Predicting Flight Delays Using Machine Learning

Bobga-Herman Gwanvoma

2025-03-31

```r
# Step 2: Extract the ZIP Files
zip_file_1 <- "C:/Users/bobi/Documents/DSC 680/ot_delaycause1_DL (1).zip"
zip_file_2 <- "C:/Users/bobi/Documents/DSC 680/flights_sample_3m.csv (2).zip"

# Extract the first zip file
unzip(zip_file_1, exdir = "C:/Users/bobi/Documents/DSC 680/ot_delaycause1_DL_1")

# Extract the second zip file
unzip(zip_file_2, exdir = "C:/Users/bobi/Documents/DSC 680/flights_sample_3m_2")
```

Load the Data

```r
# Step 3: Load the Data
delay_cause_data <- read.csv("C:/Users/bobi/Documents/DSC 680/ot_delaycause1_DL_1/Airline_Delay_Cause.cs
flights_sample_data <- read.csv("C:/Users/bobi/Documents/DSC 680/flights_sample_3m_2/flights_sample_3m.c
```

Data cleanning

```r
# Step 4: Create a Date Column for delay_cause_data
delay_cause_data$Date <- as.Date(paste(delay_cause_data$year, delay_cause_data$month, "01", sep = "-"),
```

```r
# Step 5: Remove Duplicates
# Remove duplicates based on 'FL_NUMBER' and 'AIRLINE_CODE' for flights_sample_data
flights_sample_data_unique <- flights_sample_data %>%
  distinct(FL_NUMBER, AIRLINE_CODE, .keep_all = TRUE)

# Remove duplicates based on 'carrier' for delay_cause_data
delay_cause_data_unique <- delay_cause_data %>%
  distinct(carrier, .keep_all = TRUE)
```

```r
# Step 6: Check and Convert Data Types
# Ensure both 'AIRLINE_CODE' and 'carrier' are character type
flights_sample_data_unique$AIRLINE_CODE <- as.character(flights_sample_data_unique$AIRLINE_CODE)
delay_cause_data_unique$carrier <- as.character(delay_cause_data_unique$carrier)
```

Merge the Datasets

```r
# Step 7: Merge the Datasets
# Merge datasets using 'AIRLINE_CODE' and 'carrier'
combined_data <- merge(flights_sample_data_unique, delay_cause_data_unique, by.x = "AIRLINE_CODE", by.y
```

Inspect the Combined Data

```r
# Step 8: Inspect the Combined Data
head(combined_data)
```

```
##   AIRLINE_CODE    FL_DATE           AIRLINE           AIRLINE_DOT DOT_CODE
## 1           9E 2022-03-16 Endeavor Air Inc. Endeavor Air Inc.: 9E    20363
## 2           9E 2021-06-17 Endeavor Air Inc. Endeavor Air Inc.: 9E    20363
## 3           9E 2022-10-04 Endeavor Air Inc. Endeavor Air Inc.: 9E    20363
## 4           9E 2019-10-28 Endeavor Air Inc. Endeavor Air Inc.: 9E    20363
## 5           9E 2019-06-11 Endeavor Air Inc. Endeavor Air Inc.: 9E    20363
## 6           9E 2021-11-28 Endeavor Air Inc. Endeavor Air Inc.: 9E    20363
##   FL_NUMBER ORIGIN                  ORIGIN_CITY DEST        DEST_CITY
## 1      4644    JFK                 New York, NY  CHS   Charleston, SC
## 2      4638    MBS Saginaw/Bay City/Midland, MI  DTW      Detroit, MI
## 3      5104    DTW                  Detroit, MI  IND Indianapolis, IN
## 4      3442    LAN                  Lansing, MI  MSP  Minneapolis, MN
## 5      3368    JFK                 New York, NY  RIC     Richmond, VA
## 6      5520    ATL                  Atlanta, GA  EVV   Evansville, IN
##   CRS_DEP_TIME DEP_TIME DEP_DELAY TAXI_OUT WHEELS_OFF WHEELS_ON TAXI_IN
## 1          800      751        -9       31        822      1008       4
## 2         1405     1400        -5       15       1415      1438      11
## 3         1010     1005        -5       10       1015      1056       4
## 4          700      701         1       14        715       734       5
## 5         1459     1602        63       23       1625      1720       5
## 6         2127     2122        -5       13       2135      2133       3
##   CRS_ARR_TIME ARR_TIME ARR_DELAY CANCELLED CANCELLATION_CODE DIVERTED
## 1         1017     1012        -5         0                          0
## 2         1459     1449       -10         0                          0
## 3         1120     1100       -20         0                          0
## 4          746      739        -7         0                          0
## 5         1651     1725        34         0                          0
## 6         2149     2136       -13         0                          0
##   CRS_ELAPSED_TIME ELAPSED_TIME AIR_TIME DISTANCE DELAY_DUE_CARRIER
## 1              137          141      106      636               NA
## 2               54           49       23       98               NA
## 3               70           55       41      231               NA
## 4              106           98       79      455               NA
## 5              112           83       55      288                0
## 6               82           74       58      350               NA
##   DELAY_DUE_WEATHER DELAY_DUE_NAS DELAY_DUE_SECURITY DELAY_DUE_LATE_AIRCRAFT
## 1                NA           NA                 NA                      NA
## 2                NA           NA                 NA                      NA
## 3                NA           NA                 NA                      NA
## 4                NA           NA                 NA                      NA
## 5                 1            0                  0                      33
## 6                NA           NA                 NA                      NA
##   year month      carrier_name airport
## 1 2023    12 Endeavor Air Inc.     ABE
## 2 2023    12 Endeavor Air Inc.     ABE
## 3 2023    12 Endeavor Air Inc.     ABE
## 4 2023    12 Endeavor Air Inc.     ABE
## 5 2023    12 Endeavor Air Inc.     ABE
## 6 2023    12 Endeavor Air Inc.     ABE
##                                                   airport_name arr_flights
## 1 Allentown/Bethlehem/Easton, PA: Lehigh Valley International          72
## 2 Allentown/Bethlehem/Easton, PA: Lehigh Valley International          72
## 3 Allentown/Bethlehem/Easton, PA: Lehigh Valley International          72
## 4 Allentown/Bethlehem/Easton, PA: Lehigh Valley International          72
## 5 Allentown/Bethlehem/Easton, PA: Lehigh Valley International          72
## 6 Allentown/Bethlehem/Easton, PA: Lehigh Valley International          72
##   arr_del15 carrier_ct weather_ct nas_ct security_ct late_aircraft_ct
## 1         5       2.46          1   0.73           0             0.81
## 2         5       2.46          1   0.73           0             0.81
## 3         5       2.46          1   0.73           0             0.81
## 4         5       2.46          1   0.73           0             0.81
## 5         5       2.46          1   0.73           0             0.81
## 6         5       2.46          1   0.73           0             0.81
##   arr_cancelled arr_diverted arr_delay carrier_delay weather_delay nas_delay
## 1             0            0       672            61           574        20
## 2             0            0       672            61           574        20
## 3             0            0       672            61           574        20
## 4             0            0       672            61           574        20
## 5             0            0       672            61           574        20
## 6             0            0       672            61           574        20
##   security_delay late_aircraft_delay       Date
## 1              0                  17 2023-12-01
## 2              0                  17 2023-12-01
## 3              0                  17 2023-12-01
## 4              0                  17 2023-12-01
## 5              0                  17 2023-12-01
## 6              0                  17 2023-12-01
```

Save the Combined Data to a New CSV

```r
# Load the cleaned dataset
combined_dt <- read.csv("C:/Users/bobi/Documents/DSC 680/combined_flight_data.csv")

# Check the first few rows of the dataset
head(combined_dt)
```

```
##   AIRLINE_CODE    FL_DATE           AIRLINE           AIRLINE_DOT DOT_CODE
## 1           9E 2022-03-16 Endeavor Air Inc. Endeavor Air Inc.: 9E    20363
## 2           9E 2021-06-17 Endeavor Air Inc. Endeavor Air Inc.: 9E    20363
## 3           9E 2022-10-04 Endeavor Air Inc. Endeavor Air Inc.: 9E    20363
## 4           9E 2019-10-28 Endeavor Air Inc. Endeavor Air Inc.: 9E    20363
## 5           9E 2019-06-11 Endeavor Air Inc. Endeavor Air Inc.: 9E    20363
## 6           9E 2021-11-28 Endeavor Air Inc. Endeavor Air Inc.: 9E    20363
##   FL_NUMBER ORIGIN                  ORIGIN_CITY DEST        DEST_CITY
## 1      4644    JFK                 New York, NY  CHS   Charleston, SC
## 2      4638    MBS Saginaw/Bay City/Midland, MI  DTW      Detroit, MI
## 3      5104    DTW                  Detroit, MI  IND Indianapolis, IN
## 4      3442    LAN                  Lansing, MI  MSP  Minneapolis, MN
## 5      3368    JFK                 New York, NY  RIC     Richmond, VA
## 6      5520    ATL                  Atlanta, GA  EVV   Evansville, IN
##   CRS_DEP_TIME DEP_TIME DEP_DELAY TAXI_OUT WHEELS_OFF WHEELS_ON TAXI_IN
## 1          800      751        -9       31        822      1008       4
## 2         1405     1400        -5       15       1415      1438      11
## 3         1010     1005        -5       10       1015      1056       4
## 4          700      701         1       14        715       734       5
## 5         1459     1602        63       23       1625      1720       5
## 6         2127     2122        -5       13       2135      2133       3
##   CRS_ARR_TIME ARR_TIME ARR_DELAY CANCELLED CANCELLATION_CODE DIVERTED
## 1         1017     1012        -5         0                          0
## 2         1459     1449       -10         0                          0
## 3         1120     1100       -20         0                          0
## 4          746      739        -7         0                          0
## 5         1651     1725        34         0                          0
## 6         2149     2136       -13         0                          0
##   CRS_ELAPSED_TIME ELAPSED_TIME AIR_TIME DISTANCE DELAY_DUE_CARRIER
## 1              137          141      106      636                NA
## 2               54           49       23       98                NA
## 3               70           55       41      231                NA
## 4              106           98       79      455                NA
## 5              112           83       55      288                 0
## 6               82           74       58      350                NA
##   DELAY_DUE_WEATHER DELAY_DUE_NAS DELAY_DUE_SECURITY DELAY_DUE_LATE_AIRCRAFT
## 1                NA            NA                 NA                      NA
## 2                NA            NA                 NA                      NA
## 3                NA            NA                 NA                      NA
## 4                NA            NA                 NA                      NA
## 5                 1             0                  0                      33
## 6                NA            NA                 NA                      NA
##   year month       carrier_name airport
## 1 2023    12 Endeavor Air Inc.     ABE
## 2 2023    12 Endeavor Air Inc.     ABE
## 3 2023    12 Endeavor Air Inc.     ABE
## 4 2023    12 Endeavor Air Inc.     ABE
## 5 2023    12 Endeavor Air Inc.     ABE
## 6 2023    12 Endeavor Air Inc.     ABE
##                                                   airport_name arr_flights
## 1 Allentown/Bethlehem/Easton, PA: Lehigh Valley International          72
## 2 Allentown/Bethlehem/Easton, PA: Lehigh Valley International          72
## 3 Allentown/Bethlehem/Easton, PA: Lehigh Valley International          72
## 4 Allentown/Bethlehem/Easton, PA: Lehigh Valley International          72
## 5 Allentown/Bethlehem/Easton, PA: Lehigh Valley International          72
## 6 Allentown/Bethlehem/Easton, PA: Lehigh Valley International          72
##   arr_del15 carrier_ct weather_ct nas_ct security_ct late_aircraft_ct
## 1         5       2.46          1   0.73           0             0.81
## 2         5       2.46          1   0.73           0             0.81
## 3         5       2.46          1   0.73           0             0.81
## 4         5       2.46          1   0.73           0             0.81
## 5         5       2.46          1   0.73           0             0.81
## 6         5       2.46          1   0.73           0             0.81
##   arr_cancelled arr_diverted arr_delay carrier_delay weather_delay nas_delay
## 1             0            0       672            61           574        20
## 2             0            0       672            61           574        20
## 3             0            0       672            61           574        20
## 4             0            0       672            61           574        20
## 5             0            0       672            61           574        20
## 6             0            0       672            61           574        20
##   security_delay late_aircraft_delay       Date
## 1              0                  17 2023-12-01
## 2              0                  17 2023-12-01
## 3              0                  17 2023-12-01
## 4              0                  17 2023-12-01
## 5              0                  17 2023-12-01
## 6              0                  17 2023-12-01
```

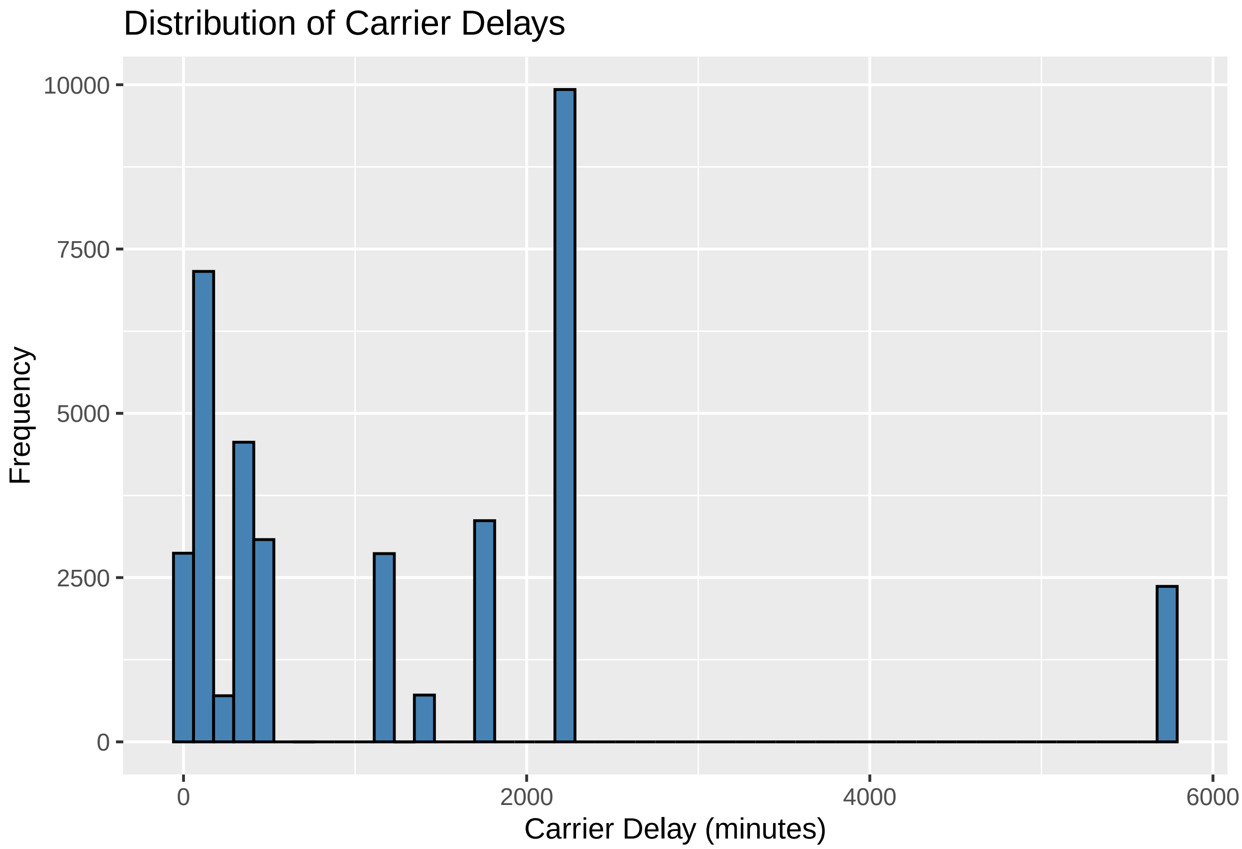
Data Preprocessing

```
# Convert columns to proper types (if needed)
combined_dt$Date <- as.Date(combined_dt$Date)
combined_dt$AIRLINE_CODE <- as.factor(combined_dt$AIRLINE_CODE)
combined_dt$FL_NUMBER <- as.factor(combined_dt$FL_NUMBER)

# Handle missing values: we can impute with the mean for simplicity (or remove rows with NAs)
combined_dt <- combined_dt %>%
  mutate(across(where(is.numeric), ~ ifelse(is.na(.), mean(., na.rm = TRUE), .)))

# Normalize the numerical columns for modeling
combined_dt_scaled <- combined_dt %>%
  mutate(across(c(arr_delay, carrier_delay, weather_delay, nas_delay, security_delay, late_aircraft_dela
                scale, .names = "scaled_{.col}"))

# Inspect the scaled data
head(combined_dt_scaled)
```

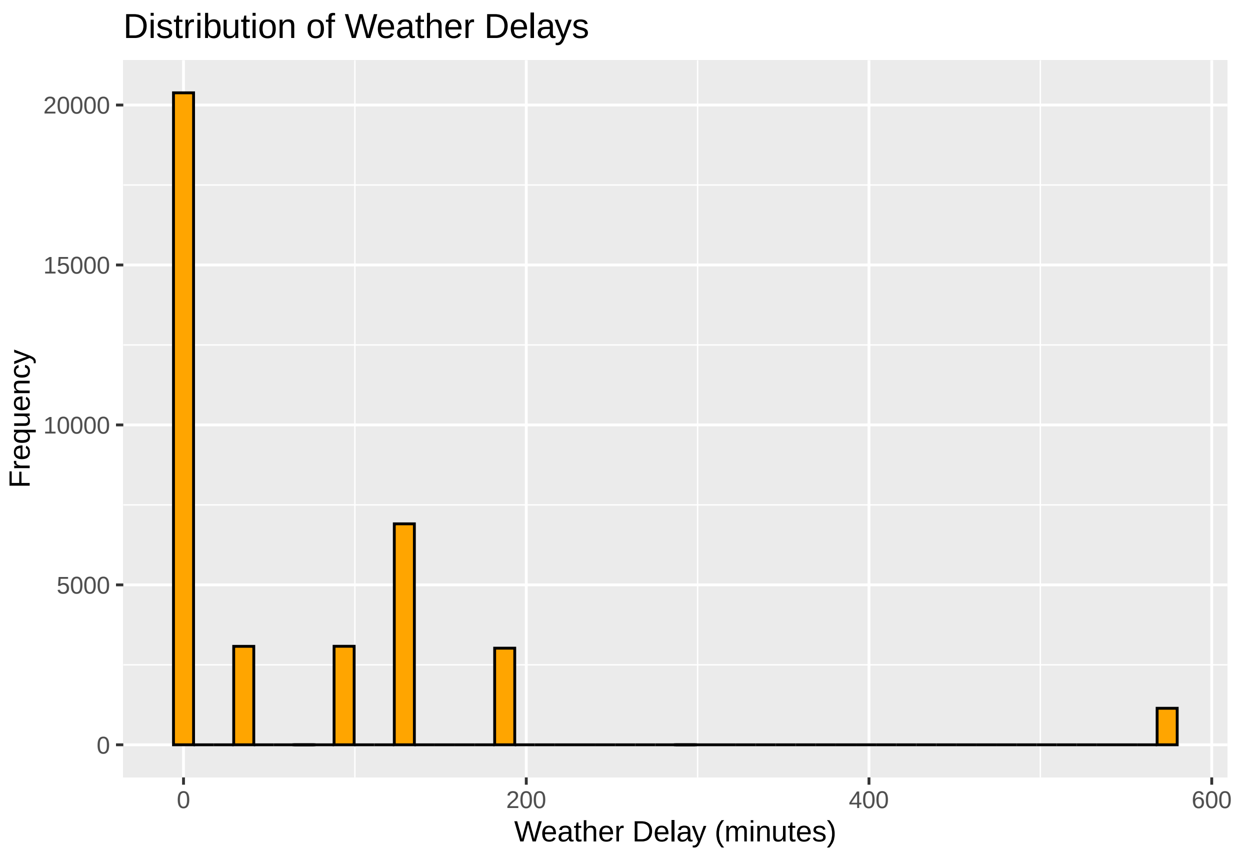
```
##   AIRLINE_CODE    FL_DATE           AIRLINE           AIRLINE_DOT DOT_CODE
## 1           9E 2022-03-16 Endeavor Air Inc. Endeavor Air Inc.: 9E    20363
## 2           9E 2021-06-17 Endeavor Air Inc. Endeavor Air Inc.: 9E    20363
## 3           9E 2022-10-04 Endeavor Air Inc. Endeavor Air Inc.: 9E    20363
## 4           9E 2019-10-28 Endeavor Air Inc. Endeavor Air Inc.: 9E    20363
## 5           9E 2019-06-11 Endeavor Air Inc. Endeavor Air Inc.: 9E    20363
## 6           9E 2021-11-28 Endeavor Air Inc. Endeavor Air Inc.: 9E    20363
##   FL_NUMBER ORIGIN                  ORIGIN_CITY DEST        DEST_CITY
## 1      4644    JFK                 New York, NY  CHS   Charleston, SC
## 2      4638    MBS Saginaw/Bay City/Midland, MI  DTW      Detroit, MI
## 3      5104    DTW                  Detroit, MI  IND Indianapolis, IN
## 4      3442    LAN                  Lansing, MI  MSP  Minneapolis, MN
## 5      3368    JFK                 New York, NY  RIC     Richmond, VA
## 6      5520    ATL                  Atlanta, GA  EVV   Evansville, IN
##   CRS_DEP_TIME DEP_TIME DEP_DELAY TAXI_OUT WHEELS_OFF WHEELS_ON TAXI_IN
## 1          800      751        -9       31        822      1008       4
## 2         1405     1400        -5       15       1415      1438      11
## 3         1010     1005        -5       10       1015      1056       4
## 4          700      701         1       14        715       734       5
## 5         1459     1602        63       23       1625      1720       5
## 6         2127     2122        -5       13       2135      2133       3
##   CRS_ARR_TIME ARR_TIME ARR_DELAY CANCELLED CANCELLATION_CODE DIVERTED
## 1         1017     1012        -5         0                          0
## 2         1459     1449       -10         0                          0
## 3         1120     1100       -20         0                          0
## 4          746      739        -7         0                          0
## 5         1651     1725        34         0                       0
## 6         2149     2136       -13         0                       0
##   CRS_ELAPSED_TIME ELAPSED_TIME AIR_TIME DISTANCE DELAY_DUE_CARRIER
## 1              137          141      106      636          23.95602
## 2               54           49       23       98          23.95602
## 3               70           55       41      231          23.95602
## 4              106           98       79      455          23.95602
## 5              112           83       55      288           0.00000
## 6               82           74       58      350          23.95602
##   DELAY_DUE_WEATHER DELAY_DUE_NAS DELAY_DUE_SECURITY DELAY_DUE_LATE_AIRCRAFT
## 1          4.236307      13.44598          0.1644584                25.45021
## 2          4.236307      13.44598          0.1644584                25.45021
## 3          4.236307      13.44598          0.1644584                25.45021
## 4          4.236307      13.44598          0.1644584                25.45021
## 5          1.000000       0.00000          0.0000000                33.00000
## 6          4.236307      13.44598          0.1644584                25.45021
##   year month      carrier_name airport
## 1 2023    12 Endeavor Air Inc.     ABE
## 2 2023    12 Endeavor Air Inc.     ABE
## 3 2023    12 Endeavor Air Inc.     ABE
## 4 2023    12 Endeavor Air Inc.     ABE
## 5 2023    12 Endeavor Air Inc.     ABE
## 6 2023    12 Endeavor Air Inc.     ABE
##                                                  airport_name arr_flights
## 1 Allentown/Bethlehem/Easton, PA: Lehigh Valley International          72
## 2 Allentown/Bethlehem/Easton, PA: Lehigh Valley International          72
## 3 Allentown/Bethlehem/Easton, PA: Lehigh Valley International          72
## 4 Allentown/Bethlehem/Easton, PA: Lehigh Valley International          72
## 5 Allentown/Bethlehem/Easton, PA: Lehigh Valley International          72
## 6 Allentown/Bethlehem/Easton, PA: Lehigh Valley International          72
##   arr_del15 carrier_ct weather_ct nas_ct security_ct late_aircraft_ct
## 1         5       2.46          1   0.73           0             0.81
## 2         5       2.46          1   0.73           0             0.81
## 3         5       2.46          1   0.73           0             0.81
## 4         5       2.46          1   0.73           0             0.81
## 5         5       2.46          1   0.73           0             0.81
## 6         5       2.46          1   0.73           0             0.81
##   arr_cancelled arr_diverted arr_delay carrier_delay weather_delay nas_delay
## 1             0            0       672            61           574        20
## 2             0            0       672            61           574        20
## 3             0            0       672            61           574        20
## 4             0            0       672            61           574        20
## 5             0            0       672            61           574        20
## 6             0            0       672            61           574        20
##   security_delay late_aircraft_delay       Date scaled_arr_delay
## 1              0                  17 2023-12-01       -0.6200673
## 2              0                  17 2023-12-01       -0.6200673
## 3              0                  17 2023-12-01       -0.6200673
## 4              0                  17 2023-12-01       -0.6200673
## 5              0                  17 2023-12-01       -0.6200673
## 6              0                  17 2023-12-01       -0.6200673
##   scaled_carrier_delay scaled_weather_delay scaled_nas_delay
## 1           -0.8823168             4.572304       -0.6140542
## 2           -0.8823168             4.572304       -0.6140542
## 3           -0.8823168             4.572304       -0.6140542
## 4           -0.8823168             4.572304       -0.6140542
## 5           -0.8823168             4.572304       -0.6140542
## 6           -0.8823168             4.572304       -0.6140542
##   scaled_security_delay scaled_late_aircraft_delay
## 1            -0.4743367                 -0.6632303
## 2            -0.4743367                 -0.6632303
## 3            -0.4743367                 -0.6632303
## 4            -0.4743367                 -0.6632303
## 5            -0.4743367                 -0.6632303
## 6            -0.4743367                 -0.6632303
```

Exploratory Data Analysis (EDA)

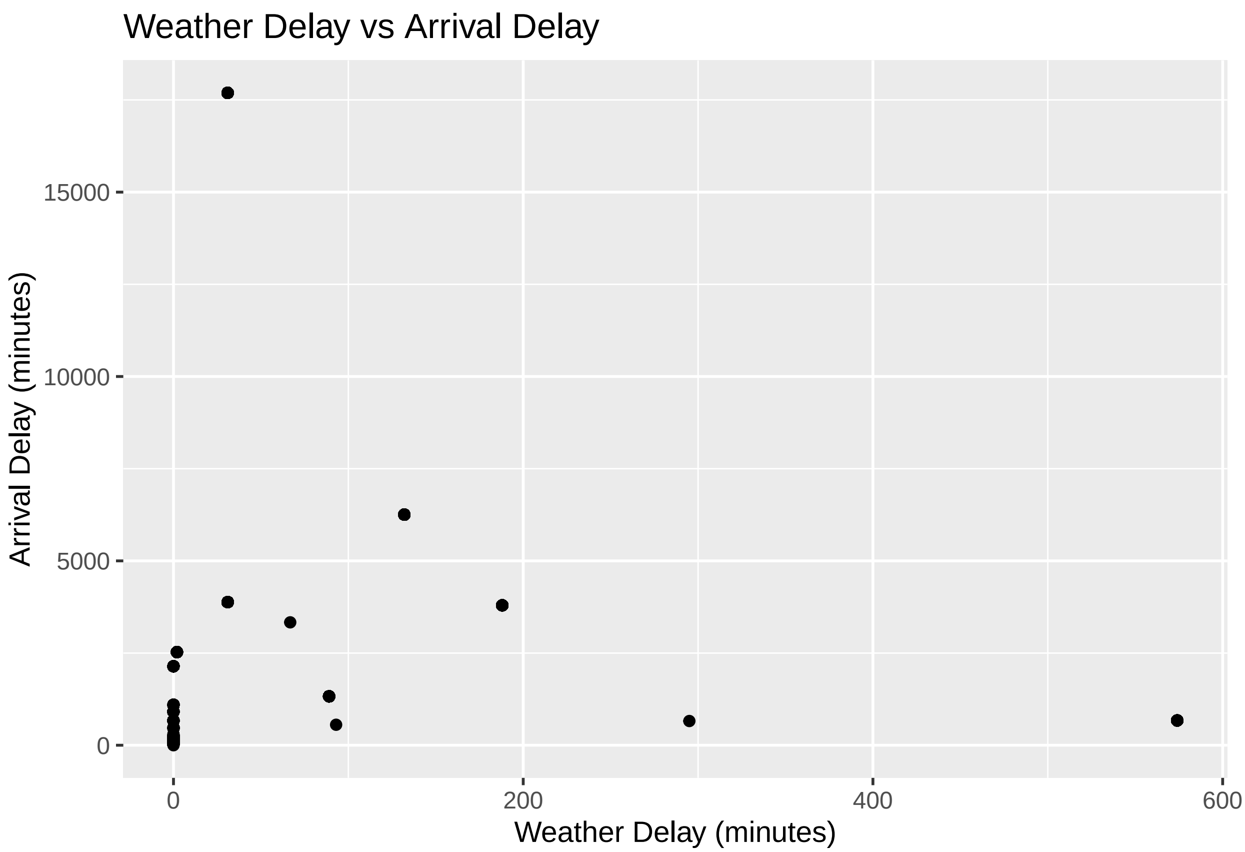
```
# Visualize the distribution of delays by cause
ggplot(combined_dt, aes(x = carrier_delay)) +
  geom_histogram(bins = 50, fill = "steelblue", color = "black") +
  ggtitle("Distribution of Carrier Delays") +
  xlab("Carrier Delay (minutes)") + ylab("Frequency")
```



```
# Visualize the distribution of weather delays
ggplot(combined_dt, aes(x = weather_delay)) +
  geom_histogram(bins = 50, fill = "orange", color = "black") +
  ggtitle("Distribution of Weather Delays") +
  xlab("Weather Delay (minutes)") + ylab("Frequency")
```



```
# Scatter plot of delay causes vs arrival delay
ggplot(combined_dt, aes(x = weather_delay, y = arr_delay)) +
  geom_point() +
  ggtitle("Weather Delay vs Arrival Delay") +
  xlab("Weather Delay (minutes)") + ylab("Arrival Delay (minutes)")
```



Feature Engineering

```
# Create a new feature for delay severity (could be a combination of delays or an indicator)
combined_dt$high_severity_delay <- ifelse(combined_dt$arr_delay > 30, 1, 0)

# Check the new feature
head(combined_dt)
```

```
##   AIRLINE_CODE    FL_DATE           AIRLINE           AIRLINE_DOT DOT_CODE
## 1           9E 2022-03-16 Endeavor Air Inc. Endeavor Air Inc.: 9E    20363
## 2           9E 2021-06-17 Endeavor Air Inc. Endeavor Air Inc.: 9E    20363
## 3           9E 2022-10-04 Endeavor Air Inc. Endeavor Air Inc.: 9E    20363
## 4           9E 2019-10-28 Endeavor Air Inc. Endeavor Air Inc.: 9E    20363
## 5           9E 2019-06-11 Endeavor Air Inc. Endeavor Air Inc.: 9E    20363
## 6           9E 2021-11-28 Endeavor Air Inc. Endeavor Air Inc.: 9E    20363
##   FL_NUMBER ORIGIN                  ORIGIN_CITY DEST        DEST_CITY
## 1      4644    JFK                 New York, NY  CHS   Charleston, SC
## 2      4638    MBS Saginaw/Bay City/Midland, MI  DTW      Detroit, MI
## 3      5104    DTW                  Detroit, MI  IND Indianapolis, IN
## 4      3442    LAN                  Lansing, MI  MSP  Minneapolis, MN
## 5      3368    JFK                 New York, NY  RIC     Richmond, VA
## 6      5520    ATL                  Atlanta, GA  EVV   Evansville, IN
##   CRS_DEP_TIME DEP_TIME DEP_DELAY TAXI_OUT WHEELS_OFF WHEELS_ON TAXI_IN
## 1          800      751        -9       31        822      1008       4
## 2         1405     1400        -5       15       1415      1438      11
## 3         1010     1005        -5       10       1015      1056       4
## 4          700      701         1       14        715       734       5
## 5         1459     1602        63       23       1625      1720       5
## 6         2127     2122        -5       13       2135      2133       3
##   CRS_ARR_TIME ARR_TIME ARR_DELAY CANCELLED CANCELLATION_CODE DIVERTED
## 1         1017     1012        -5         0                          0
## 2         1459     1449       -10         0                          0
## 3         1120     1100       -20         0                          0
## 4          746      739        -7         0                          0
## 5         1651     1725        34         0                          0
## 6         2149     2136       -13         0                          0
##   CRS_ELAPSED_TIME ELAPSED_TIME AIR_TIME DISTANCE DELAY_DUE_CARRIER
## 1              137          141      106      636          23.95602
## 2               54           49       23       98          23.95602
## 3               70           55       41      231          23.95602
## 4              106           98       79      455          23.95602
## 5              112           83       55      288           0.00000
## 6               82           74       58      350          23.95602
##   DELAY_DUE_WEATHER DELAY_DUE_NAS DELAY_DUE_SECURITY DELAY_DUE_LATE_AIRCRAFT
## 1          4.236307      13.44598          0.1644584                25.45021
## 2          4.236307      13.44598          0.1644584                25.45021
## 3          4.236307      13.44598          0.1644584                25.45021
## 4          4.236307      13.44598          0.1644584                25.45021
## 5          1.000000       0.00000          0.0000000                33.00000
## 6          4.236307      13.44598          0.1644584                25.45021
##   year month      carrier_name airport
## 1 2023    12 Endeavor Air Inc.     ABE
## 2 2023    12 Endeavor Air Inc.     ABE
## 3 2023    12 Endeavor Air Inc.     ABE
## 4 2023    12 Endeavor Air Inc.     ABE
## 5 2023    12 Endeavor Air Inc.     ABE
## 6 2023    12 Endeavor Air Inc.     ABE
##                                                   airport_name arr_flights
## 1 Allentown/Bethlehem/Easton, PA: Lehigh Valley International          72
## 2 Allentown/Bethlehem/Easton, PA: Lehigh Valley International          72
## 3 Allentown/Bethlehem/Easton, PA: Lehigh Valley International          72
## 4 Allentown/Bethlehem/Easton, PA: Lehigh Valley International          72
## 5 Allentown/Bethlehem/Easton, PA: Lehigh Valley International          72
## 6 Allentown/Bethlehem/Easton, PA: Lehigh Valley International          72
##   arr_del15 carrier_ct weather_ct nas_ct security_ct late_aircraft_ct
## 1         5       2.46          1   0.73           0             0.81
## 2         5       2.46          1   0.73           0             0.81
## 3         5       2.46          1   0.73           0             0.81
## 4         5       2.46          1   0.73           0             0.81
## 5         5       2.46          1   0.73           0             0.81
## 6         5       2.46          1   0.73           0             0.81
##   arr_cancelled arr_diverted arr_delay carrier_delay weather_delay nas_delay
## 1             0            0       672            61           574        20
## 2             0            0       672            61           574        20
## 3             0            0       672            61           574        20
## 4             0            0       672            61           574        20
## 5             0            0       672            61           574        20
## 6             0            0       672            61           574        20
##   security_delay late_aircraft_delay       Date high_severity_delay
## 1              0                  17 2023-12-01                   1
## 2              0                  17 2023-12-01                   1
## 3              0                  17 2023-12-01                   1
## 4              0                  17 2023-12-01                   1
## 5              0                  17 2023-12-01                   1
## 6              0                  17 2023-12-01                   1
```

Split the Data into Training and Testing Sets

```
# Remove rows with missing values in the target variable 'high_severity_delay'
combined_data_clean <- combined_dt %>%
  filter(!is.na(high_severity_delay))

# Check if any missing values remain
sum(is.na(combined_data_clean))
```

```
## [1] 72
```

```
# Split data into training (80%) and testing (20%) sets
set.seed(123)  # Set seed for reproducibility
train_index <- createDataPartition(combined_data_clean$high_severity_delay, p = 0.8, list = FALSE)

train_data <- combined_data_clean[train_index, ]
test_data <- combined_data_clean[-train_index, ]

# Check the dimensions of the split data
dim(train_data)
```

```
## [1] 30087    54
```

```
dim(test_data)
```

```
## [1] 7521   54
```

Train the Models

1) Random Forest Model

```
# Step 1: Ensure high_severity_delay is a factor in both train_data and test_data
train_data$high_severity_delay <- factor(train_data$high_severity_delay)
test_data$high_severity_delay <- factor(test_data$high_severity_delay)

# Step 2: Balance the data by undersampling
train_data_balanced <- train_data %>%
  group_by(high_severity_delay) %>%
  sample_n(min(table(train_data$high_severity_delay)))

# Check the new distribution
table(train_data_balanced$high_severity_delay)
```

```
##
##   0   1
## 488 488
```

```r
# Step 3: Train the Random Forest model on the balanced dataset
rf_model_balanced <- randomForest(high_severity_delay ~ carrier_delay + weather_delay + nas_delay + arr_
                                  data = train_data_balanced, importance = TRUE, ntree = 100)

# Print the results of the re-trained model
print(rf_model_balanced)
```

```
##
## Call:
##  randomForest(formula = high_severity_delay ~ carrier_delay +      weather_delay + nas_delay + arr_d
##                Type of random forest: classification
##                      Number of trees: 100
## No. of variables tried at each split: 2
##
##         OOB estimate of  error rate: 0%
## Confusion matrix:
##     0   1 class.error
## 0 488   0           0
## 1   0 488           0
```

```r
# Step 4: Predict on the test set
rf_pred_balanced <- predict(rf_model_balanced, test_data)

# Check the first few predictions
head(rf_pred_balanced)
```

```
## 14 15 23 34 44 61
##  1  1  1  1  1  1
## Levels: 0 1
```

```r
# Step 5: Ensure both rf_pred_balanced and test_data$high_severity_delay have the same factor levels
rf_pred_balanced <- factor(rf_pred_balanced, levels = levels(test_data$high_severity_delay))

# Step 6: Calculate the confusion matrix
rf_cm_balanced <- confusionMatrix(rf_pred_balanced, test_data$high_severity_delay)

# Print the confusion matrix results
print(rf_cm_balanced)
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction    0    1
##          0  114    0
##          1    0 7407
##
##                Accuracy : 1
##                  95% CI : (0.9995, 1)
##     No Information Rate : 0.9848
##     P-Value [Acc > NIR] : < 2.2e-16
##
##                   Kappa : 1
```

```
##
##  Mcnemar’s Test P-Value : NA
##
##             Sensitivity : 1.00000
##             Specificity : 1.00000
##          Pos Pred Value : 1.00000
##          Neg Pred Value : 1.00000
##              Prevalence : 0.01516
##          Detection Rate : 0.01516
##    Detection Prevalence : 0.01516
##       Balanced Accuracy : 1.00000
##
##        ’Positive’ Class : 0
##
```

2) Train the XGBoost Model

```r
# Step 2: Prepare data for XGBoost
# Convert predictor variables to a matrix (exclude the first column which is the intercept)
train_matrix <- model.matrix(high_severity_delay ~ carrier_delay + weather_delay + nas_delay + arr_dela
                             data = train_data)[,-1]
test_matrix <- model.matrix(high_severity_delay ~ carrier_delay + weather_delay + nas_delay + arr_delay
                            data = test_data)[,-1]

# Step 3: Ensure the target variable is numeric (0 and 1)
train_label <- as.numeric(train_data$high_severity_delay) - 1
test_label <- as.numeric(test_data$high_severity_delay) - 1

# Step 4: Train the XGBoost model
xgb_model <- xgboost(data = train_matrix, label = train_label, nrounds = 100, objective = "binary:logis
                     eval_metric = "logloss", verbose = 0)

# Step 5: Predict on the test set
xgb_pred <- predict(xgb_model, test_matrix)

# Convert predictions to binary labels (0 or 1)
xgb_pred_class <- ifelse(xgb_pred > 0.5, 1, 0)

# Step 6: Evaluate the model performance using confusion matrix
library(caret)
xgb_cm_balanced <- confusionMatrix(factor(xgb_pred_class), factor(test_label))

# Print confusion matrix results
print(xgb_cm_balanced)
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction    0    1
##          0  114    0
##          1    0 7407
##
##                Accuracy : 1
##                  95% CI : (0.9995, 1)
##     No Information Rate : 0.9848
##     P-Value [Acc > NIR] : < 2.2e-16
##
##                   Kappa : 1
##
##  Mcnemar’s Test P-Value : NA
##
##             Sensitivity : 1.00000
##             Specificity : 1.00000
##          Pos Pred Value : 1.00000
##          Neg Pred Value : 1.00000
##              Prevalence : 0.01516
##          Detection Rate : 0.01516
##    Detection Prevalence : 0.01516
##       Balanced Accuracy : 1.00000
##
##        ’Positive’ Class : 0
##
```

Model Evaluation

```r
# Random Forest model performance
rf_cm <- confusionMatrix(rf_pred_balanced, test_data$high_severity_delay)
print(rf_cm)
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction    0    1
##          0  114    0
##          1    0 7407
##
##                Accuracy : 1
##                  95% CI : (0.9995, 1)
##     No Information Rate : 0.9848
##     P-Value [Acc > NIR] : < 2.2e-16
##
##                   Kappa : 1
##
##  Mcnemar’s Test P-Value : NA
##
##             Sensitivity : 1.00000
##             Specificity : 1.00000
##          Pos Pred Value : 1.00000
##          Neg Pred Value : 1.00000
##              Prevalence : 0.01516
##          Detection Rate : 0.01516
##    Detection Prevalence : 0.01516
##       Balanced Accuracy : 1.00000
##
##        ’Positive’ Class : 0
##
```

```r
# XGBoost model performance
xgb_cm <- confusionMatrix(factor(xgb_pred_class), factor(test_label))
print(xgb_cm)
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction    0    1
##          0  114    0
##          1    0 7407
##
##                Accuracy : 1
##                  95% CI : (0.9995, 1)
##     No Information Rate : 0.9848
##     P-Value [Acc > NIR] : < 2.2e-16
##
##                   Kappa : 1
##
##  Mcnemar’s Test P-Value : NA
##
##             Sensitivity : 1.00000
##             Specificity : 1.00000
##          Pos Pred Value : 1.00000
##          Neg Pred Value : 1.00000
##              Prevalence : 0.01516
##          Detection Rate : 0.01516
##    Detection Prevalence : 0.01516
##       Balanced Accuracy : 1.00000
##
##        ’Positive’ Class : 0
##
```

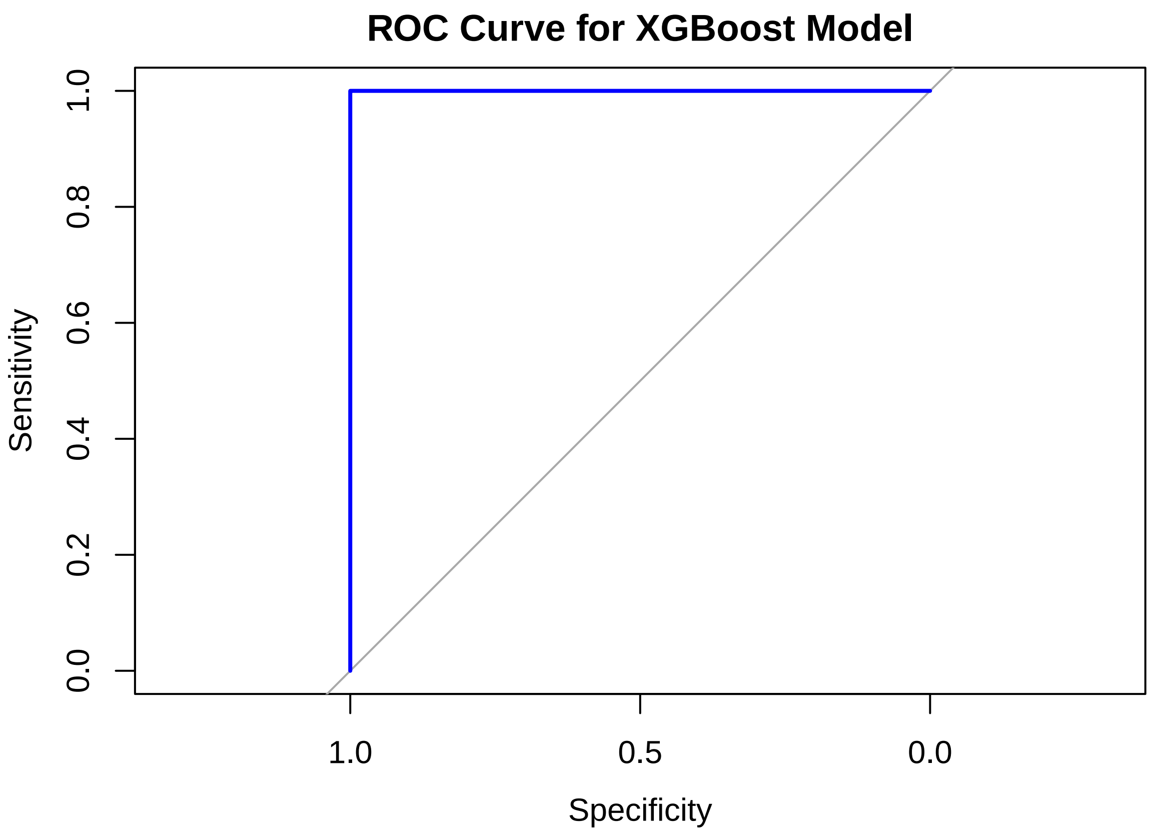
Visualizing Model Performance

```r
# Step 1: XGBoost Model ROC Curve
# For XGBoost, the predictions are probabilities, not classes
xgb_pred_prob <- predict(xgb_model, test_matrix)  # predicted probabilities for XGBoost
roc_xgb <- roc(test_label, xgb_pred_prob)  # calculate ROC for XGBoost
```

```
## Setting levels: control = 0, case = 1
```

```
## Setting direction: controls < cases
```

```r
# Plot ROC curve for XGBoost
plot(roc_xgb, main = "ROC Curve for XGBoost Model", col = "blue", lwd = 2)
```

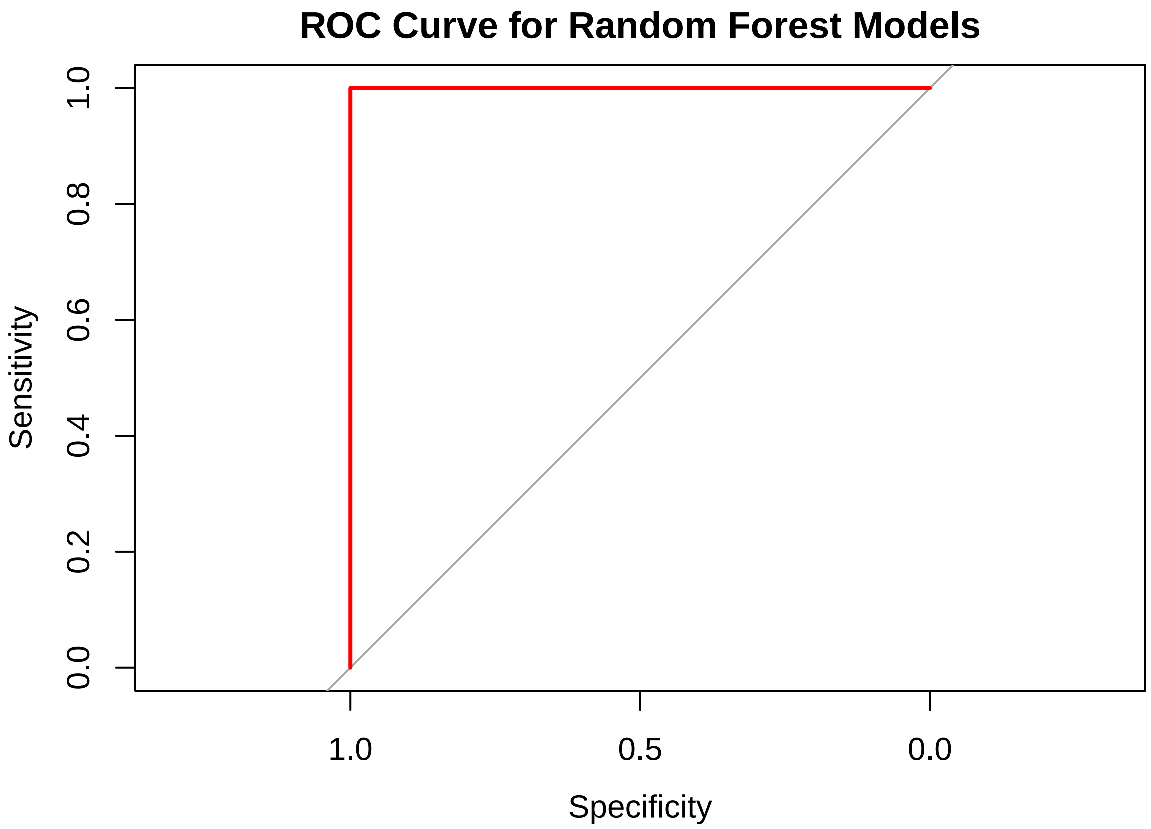


```r
# Step 2: Random Forest Model ROC Curve
# For Random Forest, you need to use type = "prob" to get predicted probabilities
rf_pred_prob <- predict(rf_model_balanced, test_data, type = "prob")[, 2]  # probabilities for class 1

# Calculate ROC for Random Forest
roc_rf <- roc(test_label, rf_pred_prob)  # calculate ROC for Random Forest
```

```
## Setting levels: control = 0, case = 1
```

```
## Setting direction: controls < cases
```

```r
# Plot ROC curve for Random Forest
plot(roc_rf, main = "ROC Curve for Random Forest Models", col = "red", lwd = 2)
```



```r
# Calculate and print AUC for both models
auc_xgb <- auc(roc_xgb)
auc_rf <- auc(roc_rf)

print(paste("XGBoost AUC:", auc_xgb))
```

```
## [1] "XGBoost AUC: 1"
```

```r
print(paste("Random Forest AUC:", auc_rf))
```

```
## [1] "Random Forest AUC: 1"
```