UOL ID: 220460044

ST2195 Programming for Data Science

Part 2 Report

Table of Contents

1. Introduction

This reports utilises flight data for all commercial flights on major carriers within the US from 1987 to 2008 provided by the Data Expo to answer the following questions:

(a) What are the best times and days of the week to minimise delays for each year?
(b) Evaluate whether older planes suffer more delays on a year-on-year basis.
(c) For each year, fit a logistic regression model for the probability of diverted US flights using as many features as possible from attributes of the departure date, the scheduled departure and arrival times, the coordinates and distance between departure and planned arrival airports, and the carrier. Visualize the coefficients across years.

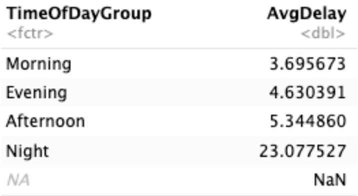
2. Addressing Questions

Before we delve into the differing parts of the questions, we load in the necessary packages in order to do data pipelining.

a. We split the question into 3 parts, focusing individually on average delay by time of day, average delay by day of week and average delay by month. Thereafter, combining the least delay in all 3 parts to conclude our best time to minimise delay for each year.

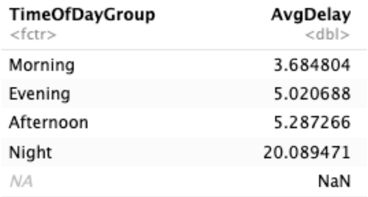
When looking at when to fly to minimise delay, we will use arrival delay as a measure for average delay instead of looking at departure delay or both. This is due to the fact that as a passenger flying, the delay they are most concerned about at the end of the day is what time they would arrive at their destination airport. We created a data frame, "ontime" to consolidate commercial flight data from 1991 to 2000. We do not take into account diverted and cancelled flights the delay from that is time spent on ground and not time spent on the plane. We remove diverted and cancelled flights have no arrival by removing all flights with NA as their arrival time.

To find the best time of day, we aggregated a day into 4 time intervals: midnight to 6am, 6am to 12 noon, 12 noon to 6pm and 6pm to midnight labelled Night, Morning, Afternoon, Evening respectively. We calculate the average delay for each time of day per year to find the time interval with the least delay.



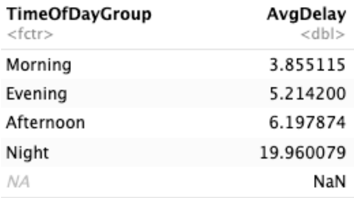
TimeOfDayGroup	AvgDelay
<fctr>	<dbl>
Morning	3.695673
Evening	4.630391
Afternoon	5.344860
Night	23.077527
NA	NaN

Table 1: 1991 TimeOfDay Average Delay



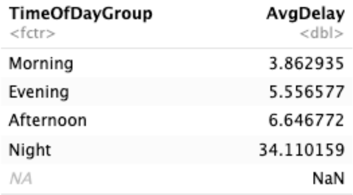
TimeOfDayGroup	AvgDelay
<fctr>	<dbl>
Morning	3.684804
Evening	5.020688
Afternoon	5.287266
Night	20.089471
NA	NaN

Table 2: 1992 TimeOfDay Average Delay



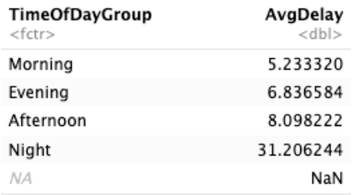
TimeOfDayGroup	AvgDelay
<fctr>	<dbl>
Morning	3.855115
Evening	5.214200
Afternoon	6.197874
Night	19.960079
NA	NaN

Table 3: 1993 TimeOfDay Average Delay



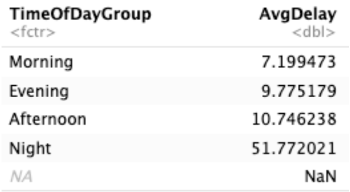
TimeOfDayGroup	AvgDelay
<fctr>	<dbl>
Morning	3.862935
Evening	5.556577
Afternoon	6.646772
Night	34.110159
NA	NaN

Table 4: 1994 TimeOfDay Average Delay



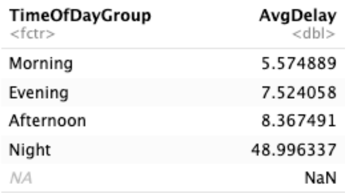
TimeOfDayGroup	AvgDelay
<fctr>	<dbl>
Morning	5.233320
Evening	6.836584
Afternoon	8.098222
Night	31.206244
NA	NaN

Table 5: 1995 TimeOfDay Average Delay



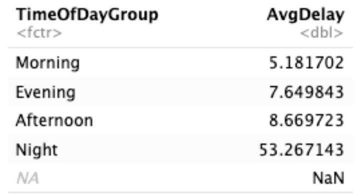
TimeOfDayGroup	AvgDelay
<fctr>	<dbl>
Morning	7.199473
Evening	9.775179
Afternoon	10.746238
Night	51.772021
NA	NaN

Table 6: 1996 TimeOfDay Average Delay



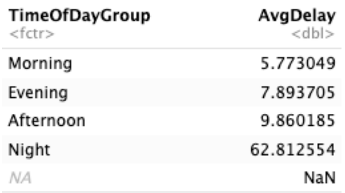
TimeOfDayGroup	AvgDelay
<fctr>	<dbl>
Morning	5.574889
Evening	7.524058
Afternoon	8.367491
Night	48.996337
NA	NaN

Table 7: 1997 TimeOfDay Average Delay



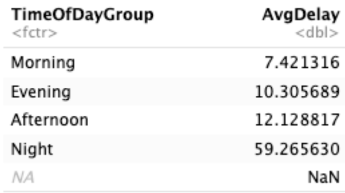
TimeOfDayGroup	AvgDelay
<fctr>	<dbl>
Morning	5.181702
Evening	7.649843
Afternoon	8.669723
Night	53.267143
NA	NaN

Table 8: 1998 TimeOfDay Average Delay



TimeOfDayGroup	AvgDelay
<fctr>	<dbl>
Morning	5.773049
Evening	7.893705
Afternoon	9.860185
Night	62.812554
NA	NaN

Table 9: 1999 TimeOfDay Average Delay



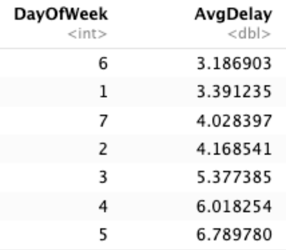
TimeOfDayGroup	AvgDelay
<fctr>	<dbl>
Morning	7.421316
Evening	10.305689
Afternoon	12.128817
Night	59.265630
NA	NaN

Table 10: 2000 TimeOfDay Average Delay

The tables show the average arrival delay (in minutes) for each time interval over the years. It can be concluded that the morning is the best time to fly as it experiences the least delay and night is the worst time to fly as it experiences the most delays. This observation is consistent through the 10 years.

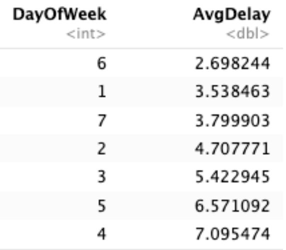
In general, the average delay in the morning ranges from 5-8 mins and night has a drastically higher range of 19 to 62. Taking year 2000 as an example, there is almost 800% increase from the average delay in the morning to the average delay at night.

We use something similar to calculate the average delay by day of week. However, instead of aggregating into 4 time intervals, we organise the delay by day of week, 1 being Monday and 7 being Sunday. The data is then ranked from least to most delay for ease of viewing.



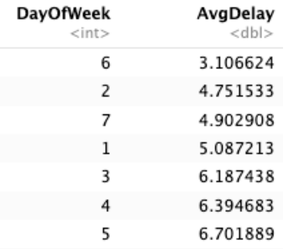
DayOfWeek	AvgDelay
<int>	<dbl>
6	3.186903
1	3.391235
7	4.028397
2	4.168541
3	5.377385
4	6.018254
5	6.789780

Table 11: 1991 DayOfWeek Average Delay



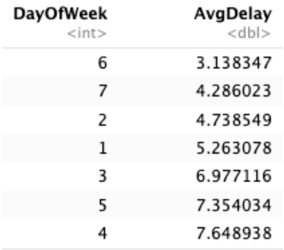
DayOfWeek	AvgDelay
<int>	<dbl>
6	2.698244
1	3.538463
7	3.799903
2	4.707771
3	5.422945
5	6.571092
4	7.095474

Table 12: 1992 DayOfWeek Average Delay



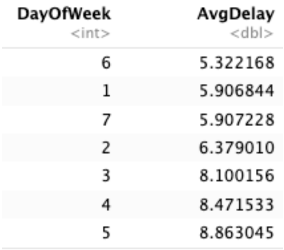
DayOfWeek	AvgDelay
<int>	<dbl>
6	3.106624
2	4.751533
7	4.902908
1	5.087213
3	6.187438
4	6.394683
5	6.701889

Table 13: 1993 DayOfWeek Average Delay



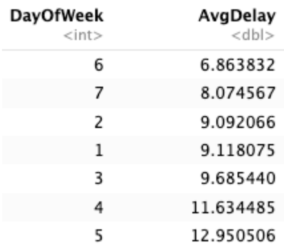
DayOfWeek	AvgDelay
<int>	<dbl>
6	3.138347
7	4.286023
2	4.738549
1	5.263078
3	6.977116
5	7.354034
4	7.648938

Table 14: 1994 DayOfWeek Average Delay



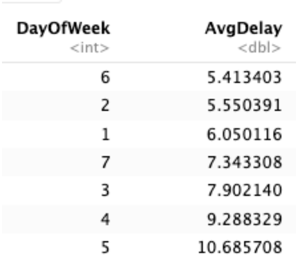
DayOfWeek	AvgDelay
<int>	<dbl>
6	5.322168
1	5.906844
7	5.907228
2	6.379010
3	8.100156
4	8.471533
5	8.863045

Table 15: 1995 DayOfWeek Average Delay



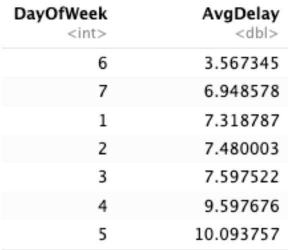
DayOfWeek	AvgDelay
<int>	<dbl>
6	6.863832
7	8.074567
2	9.092066
1	9.118075
3	9.685440
4	11.634485
5	12.950506

Table 16: 1996 DayOfWeek Average Delay



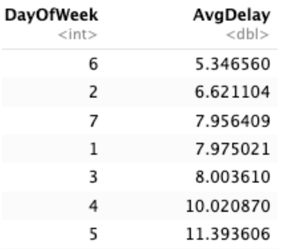
DayOfWeek	AvgDelay
<int>	<dbl>
6	5.413403
2	5.550391
1	6.050116
7	7.343308
3	7.902140
4	9.288329
5	10.685708

Table 17: 1997 DayOfWeek Average Delay



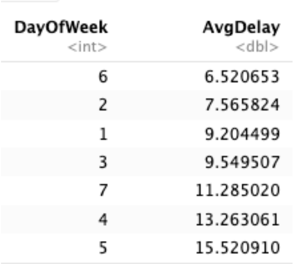
DayOfWeek	AvgDelay
<int>	<dbl>
6	3.567345
7	6.948578
1	7.318787
2	7.480003
3	7.597522
4	9.597676
5	10.093757

Table 18: 1998 DayOfWeek Average Delay



DayOfWeek	AvgDelay
<int>	<dbl>
6	5.346560
2	6.621104
7	7.956409
1	7.975021
3	8.003610
4	10.020870
5	11.393606

Table 19: 1999 DayOfWeek Average Delay

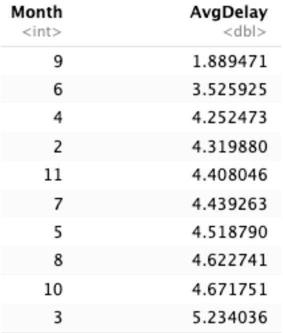


DayOfWeek	AvgDelay
<int>	<dbl>
6	6.520653
2	7.565824
1	9.204499
3	9.549507
7	11.285020
4	13.263061
5	15.520910

Table 20: 2000 DayOfWeek Average Delay

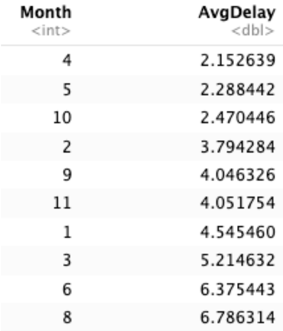
The tables show the average delay per day of the week from 1991 to 2000. In general, Saturdays has the least delay of 2 to 7 minutes while Thursdays and Fridays having highest of 6 to 15 minutes. The delays throughout the days of week stay relatively consistent and has a smaller difference compared to the time of day. Looking at year 2000, it has the largest difference in delay between days, increasing slightly more than 200% from Saturday to Friday as compared to the 800% increase between time intervals with highest to lowest delay.

We want to determine the average arrival delay for each month from 1991 to 2000 by using similar calculations to the average delay from Day of Week.



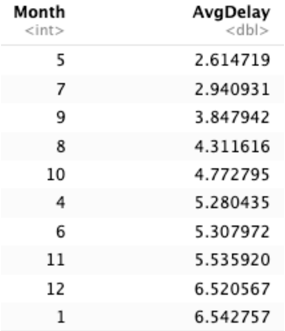
Month	AvgDelay
<int>	<dbl>
9	1.889471
6	3.525925
4	4.252473
2	4.319880
11	4.408046
7	4.439263
5	4.518790
8	4.622741
10	4.671751
3	5.234036

Table 21: 1991 Monthly Average Delay



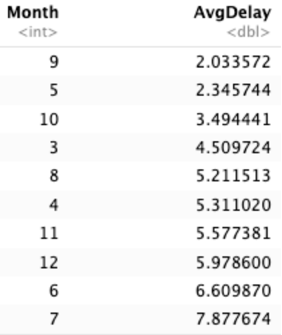
Month	AvgDelay
<int>	<dbl>
4	2.152639
5	2.288442
10	2.470446
2	3.794284
9	4.046326
11	4.051754
1	4.545460
3	5.214632
6	6.375443
8	6.786314

Table 22: 1992 Monthly Average Delay



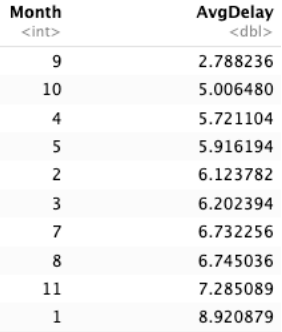
Month	AvgDelay
<int>	<dbl>
5	2.614719
7	2.940931
9	3.847942
8	4.311616
10	4.772795
4	5.280435
6	5.307972
11	5.535920
12	6.520567
1	6.542757

Table 23: 1993 Monthly Average Delay



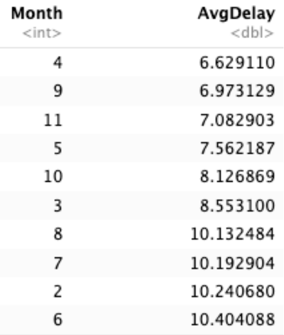
Month	AvgDelay
<int>	<dbl>
9	2.033572
5	2.345744
10	3.494441
3	4.509724
8	5.211513
4	5.311020
11	5.577381
12	5.978600
6	6.609870
7	7.877674

Table 24: 1994 Monthly Average Delay



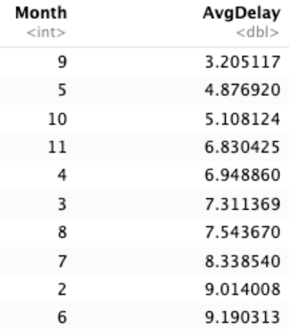
Month	AvgDelay
<int>	<dbl>
9	2.788236
10	5.006480
4	5.721104
5	5.916194
2	6.123782
3	6.202394
7	6.732256
8	6.745036
11	7.285089
1	8.920879

Table 25: 1995 Monthly Average Delay



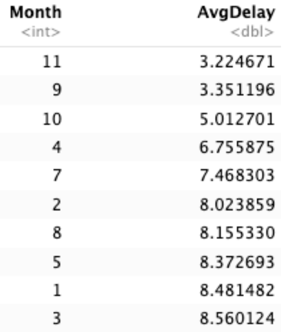
Month	AvgDelay
<int>	<dbl>
4	6.629110
9	6.973129
11	7.082903
5	7.562187
10	8.126869
3	8.553100
8	10.132484
7	10.192904
2	10.240680
6	10.404088

Table 26: 1996 Monthly Average Delay



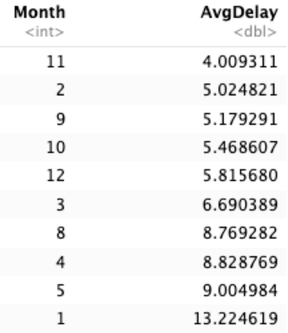
Month	AvgDelay
<int>	<dbl>
9	3.205117
5	4.876920
10	5.108124
11	6.830425
4	6.948860
3	7.311369
8	7.543670
7	8.338540
2	9.014008
6	9.190313

Table 27: 1997 Monthly Average Delay



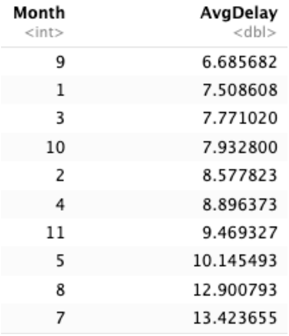
Month	AvgDelay
<int>	<dbl>
11	3.224671
9	3.351196
10	5.012701
4	6.755875
7	7.468303
2	8.023859
8	8.155330
5	8.372693
1	8.481482
3	8.560124

Table 28: 1998 Monthly Average Delay



Month	AvgDelay
<int>	<dbl>
11	4.009311
2	5.024821
9	5.179291
10	5.468607
12	5.815680
3	6.690389
8	8.769282
4	8.828769
5	9.004984
1	13.224619

Table 29: 1999 Monthly Average Delay



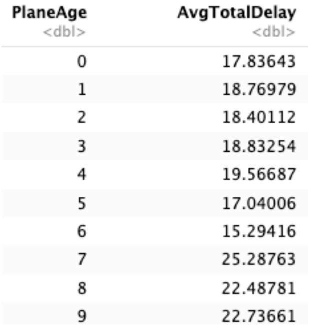
Month	AvgDelay
<int>	<dbl>
9	6.685682
1	7.508608
3	7.771020
10	7.932800
2	8.577823
4	8.896373
11	9.469327
5	10.145493
8	12.900793
7	13.423655

Table 30: 2000 Monthly Average Delay

In the tables above we look at the top 10 least average monthly delay for each year to determine which is the best month to fly. In general, September is a good month to fly to minimise delays as it consistently ranks on the lowest 6 delays historically. October ranked in the lowest 6 delays 7 times and May 6 times from 1991 to 2000. Not included in the tables above are the last 2 rankings of average monthly delays which indicates that these months experience the most and out of the 12 months, December ranked in the lowest 6 times in 10 years making it the worst month to travel in to minimise delay.

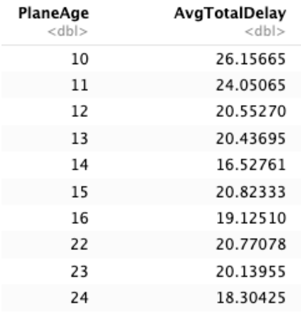
Based on the data and its analysis, we can conclude 3 things:
1. A morning flight with departure time from 6 am to 12 noon has the least delays and is the best time to fly in the day.
2. The best day of week to fly and minimise delays is Saturday.
3. Passengers should choose to fly in September as it is the best month to minimise delays.
Overall, the best time to fly in each year is usually a Saturday morning in September.

b. In determining the plane age, we assume that issue date is the date whereby the plane is certified to be safe and is able to fly and minus that from 1992 to 2000 from combining “carriers_df” and “ontime” by tailnumber. We make another dataframe “ontime_age” consisting of talinum and PlaneAge in order to simplify the data input. It is observed that we can have similar results by looking at the plane year as at year 2000. An simple illustration of this would be a plane that is 24 years old in 2000 would be the same plane with the same average delay in 1991 but is 15 years old. A difference between this part and part a is that the delay we are concerned about is the average total delay = departure delay + arrival delay instead of the average arrival delay in (a). This is because it is more important for the airline to determine the correlation between the total and plane age.



PlaneAge	AvgTotalDelay
<dbl>	<dbl>
0	17.83643
1	18.76979
2	18.40112
3	18.83254
4	19.56687
5	17.04006
6	15.29416
7	25.28763
8	22.48781
9	22.73661

Table 31: 2000 PlaneAge and its AvgTotalDelay



PlaneAge	AvgTotalDelay
<dbl>	<dbl>
10	26.15665
11	24.05065
12	20.55270
13	20.43695
14	16.52761
15	20.82333
16	19.12510
22	20.77078
23	20.13955
24	18.30425

Table 32: 2000 PlaneAge and its AvgTotalDelay cont

We remove rows with NA in DepTime, ArrTime and PlaneAge to synchronise the data between R and Python and ensure that it is not used in calculating the Average Total Delay Time.

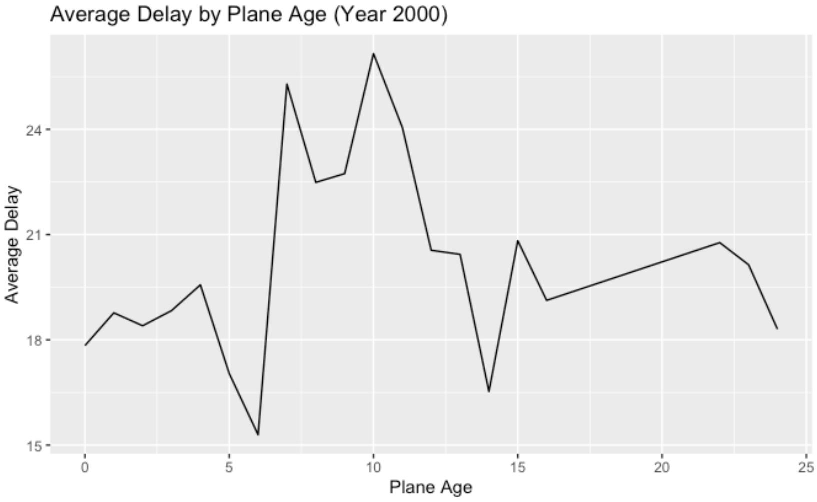


Figure 1: Graph depicting the Average Delay with Plane Age in year 2000

The table and graph indicates the average total delay for each plane age in year 2000. PlaneAge 0 means that it has only been less than 12 months since it has been issued its certification.

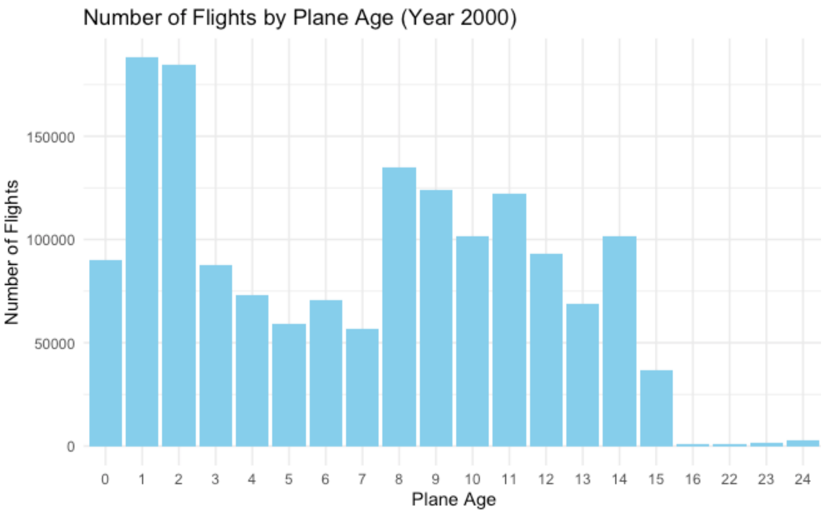


Figure 2: Number of Flights by Plane Age in year 2000

From figure 2, it is observed that planes that are 8 years-old has the most number of flight followed by that of 1 and 2 and planes that are 18 to 24 years-old has the least number of flights in year 2000. We look at the number of flights by plane age to determine whether it affects representativeness of the data of average delay total due to difference in sample size and conclude that the sample size for 16 to 24 year-old planes is too small in relation.

From the data extracted and disregarding the total average delay for 16 and 24 year-old planes, we can observe that a 6 year-old plane experiences the least delay while 9 and 14 year-old planes experiences the most delay. We are able to conclude that plane age is not a significant reason why older planes suffer more delays.

c. we import the carriers.csv into a data frame and create and empty list to store the coefficients of features in the logistic regression as well as a data frame to store the probabilities of diverted US flights from 1991 to 2000.

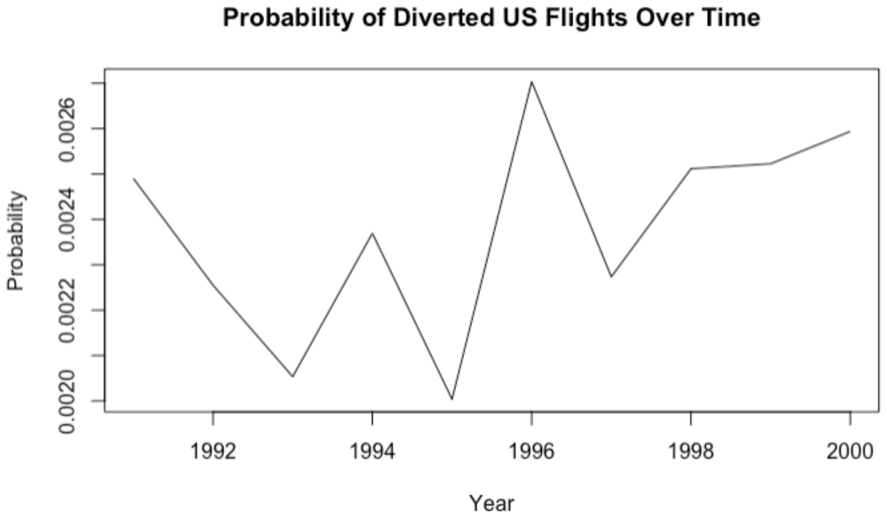


Figure 3: Probability of Diverted US Flights Over Time

We plot the probability of diverted US flights against total flight over time. Thereafter, we determine the features that we want to consider for the logistic regression model to be Departure Delay, CRS Departure Time, CRS Arrival Time, Distance between airports, Unique Carrier with target Diverted. We convert Unique carrier from character to factor for the analysis. However as Unique Carrier has 11 levels, it would introduce to many parameters into the model if we treat it as a single categorical variable. We will look at the coefficients of each individual year to see how each variable affects the probability of plane diversion. The coefficients in the model indicates the change in log-odds of the probability of plane being diverted when a predictor variable changes, ceteris paribus.

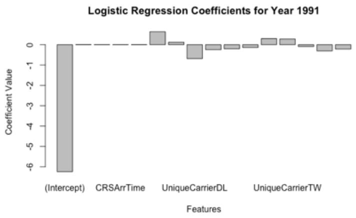


Figure 4: 1991 Logistic Regression Coefficients

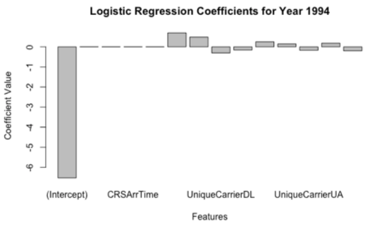


Figure 7: 1994 Logistic Regression Coefficients

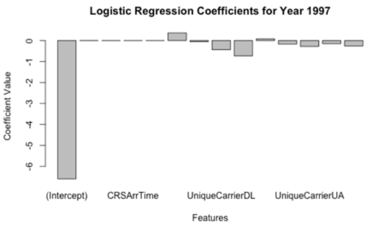


Figure 10: 1997 Logistic Regression Coefficients

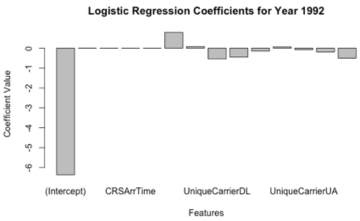


Figure 5: 1992 Logistic Regression Coefficients

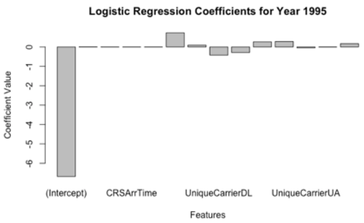


Figure 8: 1995 Logistic Regression Coefficients

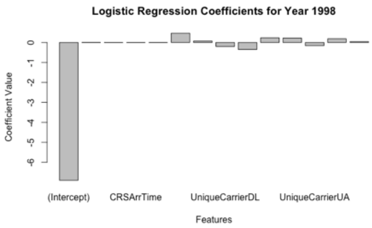


Figure 11: 1998 Logistic Regression Coefficients

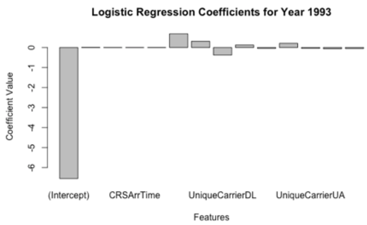


Figure 6: 1993 Logistic Regression Coefficients

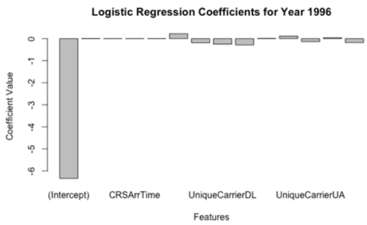


Figure 9: 1996 Logistic Regression Coefficients

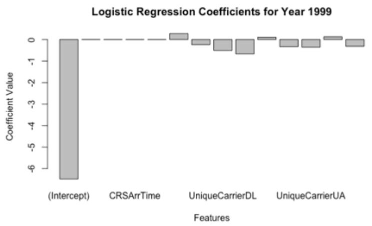


Figure 12: 1999 Logistic Regression Coefficients

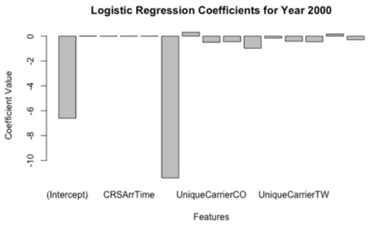


Figure 13: 2000 Logistic Regression Coefficient

The intercept is the log-odd probability of flight being diverted when all other variables are zero. In the tables, it can be seen that the intercept coefficient is negative for all years, however it might not directly relate to something in reality, for example departure time which is a time related variable that cannot be zero. Thus, instead of focusing on the intercept, it is better to look at the coefficients of features in order to have a more meaningful analysis. The absolute coefficient is the change is log-odds probability of diversion when the feature/ predictor variable is changes by 1 unit, ceteris paribus.

From the logistic regression coefficients, we are able to identify that factors affect the probability of diversion at differing weights. For example, looking at 1991, we can see that the extent of change due to change in UniqueCarrierAS and UniqueCarrierDL is about the same even though one is positive and the other is negative as the absolute value of coefficient being around the same and that their extent of change is more than CRSDepTime and CRSArrTime. This is further emphasised in 2000 where the absolute values of some are significantly larger than others, resulting in the change in probability of diversion being changed to a larger extent. Through this, we can conclude that carriers typically has a larger impact on probability of the plane being diverted as compared to DepDelay, CRSDepTime, CRSArrTime and Distance and this is being uniformly observed throughout 1991 to 2000.