Airline Flight Delay Prediction Using Machine Learning Models

DA5030

Janiel Thompson

Spring 2024

Introduction

Within the transportation industry, the airline industry is one of the most rapidly growing sectors and the size of the global market was estimated at 814.5 billion US dollars in 2023[1]. Flight delays negatively impact the industry causing significant financial losses and customer dissatisfaction[2]. The United States Bureau of Transportation Statistics estimated a 41 billion dollar cost to travelers and the airline industry caused by over 20% of US flights being delayed in 2018[3]. Reducing the losses and negative economic impact caused by flight delays can be achieved by improving airline operations and passenger satisfaction, and predicting flight delays is one step that can be taken to do so[4]. The goal of this project is to utilize machine learning algorithms to predict whether a flight will be delayed and accuracy of 70% or greater will be considered a successful predictive outcome.

Data Exploration

The data for this project is from the Airlines Delay Kaggle dataset[5] contributed by Ulrik Pedersen.

Data Structure

There are 539,382 observations of 8 features in the dataset. Most features are numerical, with *Airline*, *AirportTo*, and *AirportFrom* being categorical.

```
'data.frame':   539382 obs. of  8 variables:
 $ Flight     : num  2313 6948 1247 31 563 ...
 $ Time       : num  1296 360 1170 1410 692 ...
 $ Length     : num  141 146 143 344 98 60 239 80 105 108 ...
 $ Airline    : chr  "DL" "OO" "B6" "US" ...
 $ AirportFrom: chr  "ATL" "COS" "BOS" "OGG" ...
 $ AirportTo  : chr  "HOU" "ORD" "CLT" "PHX" ...
 $ DayOfWeek  : int  1 4 3 6 4 4 4 3 7 3 ...
 $ Class      : int  0 0 0 0 0 0 0 0 0 0 ...
```

Data Summary

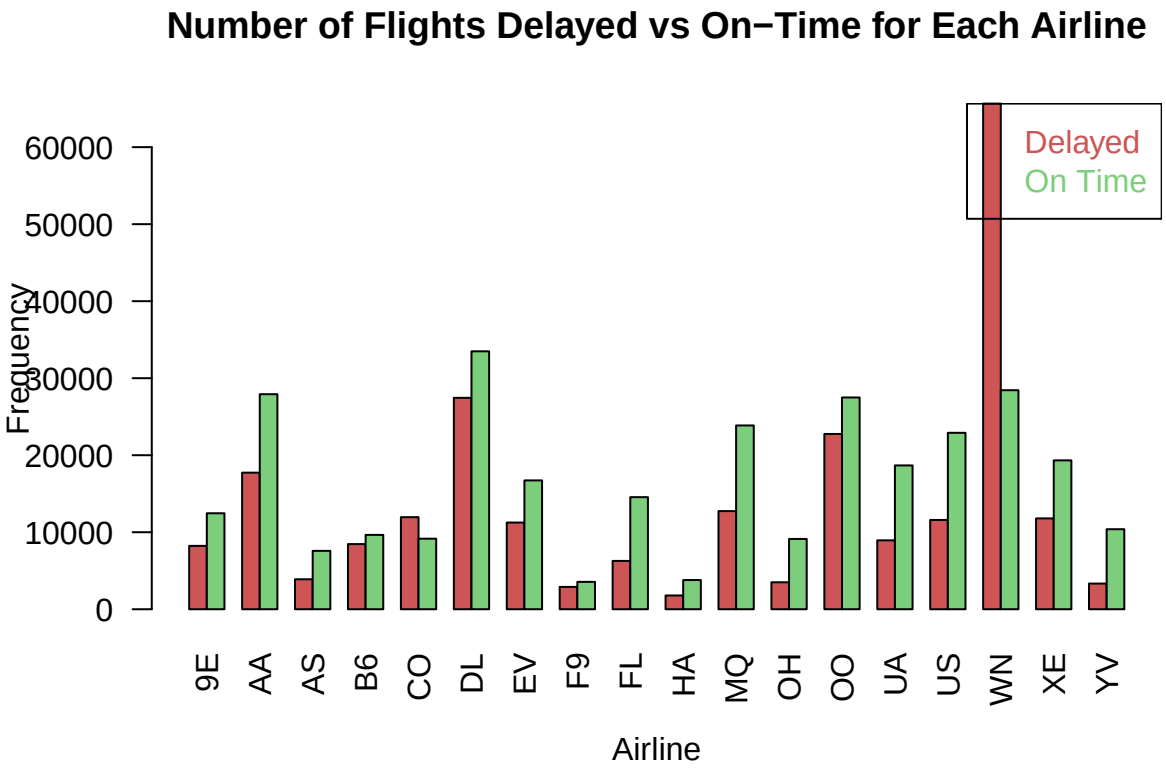
- Flight represents the flight ID for each observation and therefore does not affect whether or not the flight will be delayed
- Time is the time of departure ranging from 10 (12:10 am) to 1439 (11:59 pm)

[1] https://www.statista.com/markets/419/topic/490/aviation/#overview
[2] https://dl.acm.org/doi/fullHtml/10.1145/3497701.3497725
[3] https://medium.com/analytics-vidhya/using-machine-learning-to-predict-flight-delays-e8a50b0bb64c
[4] https://dl.acm.org/doi/fullHtml/10.1145/3497701.3497725
[5] https://www.kaggle.com/datasets/ulrikthygepedersen/airlines-delay/data

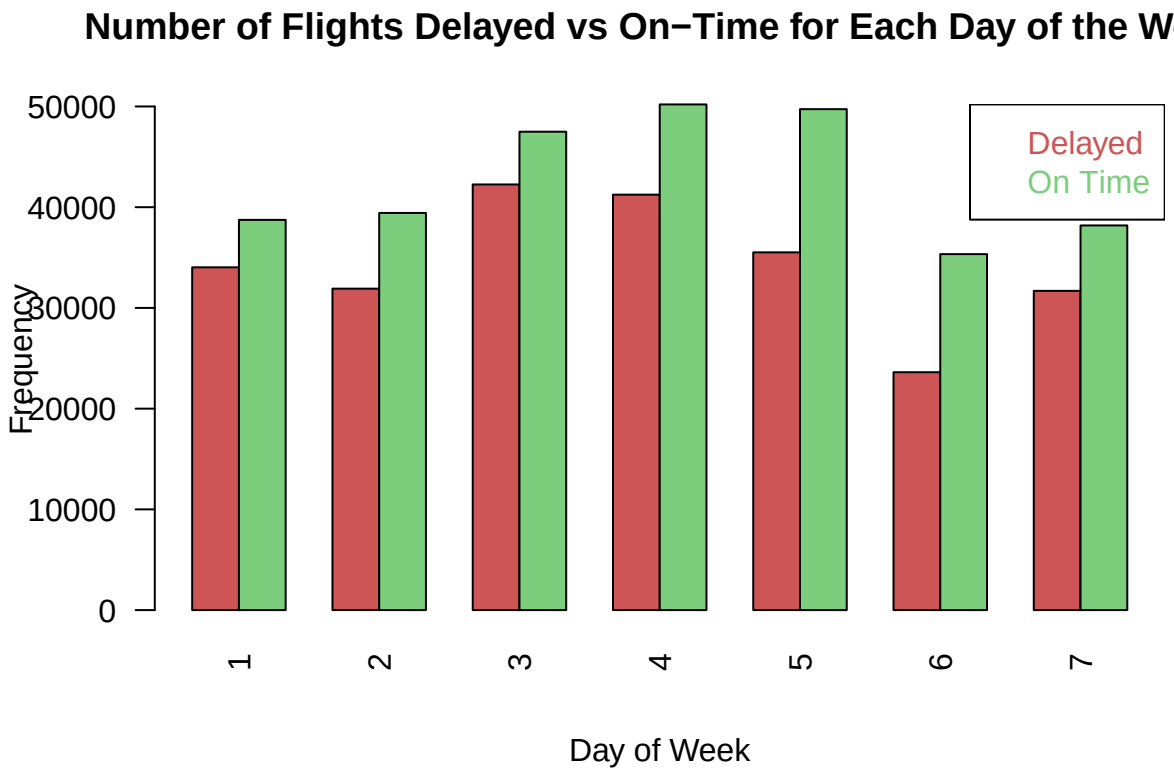
- Length is the duration of the flight in minutes
- Airline is the code for the airline taking flight
- AirportFrom is the three-letter code for the airport from which the airline departed
- AirportTo is the three-letter code for the destination airport
- DayOfWeek is the the day of the week on which the flight took place, ranging from 1-7
- Class is a binary feature where 0 indicates the flight was not delayed and 1 indicates that the flight was delayed. This will be the target variable for classification.

```
    Flight           Time            Length        Airline
Min.   :   1   Min.   :  10.0   Min.   :  0.0   Length:539382
1st Qu.: 712   1st Qu.: 565.0   1st Qu.: 81.0   Class :character
Median :1809   Median : 795.0   Median :115.0   Mode  :character
Mean   :2428   Mean   : 802.7   Mean   :132.2
3rd Qu.:3745   3rd Qu.:1035.0   3rd Qu.:162.0
Max.   :7814   Max.   :1439.0   Max.   :655.0
AirportFrom         AirportTo          DayOfWeek        Class
Length:539382      Length:539382      Min.   :1.00   Min.   :0.0000
Class :character   Class :character   1st Qu.:2.00   1st Qu.:0.0000
Mode  :character   Mode  :character   Median :4.00   Median :0.0000
                                      Mean   :3.93   Mean   :0.4454
                                      3rd Qu.:5.00   3rd Qu.:1.0000
                                      Max.   :7.00   Max.   :1.0000
```

Exploratory Data Plots



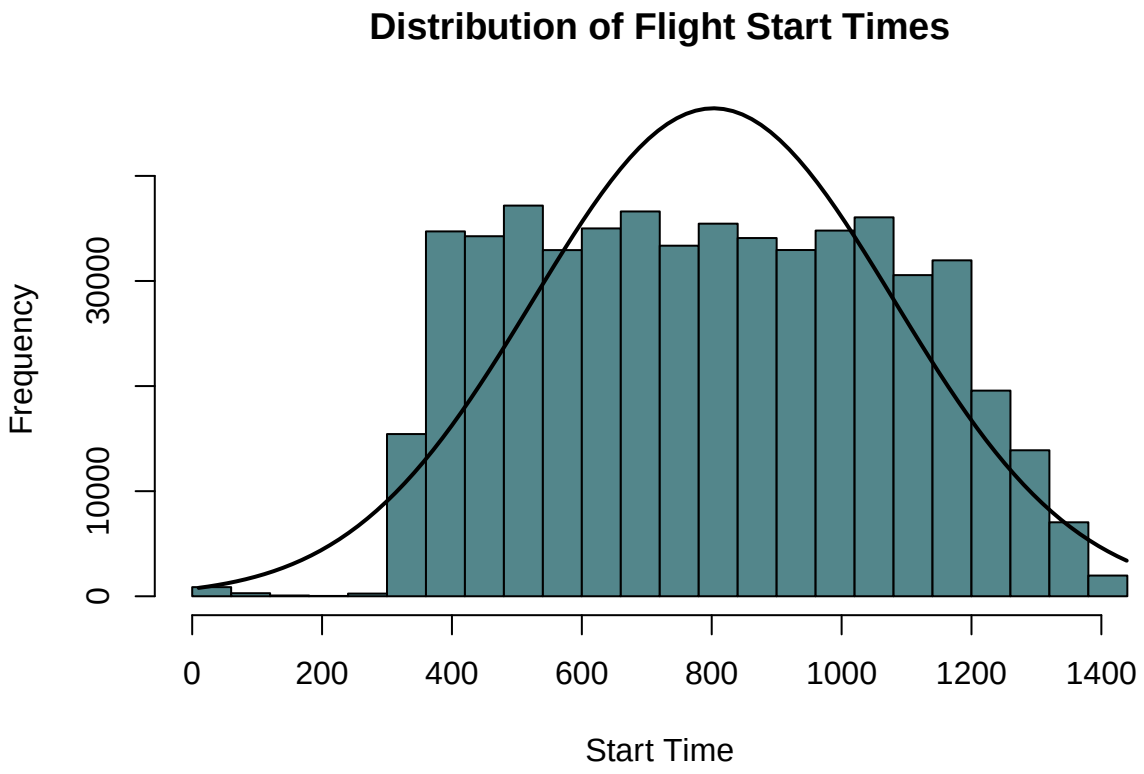
Most airlines had more flights leaving on time than being delayed, but not by a wide margin. Airline 'WN' had the highest number of delayed flights out of all airlines and was more than twice the number of their on-time flights. Having such a large number of delayed flights attributed to a single airline may affect the performance of the classification models.



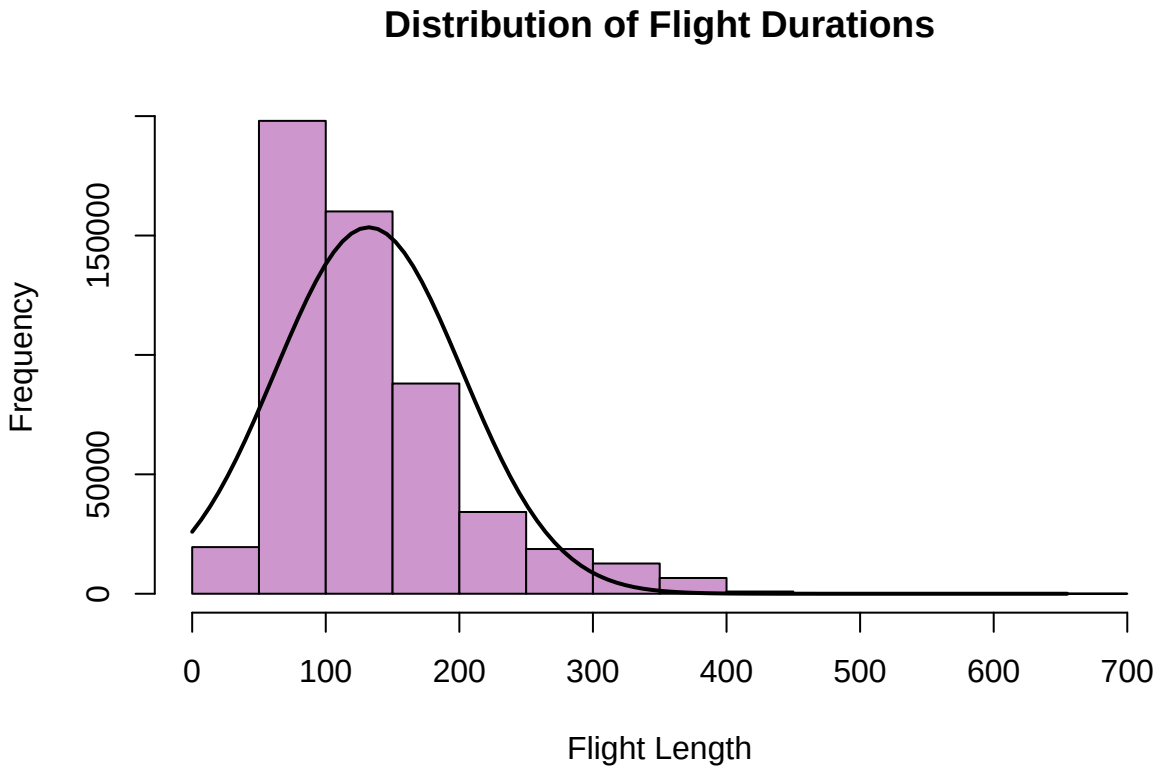
For all days of the week, the number of on-time flight marginally exceeds the number of delayed flights. The least number of flights take place on day 6.

Outlier Detection

For this dataset, any departure time or flight length more than 3 standard deviations away from the mean will be considered an outlier. There are 0 observations that are outliers with respect to time of departure, and 9910 observations that are outliers with respect to flight length. The distribution of start time and flight length are shown below.



All the data points for start time fall within the normal curve overlay and there is no skewness to the distribution. Each bar in the histogram represents one hour out of the day.



The distribution of flight durations is right-skewed and outlying flight lengths, outside the normal curve, are ~350 minutes and greater.

Data Shaping

Handling Missing Data

The first step of data preparation will be removing missing or invalid data. The *is.na()* function will be used to find missing values.

No missing values were found in the dataset so missing value imputation is not required. However, the output of the summary function above shows that the lowest flight length in the dataset is 0 minutes. Since a flight length of zero isn't possible, these can be treated as missing values and removed since imputation is not appropriate without industry knowledge.

Having removed rows where Length = 0, 539,378 observations remain.

Training/Validation Split

Prior to processing, the dataset will be split into a training set and a validation set to prevent data leakage. The training set will be composed of a random sampling of 80% of the dataset. The remaining observations will be placed into the validation set.

The training set consists of 431,502 observations, while the validation set contains 107,876 observations.

Log-Transformation and Normalization

The outliers for flight length range from 27 on the low end to 655 on the high end. Since outliers comprise 2% of the Length observations, the Length feature will be log-transformed to preserve the entire dataset and reduce the impact of outliers. There are no outliers for Time, but since the range is quite large, Min-Max normalization will be applied to the Time feature. This is to ensure the kNN algorithm is not impacted heavily by the scale. The normalization formula is:

$$\frac{x_i - min(x)}{max(x) - min(x)}$$

Encoding Categorical Variables

Since kNN is one of the algorithms that will be used for classification, encoding categorical variables is necessary. The categorical variables in the dataset are Airline, AirportFrom and AirportTo. Count Encoding will be applied to all 3 variables and they will be subsequently normalized using Min-Max Normalization.

Taking a look at the training and validation sets we can see a similar spread of data for all variables below.

Training Set Summary

```
      Time             Length        Airline           AirportFrom
 Min.   :0.0000   Min.   :3.135   Length:431502      Length:431502
 1st Qu.:0.3884   1st Qu.:4.394   Class :character   Class :character
 Median :0.5493   Median :4.745   Mode  :character   Mode  :character
 Mean   :0.5548   Mean   :4.762
 3rd Qu.:0.7173   3rd Qu.:5.088
 Max.   :1.0000   Max.   :6.485
  AirportTo           DayOfWeek         Class         AirlineEncode
 Length:431502      Min.   :1.00   Min.   :0.0000   Min.   :0.0000
 Class :character   1st Qu.:2.00   1st Qu.:0.0000   1st Qu.:0.2495
 Mode  :character   Median :4.00   Median :0.0000   Median :0.3511
                    Mean   :3.93   Mean   :0.4456   Mean   :0.4480
                    3rd Qu.:5.00   3rd Qu.:1.0000   3rd Qu.:0.6256
                    Max.   :7.00   Max.   :1.0000   Max.   :1.0000
 AirportFromEncode AirportToEncode
```

```
 Min.   :0.0000    Min.   :0.00000
 1st Qu.:0.1009    1st Qu.:0.09908
 Median :0.2496    Median :0.24857
 Mean   :0.3071    Mean   :0.30688
 3rd Qu.:0.4541    3rd Qu.:0.45271
 Max.   :1.0000    Max.   :1.00000
```

Validation Set Summary

```
      Time            Length        Airline           AirportFrom
 Min.   :0.0000   Min.   :3.219   Length:107876      Length:107876
 1st Qu.:0.3884   1st Qu.:4.394   Class :character   Class :character
 Median :0.5493   Median :4.745   Mode  :character   Mode  :character
 Mean   :0.5543   Mean   :4.761
 3rd Qu.:0.7173   3rd Qu.:5.081
 Max.   :1.0000   Max.   :6.485
  AirportTo           DayOfWeek         Class       AirlineEncode
 Length:107876      Min.   :1.000   Min.   :0.000   Min.   :0.0000
 Class :character   1st Qu.:2.000   1st Qu.:0.000   1st Qu.:0.2469
 Mode  :character   Median :4.000   Median :0.000   Median :0.3480
                    Mean   :3.927   Mean   :0.445   Mean   :0.4492
                    3rd Qu.:5.000   3rd Qu.:1.000   3rd Qu.:0.6246
                    Max.   :7.000   Max.   :1.000   Max.   :1.0000
 AirportFromEncode AirportToEncode
 Min.   :0.0000    Min.   :0.0000
 1st Qu.:0.1017    1st Qu.:0.1014
 Median :0.2390    Median :0.2412
 Mean   :0.3053    Mean   :0.3070
 3rd Qu.:0.4412    3rd Qu.:0.4474
 Max.   :1.0000    Max.   :1.0000
```

Model Construction

Before constructing the models, it's important to know whether the proportion of each classification is comparable between the original dataset and the training/validation sets.

```
                   0         1
Original   0.5545554 0.4454446
Training   0.5544493 0.4455507
Validation 0.5549798 0.4450202
```

From the table above, we can see that ~55% of the observations in all 3 datasets belong to the negative class (not delayed) and ~45% to the positive class (delayed).

Airline Proportions

```
                   9E         AA         AS         B6         CO         DL
Training   0.03858151 0.08485013 0.02141357 0.03345060 0.03901025 0.1129334
Validation 0.03743187 0.08382773 0.02068115 0.03408543 0.03972153 0.1131762
                   EV         F9         FL         HA         MQ         OH
Training   0.05204147 0.01197909 0.03864872 0.0104171 0.06795102 0.02338807
Validation 0.05123475 0.01190255 0.03847010 0.0100393 0.06751270 0.02352701
                   OO         UA         US         WN         XE         YV
Training   0.09275044 0.05130220 0.06401129 0.1742750 0.05766370 0.02533244
Validation 0.09484964 0.05081761 0.06376766 0.1751733 0.05788127 0.02590011
```

Day of Week Proportions

```
                  1         2         3         4         5         6
Training   0.1347271 0.1322798 0.1661360 0.1699111 0.1583979 0.1088129
Validation 0.1356558 0.1321981 0.1673959 0.1680448 0.1566428 0.1112481
                  7
Training   0.1297352
Validation 0.1288146
```

Similar to the proportions of the target variable in each dataset, the proportion of each airline and day of week is comparable between the training and validation set.

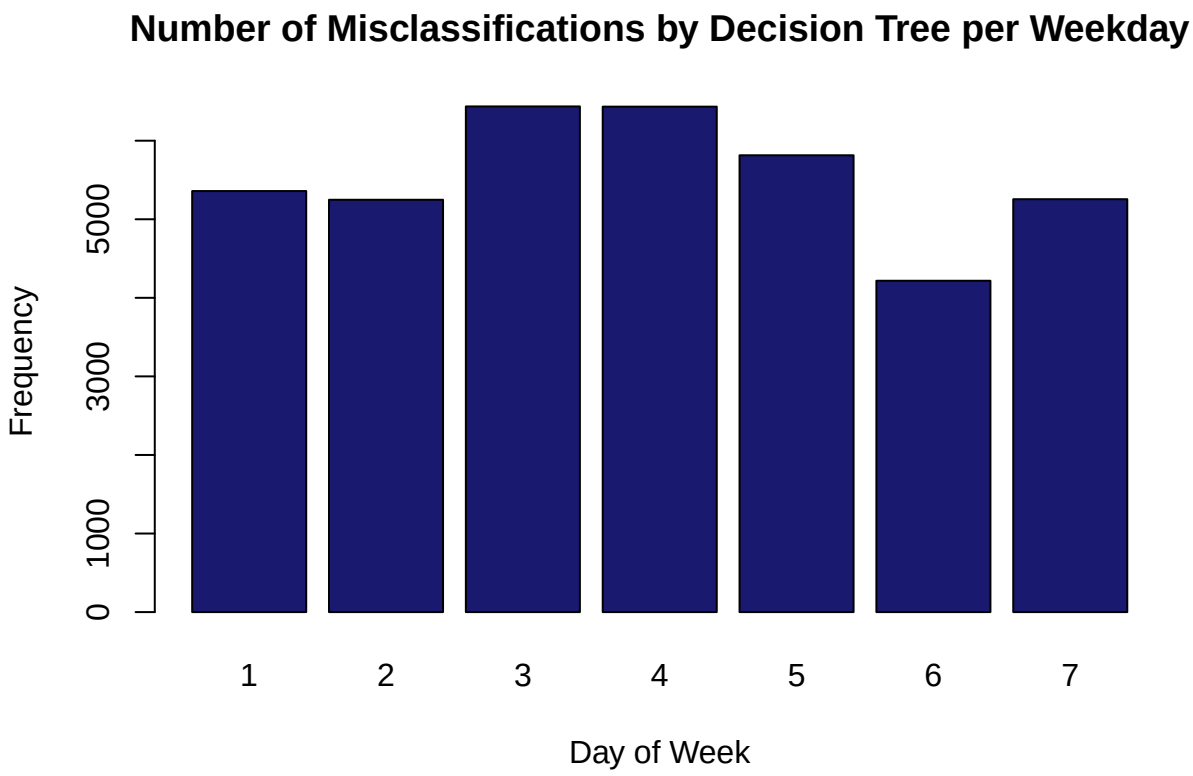
Decision Trees

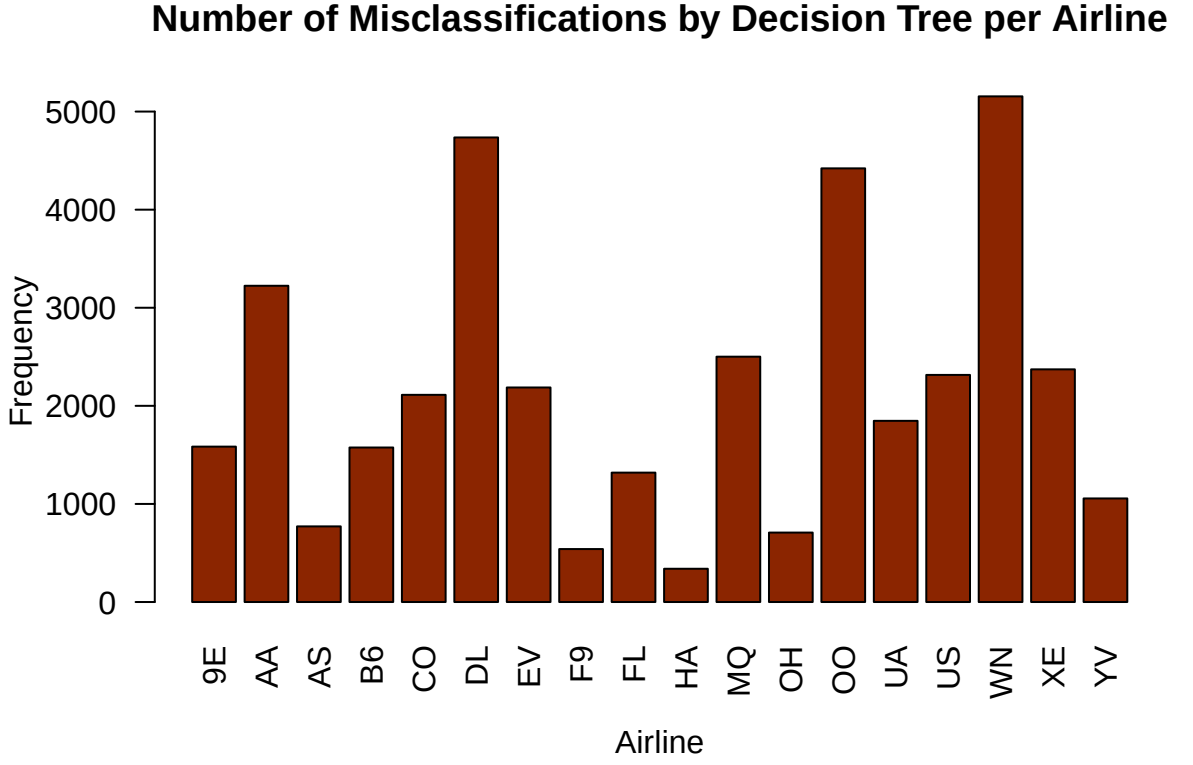
The **C50** package will be used to construct the decision trees for classifications. Class will be predicted as a function of all other features.

```
Confusion Matrix and Statistics

         Actual
Predicted     0     1
        0 47052 25945
        1 12817 22062

               Accuracy : 0.6407
                 95% CI : (0.6378, 0.6435)
    No Information Rate : 0.555
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.2523

 Mcnemar's Test P-Value : < 2.2e-16

              Precision : 0.6325
                 Recall : 0.4596
                     F1 : 0.5323
             Prevalence : 0.4450
         Detection Rate : 0.2045
   Detection Prevalence : 0.3233
      Balanced Accuracy : 0.6227

       'Positive' Class : 1
```

The confusion matrix above shows that more false negatives exceeded false positives by more than 2X.

Failure Analysis The performance metrics for the model are not high. It's difficult to establish what caused poor performance since the training and validations sets have similar proportions of each variable. This would have ideally resulted in a properly trained model and a validation set that doesn't have test cases with never-before-seen values. The low performance of the model could likely be due to the fact that each day of the week and each airline, for the most part, have a roughly even split of on-time and delayed flights. The delays are more likely due to factors not accounted for in the dataset.





As shown in the plots above, there is little variation in the number of incorrect predictions for each day of the week. The variation in incorrect classifications for airline is similar to the distribution of the total number of flights for each airline.

C50 Hyperparameter Tuning

To improve the performance of the model, it will be boosted by setting the *trials* argument to 10.

For further optimization, error costs will be applied. For this problem, classifying a delayed flight (1, positive) as not delayed (0, negative) will be given a higher error cost.

Logistic Regression

A binomial logistic regression model will be used since the target variable can be one of two outcomes, 0 or 1. For the first iteration, a threshold probability of 0.5 will be used to classify flights in the positive class.

A similar logistic regression model will be constructed with a threshold of 0.56.

kNN

The *knn()* function from **class** package will be used to build a kNN model.

For the first iteration, a k value of 85 will be used for the algorithm.

kNN Hyperparameter Tuning

A similar model will be constructed with a higher k of 100.

Ensemble Model

Each of the 3 types of models with the highest accuracy will be used to build an ensemble learner. A positive classification will be made if at least 2 of the individual models make a positive prediction. For the Decision Tree and kNN models, since the target variable is a factor with 2 levels, the levels are 1 and 2. Whereas the predictions for the Logistic Regression model are either 0 or 1. Therefore, a sum of at least 3 is required for the ensemble model to predict that a flight will be delayed.

Model Comparison

A confusion matrix was constructed for each model's predictions using the **caret** package. A table displaying evaluation metrics is shown below.

	Accuracy	Kappa	F1	Precision	Recall
C50 Default	0.641	0.252	0.532	0.633	0.460
C50 10 Trials	0.645	0.256	0.518	0.655	0.428
C50 w/ Error Cost	0.646	0.259	0.523	0.654	0.435
Logistic Regression	0.625	0.220	0.515	0.606	0.447
Logistic Regression w/ 0.56 Threshold	0.630	0.210	0.442	0.671	0.330
kNN w/ k=85	0.645	0.260	0.532	0.645	0.453
kNN w/ k=100	0.646	0.262	0.531	0.648	0.449
Ensemble	0.631	0.255	0.589	0.585	0.592

The Logistic Regression model with 0.56 probability threshold yielded highest precision but lowest Kappa and lowest recall. Only 1/3 of delayed flights were predicted to be delayed by this model, but of those predictions ~67% were truly delayed. The accuracy of this model was comparable to the others.

The Decision Tree model with 11 trials and error costs had the same accuracy as the kNN model with a k of 100, which was the highest overall at 64.6%. All other performance metrics for these two models are similar, with Kappa, F1 and recall being marginally higher for the kNN model. For a classification problem such as this, recall would be more important than precision because we want to maximize the percentage of delayed flights predicted as delayed. Therefore, kNN (k=100) outperformed the Decision Tree with error costs.

The Ensemble model had the highest recall but the lowest precision. That is, it was able to identify the most delayed flights out of all delayed flights, but made more false positive predictions (predicted non-delayed flights as delayed) than all the other models.

The performance metrics are all similar for all the models, so it can be concluded that the failures are due to the same reason. This indicates that additional variables are required to more accurately predict flight delays. Factors such as weather, staffing resources, air traffic, among others, factor in to a flight being delayed.