

TAMIDS 2020 Data Science Competition

Is My Flight Delayed?

Team: Pi-star Skyblazers

Sheelabhadra Dey, Sumedh Pendurkar

1. Introduction

1.1. Goals

- Predicting departure delays given the origin airport, destination airport, airline, date of travel, and time of departure.
- Analyzing the main factors contributing towards flight delays.
- Creating a web application in which a user can provide information regarding her planned travel and obtain an estimate of the flight delay.

1.2. Approaches

- We visualize the data, with various charts, graphs, to get an overall overview of the data.
- We find patterns from the observations and plots, and analyse the importance of each feature on the delays to greater depth.
- We use tree-based machine learning models such as random forests and gradient-boosted trees to predict flight delays.
- We analyze the importance of features to identify their amount of contribution to our predictions.
- We use partial dependence plots to analyze the impact of individual features, and feature interactions on delays.
- Using tree interpreters and waterfall charts we identify how different features impact our model's prediction of delays.

1.3. Benefits

- We provide data-driven predictions of delays that will help passengers schedule their travel better.
- We identify the key factors that contribute toward delays in flights.
- We facilitate a web application which provides an estimated delay for any given flight and date of travel which could be helpful in travel planning.

2. Data Summary Exploratory Data Analysis

2.1. Provided Data

The data-set containing flight logs has 46 variables as descriptors of the flight trips from January 2018 to June 2019. Table 2.1 contains a summary of the types of variables. Details regarding all the variables can be found in appendix A of this report.

Numeric Variables	28
Ordinal Variables	9
Nominal Variables	8
Time Variables	1

Table 2.1

Another data-set contained information regarding the airfares along the routes. In our analysis, we did not find the data in this table to be particularly useful in predicting flight delays. So we restrict our data analysis and modeling on the flight-log data.

2.2. Literature Review

We found a few interesting articles that earmarked factors that led to flight delays. A particular article, <https://www.claimcompass.eu/blog/why-is-my-flight-delayed/> listed the 20 major reasons for flight delays. Below are a few excerpts from the article.

“On the one hand, there are factors that are under the direct control of the carrier, such as aircraft turnarounds between flights, passenger punctuality, technical and crew performance, etc.”

“On the other hand, there are perhaps even more factors that are outside of the airline’s control, such as weather, air traffic control, security, airport conditions, etc.”

“The reality is such that so long as airplanes continue flying, flight delays will be a part of the experience. According to the Bureau of Statistics, about 20% of all flights are delayed by 15 minutes or more.”

The following are a few of the factors that the article listed as the major reason for flight delays:

- Air Traffic Control (ATC) restrictions
- Adverse weather conditions
- Knock-on effect due to a delayed aircraft
- Waiting for cargo
- Waiting for crew

The same article goes on to claim the following:

“Worse: the airline sometimes lie about the cause of the delay to avoid paying compensation.”

These reasons make the identification of the cause behind flight delays all the more important.

Another article, <https://www.businessinsider.com/why-your-flight-delayed-2016-12> derived helpful statistics from the flight log data released by the U.S. Bureau of Transportation Statistics (BTS) for trips between June 2015 and June 2016. Based on their analysis they made the following claims.

“According to the BTS, about 50% of all tardy flights between June of 2015 and June 2016 are attributed to circumstances within the airline's control such as aircraft maintenance, crew scheduling, refueling etc. This includes the nearly 210,000 flights there were delayed during this period due to late arriving inbound flights caused by airline-related issues.”

“At the same time, 30% of the delays can be blamed directly on weather. According to the BTS, this includes roughly half of the flights delayed due to the US aviation system, as well as 33% of delays due to late arriving inbound aircraft.”

“Finally, about 15% of the delays are caused by broad issues within the US aviation network. According to the BTS, this means problems with air traffic control, airport operations, and heavy traffic volume.”

That being said, they conclude their analysis on a positive note stating the following.

“As bad as delayed flights may be, they don't happen as often as one would expect once you consider America's aging aviation infrastructure coupled with surging demand for air travel and unpredictable weather. Even with these issues, 81.5% of flights did arrive on time. That's up from the 79.7% on-time arrival rate in 2003-2004 when the BTS first began publishing this data.

2.3. Data Exploration

There are a total of 17 airlines in the dataset. These 17 airlines run a total of 7148 flights along 6684 different routes across the U.S. 362 airports served in total by these flights.

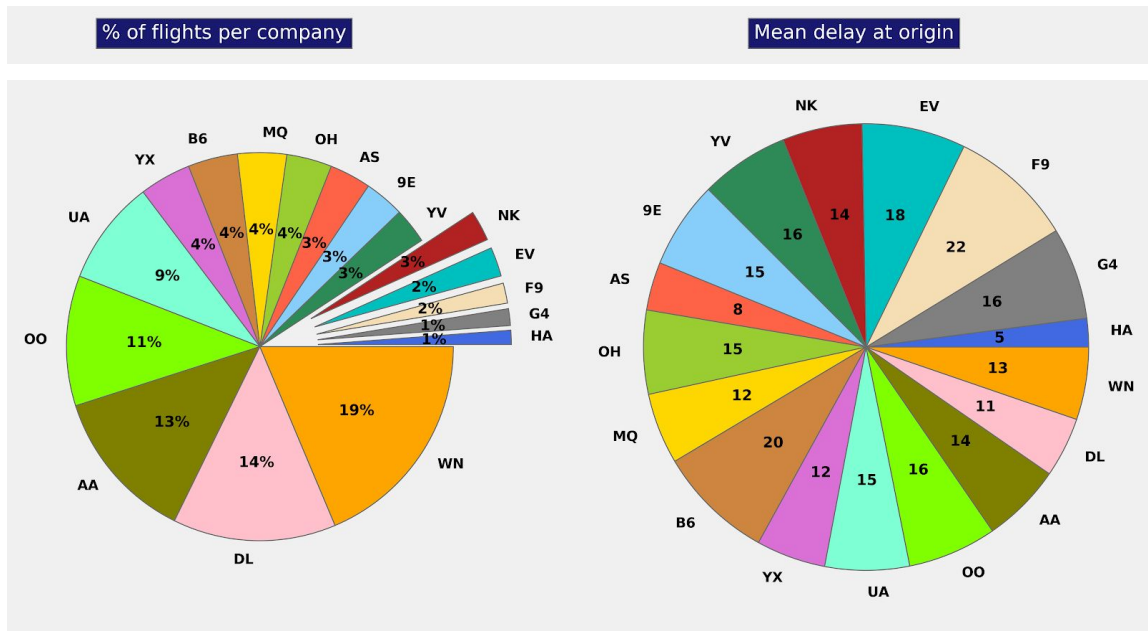


Fig 2.1 Pie Charts

The pie chart on the left in Fig 1.1 shows the market share of different airlines. It can be observed that there's some disparity between the percentage of flights per airline. Airline "WN" accounts for almost 20% of the flights which is similar to the percentage of flights run by the 7 tiniest airlines combined.

The pie chart on the right of Fig 1.1 shows the differences in mean delays between the flights is less pronounced. "HA" and "AS" have low (less than 10 min) mean delays. "F9" and "B6" show a high mean delay (greater than 20 min) even though they occupy a tiny portion of the total air traffic.

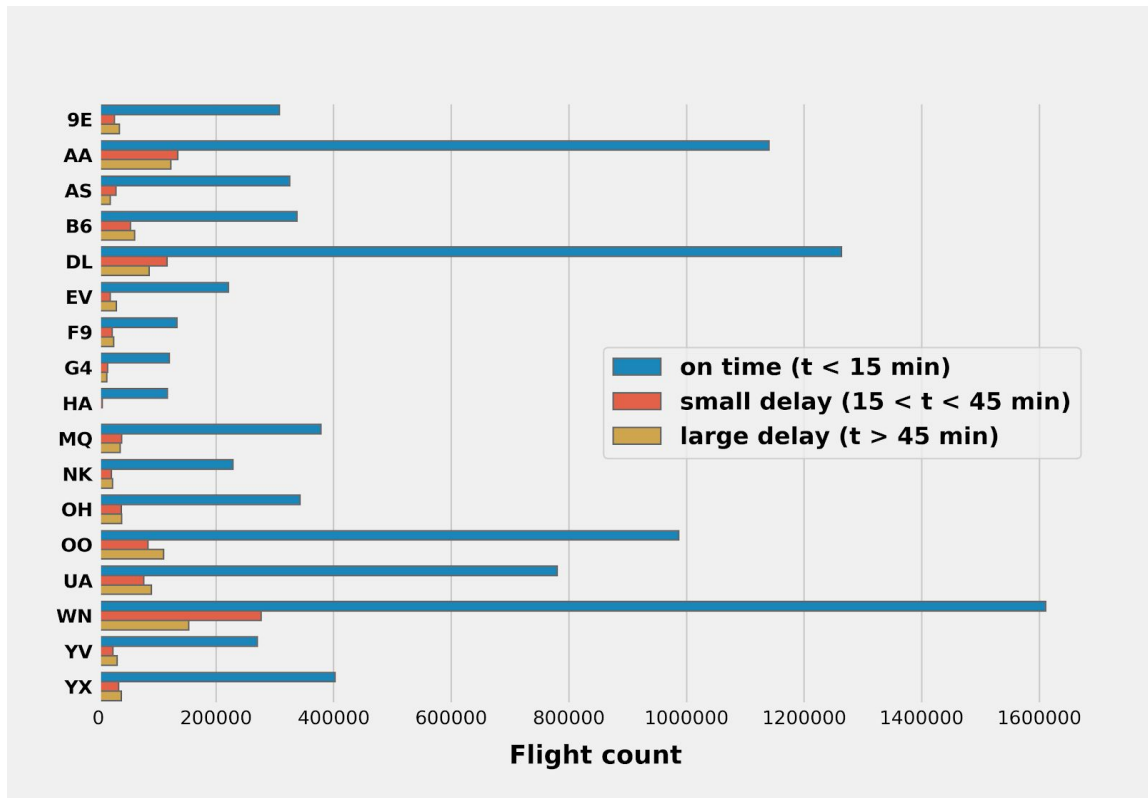


Fig 2.2 Total number of delays per carrier

From Fig 1.2 we can observe that Independent of the airline, large delays i.e. delays greater than 45 minutes constitute only a small portion of the total number of delays.

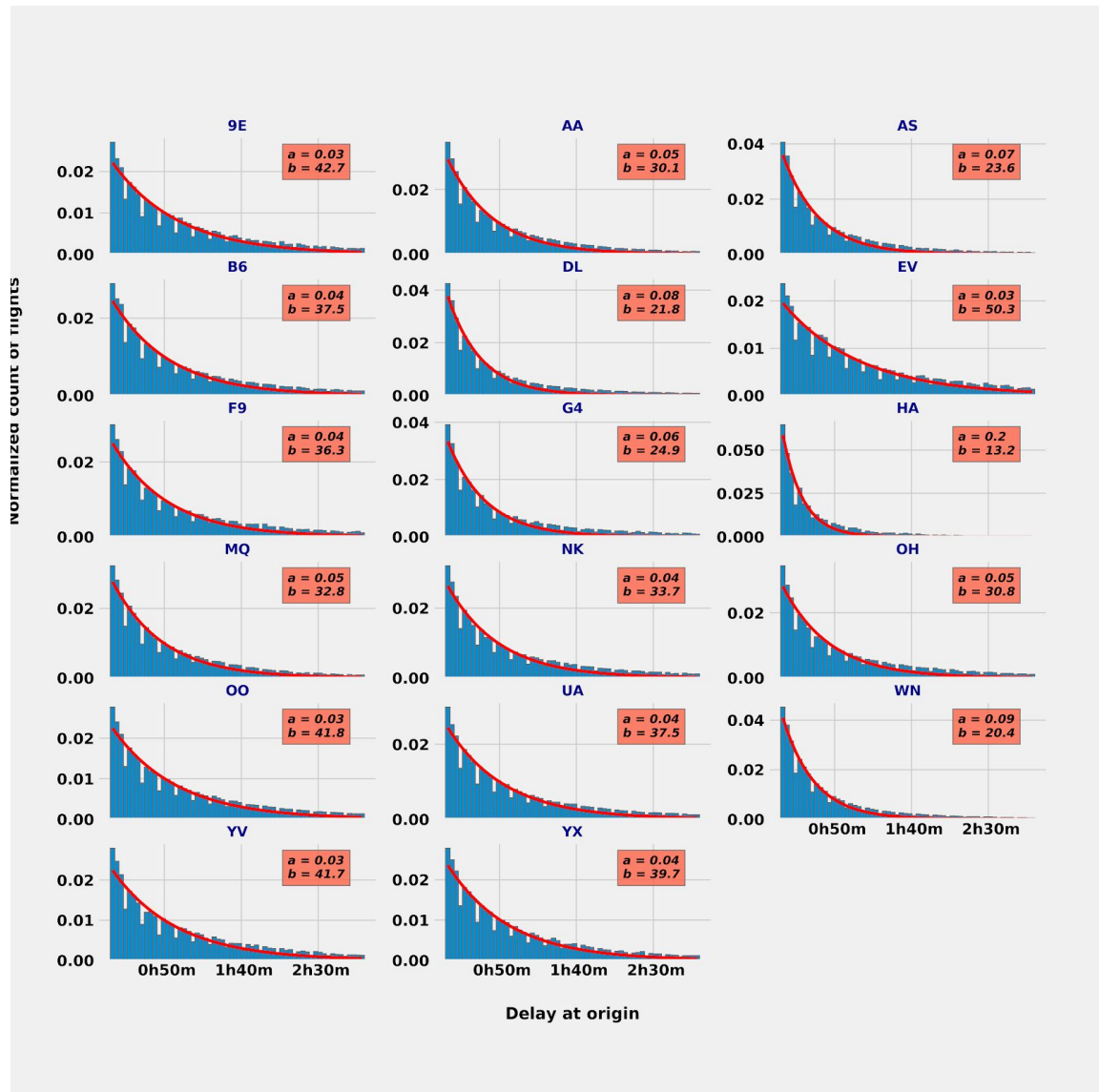


Fig 2.3 Fitting exponential distribution to delays

Next, we fit the delay distribution for each airline with exponential distributions as shown in Fig 1.3. This will allow us to get a rough idea regarding the punctuality of different airlines. Low values of “a” will correspond to airlines with a large proportion of delays. The plot below shows all the airlines ranked on the basis of their punctuality.

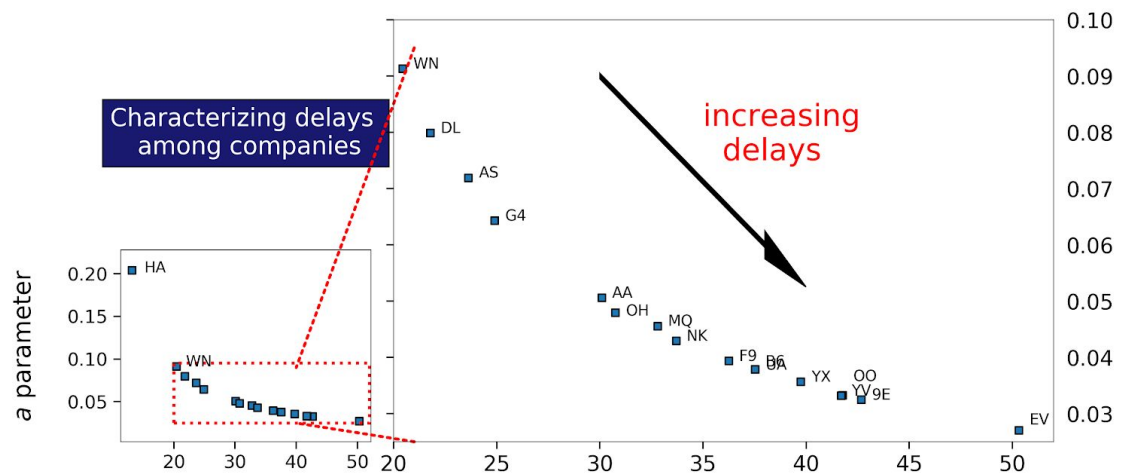


Fig 2.4 Characterizing delays between airlines

From the plot in Fig 1.4, we can observe that “EV” is the worst carrier since it has the worst “a” value and also from the pie chart we can see that it has a mean delay of 18 mins. “WN” is the carrier that has the maximum portion of flights (19%). At the same time it also sits 2nd in the delays plot which shows that it is also quite punctual. This could be the reason why it is the most widely used airline.

We also try to visualize the relationship origin airports and delays at departures. Fig 1.5 shows the number of airports covered by different airlines.

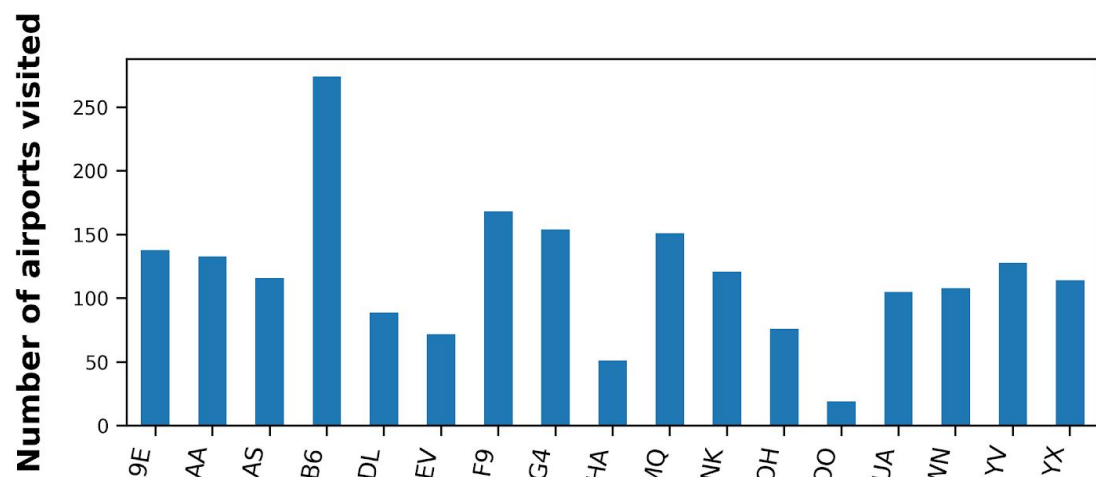


Fig 2.5 Number of airports per airline



Fig 2.6 Airline wise mean delays at origin airport

Fig 1.6 shows a subset of the average departure delay per airline at each airport. We can deduce from the infographic that there is high variability in the average delays, both between different airports and different airlines. So it will be necessary to adopt a model that is specific to the airline and the airport of origin.

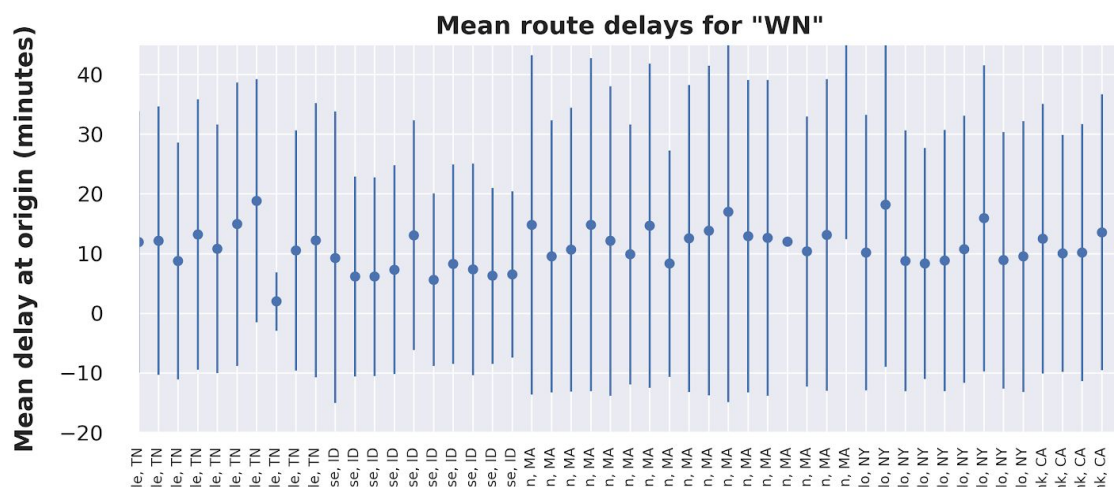
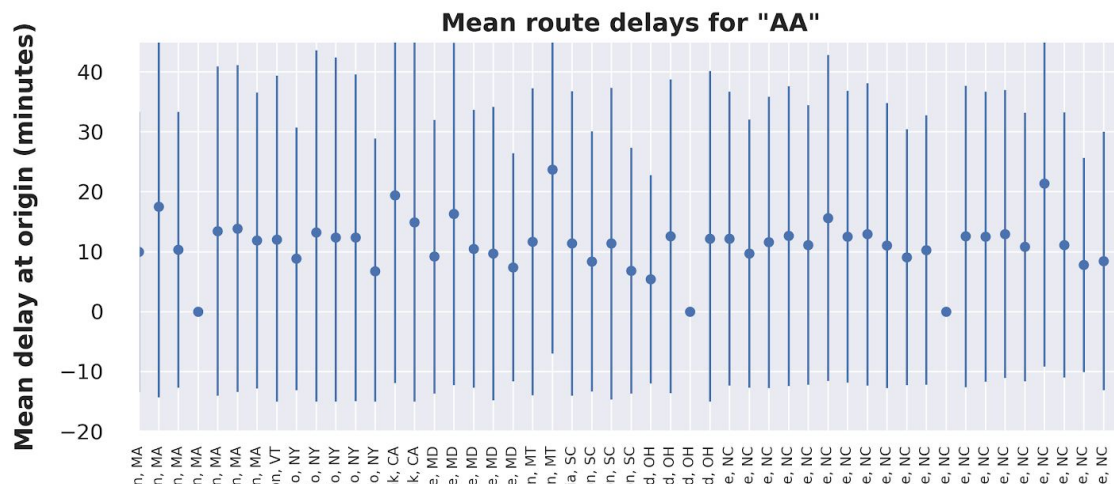


Fig 1.7 and Fig 1.8 contain the mean delays for airlines “AA” and “WN”, according to the city of origin and a destination city for a selected small portion from all the possible routes. Only the city of origin has been indicated in the plot. It can be observed that for a given airport of origin, delays fluctuate depending on the destination. Hence, in order to predict delays we will need to consider both the airline as well as the route in which it flies.

3. Delay Prediction

3.1. Feature Engineering

To take into account the historical traffic and delay information, we engineered the following features after careful thought.

- Hour extracted from the scheduled time of departure and arrival of the flight (CRS DEP TIME HR, CRS ARR TIME HR)

- Minute extracted from the scheduled time of departure and arrival of the flight (CRS_DEP_TIME_MIN, CRS_ARR_TIME_MIN)
- Number of flights along a route each hour (FLIGHTS_ROUTE)
- Number of passengers for each airline along a route in each hour (PASSENGERS_ROUTE_CARRIER)
- Mean delay to carrier-specific issues for each airline along a route in each hour (MEAN_CARRIER_DELAY)
- Mean delay due to weather for each airline along a route in each hour (MEAN_WEATHER_DELAY)
- Mean delay due to NAS (National Air System) for each airline along a route in each hour (MEAN_NAS_DELAY)
- Mean delay due to security checks for each airline along a route in each hour (MEAN_SECURITY_DELAY)
- Mean delay due to late arrival of aircraft for each airline along a route in each hour (MEAN_LATE_AIRCRAFT_DELAY)

These features were used along with the existing features in the dataset as input to our model. The nice thing about these features is that they can be updated as time passes making the model more and more robust and accurate.

3.2. Will there be a delay?

The simplest possible model that we decided to come up with was one that predicts whether there would be a delay for a particular flight on a particular date in the future. We used the flight log dataset to create our model. The dataset contained missing values for a number of fields which prompted us to use tree-based modeling approaches since they can handle missing values by default. Hence we decided to use a random forest for this task since it is a fairly powerful model for tabular data and has nice properties that could be exploited for interpretation of its predictions.

We obtained an accuracy close to 81% on our held out validation set which contained 100,000 trips. Our trained model predicts whether a particular flight leaves its origin airport before time, on time or is delayed. Fig 3.1 shows the confusion matrix for our model's prediction on our validation set. A complete description of our model and the associated hyper-parameters can be found in appendix B of this report.

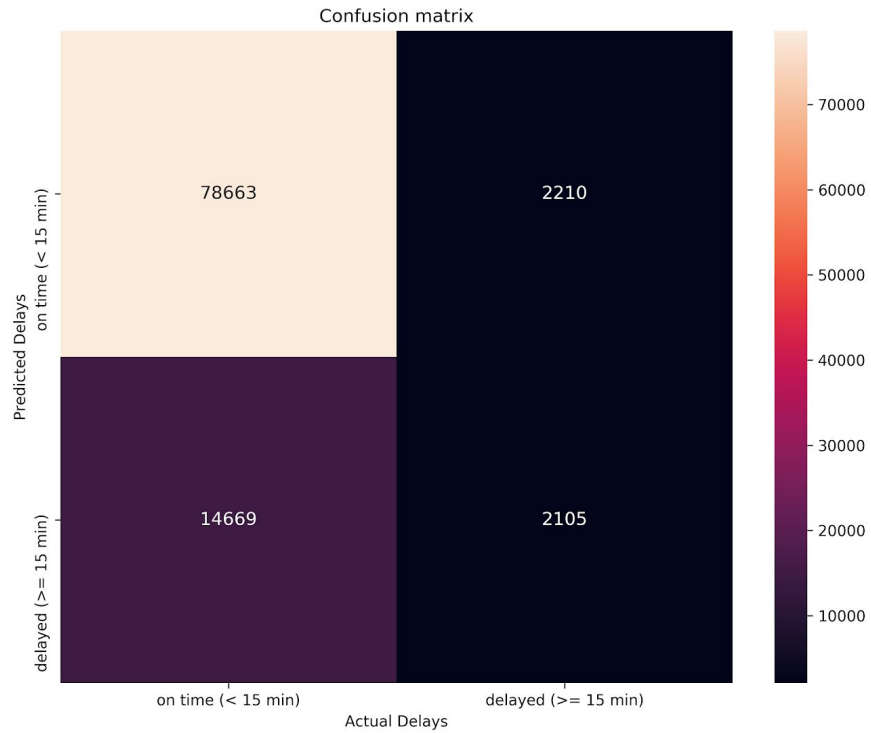


Fig 3.1 Confusion Matrix

Figure 3.2 contains the most important features found by our model. The vertical axis contains the IDs for the features and the horizontal axis provides a qualitative measure of a feature's importance towards our model's prediction.

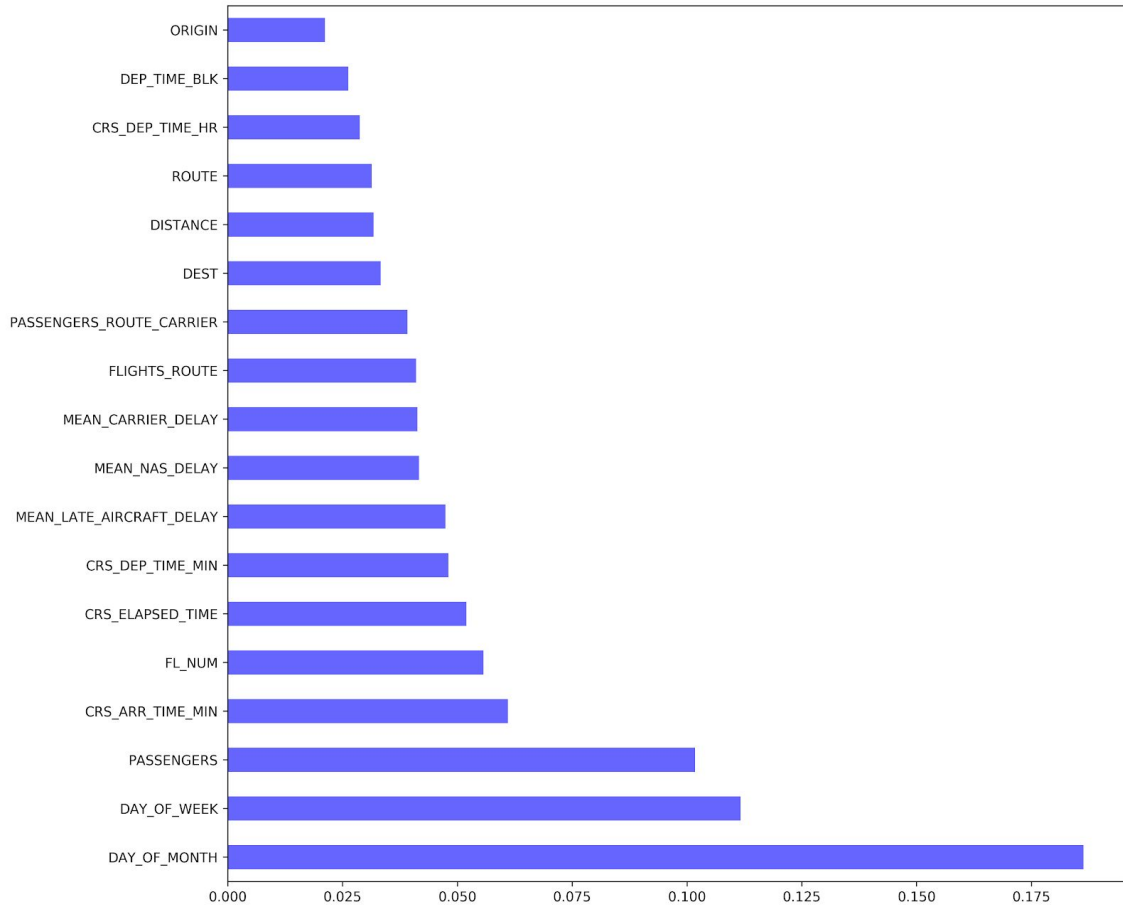


Fig 3.2 Feature Importance

The importance plot in Fig 3.2 suggests that the day of the month and day of the week in which a flight departs, the number of passengers in the flight, and the flight number are the most important factors contributing towards flight delays. This somewhat aligns with our analysis from the data in the previous section since flight number implicitly contains information regarding the airline and the route along which it operates. A more thorough analysis of these features and how they affect the delay prediction is presented in a later section. For a detailed description regarding each of the features please refer to appendix A of this report.

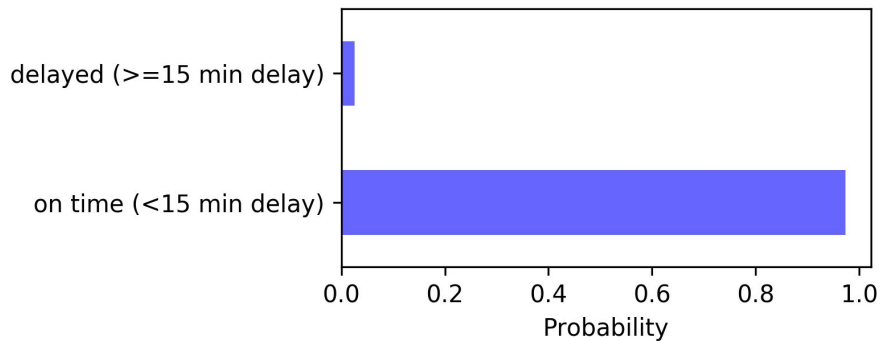


Fig 3.3 Prediction Probability Distribution

As shown in Fig 3.3, for any given input data sample, our model provides information regarding how confident it is about its prediction. We feel that it provides a user a better understanding of the other possible events that can occur and plan for her trip much better.

3.3. How much delay can be expected?

Initially, we tried to build a random forest-based regression model using the provided features in the dataset and features that we engineered. The high variability in delays resulted in a model that did not perform well on the held out validation set. We obtained a root mean squared error close to 11 minutes on our validation set.

These hurdles egged us to rethink our modeling strategy. What we felt is that typically passengers don't really need an exact number for the delay they can expect when planning to take a particular flight. A delay within a range is also completely acceptable. Additionally, we decided to output the confidence of our model's prediction due to the same reasons we mentioned in the previous section.

We obtained an accuracy close to 62% on our held out validation set which contained about 100,000 trips. Our trained model predicts an interval in which the delay for a particular flight will fall. Specifically, we provide a distribution over the delay intervals. The interval of the delays is 15 minutes and the minimum delay predicted is -30 minutes i.e the flight leaves before time while delays greater than 80 minutes have been grouped into a single interval block. Shown in fig 3.4 is the confusion matrix for our model's prediction on our validation set. A complete description of our model and the associated hyper-parameters can be found in the appendix B of this report.

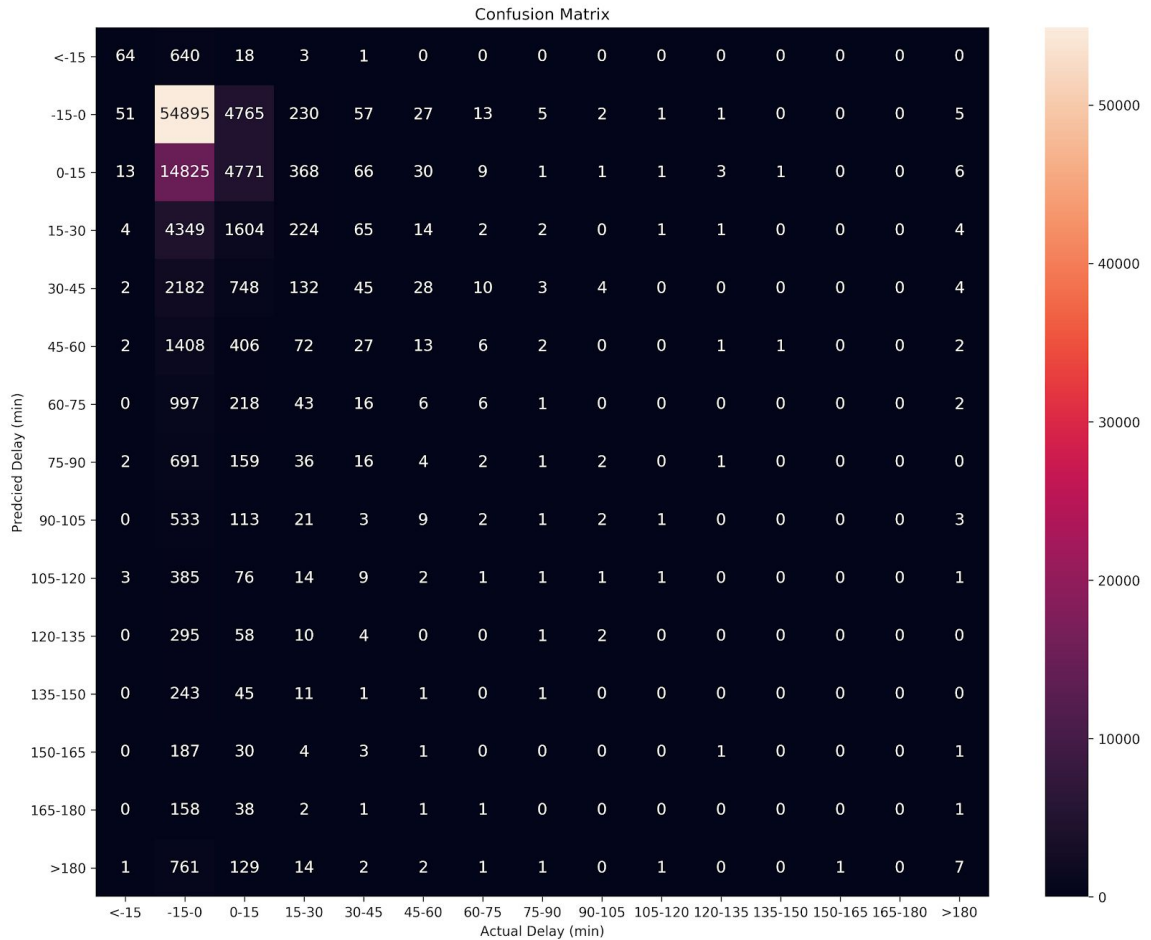


Fig 3.4 Confusion Matrix

Followed by the confusion matrix is the feature importance plot in Fig 3.5 which quantifies the effect of each feature in our model on its predictions. The top contributing features are similar to what we obtained for the binary delay prediction model. But, an interesting feature here is the net income for the current year, month, route and carrier. This particularly emphasizes the significance of machine learning models to find patterns in the data as it seems a difficult relationship to capture using just visualizations and heuristics.

Fig 3.6 is the plot for the prediction probability distribution for a given input data sample from our validation set.

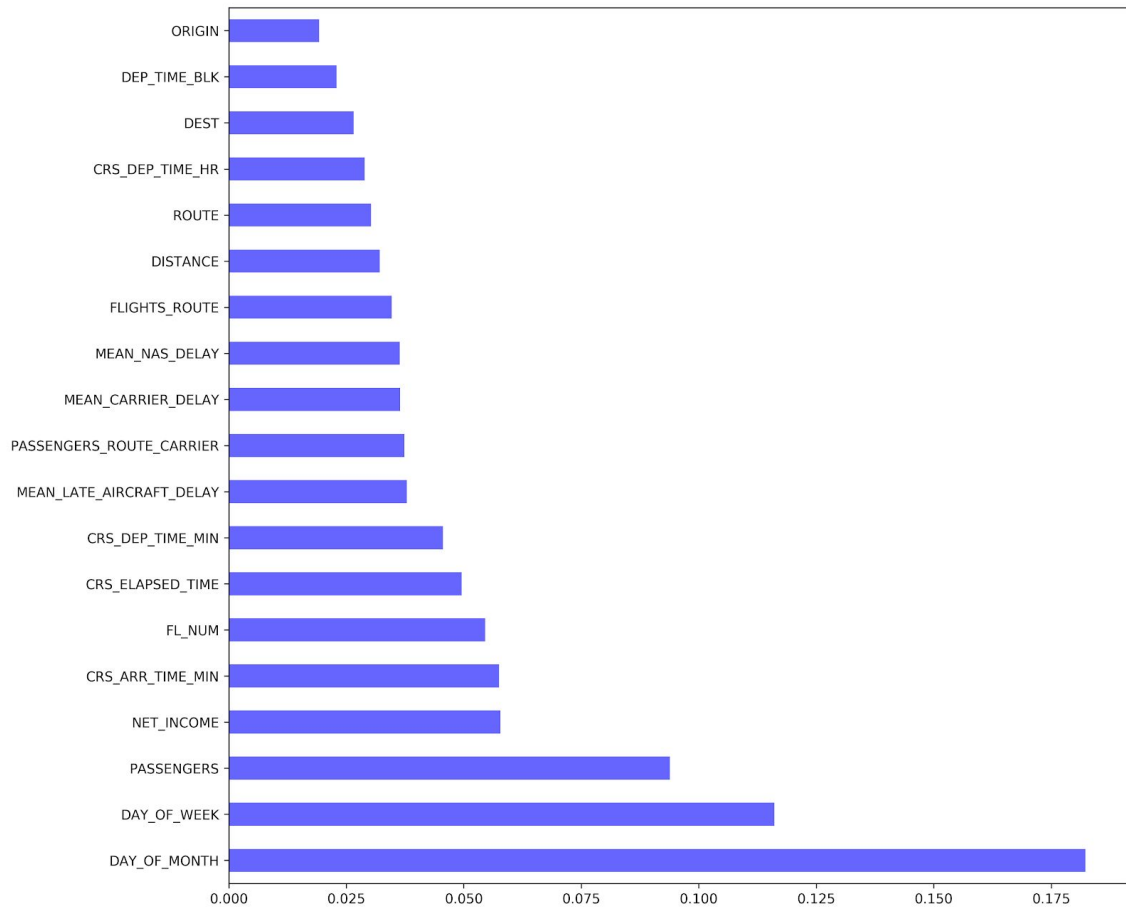


Fig 3.5 Feature Importance

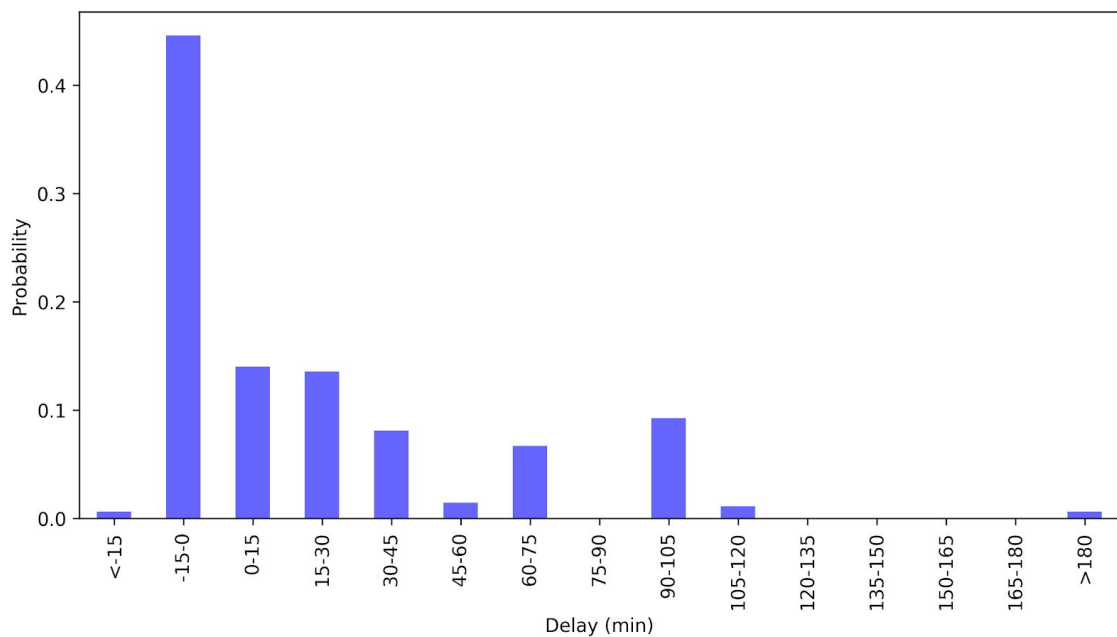


Fig 3.6 Prediction Probability Distribution

3.4. Tree Interpreter

To estimate the amount by which the independent variables impact the prediction of our model, we take the help of a tree interpreter. It is helpful in calculating the individual contribution of each of the independent variables towards the mean prediction over the entire dataset. The combination or sum of these effects together results in the final prediction of the model.

For illustration, we choose a sample trip from our validation set and analyze how the values of the independent variables affect the final probability distribution of the model's prediction. The bias of model i.e. the mean prediction of our model is the plot below.

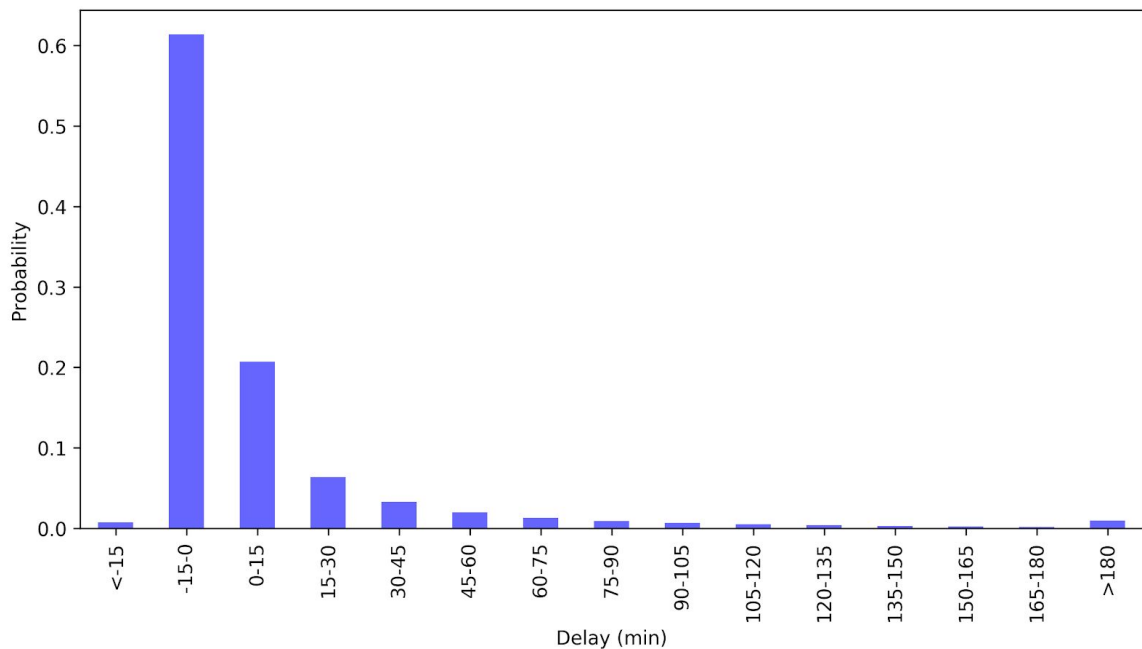


Fig 3.7 Average Prediction of the model

Table 3.1 below shows the contribution of each of the features of the sample trip from our validation set on the final probability value in the (-15 - 0) minutes interval. The mean probability prediction of the model for the interval (-15 - 0) minutes is 0.61343 while the probability prediction of the model on the sample trip for the same interval is 0.44575. If we add up the contributions on the right-most column, the result is -0.16768 which is exactly also the result of the difference between 0.44575 and 0.61343. It is also interesting to observe that the flight number being “5237” most negatively impacts the probability i.e. contributes towards delays while the net income being “41490” most positively impacts the probability i.e. helps it avoid delays and stay on time.

Feature	Value	Contribution: (-15 - 0) min interval
DAY_OF_MONTH	13	-0.00796
DAY_OF_WEEK	2	0.02477
FL_NUM	5237	-0.09448
ROUTE	1678	0.01894
ORIGIN	DEN	0.00909
DEST	SLC	-0.01273
CRS_DEP_TIME_HR	19	-0.02446
CRS_ELAPSED_TIME	99	-0.03003
CRS_DEP_TIME_MIN	5	-0.02876
CRS_ARR_TIME_MIN	44	-0.01387
DEP_TIME_BLK	1900-1959	-0.01975
DISTANCE	391	-0.00244
PASSENGERS	6620	-0.0023
NET_INCOME	41490	0.08921
FLIGHTS_ROUTE	115	0.00005
PASSENGERS_ROUTE_CARRIER	622460	-0.01842
MEAN_CARRIER_DELAY	58.93	-0.02355
MEAN_NAS_DELAY	11.4	0.0021
MEAN_LATE_AIRCRAFT_DELAY	27.47	-0.03309

Table 3.1 Sample observation from the validation set

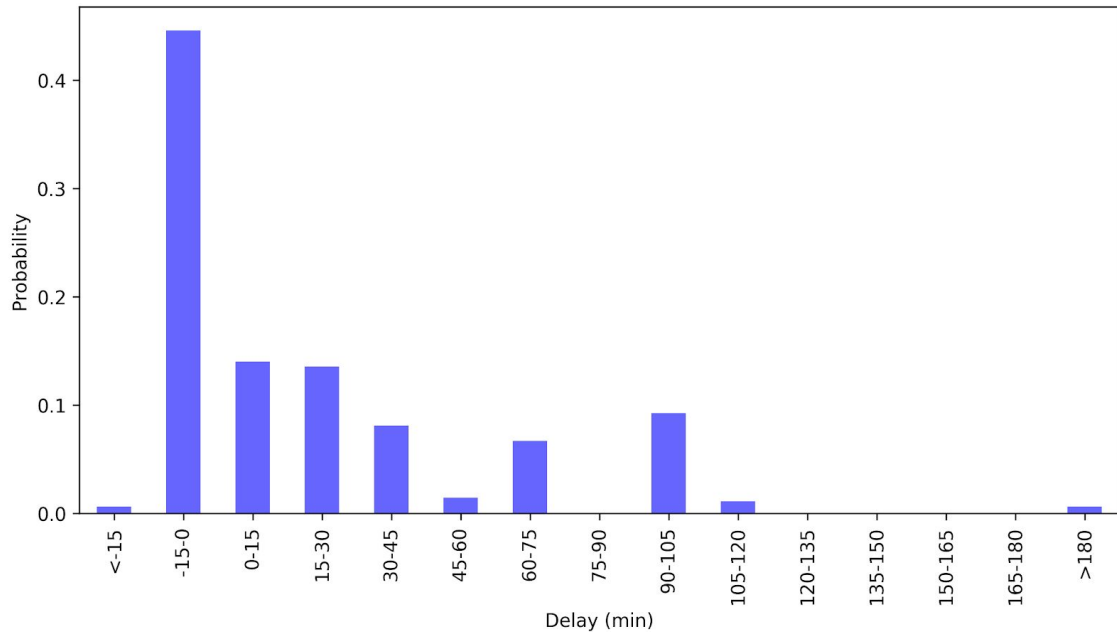


Fig 3.8 Final prediction of the model

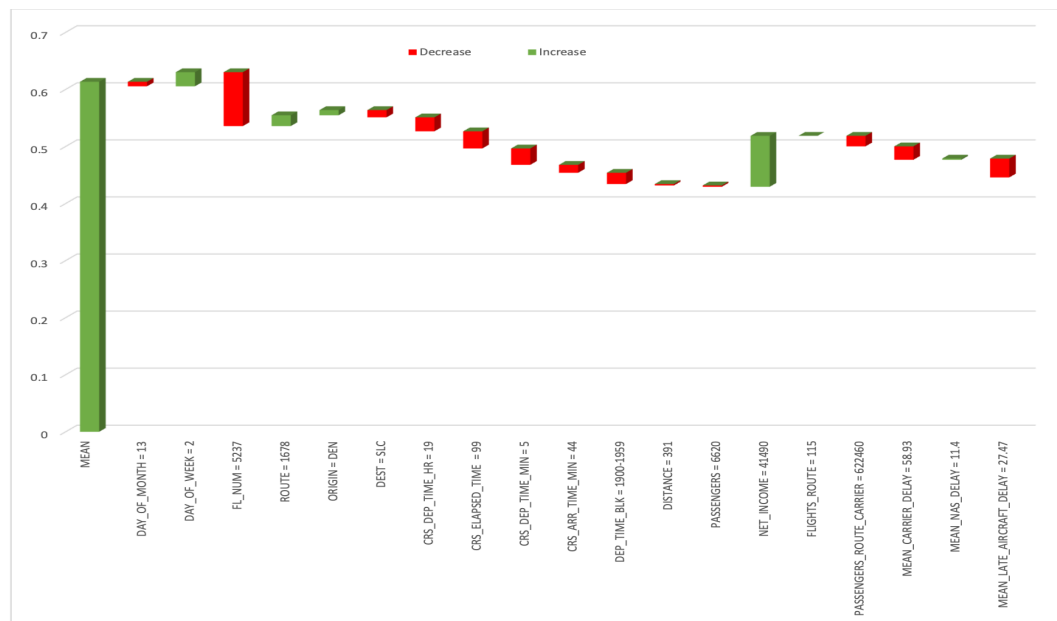
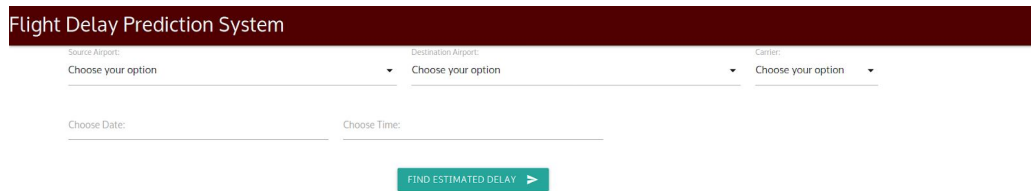


Fig 3.9 Waterfall chart depicting the impact of features on probability

4. A web application to estimate delay

We create a web application which would allow users to predict the delay based on their flight details, and time of flights. Our application is specifically intended for flyers to get an estimate on the delay which might help to manage their itineraries accordingly. Here are some of the screenshots of the application that we have deployed based on the methodologies discussed above.

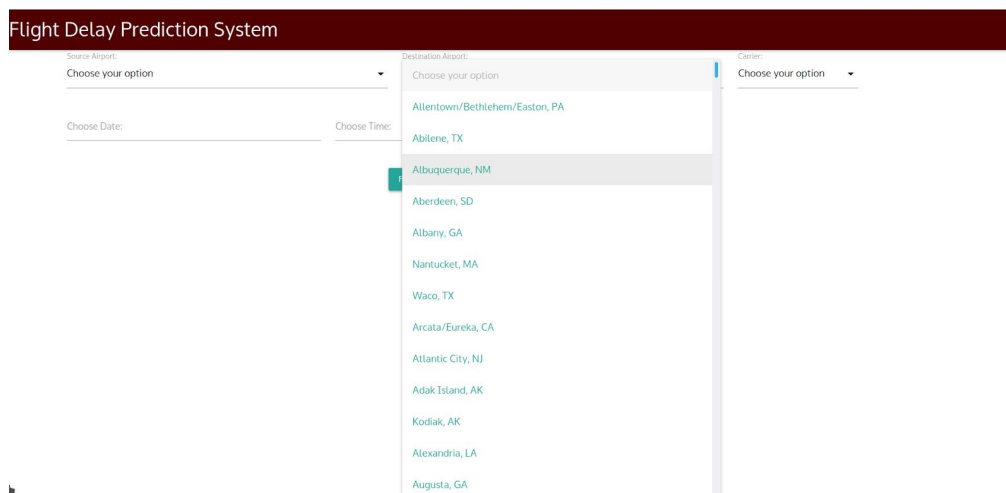
The image below shows an overall look of our product where we take the inputs.



The image shows the overall UI of the Flight Delay Prediction System. It features a dark red header with the title "Flight Delay Prediction System". Below the header, there are three dropdown menus for "Source Airport:", "Destination Airport:", and "Carrier:". Each dropdown menu has a placeholder text "Choose your option". Below these dropdowns, there are two input fields for "Choose Date:" and "Choose Time:". At the bottom, there is a green button labeled "FIND ESTIMATED DELAY" with a right-pointing arrow.

Fig 4.1 Overall UI

The drop down allows users to select their destination for the airports that are known to us based on the dataset provided.



The image shows a close-up of the "Destination Airport:" dropdown menu. The dropdown is open, displaying a list of airport names. The list includes: Allentown/Bethlehem/Easton, PA; Abilene, TX; Albuquerque, NM; Aberdeen, SD; Albany, GA; Nantucket, MA; Waco, TX; Arcata/Eureka, CA; Atlantic City, NJ; Adak Island, AK; Kodiak, AK; Alexandria, LA; and Augusta, GA. The "Albuquerque, NM" option is highlighted with a green background.

Fig 4.2 Drop Down for Airports

Based on these inputs, we select the carrier from a similar drop down menu. We also Pick up dates and time using a date time picker UI to ensure validation of the timings provided. Along with that, it also helps the user to pick the dates very quickly due to a very simple to use UI. Below are a couple of screenshots from the UI.

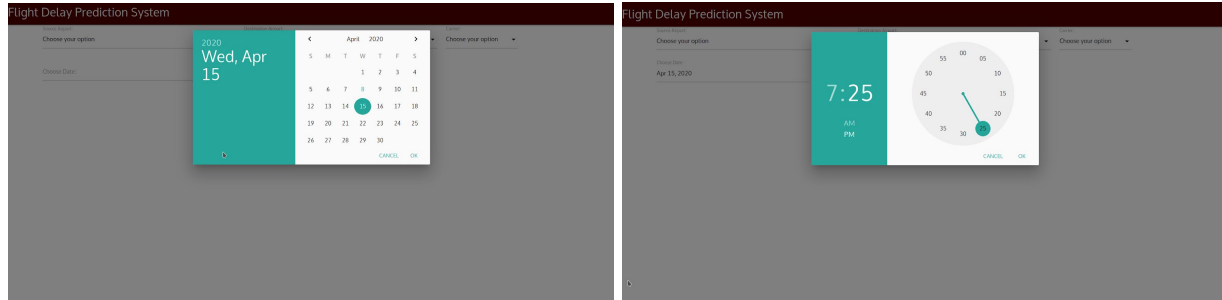


Fig 4.3 DateTime selection

After you fill the information about your source destination, carrier and time, we find the flight that is travelling in this time, and submit our query, we find information about the specific flight and delays associated with it and it then shows a pop up message showing the status of the flight.

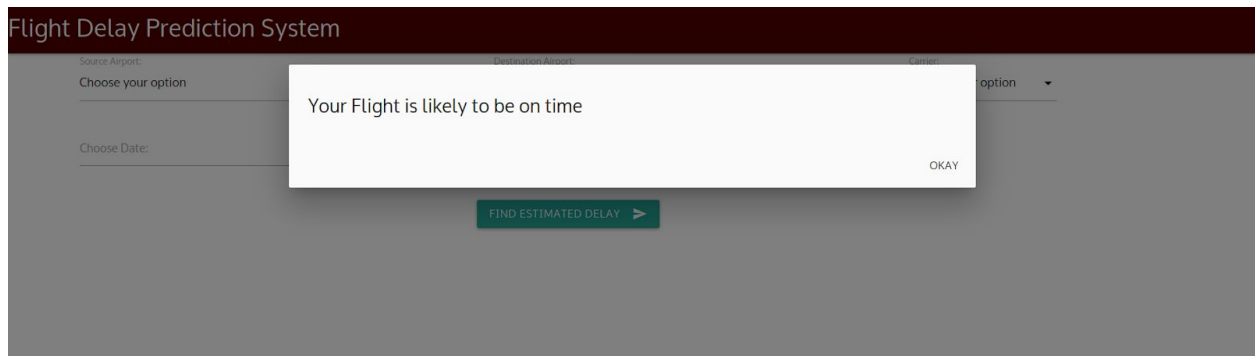


Fig 4.4 Result of the query.

We believe this application would help users to have a rough idea about delays before head based on airline/airport characteristics as well condition of the carrier, and passengers of flight. The link for the application is given [here](#).

5. Conclusion

In conclusion we found the following insights:

- The departure delay is correlated to the airline, day of week of travel, day of month of travel, and number of passengers in the flight.
- Important features that could be derived from the dataset are the historical delay specifically the NAS delay and late aircraft delay and traffic information along a route
- The application that we have built would help flyers to estimate predictions of delay based on the history of their route, and carrier

Appendix

A. Feature Description

Given below is the description of the features in the flight log dataset.

ATTRIBUTE	TYPE	DESCRIPTION
YEAR	Binary	Year 2018 or 2019
FL_DATE	yyyymmdd	Flight Date as yyyymmdd
CARRIER	Nominal	Airline Identifier 2-3 characters
FL_NUM	Nominal	Flight Number
ROUTE	Nominal	Route number assigned alphabetically by origin and dest
ORIGIN	Nominal	Airport Identifier for Flight Origin
DEST	Nominal	Airport Identifier for Flight Destination
DEST_CITY	Nominal	Destination City Name and State
DEST_STATE	Nominal	Destination State Abbreviation
CRS_DEP_TIME	Interval	CRS Departure Time (hhmm)
DEP_TIME	Interval	Actual Departure Time (hhmm)
DEP_DELAY	Interval	Difference in minutes between scheduled and actual departure times
DEP_DELAY_NEW	Interval	Same as DEP_DELAY except early arrivals are set to zero
DEP_DEL15	Binary	Departure Delay Indicator: 0=less than 15 min, 1=more than 15 min
DEP_DELAY_GROUP	Ordinal	Departure Delay Intervals, every 15 min from -15 to 180 min
DEP_TIME_BLK	Ordinal	CRS Departure Time Block Hourly Intervals
TAXI_OUT	Interval	Taxi Out Time
WHEELS_OFF	Interval	Wheels Off Time (hhmm)
WHEELS_ON	Interval	Wheels On Time (hhmm)
TAXI_IN	Interval	Taxi In Time in Minutes
CRS_ARR_TIME	Interval	CRS Arrival Time
ARR_TIME	Interval	Actual Arrival Time
ARR_DELAY	Interval	Difference in minutes between scheduled and actual arrival time.
ARR_DELAY_NEW	Interval	Difference in minutes between scheduled and actual arrival time, early arrivals set to zero.
ARR_DEL15	Binary	Arrival Delay Indicator: 0=less than 15 min, 1=more than 15 min
ARR_DELAY_GROUP	Ordinal	Arrival Delay Intervals, every 15 min from -15 to 180
ARR_TIME_BLK	Ordinal	CRS Arrival Time Block, hourly intervals
CANCELED	Binary	Flight Cancelled Indicator: 0=not cancelled, 1=cancelled flight
CANCELLATION_CODE	Nominal	A, B, C or D
DIVERTED	Binary	0=Was not diverted, 1=Was diverted
CRS_ELAPSED_TIME	Interval	CRS Elapsed Time of Flight in Minutes
ACTUAL_ELAPSED_TIME	Interval	Elapsed Time of Flight in Minutes
AIR_TIME	Interval	Elapsed Time of Flight in Minutes
DISTANCE	Interval	Distance between Airports
CARRIER_DELAY	Interval	Carrier Delay in Minutes
WEATHER_DELAY	Interval	Weather Delay in Minutes
NAS_DELAY	Interval	National Air System Delay in Minutes
SECURITY_DELAY	Interval	Security Delay in Minutes
LATE_AIRCRAFT_DELAY	Interval	Late Aircraft Delay in Minutes
PASSENGERS	Interval	Total number of passengers for this year, month, route and carrier
EMPFULL	Interval	Full time employees for this year, month, route and carrier
EMPPART	Interval	Part time employees for this year, month, route and carrier
EMPTOTAL	Interval	Total employees for this year, month, route and carrier
EMPfte	Interval	Full time equivalent employees for this year, month, route and carrier
NET_INCOME	Interval	Net income for this year, month, route and carrier
OP_REVENUES	Interval	Operating revenues for this year, month, route and carrier

Table 5.1 Feature description

B. Model specifications

The binary classifier is a random forest with 40 estimators, has a minimum of 3 samples in the leaf nodes, and considers only 50% of all the features at each split of the trees. This reduces correlation between the individual trees helping the model capture more variance in the data.

The multi-class classifier is also a random forest with 40 estimators, has a minimum of 3 samples in the leaf nodes, and considers only 50% of all the features at each split of the trees.

C. Partial Dependence Plots

In this section, we analyze why simple univariate plots can sometimes be misleading and provide a principled way to analyze relationships between the independent and dependent variables.

The delay in the provided dataset is related to the day of the month, day of the week, the flight number, number of passengers, etc. Here, we try to analyze the impact of the day of the month and the day of the week on the delay.

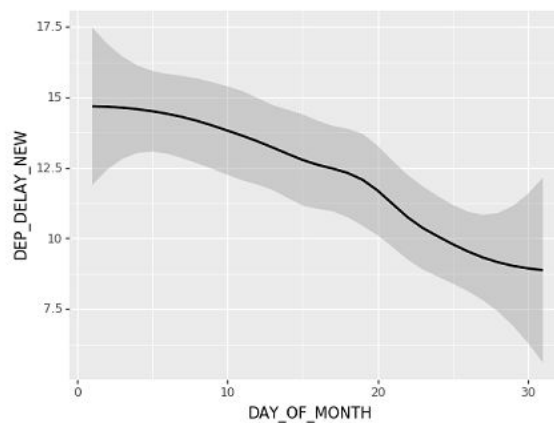


Fig 5.1 Departure delay vs day of month

Analyzing the mean departure delay with respect to the day of the month on a small subset of the dataset suggests that the mean departure delay decreases. In the following plot, visualizing the mean departure delay with respect to the day of the week suggests that the mean departure delay decreases mid-week.

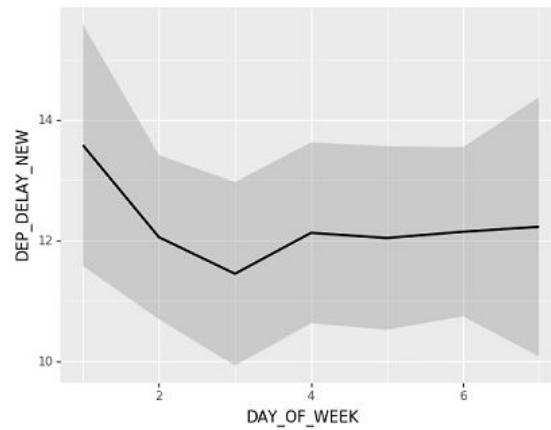


Fig 5.2 Departure delay vs day of week

This type of analysis sometimes can be detrimental to the performance of the model since we are unknowingly ignoring the effect of interaction between the independent variables.

To address this issue, we took a look at their partial dependence plots in Fig 5.3 and Fig 5.4. The plots suggest that there are other factors affecting the delays as there is no trend in general seen across a few of the interval categories.

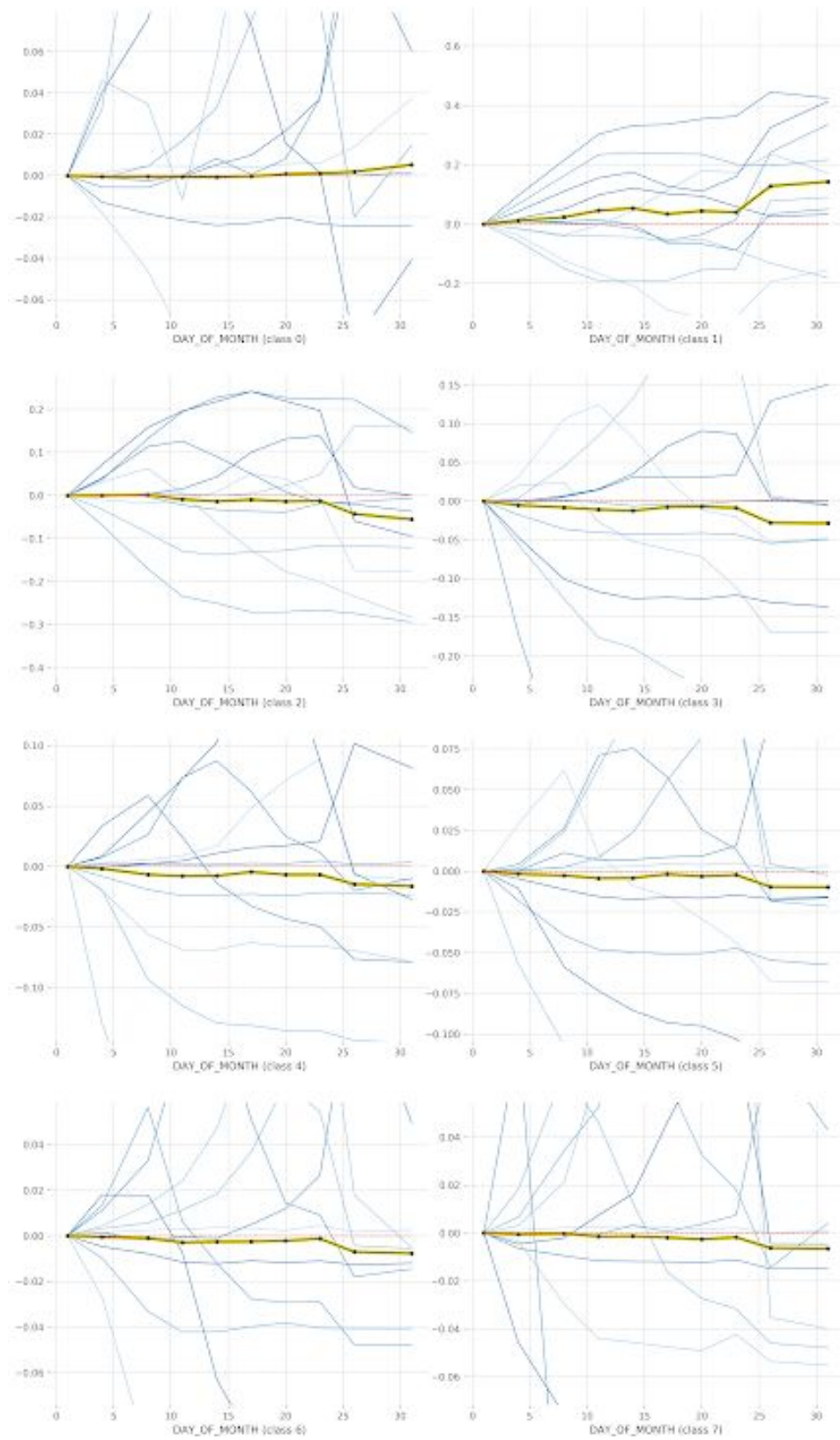


Fig 5.3 Partial Dependence plot for mean departure delay vs day of month

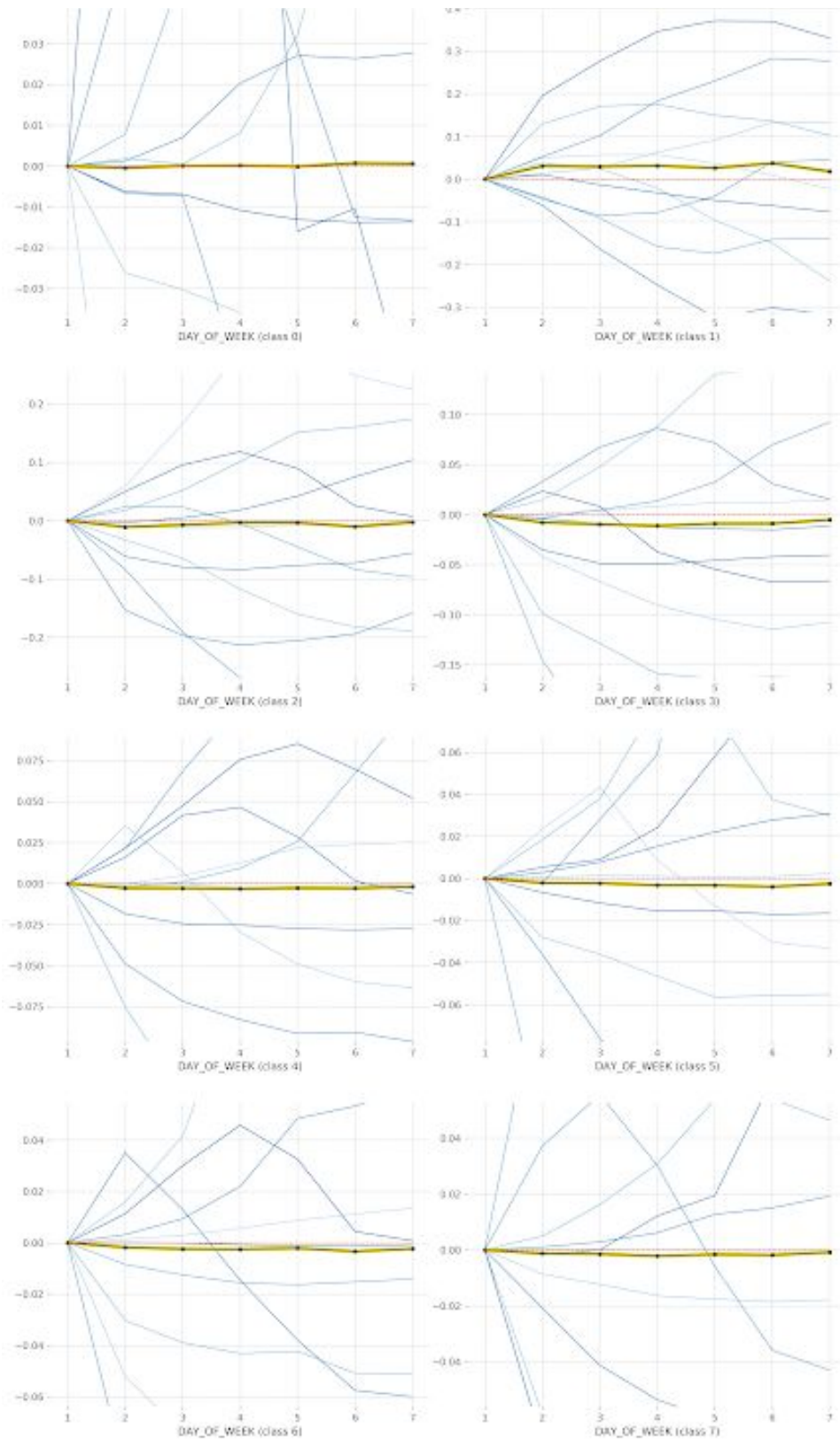


Fig 5.3 Partial Dependence plot for mean departure delay vs day of week