ECO395M Final Project: Impact of Covid-19 on the flight delays and cancellation in California and Texas

Steven Kim and Shreekara Shastry

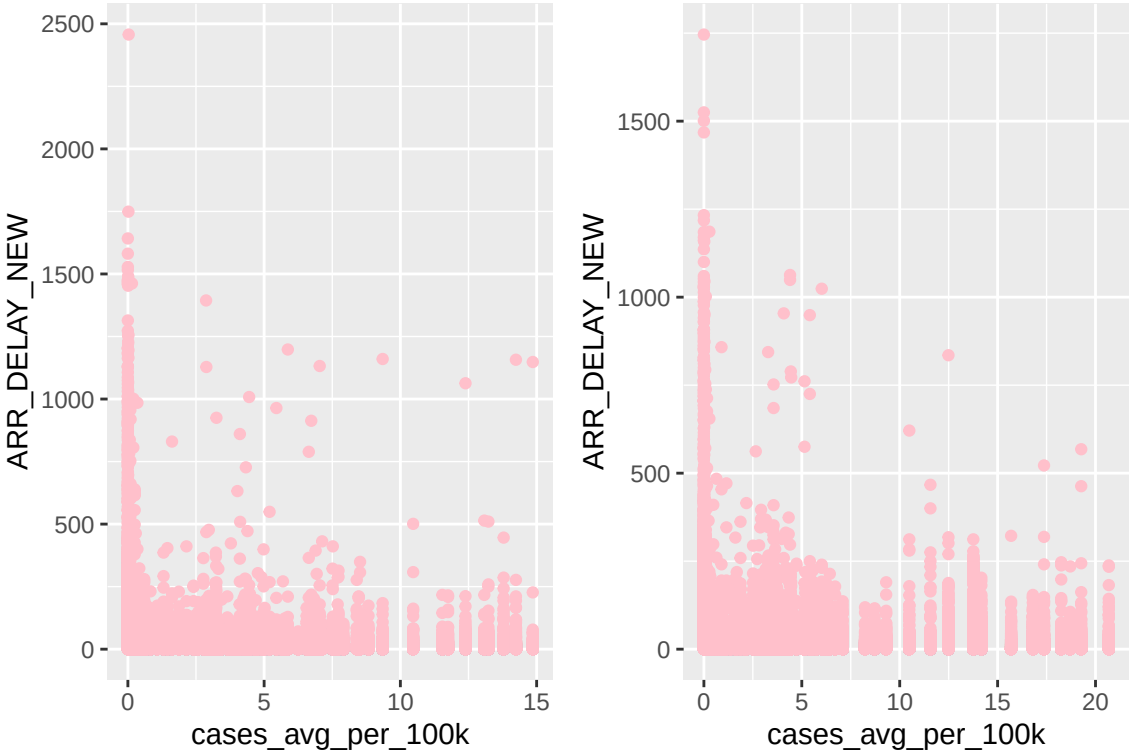
Abstract

This is our abstract.

Introduction

COVID-19 has affected our society in various ways. Among all of them, we decided to look at its effect on the flight delay and cancellations. Not only COVID-19 would have reduced the demand for the air travel, but also the counter-measures such as lockdowns are expected to negatively impacted the flights' on-time performance. Plus, if crew members had been exposed to the virus, they would have been required to quarantine which would likely cause the flight to be delayed or cancelled. Therefore, we would like to analyze the relationship between the COVID-19 cases and the delay and cancellation of the flights.

Figure 1: Average cases per 100k vs Arrival Delay for Flights departing from California (left) and Texas (right)



From the graph above, the relationship between the number COVID cases and the flight delay is not apparent in both California and Texas cases. This motivated us to conduct more rigorous analyses with controlling other factors that could be associated with flight delays, so see partial effect (or dependence) of the number of COVID-19 cases. We do this by asking three separate but related questions: "Is a flight more likely to be delayed with a higher number of cases of COVID-19?", "Is a flight more likely to be cancelled with a higher number of cases of COVID-19?", and "If a flight is delayed, is it expected to be delayed for longer with a higher number of cases of COVID-19?". The rest of our report goes in detail to explain how we estimated the effect.

Methods

Data Description

We have used 2 separate datasets and combined them to form a data frame that we used in the further analysis. The first dataset is from The United States Department of Transportation's (DOT) Bureau of Transportation Statistics tracks the on-time performance of domestic flights operated by large air carriers. The data collected is from January to June 2020 and contains relevant flight information (on-time, delayed, canceled, diverted flights) from the Top 10 United States flight carriers for 11 million flights. The second dataset is from the New York Times[2] which contains the state-wise data on the daily number of new cases and deaths, the seven-day rolling average, and the seven-day rolling average per 100,000 residents. We merged these two datasets based on the date and state to create a new dataset that we used in all the models. This combined dataset has in total of 2745847 observations with data from 375 different airports.

Data Wrangling

Data cleaning and preprocessing for the dataset was a four-step process.

1. Formatting the date field in both the individual datasets to match before performing a left join.
2. Merging the covid case dataset into the flight data based on date and state.
3. Factorizing the categorical variables from this combined dataset. `MONTH`, `DAY_OF_MONTH`, `DAY_OF_WEEK`, `MKT_UNIQUE_CARRIER`, `TAIL_NUM`, `ORIGIN`, `ORIGIN_STATE_NM`, `DEST`, `DEST_STATE_NM`, `ARR_DEL15`, `CANCELLED`, `CANCELLATION_CODE`, these are the categorical variables in the dataset.
4. Removing the variables that are not used in the analysis to have a cleaner dataset.

The cleaned data set is split into training and testing sets for California and Texas separately. There were 284941 observations with California as the origin state. So California data set has 227952 observations in the training set and 56989 observations in the testing set. Whereas, there were 287693 observations with Texas as the origin state and there are 230154 observations in the training set and 57539 observations in the testing set for Texas. In terms of the air traffic both California and Texas are comparable. The list of variables used in our models are below:

Variable Name	Description
`MONTH`	Month of Year
`DAY_OF_MONTH`	Text
`DAY_OF_WEEK`	Date of Month

Variable Name	Description
FL_DATE	Day of Week (1: Monday, 7: Sunday)
MKT_UNIQUE_CARRIER	Airline Carrier Code (Look at appendix for detail)
ORIGIN	Flight Departure 3-Letter Airport Abbreviation
ORIGIN_STATE_NM.	Flight Departure State Name
DEST	Flight Arrival 3-Letter Airport Abbreviation
DEST_STATE_NM	Flight Arrival State Name
ARR_DELAY_NEW	Departure Delay Ignoring Early Departures (Listed as 0)
ARR_DEL15	Departure Delay Greater Than 15 Minutes (0: Not Greater Than 15, 1: Greater Than 15)
CANCELLED	0: Flight Not Cancelled, 1: Flight Cancelled
DISTANCE	Distance Between Departure and Arrival Airports (in Miles)
cases_avg_per_100k	The average number of new cases per 100,000 people, reported over the most recentseven days of data. In other words, the seven-day trailing average.

Models of choice

With the dataset, we used various models to answer the questions. The dependent variables of interest are ARR_DEL15, ARR_DELAY_NEW, CANCELLED. First, we look at ARR_DEL15 and CANCELLED. They are binary variables, which call for classification models. We used a logit model to predict ARR_DEL15 and logit and tree models to predict CANCELLED. We used the tree model with a complexity parameter of 0.005 and minsplit of 30 since we have a lot of observations in both the cases of California and Texas. We adjusted the parameters by looking at the results and picking the adequate looking trees. They all have MONTH, DAY_OF_WEEK, MKT_UNIQUE_CARRIER, ORIGIN, DEST_STATE_NM, DISTANCE, and cases_avg_per_100k as features. We wanted to control for the origin and destination airports, but that would involve too many levels for categorical variables. Therefore, we included the destination state instead of the destination airport. We added month, day of week, carrier and distance to control for other causes than COVID cases.

Next, we look at ARR_DELAY_NEW. ARR_DELAY_NEW is 0 if the flight is not delayed and has the minutes of delay if the flight is delayed. Since the majority of the flights are not delayed, this variable has a lot of zeros compared to the other numbers. For this reason, we used zero-inflated poisson models, hoping that this would be more accurate. Although it is typically used for count data with many zeros and our variable of interest is time, I thought it is worth trying this model as we have discrete time values in minutes. We compare the results in this model with the basic linear model. For the zero-inflated poisson model, the first process generates zeros and the second process is governed by a Poisson distribution that generates counts, some of which may be zero. In this model building, the assumption is that the COVID cases would generate the non-zero counts. They both have MONTH, DAY_OF_WEEK, MKT_UNIQUE_CARRIER, DISTANCE, and cases_avg_per_100k for same reasons as before, with ORIGIN and DEST_STATE_NM also excluded for having too many categories. We also use tree models to predict 'ARR_DELAY_NEW. This has MONTH, DAY_OF_WEEK, MKT_UNIQUE_CARRIER, ORIGIN, DEST_STATE_NM, DISTANCE, and cases_avg_per_100kas features. We used the tree model with a complexity parameter of 0.0009 and minsplit of 30. We adjusted the parameters by looking at the results and picking the adequate looking trees.

Another way to look at the ARR_DELAY_NEW variable is to focus on the flights that are delayed. That is, we can see if the COVID cases would increase the delay time, given the flight is already delayed. We used

random forest models to predict this, as it usually performed the best for the homework questions. These also have MONTH, DAY_OF_WEEK, MKT_UNIQUE_CARRIER, DISTANCE, and cases_avg_per_100kas the features, without ORIGIN and DEST_STATE_NM, for having too many levels. We used default values for ntrees and mtry.

Results

Delay Rates - Logit Model

We looked if the Covid cases caused more delays by running a logit regression of ARR_DEL15 on MONTH, DAY_OF_WEEK, MKT_UNIQUE_CARRIER, ORIGIN, DEST_STATE_NM, DISTANCE, cases_avg_per_100k variables. In the case of California, the model has an out-of-sample accuracy of 0.9119995. Whereas the logit model built on the Texas data has an out-of-sample accuracy of 0.8900261. These accuracies were calculated for a k value of 0.5.

For California, the partial effect of cases_avg_per_100k is 0.02087 that means if the cases_avg_per_100k increases by a unit, the log odds of flights delaying by 15 mins will increase by 0.02087. This effect is not even statistically significant at 10% significance level. Whereas for Texas, the partial effect of cases_avg_per_100k is 0.03658 that means if the cases_avg_per_100k increases by a unit, the log odds of flights delaying by 15 mins will incrase by 0.03658. This effect is statistically significant at even 0.1% significance level and is slightly higher compared to Texas and this is clearly more significant compared to California.

Cancellation Rates - Logit Model

We looked at the effect of Covid cases on flight cancellations by running a logit regression of CANCELLED on MONTH, DAY_OF_WEEK, MKT_UNIQUE_CARRIER, ORIGIN, DEST_STATE_NM, DISTANCE, cases_avg_per_100k variables. In the case of California, the model has an out-of-sample accuracy of 0.8955939. Whereas the logit model built on the Texas data has an out-of-sample accuracy of 0.8903874. These accuracy rates were calculated for a k value of 0.5.

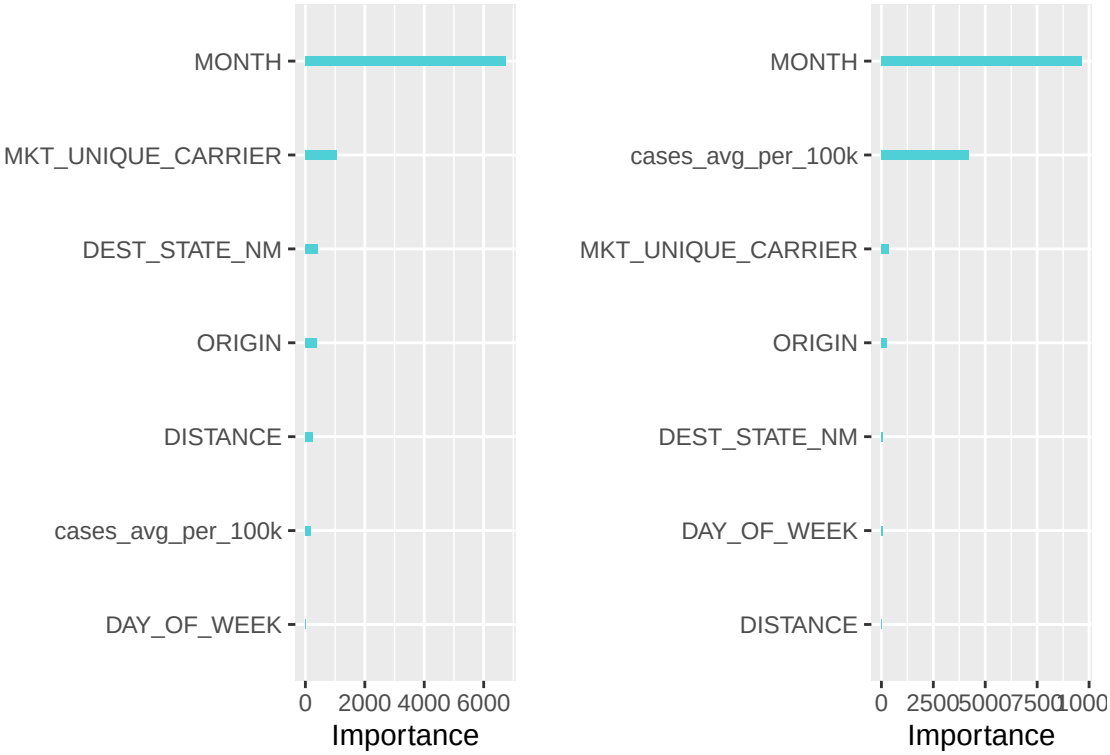
For California, the partial effect of cases_avg_per_100k is 0.9401 that means if the cases_avg_per_100k increases by a unit, the log odds of cancellation will increase by 0.9401. This effect is statistically significant at even 0.1% significance level. Whereas for Texas, the partial effect of cases_avg_per_100k is 0.9928 that means if the cases_avg_per_100k increases by a unit, the log odds of cancellation will increase by 0.9928. This effect is statistically significant at even 0.1% significance level and is slightly higher compared to California.

Cancellation Rates - Tree Model

For California, the initial decision in the tree is based on if the month was April. It is interesting to see the tree model capture the lock downs in April. Further decisions are made based on the type of carrier and distance of the flights. The importance plot shows that avg_cases_per_100k had an effect on the flight cancellations. In the case of Texas, the initial decision in the tree is based on if the month was April. It is interesting to see the tree model capture the lock downs in April even in the case of Texas. But the since the lock downs in Texas were not enforced as strictly as California, the left side of the decision tree has further classifications based on the flight carrier, day of the week and cases_avg_per_100k. When we look at the

importance plot, Texas clearly deviates from the California. The `avg_cases_per_100k` is the second most important variable, and has played more effect on the flight cancellations.

Figure 2: Variable importance plot for tree model predicting if Flights were Cancelled in California (left) and Texas (right)



We used the tree model with a complexity parameter of 0.005 and minsplit of 30 since we have a lot of observations in both the cases of California and Texas. For California, the initial decision in the tree is based on if the month was April. It is interesting to see the tree model capture the lock downs in April. Further decisions are made based on the type of carrier and distance of the flights. The importance plot shows that `avg_cases_per_100k` had an effect on the flight cancellations. In the case of Texas, the initial decision in the tree is based on if the month was April. It is interesting to see the tree model capture the lock downs in April even in the case of Texas. But the since the lock downs in Texas were not enforced as strictly as California, the left side of the decision tree has further classifications based on the flight carrier, day of the week and cases_avg_per_100k. When we look at the importance plot, Texas clearly deviates from the California. The `avg_cases_per_100k` is the second most important variable, and has played more effect on the flight cancellations. Figure 2 shows the Variable Importance Graph. The tree models can be seen at the Appendix. Figure A1 shows the tree for California and Figure A2 shows the tree for Texas.

Delay Minutes - Linear Regression

We looked at the effect of Covid cases on flight cancellations by running a linear regression of `ARR_DELAY_NEW` on `MONTH`, `DAY_OF_WEEK`, `MKT_UNIQUE_CARRIER`, `DISTANCE`, `cases_avg_per_100k` variables. In the case of California, the model has an out-of-sample accuracy of 33.80909. Whereas the linear regression model built on the Texas data has an out-of-sample accuracy of 31.96574. For California, the partial effect of `cases_avg_per_100k` is 0.0925041 that means if the `cases_avg_per_100k` increases by a unit, the

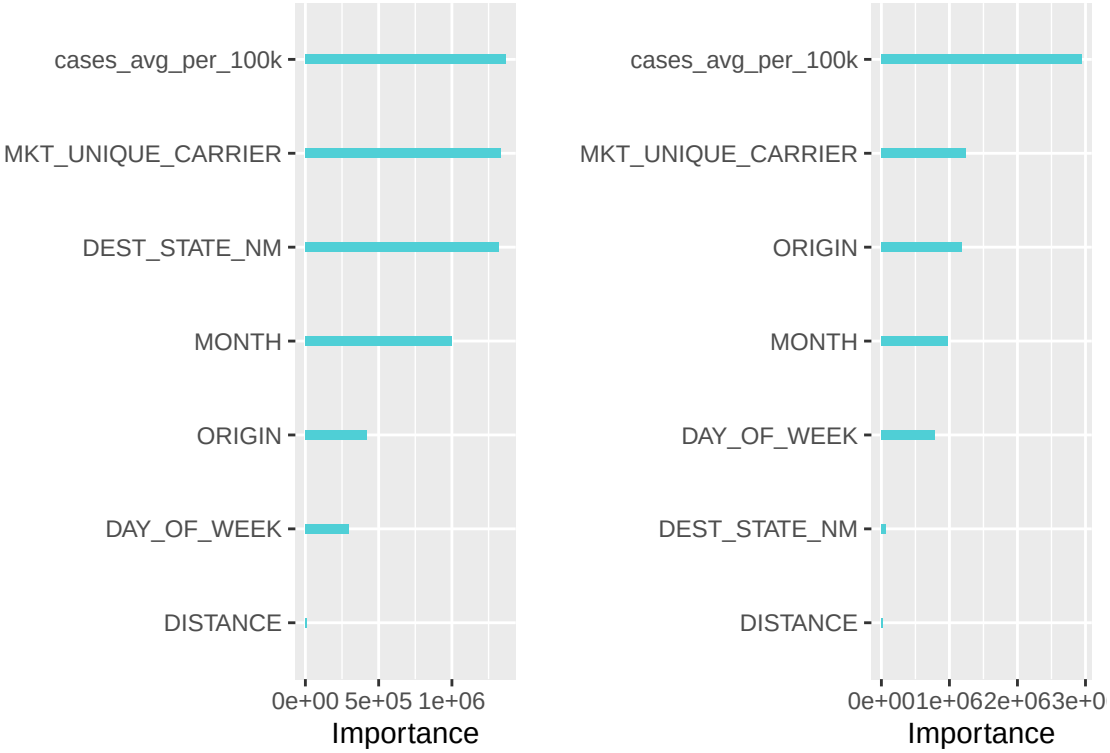
arrival delay will increase by 0.0925041. This effect is not statistically significant at even 10% significance level. Whereas for Texas, the partial effect of `cases_avg_per_100k` is 0.2228197 that means if the `cases_avg_per_100k` increases by a unit, the arrival delay will increase by 0.2228197. This effect is statistically significant at even 0.1% significance level and is significantly higher compared to California.

Delay - Zero-Inflated Poisson Regression

Next, we used zero inflated poisson regression to see the effect of covid cases on the arrival delay. In both California and Texas the effect of `cases_avg_per_100k` is small and positive but statistically significant. The RMSE in the case of California is 33.72181 and Texas is 31.96069.

Delay - Tree Model

Figure 3: Variable importance plot for tree model predicting Arrival Delay of Flights in California (left) and Texas (right)



For California, the initial decision in the tree is based on if the average covid cases per 100k was less than 0.32 or not and rest of the decisions are made based on destination states. The importance plot shows that `avg_cases_per_100k` has the biggest effect on the flight cancellations. In the case of Texas, the initial decision in the tree is based on if the average covid cases per 100k was less than 0.2 or not which is less than in the case of California and rest of the decisions are made based on destination states. When we look at the importance plot, Texas follows California with `avg_cases_per_100k` being the most important variable. The out of sample RMSE value for California is 33.38648 and for Texas is 31.45617. The tree looks to preform slightly better for Texas state data. Figure 3 shows the Variable Importance Graph. The tree models can be seen at the Appendix. Figure A3 shows the tree for California and Figure A4 shows the tree for Texas.

Delay Time Effect for Those Delayed - Random Forest Model

Figure 4: Variable Importance Plot and Partial Dependance Plot for RandomForest model predicting Arrival Delay in California and Texas

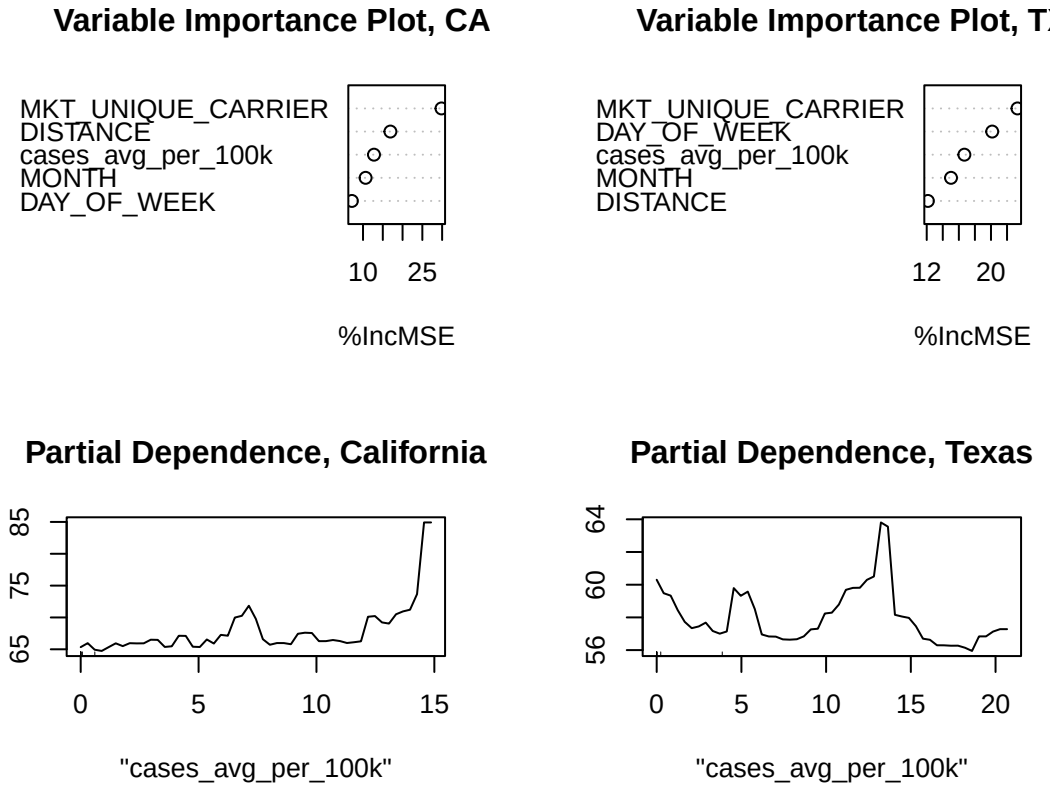


Figure 4 shows the Variable Importance Graph and Partial Dependence Graph for the random forest model. Random Forest model for the data with the delayed flights shows that for both California and Texas the `cases_avg_per_100k` is the third most important variable ahead of the `month`. Partial Dependence graph on `cases_avg_per_100k` for California seems to suggest that the delays exponentially increase from 65 mins to 85 mins when the `cases_avg_per_100k` reached 15. But in the case of Texas the range is smaller and the delays fluctuating between 54 and 64 mins. The out of sample RMSE value for California is 93.58049 and for Texas is 81.30241. This suggests that the randomForest model worked better with the Texas state data.

Conclusion

Appendix

Code	Airline Name
AA	American Airlines
AS	Alaska Airlines
B6	JetBlue
DL	Delta Air Lines
F9	Frontier Airlines
G4	Allegiant Air.

Code	Airline Name
HA	Hawaiian Airlines
NK	Spirit Airlines
UA	United Airlines
WN	Southwest Airlines

Figure A1: Tree model for Cancellation of Flights in California

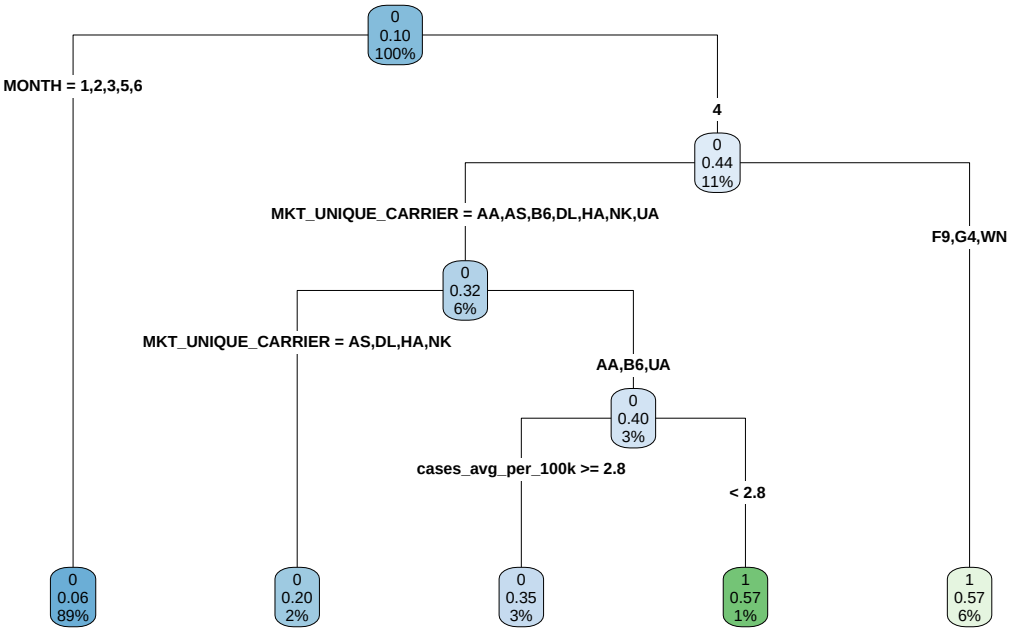


Figure A2: Tree model for Cancellation of Flights in Texas

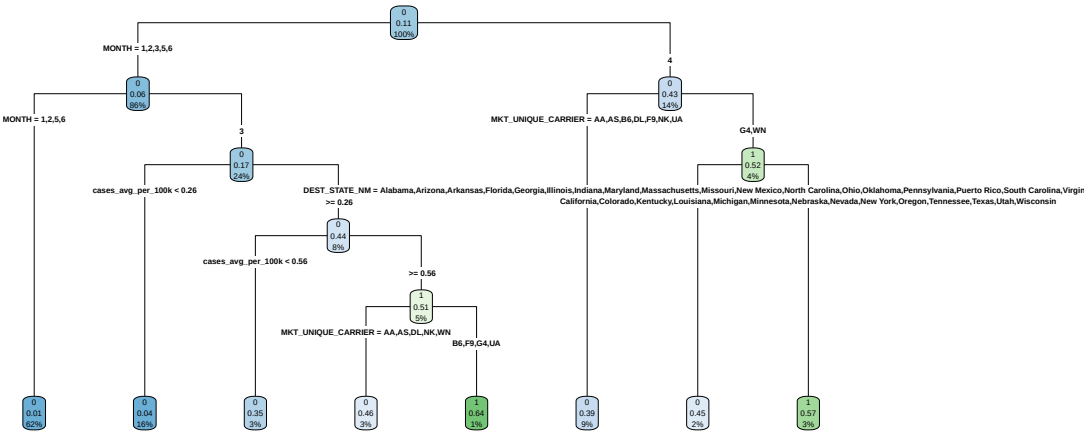


Figure A3: Tree model for Delay of Flights in California

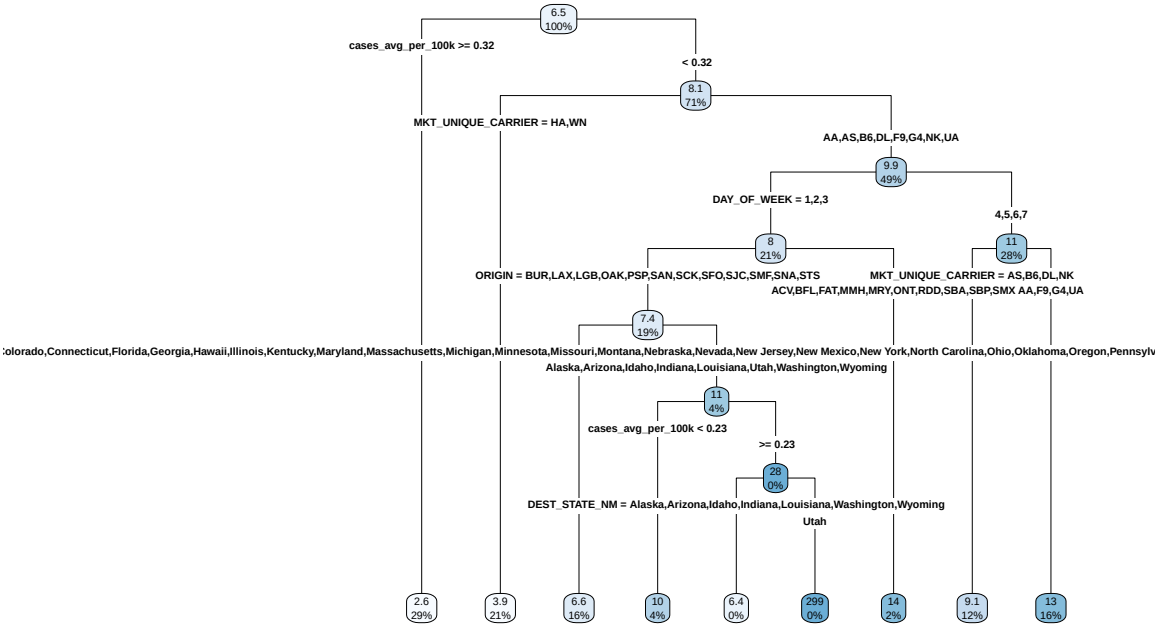


Figure A4: Tree model of Delay of Flights in Texass