monthly deflections to have a mean of zero in order to reduce the degrees of freedom of the model by one, which should help with parameter identifiability.” (Vincent, 2022), in their analysis of COVID-19 time series data.

As the dataset is large we used Automatic Differentiation Variational Inference (ADVI) to approximate the model posteriors, and evaluated posterior convergence via Evidence Lower Bound (ELBO) plots, available in the appendix (see Appendix 4a).

We evaluated the base performance of each of the three models using both prior predictive and posterior predictive checks, available in the appendix (see Appendix 4b and 4c). The plots are similar across each of the models, showing clear separation in prior and posterior predictions, with the posterior predictive means aligning with the observed means. There is slight imbalance in the prior predictions, however, this is reflected in the imbalance in the original data.

After evaluating the three models, we conducted a comparison using the Widely Applicable Information Criterion (WAIC) (see Appendix 4d). The model featuring origin and destination information without carrier details exhibited the best performance, followed by the full model. Despite having the lowest WAIC and thus possibly underfitting, the model incorporating carrier information still contributes valuable insights. Examining the proposed model weights provides further nuance to this observation. The origin-destination model received a weight of 0.69, as expected from the higher WAIC. In contrast, the full model received a weight of zero, suggesting it may be less valuable in an ensemble approach. The carrier model received a weight of 0.31, underscoring its significance as a complementary information source to the origin-destination model. Despite these findings, our interest remains in exploring the unique information provided by each of these two models, prompting a separate evaluation of results from both.

Results

Origin-Destination Model

To understand the top-level prediction capabilities of the origin-destination model, we predicted flight delays on the test set reserved from December 2019. By comparing these predictions to the observed flight delays we constructed an ROC curve (see Appendix 5a). With this ROC curve we achieved an AUROC of approximately 0.6. The ROC curve and AUROC indicate that the model does perform better than random guessing on the predictions. Furthermore, since we are interested in minimizing false negatives in our predictions, we constructed a confusion matrix with a hard classification cutoff at $p > 0.1 =$ Delayed and achieved a sensitivity of 0.76, giving a false negative rate of 0.24.

To address the fit and prediction accuracy of the model with respect to the individual origins and destinations, we evaluated the model against the observed proportion of delays. While evaluating predicted probability of delay against the observed proportion of delays is not a one-to-one comparison, we feel as though it is a sufficient measure of fit and predictive capability, given the lack of a true probability measurement. The posterior probability means and 94% High-Density-Intervals (HDIs) for origin and destination have good agreement with the observed proportion of delays when comparing to the training data. This indicates good model fit for these variables (see Appendix 5b). For the model predictions, generally the observed proportions fall within the 94% HDIs. However, there are a few origins and destinations for which our predictions do not capture the observed proportions in the HDIs (see Appendix 5c). It is possible that this is due to a limited number of observations for these airports as we are limited by the stratified sample in this case.

For the origin-destination model we found that almost all airport coefficients were significant and below zero, indicating that most origins and destinations in the model decrease the log-odds of a delay occurring. Initially this seems counterintuitive as one would expect some airports increase the odds of a delay. However, a comparison of airport delay impact is still relevant by considering which airports are decreasing the odds of delay the most. For ease of visualization and analysis, forest plots of all origins and destinations ordered by 94% HDI upper-bound are available in the appendix (see Appendix 5d). Based on 94% HDI width, there does not appear to be an apparent trend between airport volume and coefficient uncertainty. For instance, Dallas-Fort Worth International Airport was one of the busiest airports by traffic during the time period analyzed and has one of the smaller HDIs for both origin and destination, whereas Los Angeles International (having similar traffic volume) has significantly more uncertainty in its coefficient value.

For ease of interpretation, we can transform the coefficient values for origins and destinations into odds ratio plots. As an example, we can see that in comparing plots for Dallas-Fort Worth International and Las Vegas International (see Appendix 5e) the upper bound of the 95% credible interval for DFW is approximately 0.5, whereas the lower bound of the 95% credible interval for LAS is approximately 0.6. This indicates that flights originating from DFW are less likely to be delayed than flights originating from LAS.

To address fit and prediction accuracy with respect to time features, we evaluated the model against the observed proportion of delays by time. Again this is not a direct comparison, but we feel as though it is a sufficient measure of fit and predictive capability. These plots are available in the appendix (see Appendix 5f). With the exception of day of week, the model fits the training data well, with most observed proportions falling within the 95% HDIs. The day of year fit indicates decreasing probability of delay from January to March, increasing probability of delay from March into November, and increasing probability of delay from November into December, with relatively constant uncertainty throughout the year. Unfortunately, the fit for day of week is inaccurate, drastically over and under shooting the observed proportions. We believe that the inaccuracy of fit is influenced by correlation with another time predictor. The fit to hour of day is much more accurate, with most observed proportions falling within the 95% HDI. The model indicates the probability of a delay increases from roughly 5:00AM throughout the day and into the early morning of the next day. Uncertainty around this fit increases with the same periodicity.

The accuracy of model fit for the time-based variables translates well to model predictions, with most observed proportions falling within the 95% HDIs. We see similar trends for both probability of delay and uncertainty as discussed in the model fit paragraph above (see Appendix 5g).

Forest plots for the various time-based predictors are available in the appendix (see Appendix 5h). There is consistent uncertainty but varying significance around coefficient values for day of week, day of year, and month, whereas there is variable uncertainty and significance around coefficients for hour.

Carrier Model

Despite the difference in WAIC mentioned in the "Approach" section above, the model with only carrier information performed comparably to the origin-destination model in evaluation of the ROC curve (see Appendix 6a). The carrier-only ROC curve has an AUROC of approximately 0.56, compared to the origin-destination AUROC of 0.6. Nevertheless, the carrier-only ROC curve and AUROC indicate that the model does perform better than random guessing on predictions. With the same cutoff ($p > 0.1 = \text{Delayed}$) as the origin-destination model, we constructed a confusion matrix and achieved a sensitivity of 0.72, giving a false negative rate of 0.28.

Similar to the origin-destination model, we evaluated the carrier-only model fit and prediction probabilities against the observed proportion of delays by carrier. For all carriers, the observed proportion of delays were close to the fitted and predicted means of the posterior probabilities, with all observed proportions falling within the 94% HDIs. This indicates good fit and predictive performance in relation to the flight carriers (see Appendix 6b).

The model indicates all carriers have significant coefficients, shown in the forest plots in the appendix (see Appendix 6c). Despite this significance, there is little differentiation between the coefficients for Skywest, Southwest, American Airlines, and United Airlines, indicating that they have similar impact on the probability of flight delays. Notably, the coefficient for Delta does not overlap with the other airlines. As such, Delta decreases the probability of delay more than the other airlines, according to the model. This is easier to visualize with the odds ratio plots (see Appendix 6d), where we see that the upper bound of the credible interval for Delta only slightly overlaps with the lower bound of United, whereas the other airlines have significantly overlapping credible intervals.

Conclusion

Overall, the model exhibits better performance than random guessing, though there is room for improvement. Through Bayesian logistic regression we have provided a baseline of valuable information that travelers can use to optimize their flight schedules and minimize the likelihood of delays. The model incorporates measures of uncertainty on the probability of delay based on factors of time, origin, destination, and carrier, offering insights to inform decision-making. With access to a more powerful computing system, we could broaden our analysis by incorporating additional existing data, hopefully decreasing model uncertainty. With more support from additional data and computational power, we could investigate the impact of interaction effects between predictors and try to tune predictions for day of week. Furthermore, we could update our analysis with the latest data, ensuring that the model remains current and relevant in the ever-evolving landscape of flight traffic and potential delays.

Citations

Vincent, Benjamin T., “Counterfactual inference: calculating excess deaths due to COVID-19,” PtMC Team, eds., 10.5281/zenodo.5654871

Github Repo

https://github.com/rlipps/BayesML_Final_Project

Appendix 1: Data Dictionary

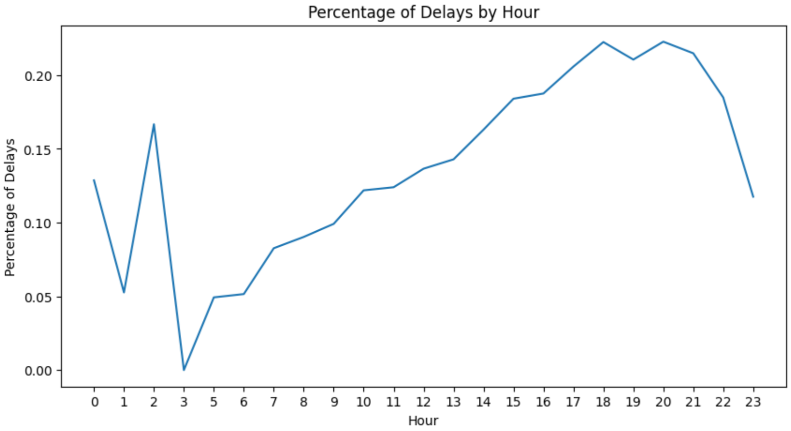
Reference: https://www.transtats.bts.gov/Fields.asp?P74_e19=EDD&P74_P1y=&n22yB_4n6r=&5146 _p1y7z0=&frn4pu_Y11x72=&sv0q=&5146_14qr4=

- FL_DATE: Date of flight
- OP_CARRIER: Abbreviation of carrier
- OP_CARRIER_FL_NUM: Carrier flight number
- ORIGIN: Origin airport code
- DEST: Destination airport code
- CRS_DEP_TIME: CRS (scheduled) departure time (local time: hhmm)
- DEP_TIME: Actual departure time (local time: hhmm)
- DEP_DELAY: Difference in minutes between scheduled and actual departure time. Early departures show negative numbers.
- TAXI_OUT: Taxi Out Time, in Minutes
- WHEELS_OFF: Wheels Off Time (local time: hhmm)
- WHEELS_ON: Wheels On Time (local time: hhmm)
- TAXI_IN: Taxi In Time, in Minutes
- CRS_ARR_TIME: CRS (scheduled) Arrival Time (local time: hhmm)
- ARR_TIME: Actual Arrival Time (local time: hhmm)
- ARR_DELAY: Difference in minutes between scheduled and actual arrival time. Early arrivals show negative numbers.
- CANCELLED: Cancelled Flight Indicator (1=Yes)
- CANCELLATION_CODE: Reason for cancellation (A = carrier, B = weather, C = NAS, D = security)
- DIVERTED: Diverted Flight Indicator (1=Yes)
- CRS_ELAPSED_TIME: CRS (scheduled) Elapsed Time of Flight, in Minutes
- ACTUAL_ELAPSED_TIME: Elapsed Time of Flight, in Minutes
- AIR_TIME: Flight Time, in Minutes
- DISTANCE: Distance between airports (miles)
- CARRIER_DELAY: Carrier Delay, in Minutes
- WEATHER_DELAY: Weather Delay, in Minutes
- NAS_DELAY: National Air System Delay, in Minutes
- SECURITY_DELAY: Security Delay, ft Delay, in Minutes
- Unnamed: 27: Unused

Appendix 2: Exploratory Data Analysis Plots

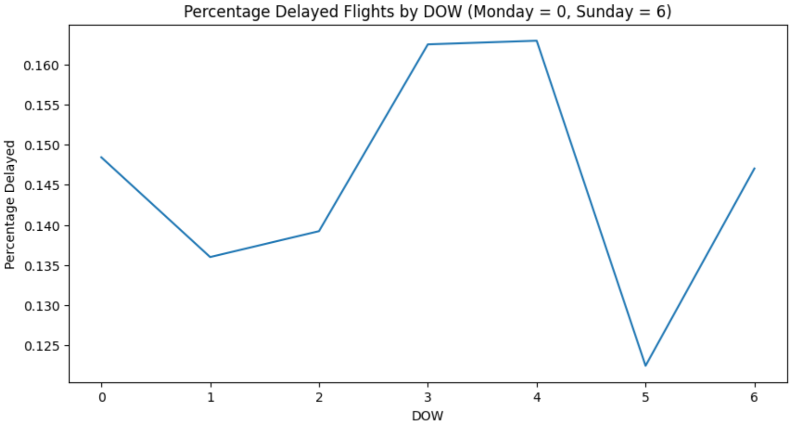
- EDA plots are given as percentage of delays as opposed to number of delays to account for overall flight volume

Appendix 2a: Plot of Percentage Delays by Hour



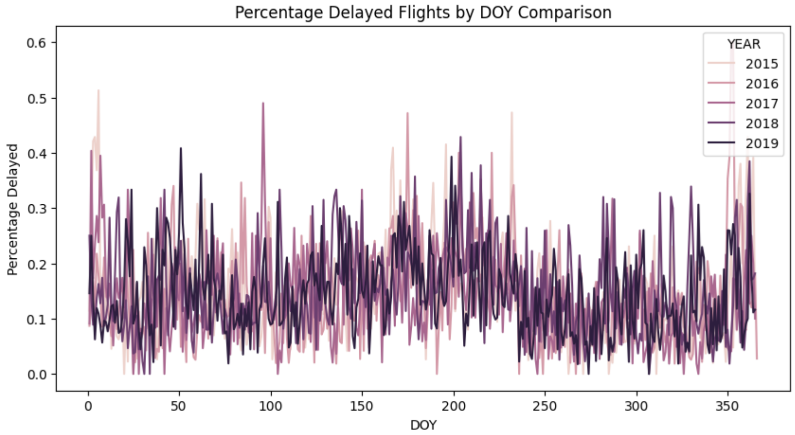
* Percentage of delays increases by hour from 3:00AM into the next morning, with a slight drop-off in the late-night/early morning

Appendix 2b: Plot of Percentage Delays by Day of Week



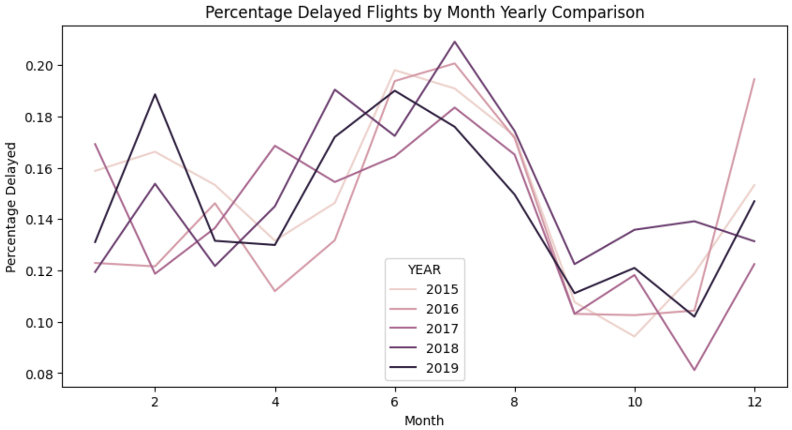
* There appears to be a slight trend in percentage of delays by day of week

Appendix 2c: Plot of Percentage Delays by Day of Year



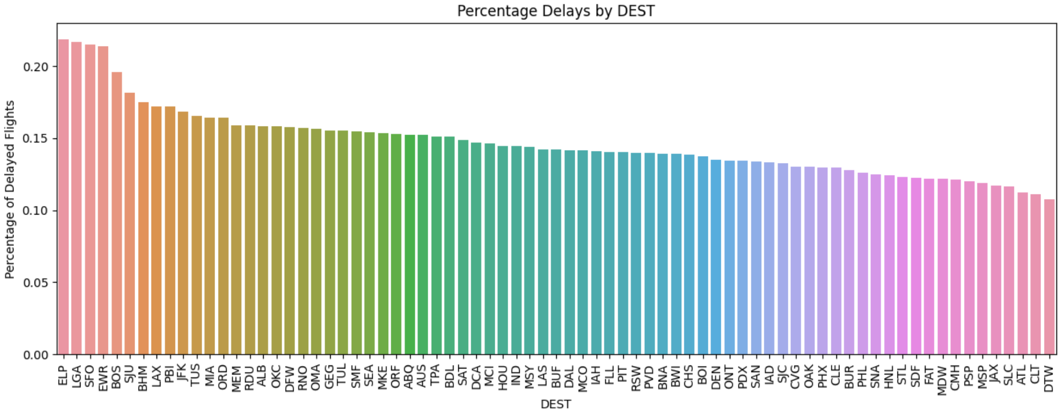
* This plot is noisy but DOY is included in models to try and account for relative holiday volume

Appendix 2d: Plot of Percentage Delays by Day of Month



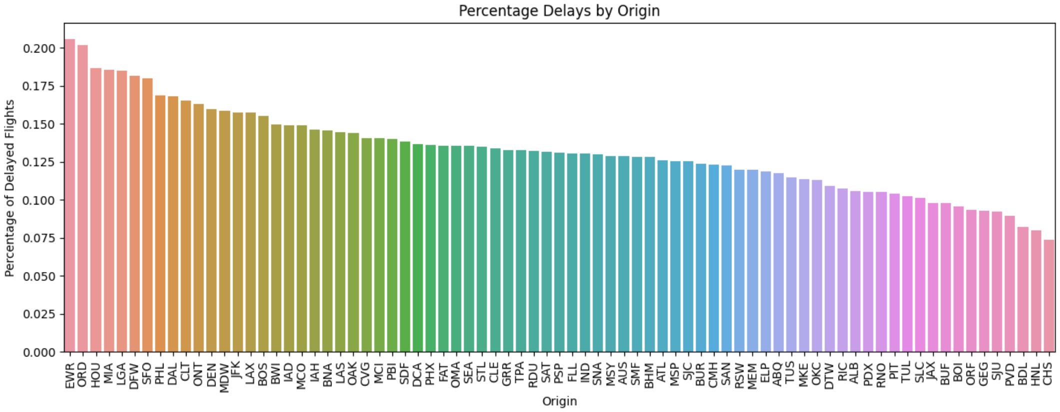
* Delays appear to increase in the summer, a drop off in fall, and increase in December

Appendix 2e: Plot of Percentage Delays by Destination



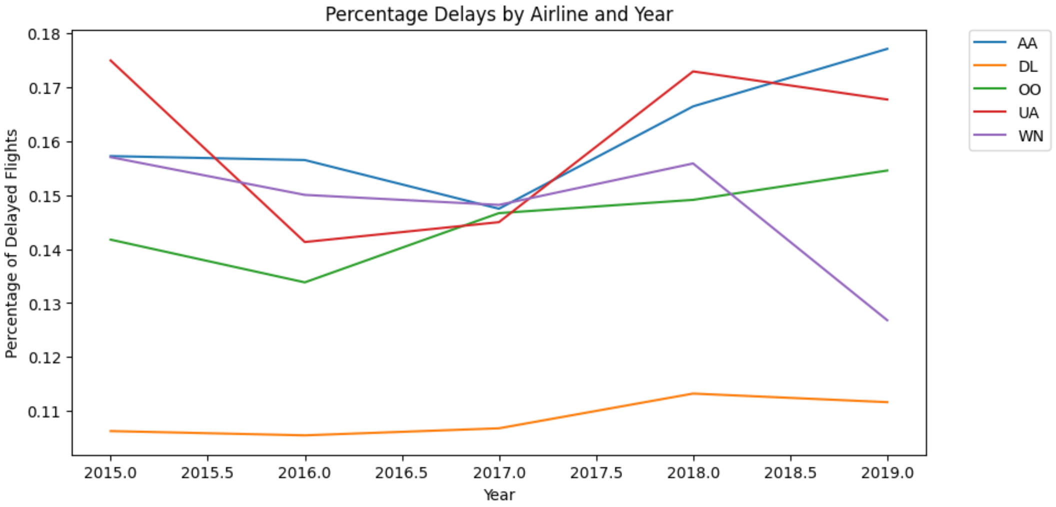
* There appears to be a relationship between delays and flight destination

Appendix 2f: Plot of Percentage Delays by Origin



* There appears to be a relationship between delays and flight origin

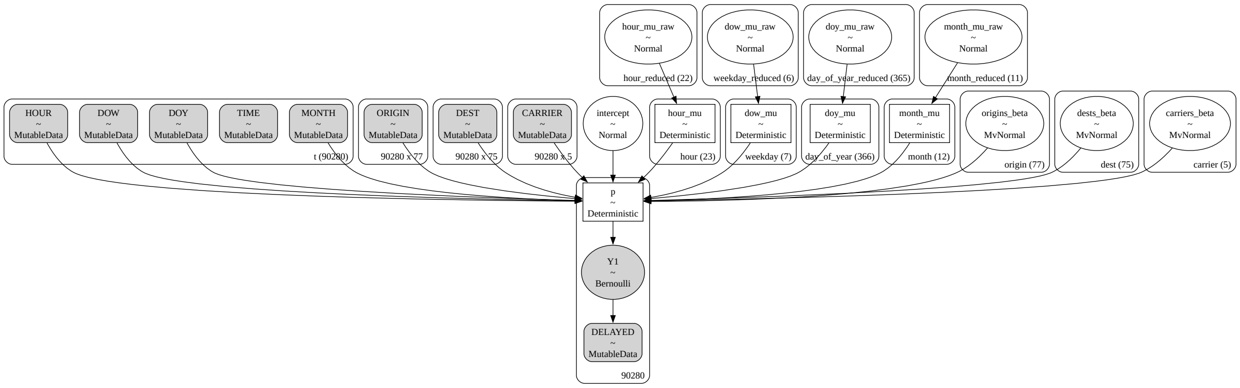
Appendix 2g: Plot of Percentage Delays by Carrier



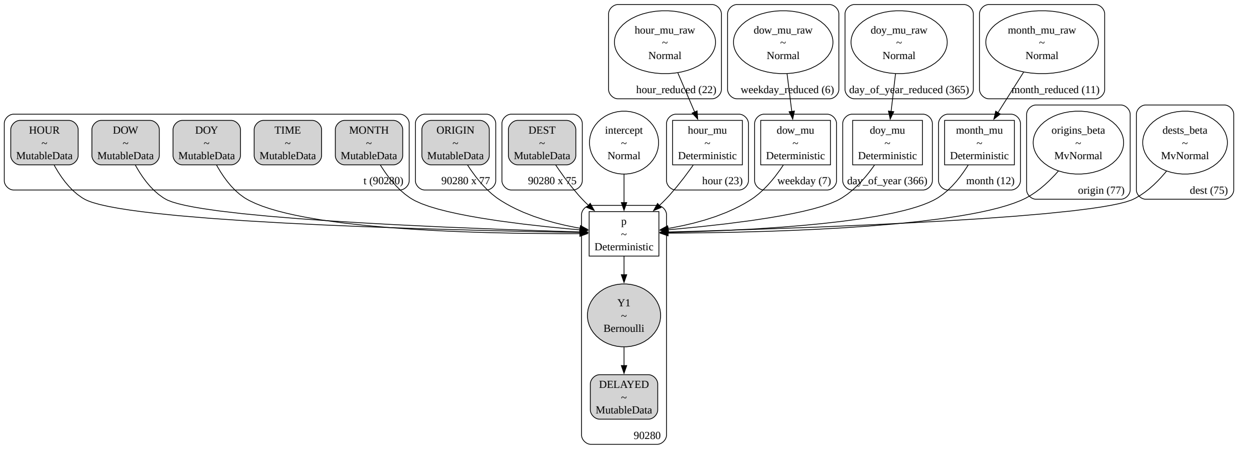
* There appears to be a relationship between delays and flight carrier

Appendix 3: Model Diagrams

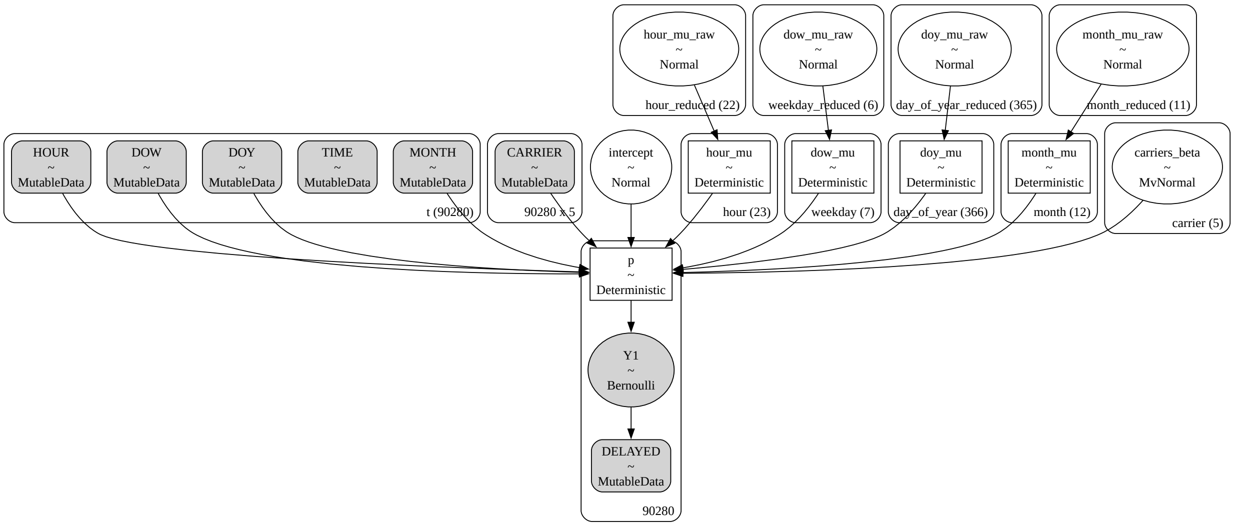
Appendix 3a: Full Model Diagram



Appendix 3b: Origin-Destination Model Diagram

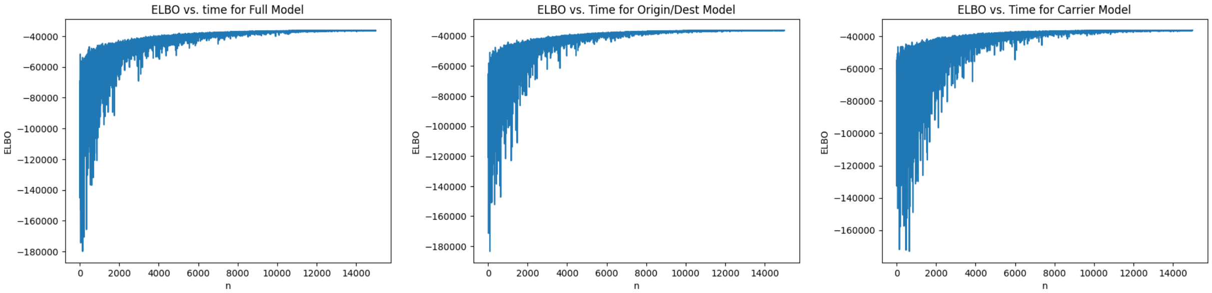


Appendix 3c: Carrier Model Diagram

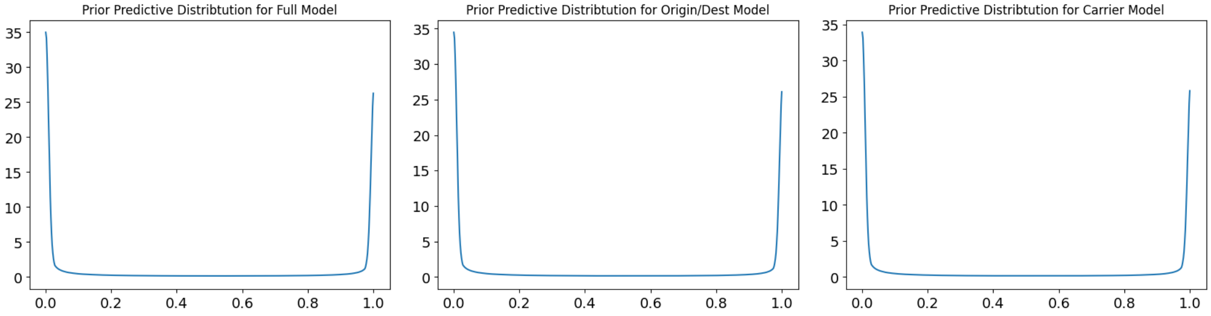


Appendix 4: Model Evaluation Plots and Information

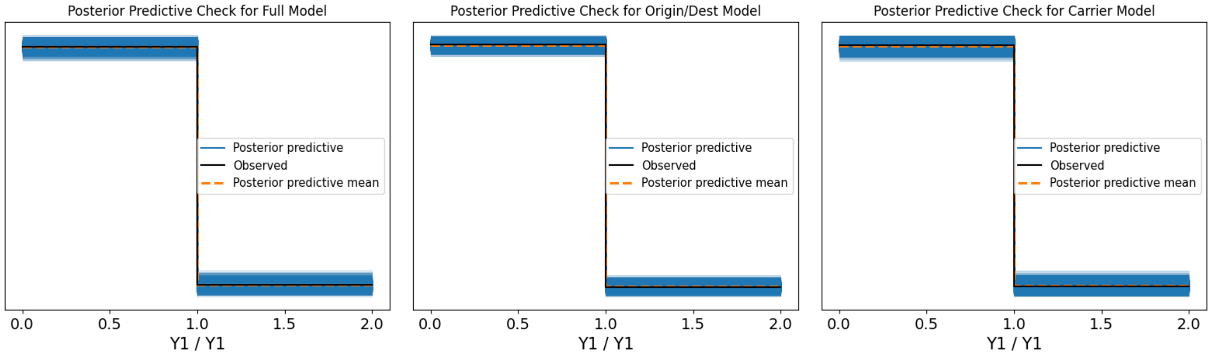
Appendix 4a: Model ELBO Plots



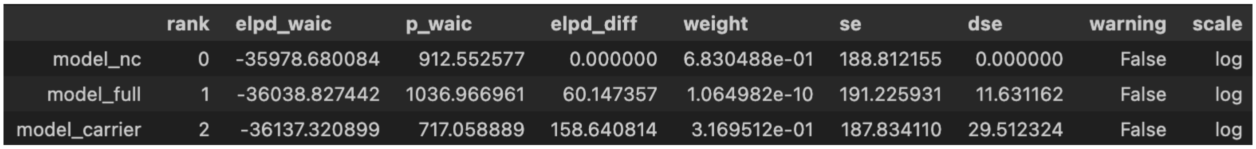
Appendix 4b: Model Prior Predictive Plots



Appendix 4c: Model Posterior Predictive Plots



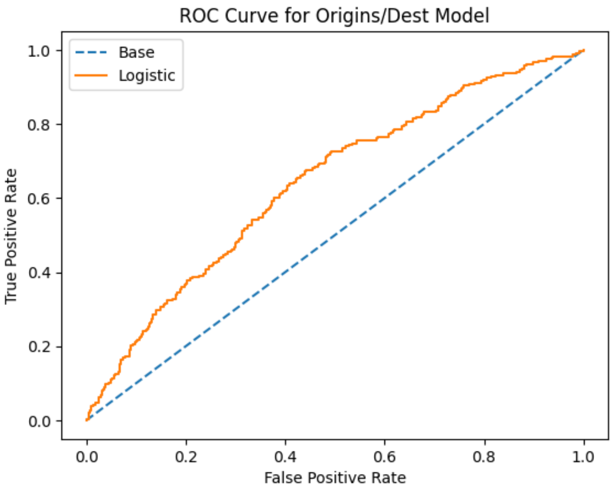
Appendix 4d: WAIC Comparison



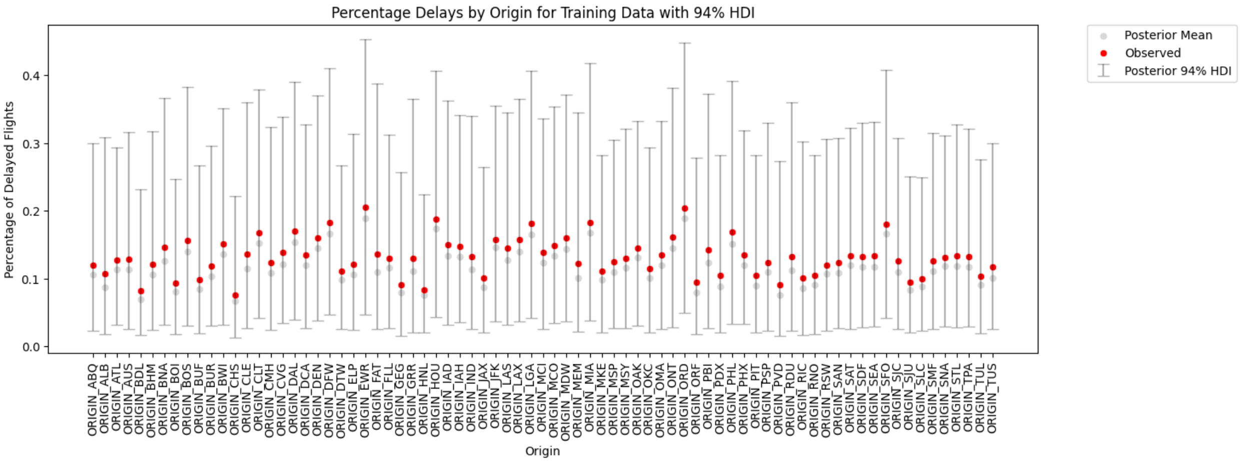
	rank	elpd_waic	p_waic	elpd_diff	weight	se	dse	warning	scale
model_nc	0	-35978.680084	912.552577	0.000000	6.830488e-01	188.812155	0.000000	False	log
model_full	1	-36038.827442	1036.966961	60.147357	1.064982e-10	191.225931	11.631162	False	log
model_carrier	2	-36137.320899	717.058889	158.640814	3.169512e-01	187.834110	29.512324	False	log

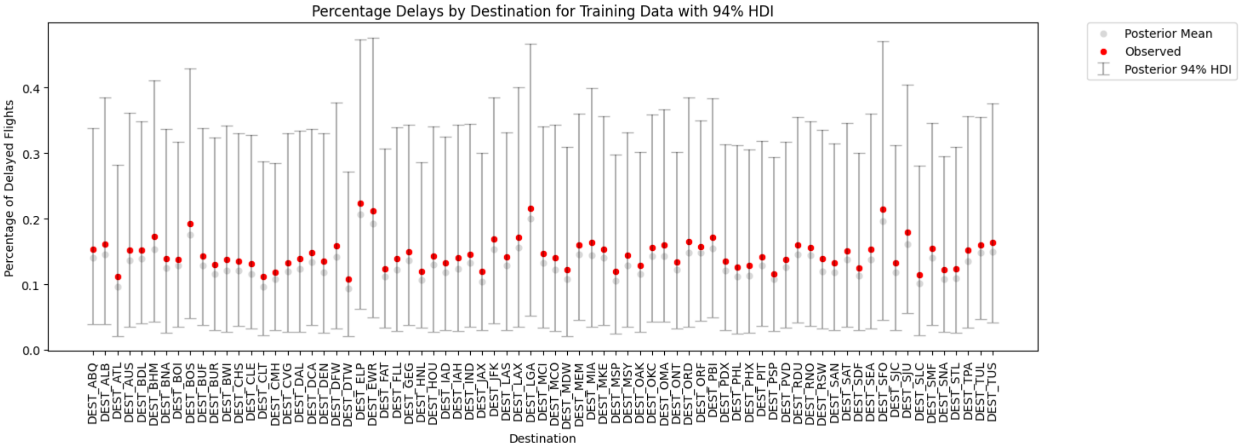
Appendix 5: Origin-Destination Model Plots

Appendix 5a: Origin-Destination ROC Curve

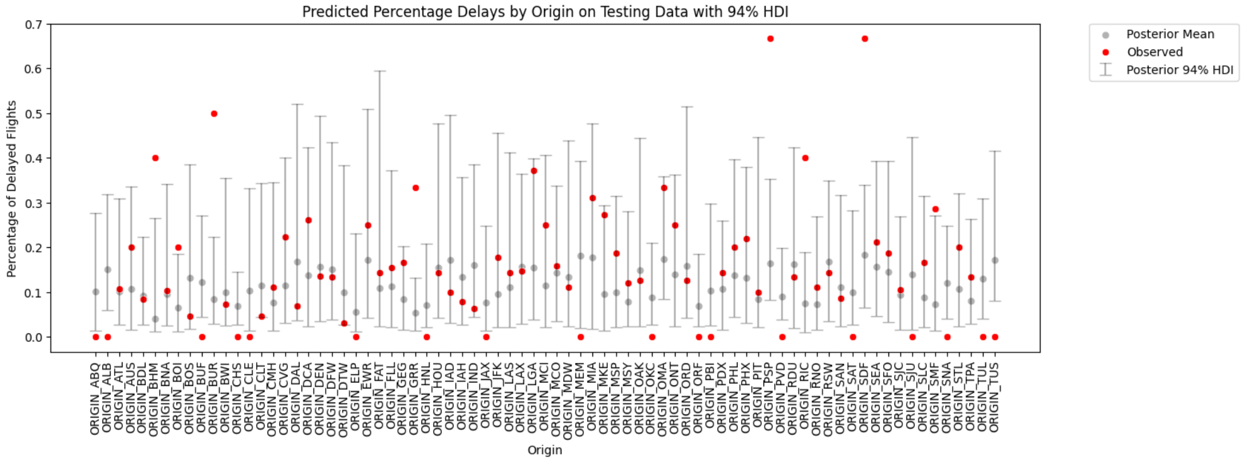


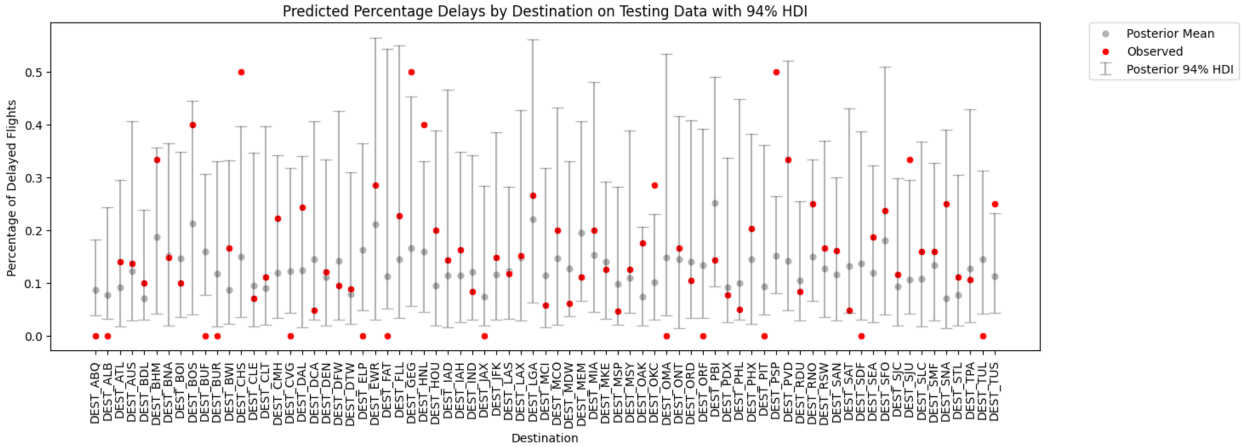
Appendix 5b: Origin-Destination Categorical Fit Plots



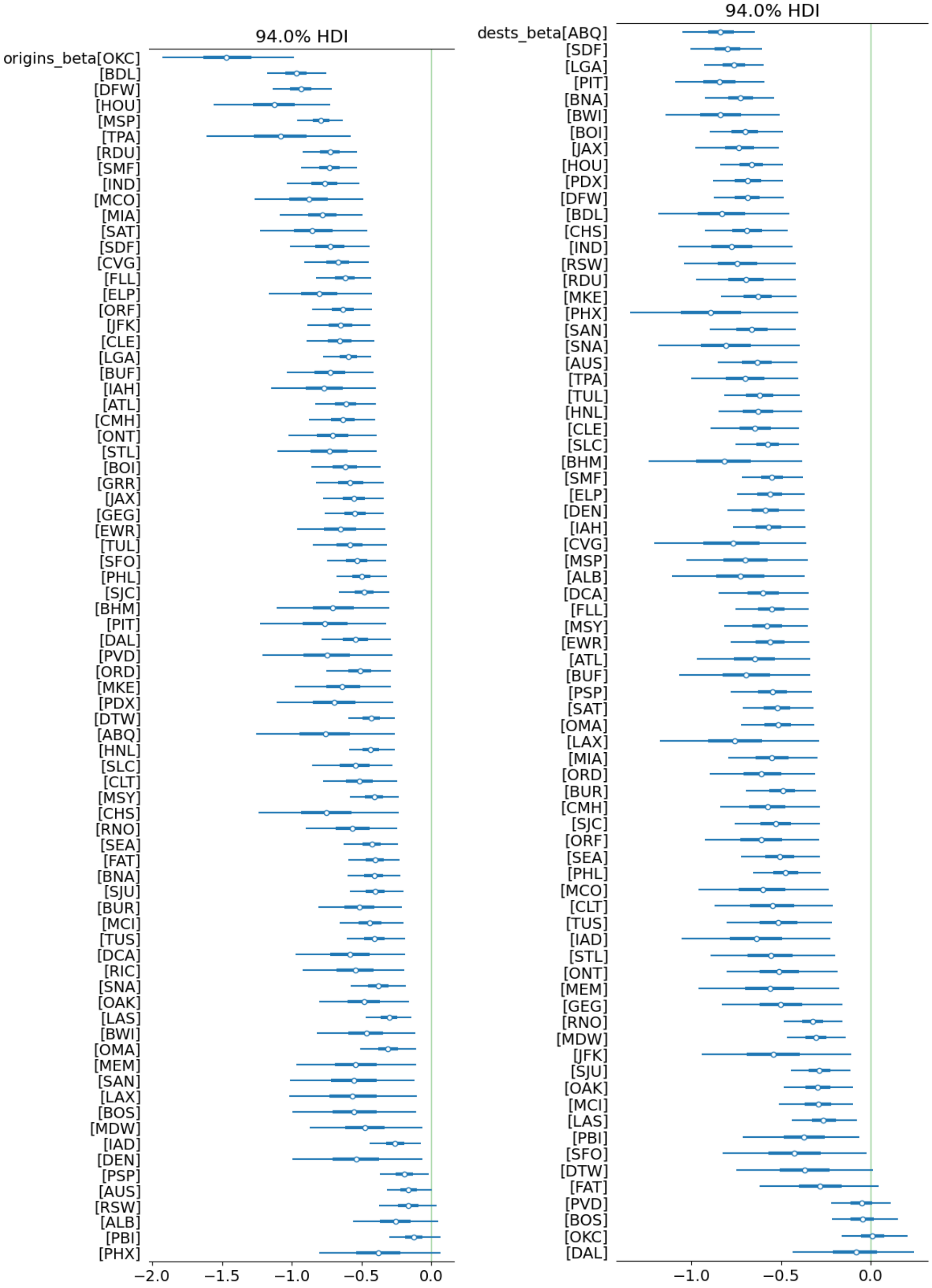


Appendix 5c: Origin-Destination Categorical Prediction Plots

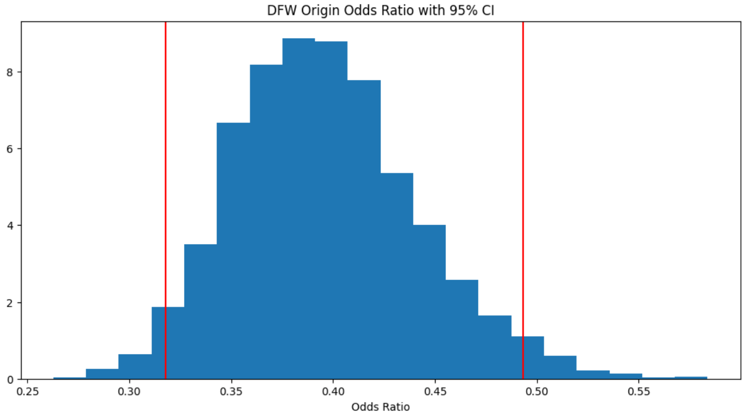


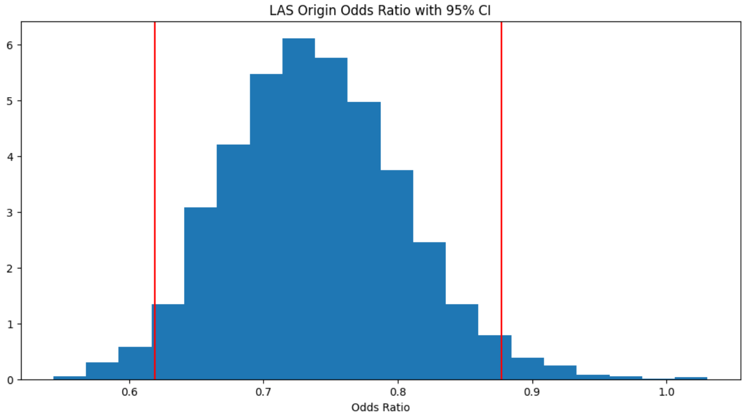


Appendix 5d: Origin-Destination Forest Plots

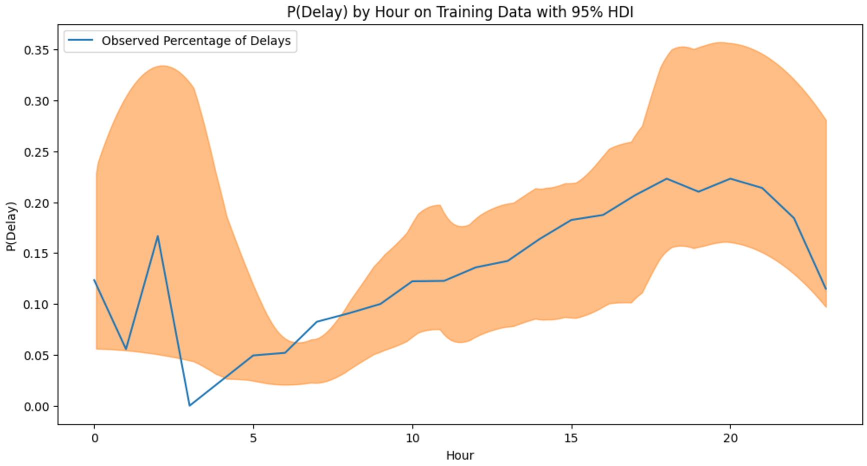


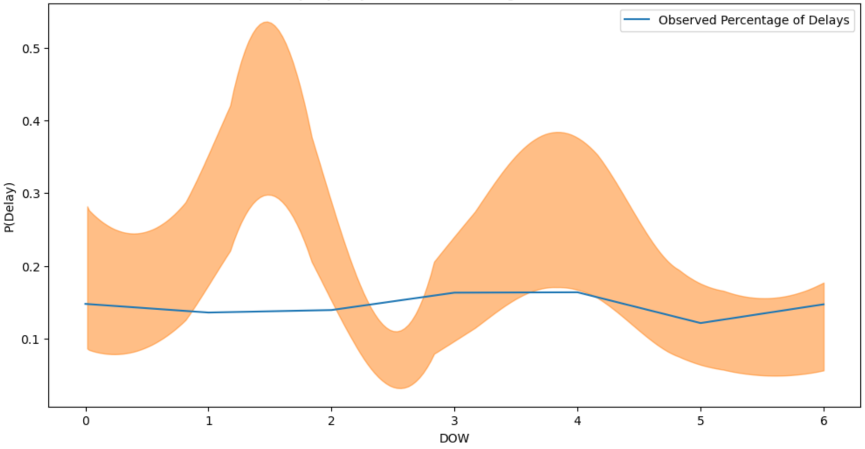
Appendix 5e: DFW and LAS Origin Odds Ratio Plots

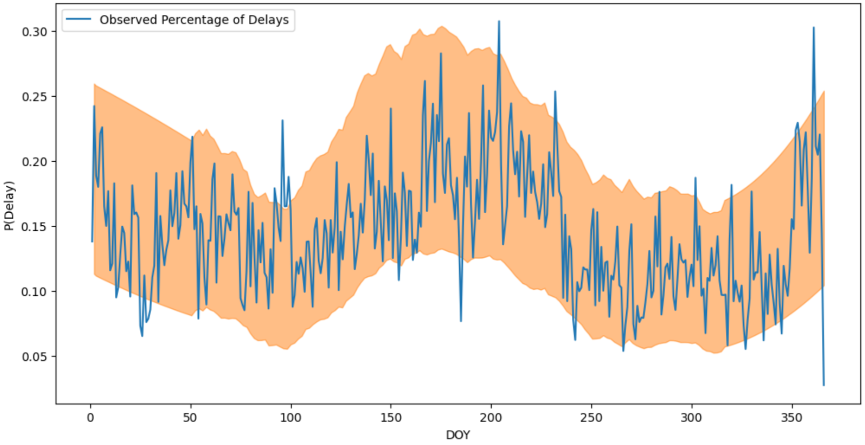

DFW Origin Odds Ratio with 95% CI

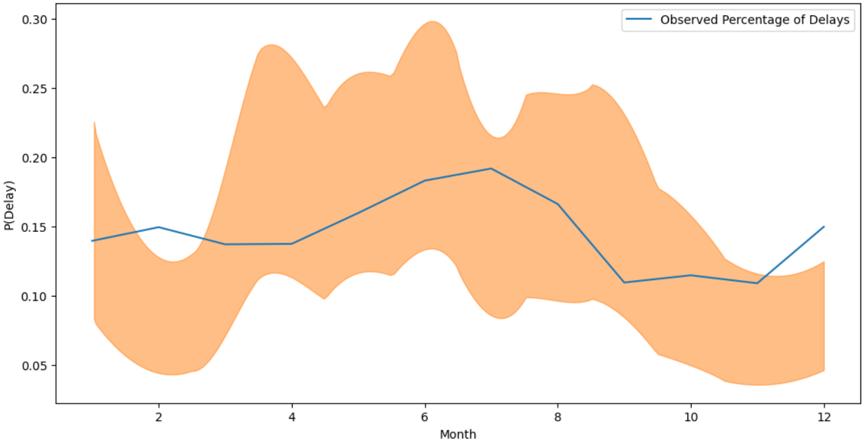

LAS Origin Odds Ratio with 95% CI

Appendix 5f: Time Feature Fit Plots

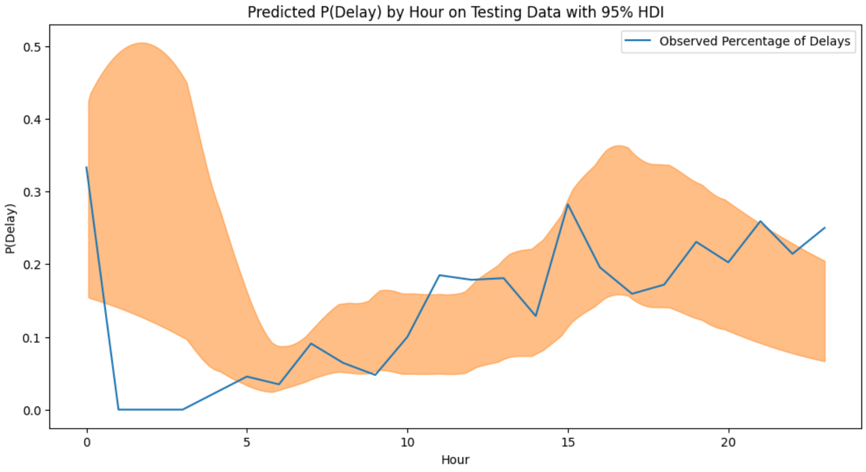

P(Delay) by Hour on Training Data with 95% HDI

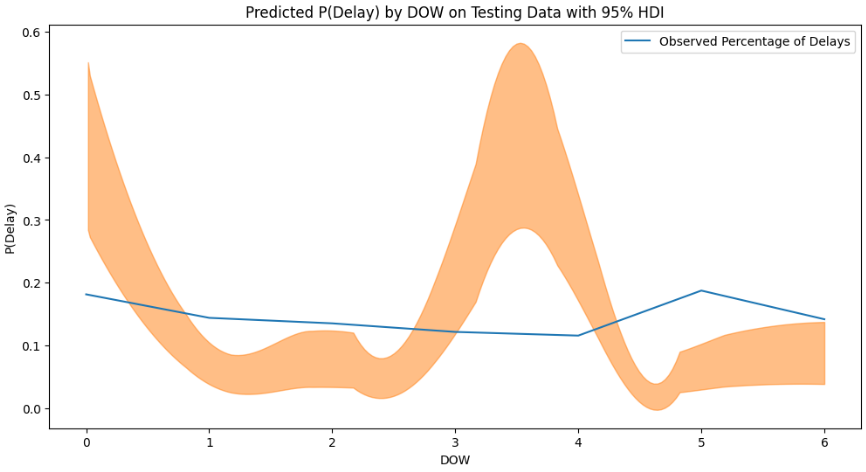

P(Delay) by Day of Week on Training Data with 95% HDI

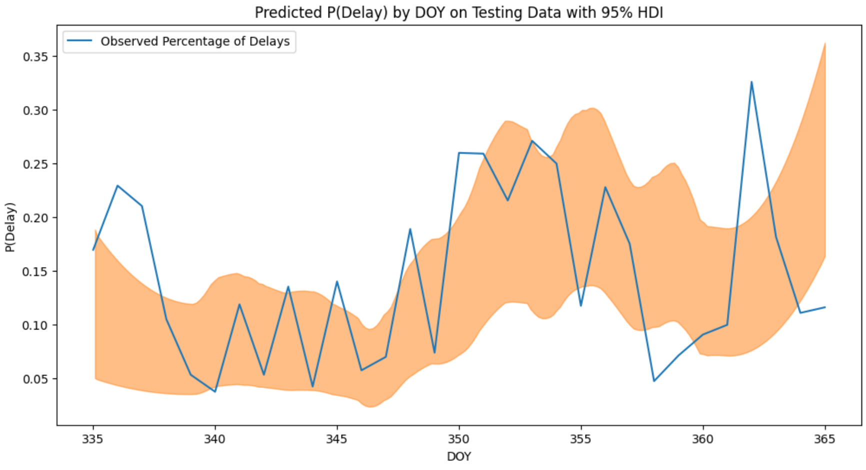

P(Delay) by Day of Year on Training Data with 95% HDI


P(Delay) by Month on Training Data with 95% HDI

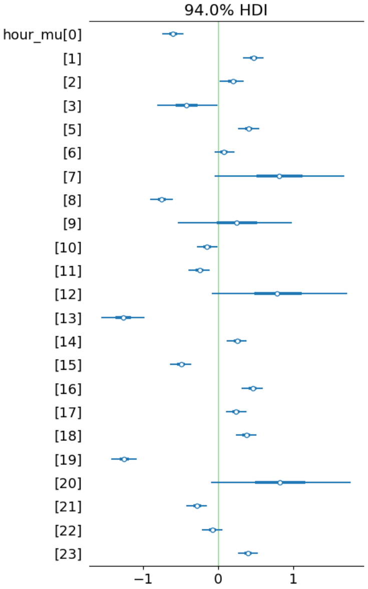
Appendix 5g: Time Feature Prediction Plots

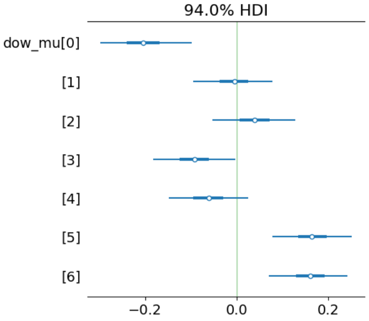


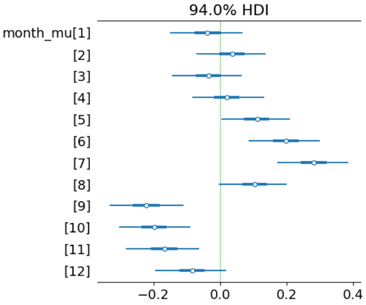




Appendix 5h: Time Feature Forest Plots

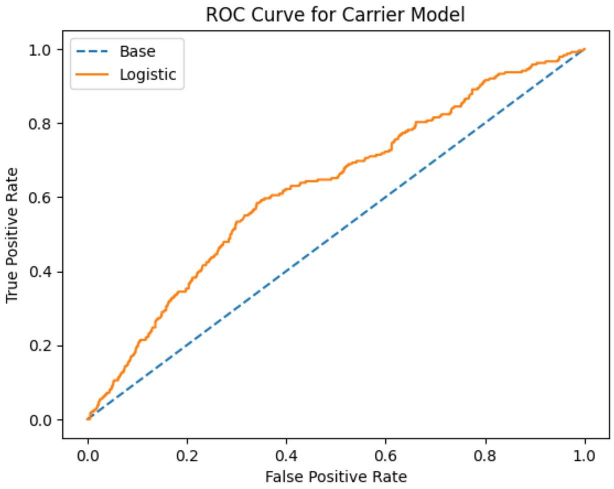




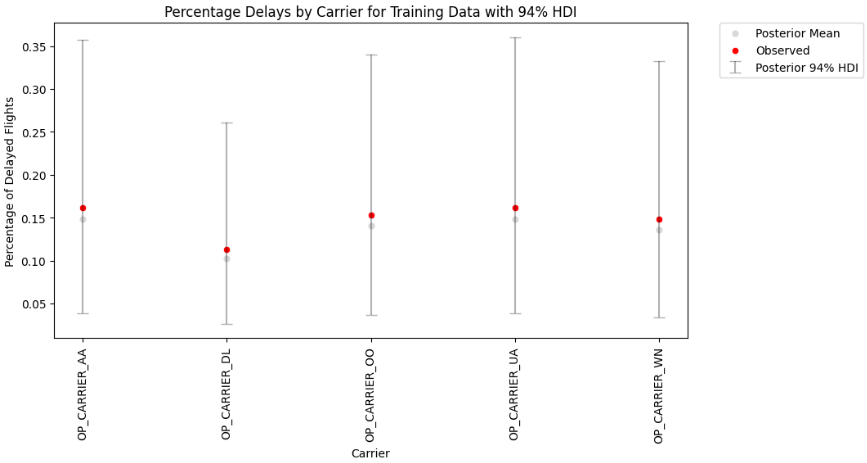


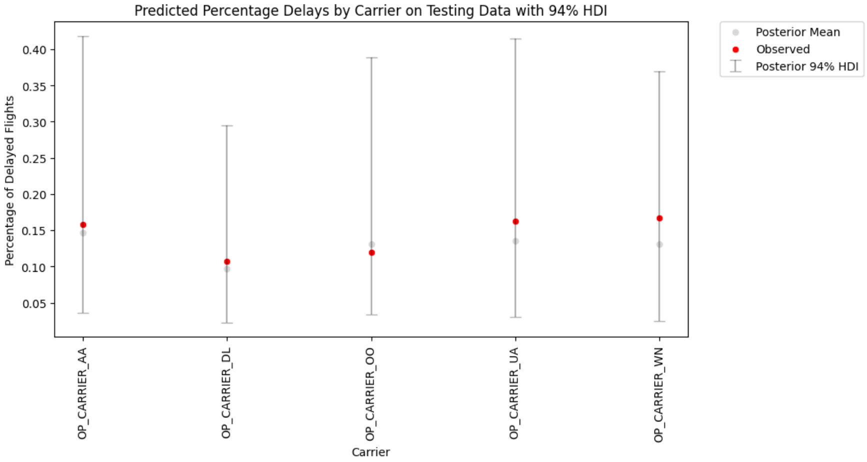
Appendix 6: Carrier Model Plots

Appendix 6a: Carrier ROC Curve

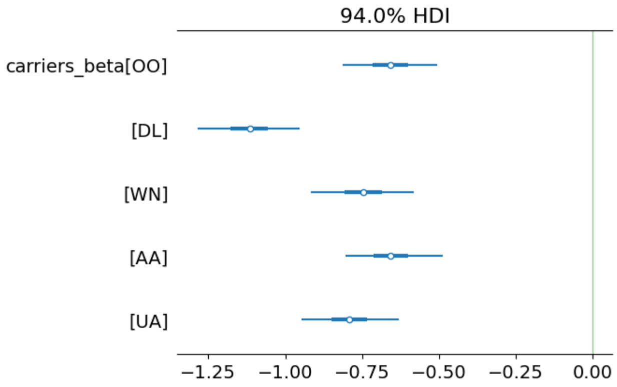


Appendix 6b: Carrier Categorical Fit and Prediction Plots

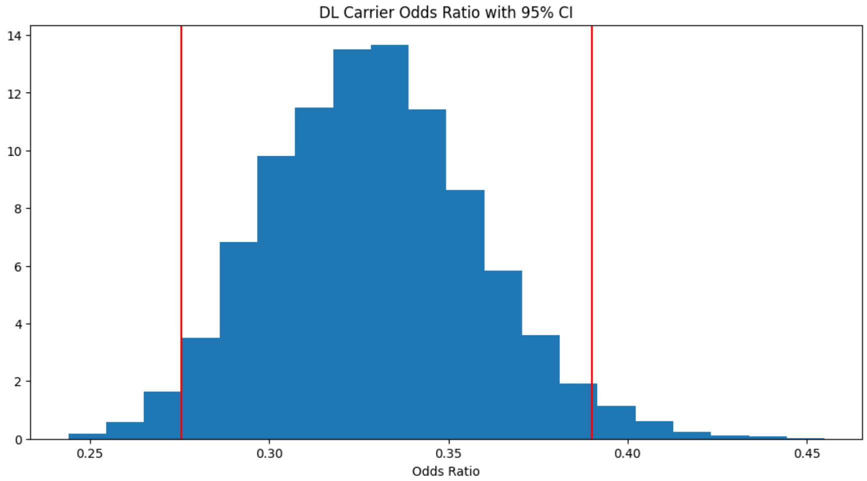


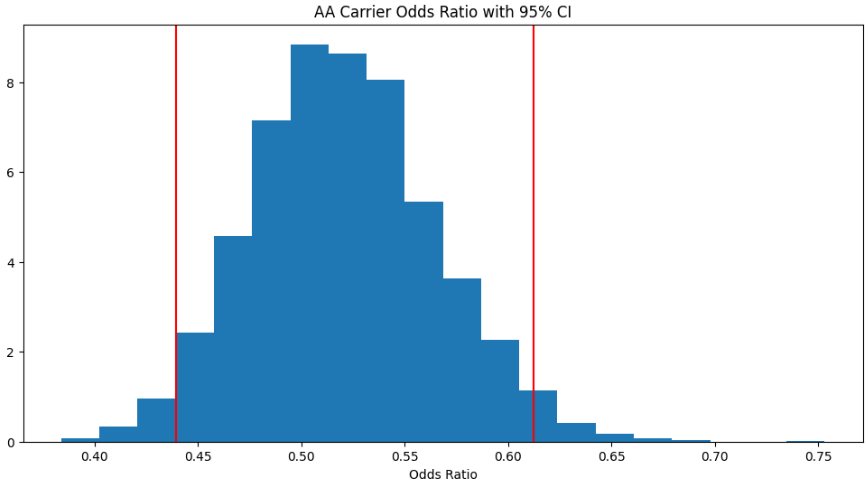


Appendix 6c: Carrier Forest Plot

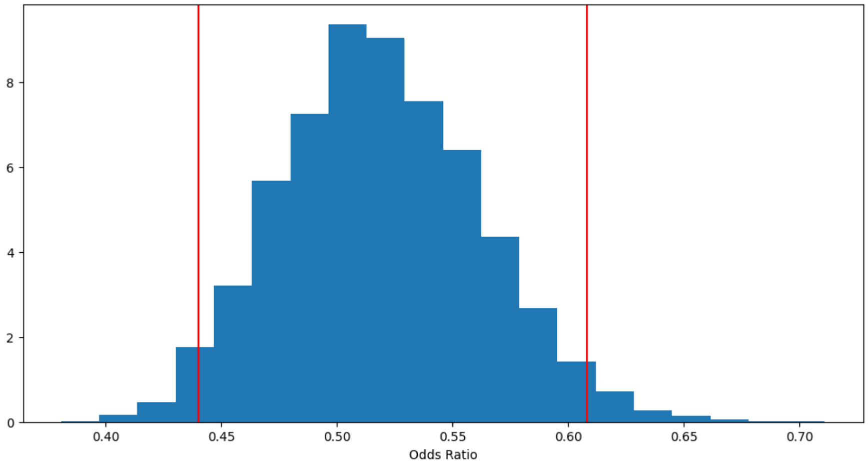


Appendix 6d: Carrier Odds Ratio Plots

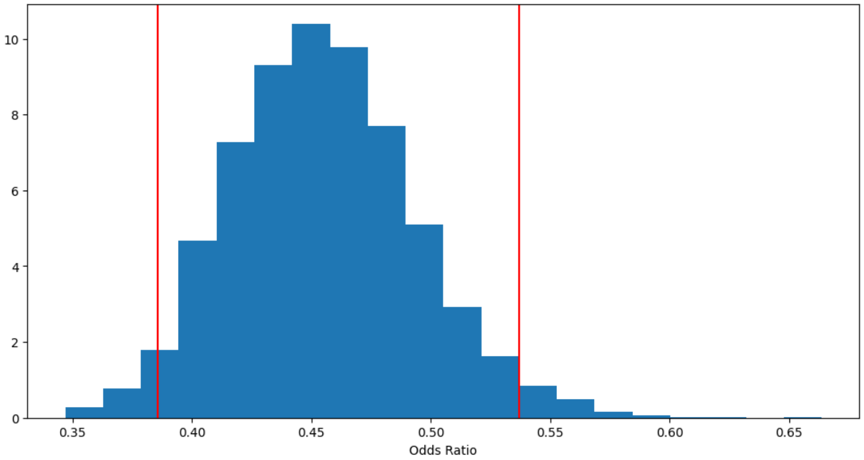




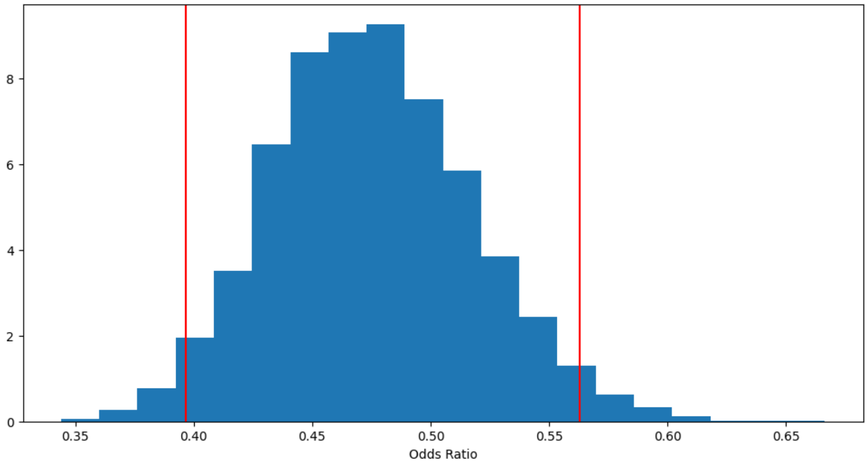
OO Carrier Odds Ratio with 95% CI



UA Carrier Odds Ratio with 95% CI



WN Carrier Odds Ratio with 95% CI


Odds Ratio