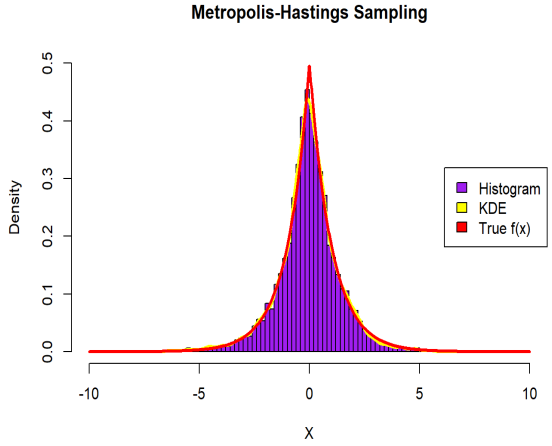
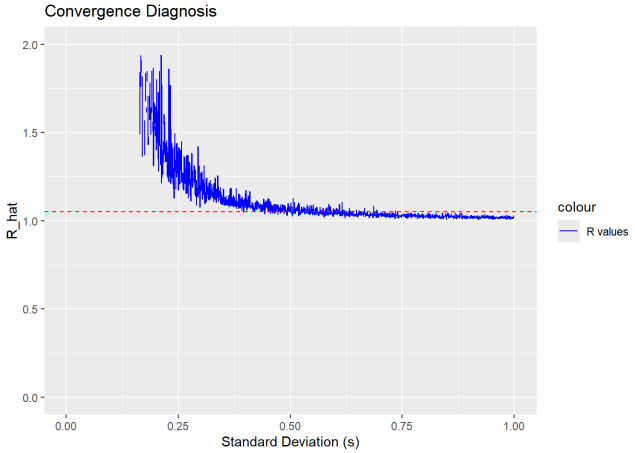
Part 1 R outputs

1a) Metropolis Hasting Algorithm, it aims to generate random samples from a probability distribution that is so high in dimension that makes direct sampling difficult to do so. By following the steps provided by the guideline we will produce similar results.






```r
print(paste("The Monte Carlo Estimate of Mean: ", mean(sample)))
print(paste("The Monte Carlo Estimate of Standard Deviation: ", sd(sample)))
```

```
[1] "The Monte Carlo Estimate of Mean:  -0.0423533175227206"
[1] "The Monte Carlo Estimate of Standard Deviation:  1.39006365471084"
```

Based on N=10000 and s=1 as defined in the question requirement, the Mean and Standard Deviation are as stated in our output

1b) The convergence diagnosis, otherwise known as Gelman-Rubin statistics, where we conclude whether our algorithm runs correctly based on whether the R hat value is less than 1.05, as stated in the guidelines. We follow the steps needed for the creation of multiple Markov Chains in the form of functional programming. Based on N=2000, J=4 and over a grid of s values, we abide by the conditional threshold of 1.05 for R hat.

```r
cat("Final R hat value:", result$latest_rhat,"\n")
```

```
Final R hat value: 1.039956
```

Based on s=0.001, the R hat value observed would be 1052.7730

s_vals <dbl>	rhat_vals <dbl>
0.0010	1052.7730
0.0015	710.4804
0.0020	433.2699
0.0025	342.5699
0.0030	366.5923
0.0035	404.7991

Part 2 R outputs

Database Creation

The database was created by running a loop over the 2001 to 2005 flight files taken from Harvard Dataverse as instructed in order to create the ontime table. Other supplementary files like airports, planes and carriers were also used to convert into tables for SQLite , joining the tables together to form the airline database.

2a) What are the best times and days that minimise delays for each year?

For clarification, interpretation is from a flight passenger's viewpoint. A flight passenger would want to be on time for their departing plane without any delays to their desired destination.

Data Preparation

A SQL query was used to extract flights from the on-time table, excluding all cancelled and diverted records. The selected columns included: CRSDepTime, departure delay, month, day of month, day of the week and flight metadata. This early filtering reduced noise and ensured only operational flights were considered for delay analysis. Year, month, day of the month, scheduled departure time(CRSDepTime), and lastly departure delay was selected. Scheduled departure time was chosen instead ofthe actual was due to the scheduled departure time being more useful to actually assess an on-time flight and the actual departure time is already influenced by delays.

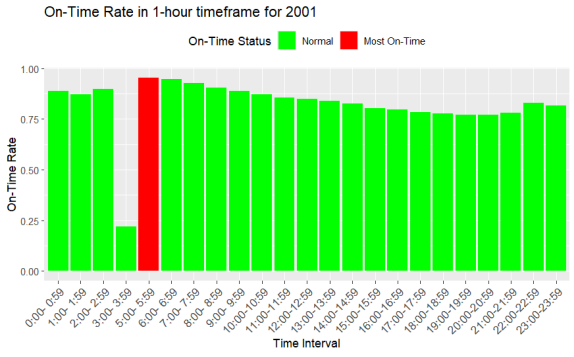
Data Wrangling And Strategy For The Best Time

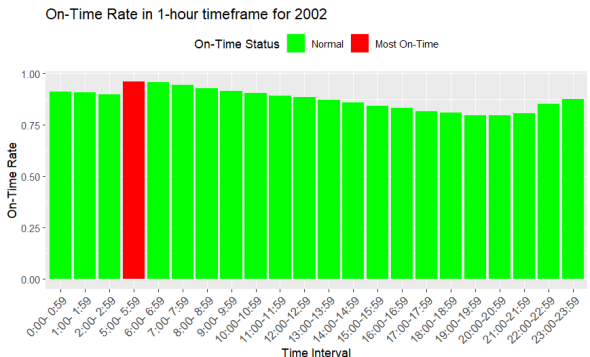
- First, we converted the scheduled departure time from HHMM to an hour
- Created time interval bins using a custom function
- Defined on-time flights where on-time=1 if departure delay is between -15 and 15 minutes(*North America on-time performance data for airlines and airports: OTP Flight Data: OAG*) using a custom function as well, then we calculate the average of on-time flights to which we gain on-time rate
- I also checked for the ratio of delayed flights to see if its evenly distributed and there is no extreme outliers
- Lastly, group by year to observe the best time that has the highest on-time rate

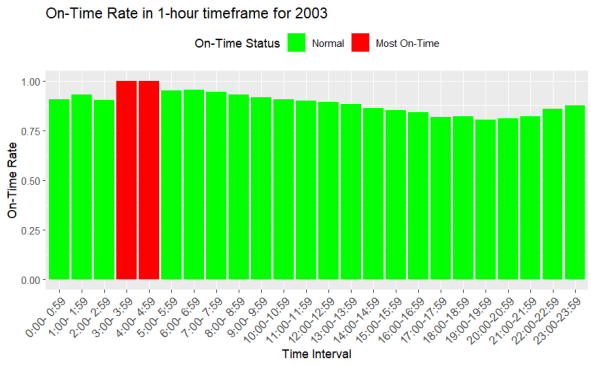
Data Wrangling And Strategy For The Best Day

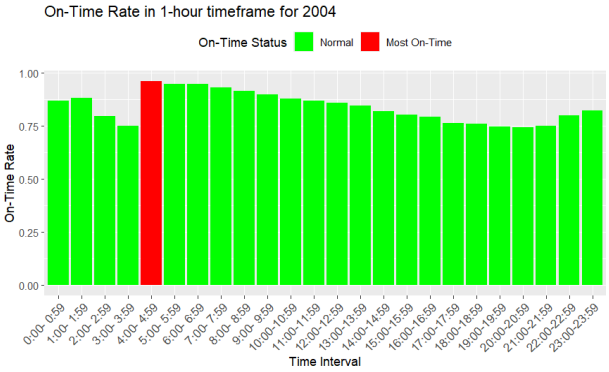
- First, we made sure to transform the day of the week into names of the days Monday, Tuesday etc
- Grouped the data by year and day of the week
- Then calculate the mean of departure delay after that
- Lastly the day that has the minimum average delay would be the best day by the logic of being on time, the days that has average delay closest to 0 would be the best day

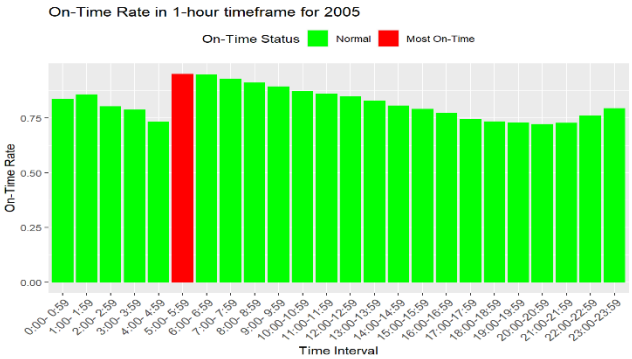
Best Times



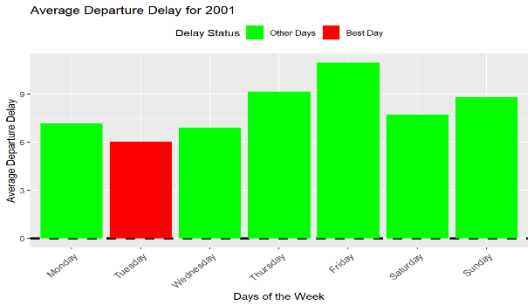


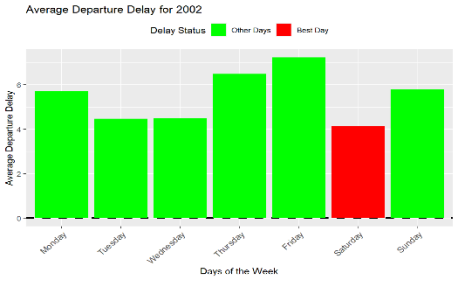


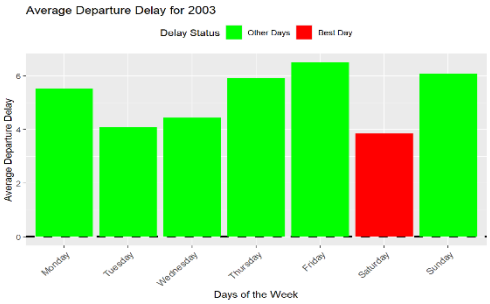


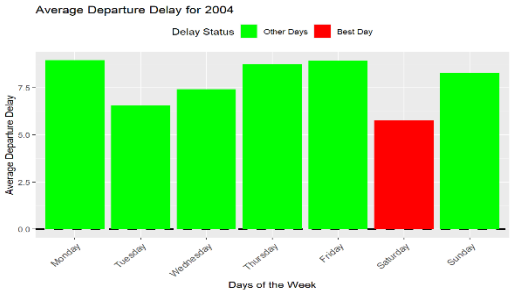


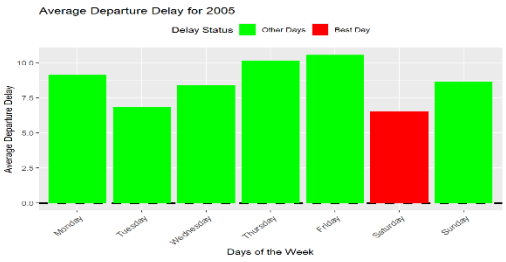
Best Days











Results Discussion

- Morning flights seem to have the highest on-time rates, especially 5-6 am. After that, the on-time rate seems to go on a downtrend till late evening, when it starts to pick up again.
- Saturdays seem to be the best days to minimise delays for most of the years, while 2001 seems to have Tuesdays as its contender.
- A pattern is observed where weekends and morning flights are generally consistent, and thus, passengers should choose to depart on morning flights during the weekend

Data Preparation

A SQL query was made to perform an inner join between the ontime and planes table for us to get the plane's manufacturing year as well as the plane models. A filter condition was also made again with the previous question's logic in mind we want rows where the destination is complete and so cancelled, along with diverted flights, must not be in the equation.

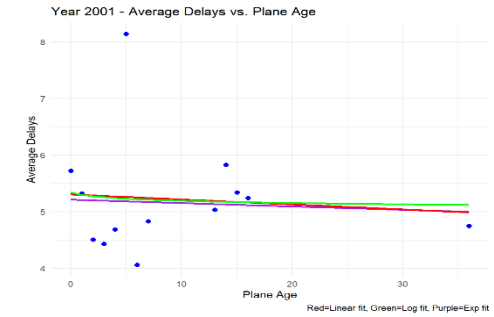
Data Wrangling

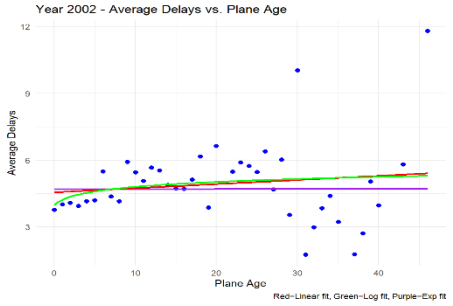
- First, we cleaned out records from the planes table that had blanks, None and 0000 as the value inside the plane models and the plane issued year, then we dropped any NAs we had.
- Computed the Plane age by subtracting the plane manufacturing year. This will result in negative plane age for some rows, we will also filter those out as those data points are impossible events, so they need to be removed.
- Group the data by plane age and year to calculate total delay, which we will average out using the total number of flights per aircraft age group

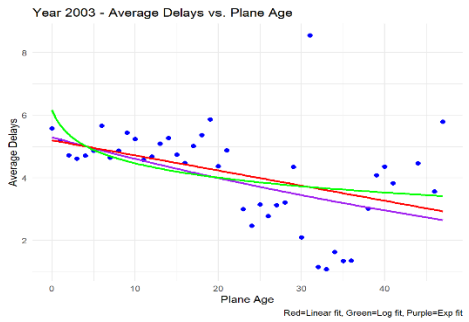
Data Manipulation Strategy

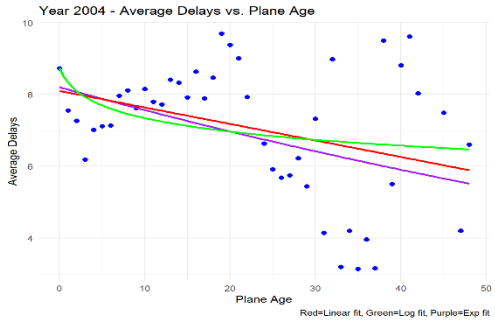
- Here we will do a correlation test to check for any sort of relationship between the 2 variables, being plane age and average delays
- Pearson, Spearman and Kendall tests are used annually. Observe if the p-value is fewer than 0.05 if so then it is statistically significant. E.g.. if Pearson test p value<0.05, there may be a linear relationship however, if its false, then we check for logarithmic and exponential relationships, which is why there are 2 other tests to do.
- Although statistically it may be significant, it may not be practically significant, so a visualisation of each model will be required to observe any sort of trends

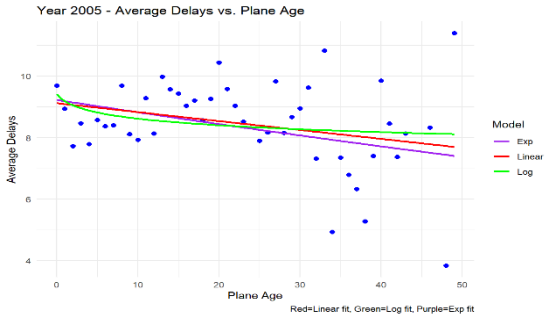
2b) In the case of whether older planes suffer more delays on a year-to-year basis











FlightYear <int>	Pearson_cor <dbl>	Spearman_cor <dbl>	Kendall_cor <dbl>	Pearson_Significant <lgl>	Spearman_Significant <lgl>	Kendall_Significant <lgl>
2001	-0.08465427	0.0879120879	0.05128205	FALSE	FALSE	FALSE
2002	0.13349433	-0.0008710801	0.05365854	FALSE	FALSE	FALSE
2003	-0.43133837	-0.5453035337	-0.36434109	TRUE	TRUE	TRUE
2004	-0.35046369	-0.2438735178	-0.14949495	TRUE	FALSE	FALSE
2005	-0.28526150	-0.2402096824	-0.15555556	FALSE	FALSE	FALSE

Results Discussion

We first observe that the statistical significance(*Correlation (Pearson, Kendall, Spearman)* 2024) suggests that there is no relationship of any kind between plane age and delays in 2001,2002,2005.
While there could be a linear relationship in 2004, there could be a linear or non-linear relationship between plane age and delays in 2003.
Negative correlation is observed a lot in the later years.

Conclusion

To conclude, plane ages and delays have an exponential relationship, which it is observed to be most prominent over the years, showing a consistent negative relationship in the later years. Although the statistics show no significance for some years, it is still practically significant given the plot we have seen. We then infer that, likely due to older planes requiring more maintenance effort, a better maintenance schedule is in place with experienced crews, which ended up making the older planes much more reliable than the newer planes. We deduce that newer planes have much better technology, and the crew members, along with technicians, are unsure of how to use and tune it, leading to more time required to get used to it, resulting in higher delays. As the year goes by, we observe that older planes suffer fewer and fewer delays, which is likely due to the planes being tuned to be more reliable after the technicians become more experienced and accumulated knowledge to reduce any possible flight delay as time goes on, there could also be an increase in investments to make older planes for reliable. The exception of 2002 could be due to a change in airline policy or the introduction of a new fleet of planes, resulting in delays increasing slightly with age, likely due to planes not old enough to require heavy maintenance , as we do observe a lot of data points in the younger age group. It could also be the year of transition to a new maintenance procedure as only after 2002 the performance of older planes got better.

Data Wrangling, Cleaning And Preparation

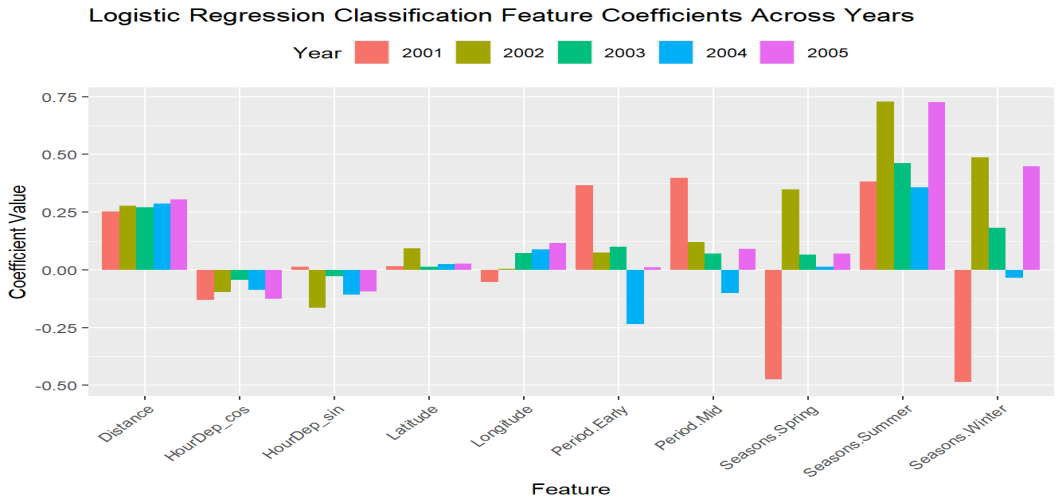
- A SQL query was made to the airline database to pull multiple data columns through an inner join between ontime, planes, airports and carriers on common columns such as tail number, IATA codes, and unique carrier codes, respectively. The condition was that flights had to be not cancelled as the flight cannot be cancelled and diverted at the same time.
- Through the filter, there are no more rows with NA values anymore due to the inner join property
- Checked for the diverted flights ratio and found that we are dealing with severe data imbalance
- Transformed the months and day of the month column extracted through the query into seasons where we assigned breaks and labelled 0 to 2 as winter, 2 to 5 as spring, 5 to 8 as summer, 8 to 11 as fall, 11 to 12 as winter again, simulating seasonal occasions. Day of month was transformed into periods similarly through breaks and labels, 0 to 10 as early, 10 to 20 as middle, and 20 to 31 as late, simulating the period within a month. This is to make it easier on the encoding process
- Transformed time features (CRSDepTime and CRSArrTime) into cyclical features using sine and cosine to simulate a 24-hour clock behavior. This is because I realised that the scheduled departure and arrival times are in an integer format where 0000 is numerically far away from 2359, despite being close in terms of time perspective, which can affect our machine learning model. The final result is HourDep_sin as well as cos, and HourArr_sin as well as cos
- The arrival times are then dropped to avoid multicollinearity, as late departure time leads to late arrival time anyway so its better to just use departure time instead

- Features prepared are then converted into factors and numeric variables where factors are {Seasons and Period} and numerics are {Latitude, Longitude, HourDep_sin, HourDep_cos and Distance}. Diverted is our target, so it has to change into a factor but levelled so that Diverted=1 is considered positive in the model's task

Data Manipulation Strategy

- For each year, we would split the data into 80% training and 20% testing
- The classification task is selected to target diverted flights
- Numerical imputation is used to handle missing numerical values through mean imputation
- Categorical imputation is used to handle missing factor values through mode imputation
- Scaling is done to standardise numeric features
- We do one-hot encoding to handle categorical features in a logistic regression model
- The SMOTE technique is used to create synthetic samples for the minority class to balance the dataset by oversampling(Likebupt, *Smote - Azure Machine Learning* 2024)
- Lastly, class balancing to undersample the majority class with a 0.3 ratio

2c)Fit a logistic regression model for diverted US flights on a year to year basis and extract coefficients for the model



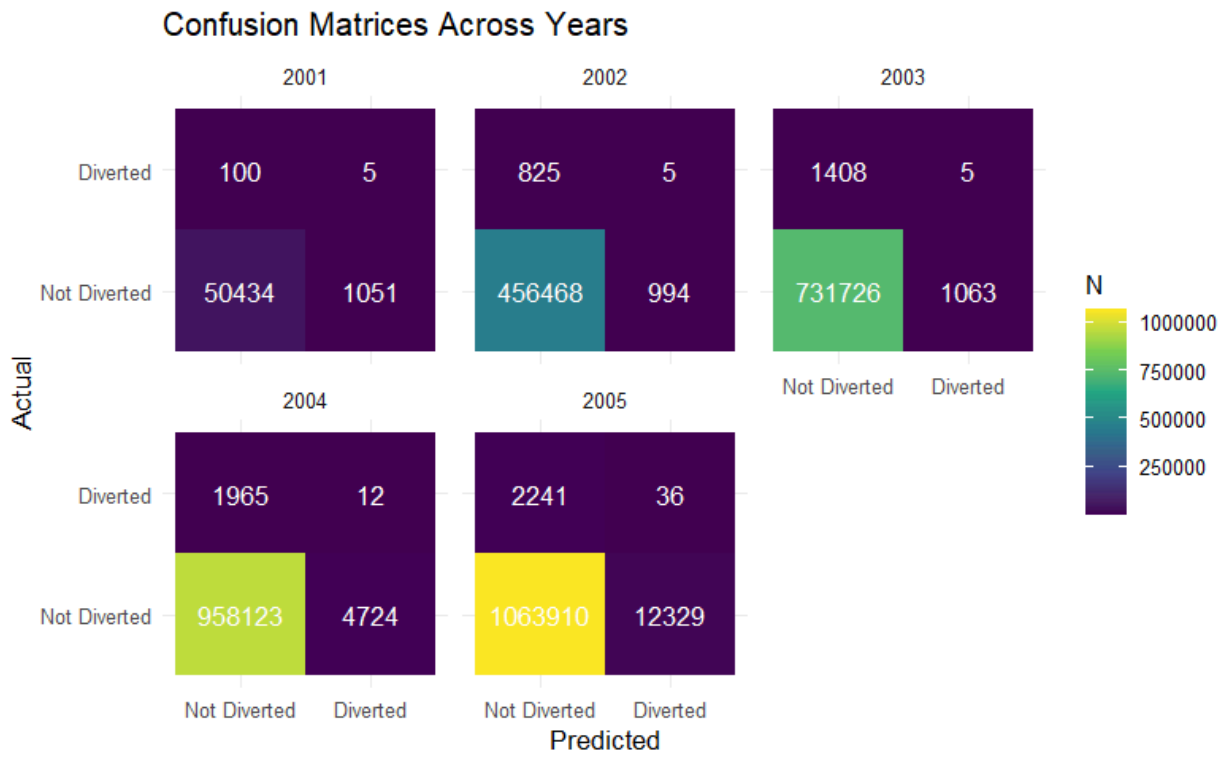
Results discussion and conclusion

- Distance has a positive coefficient, which suggests that longer flights are more likely to be diverted for all years, possibly due to longer exposure to in-flight issues.
- Summer also increases the probability of diversion, perhaps due to higher traffic, as it is a holiday occasion. Similar to winter, there could be overbooked flights as most people want to travel during such occasions and thus lead to diverted flights, or it could be due to severe weather conditions.
- Early and middle periods often have positive coefficients, especially in the earlier years, which suggests that the time of the month can affect diversion patterns, which could be due to a maintenance backlog.
- HourDep_sin and cos have negative coefficients, which means certain times of the day, diversion likelihood is reduced, which is something we found out earlier where early mornings flights are more likely to be on time in the early morning, as well as late evening perhaps leading to lower traffic and thus a reduced diversion risk.

- Latitude and Longitude have little effect on diversion, possibly due to standardised routine and traffic control that coordinates have little to no effect on diversion.

Season and distance are the biggest contributors to a flight being diverted. Confusion matrix is computed to assess the model's competence which can be viewed in the appendix. The model however, is not great as data imbalance is too severe for conventional techniques to solve.

Appendix


Confusion Matrices Across Years
2001
2002
2003
2004
2005
Diverted
Not Diverted
100
5
50434
1051
825
5
456468
994
1408
5
731726
1063
1965
12
958123
4724
2241
36
1063910
12329
Actual
Not Diverted
Diverted
Predicted
N
1000000
750000
500000
250000

References

1. ***North America on-time performance data for airlines and airports: OTP Flight Data: OAG*** **(no date)** ***oag_white*****. Available at: https://www.oag.com/nam-on-time-performance-data (Accessed: 03 April 2025).**
2. ***Correlation (Pearson, Kendall, Spearman)*** **(2024)** ***Statistics Solutions*****. Available at: https://www.statisticssolutions.com/free-resources/directory-of-statistical-analyses/correlation-pearson-kendall-spearman/ (Accessed: 03 April 2025).**

3. **Likebupt (2024a)** ***Smote - Azure Machine Learning*****,** ***Azure Machine Learning | Microsoft Learn*****. Available at: https://learn.microsoft.com/en-us/azure/machine-learning/component-reference/smote?view=azureml-api-2 (Accessed: 04 April 2025).**