Project Report

CIS 3920 FTRA

Minh Quang Duong
(MINHQUANG.DUONG@baruchmail.cuny.edu)

December 13th, 2023

Aviation Classification Report

The report provides the background of this analysis, methods behind the analysis and the findings of the Aviation Classification. This report documents both the methodology and findings of an analysis of aircraft damage severity.. We decided to use the data from the National Transportation Safety Board for my analysis. The data specifically targets flight accidents from the period of 1/1/2020 to 1/1/2023. The initial goal of this project was to find out if there any factor that majorly contributes to the occurrences of flight accidents. The data originally had 38 different variables, but after careful consideration, we have wound the data down to what I believed to be the 3 most relevant contributing factors. As stated above, our main objective is to determine which factor has the most influential impact on the severity of aircraft damage in an accident.

Before incorporating predictive models, we recognized the importance of comprehending the dataset through exploratory data processing so as to appropriately analyze the content. The dataset consists of a dimension of 4740 rows and 38 columns. After a process of cleaning all the null values of formatting, eliminating irreverent variables, the dataset ended up with a dimension of 3534 rows and 4 columns. Figure 1 provides an insight into the distribution of AirCraftDamage, the majority of the damage is Substantial across all instances, in less than 25% of the time the aircraft was Destroyed in an accident, the rest are spread across either Minor damage or Unknown.

The next step was to implement the usage of the Generalized Linear Model (GLM). The reason we chose this model was because our target variable is categorical and not continuous, hence using linear models could lead to inaccurate results. Additionally, GLM provides a better framework for analyzing dependent categorical variables. Figure 2 provides a summary of statistics of the GLM model. The figure suggests that the variable AmatureBuilt and WeatherCondition were both statistically significant while the NumOfEngine variable might not be as statistically significant.We then split the data frame into 2 sets, 80% for training and 20% for testing. We implemented the tree() function to build a decision tree model with our training data. Figure 3 provides the summary of the tree model and Figure 4 displays the visual representation of the decision-making process of the model. The tree initiates at the root node, splitting data based on the AmateurBuilt feature. If amateur-built, it further considers WeatherCondition; IMC predicts substantial damage, while Unknown conditions suggest sustainability. For non-amateur-built aircraft, the tree examines NumberOfEngines. Fewer than 1.5 engines predict substantial damage, while 1.5 or more engines suggest sustainability. The detailed interpretation underscores that amateur-built aircraft are prone to damage, especially in IMC conditions, while non-amateur-built aircraft with fewer engines also face higher risks.

Moving forward, the random forest algorithm was implemented for the classification model. Amongst the available classification methods, random forest provides the highest accuracy. Figure 6 provides the plot of the Variable Importance within the model. The Mean Decrease Gini values provided suggest that WeatherCondition is the most important predictor in the random forest, followed by AmateurBuilt and then NumberOfEngines. Initially, we built a random forest model with the training data set, we then used said model to generate a prediction model using the testing dataset. Figure 5 portrays the confusion matrix for the prediction model. The figure indicated that the model performed well with an accuracy of 86.56%. There were some instances of misclassification. Following this the detailed summary of the matrix:

- True positive (TP): 14 aircraft were correctly predicted to suffer substantial damage.
- False positive (FP): 16 aircraft were incorrectly predicted to suffer substantial damage.
- True negative (TN): 590 aircraft were correctly predicted to be sustainable.
- False negative (FN): 5 aircraft were incorrectly predicted to be sustainable.

The precision of the model is TP/(TP+FP) equates to 46.66%. The recall of the model is formulated by TP/(TP+FN) which equates to 73.68%. These results mean that 46.66% of the aircraft that were predicted to suffer substantial damage, actually suffered substantial damage and 73.684 % of the aircraft that actually suffered substantial damages were correctly predicted by the model.

This report investigated the factors influencing aircraft damage severity in accident scenarios. By analyzing data from the National Transportation Safety Board (NTSB) for accidents from January 1st, 2020, to January 1st, 2023, the report identified three key contributing factors: AmateurBuilt, WeatherCondition, and NumberOfEngines. Overall, this analysis provides valuable insights into the factors influencing aircraft damage severity. These insights can inform aviation safety efforts and guide the development of preventative measures. Moving forward, this analysis could be expanded by extending the dataset by including a broader timeframe, including additional models and analyzing special types of accidents.

Appendix

Figure 1. Distribution Plot

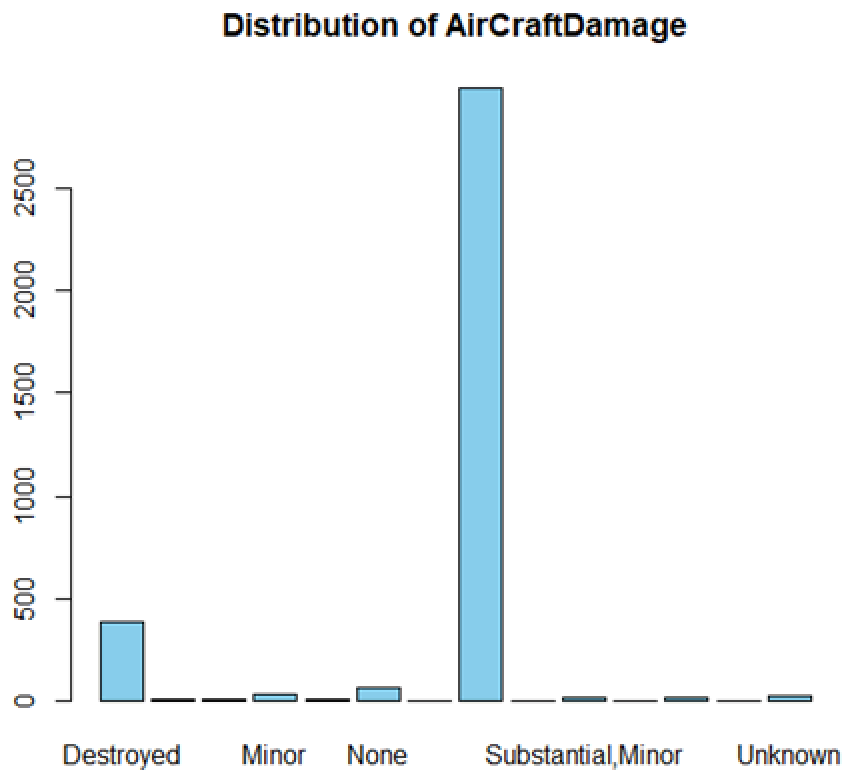

Figure 2. Summary of the logistic regression model

```
Coefficients:
                          Estimate Std. Error z value Pr(>|z|)    
(Intercept)             -1.097e+14  2.225e+14  -0.493 0.622151    
AmateurBuiltFALSE,FALSE  4.613e+15  2.225e+14  20.732  < 2e-16 ***
AmateurBuiltFALSE,TRUE   4.613e+15  2.225e+14  20.732  < 2e-16 ***
AmateurBuiltTRUE        -4.608e-01  1.568e-01  -2.938 0.003301 ** 
AmateurBuiltTRUE,TRUE    4.613e+15  2.225e+14  20.732  < 2e-16 ***
WeatherConditionUnknown  1.845e+00  5.024e-01   3.672 0.000241 ***
WeatherConditionVMC      2.251e+00  1.904e-01  11.824  < 2e-16 ***
NumberOfEngines,1        2.676e+06  8.220e+07   0.033 0.974032    
NumberOfEngines,2       -1.198e+07  9.728e+07  -0.123 0.901997    
NumberOfEngines0         1.097e+14  2.225e+14   0.493 0.622151    
NumberOfEngines0,0       6.688e+05  9.491e+07   0.007 0.994377    
NumberOfEngines1         1.097e+14  2.225e+14   0.493 0.622151    
NumberOfEngines1,        8.668e-01  7.749e+07   0.000 1.000000    
NumberOfEngines1,1      -3.023e+04  6.885e+07   0.000 0.999650    
NumberOfEngines1,2       1.938e+00  7.503e+07   0.000 1.000000    
NumberOfEngines2         1.097e+14  2.225e+14   0.493 0.622151    
NumberOfEngines2,1       1.936e+00  8.219e+07   0.000 1.000000    
NumberOfEngines2,2       1.366e+00  7.174e+07   0.000 1.000000    
NumberOfEngines3         1.097e+14  2.225e+14   0.493 0.622151    
NumberOfEngines4         1.097e+14  2.225e+14   0.493 0.622151    
NumberOfEngines4,1       1.048e+00  9.491e+07   0.000 1.000000    
NumberOfEngines8         1.097e+14  2.225e+14   0.493 0.622151    
Latitude                 2.546e-02  6.886e-03   3.698 0.000218 ***
Longitude               -5.721e-04  2.885e-03  -0.198 0.842817    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 2437.7  on 3533  degrees of freedom
Residual deviance: 2261.2  on 3510  degrees of freedom
AIC: 2309.2

Number of Fisher Scoring iterations: 25
```

```
 AmateurBuilt       WeatherCondition   NumberOfEngines    AirCraftDamage    
 Length:3534        Length:3534        Length:3534        Length:3534       
 Class :character   Class :character   Class :character   Class :character  
 Mode  :character   Mode  :character   Mode  :character   Mode  :character  
```

Figure 3. Summary of the tree model

```
Classification tree:
tree(formula = AirCraftDamage ~ ., data = cleaned_AviationData, 
    subset = train_indices)
Variables actually used in tree construction:
[1] "AmateurBuilt"     "NumberOfEngines"  "WeatherCondition"
Number of terminal nodes:  5 
Residual mean deviance:  0.9716 = 2742 / 2822 
Misclassification error rate: 0.1496 = 423 / 2827 
```

Figure 4. Classification Tree

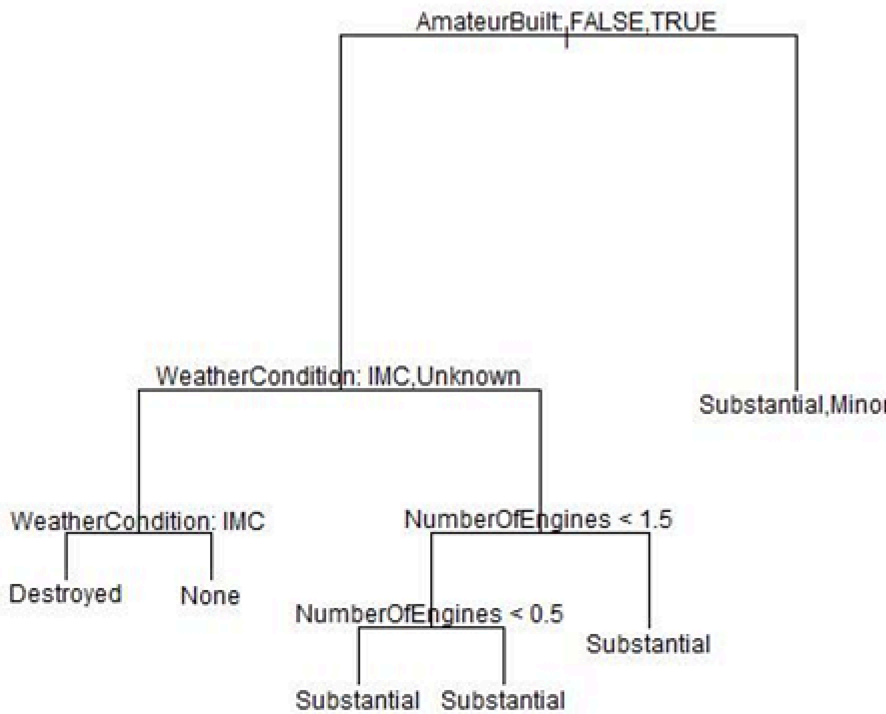

Figure 5. Confusion Matrix of the prediction model

```
Confusion Matrix and Statistics

                        Reference
Prediction               Destroyed Destroyed,Destroyed Destroyed,Substantial Minor Minor,Substantial None None,None Substantial Substantial,Destroyed
  Destroyed                     12                   0                     0     0                 1    1         0          14                     0
  Destroyed,Destroyed            0                   0                     0     0                 0    0         0           0                     0
  Destroyed,Substantial          0                   0                     0     0                 0    0         0           0                     0
  Minor                          0                   0                     0     0                 0    0         0           0                     0
  Minor,Substantial              0                   0                     0     0                 0    0         0           0                     0
  None                           1                   0                     0     0                 0    0         0           0                     0
  None,None                      0                   0                     0     0                 0    0         0           0                     0
  Substantial                   55                   0                     0     5                 0   16         0         590                     0
  Substantial,Destroyed          0                   0                     0     0                 0    0         0           0                     0
  Substantial,Minor              0                   0                     0     0                 0    0         0           0                     0
  Substantial,None               0                   0                     0     0                 0    0         0           0                     0
  Substantial,Substantial        0                   3                     0     0                 0    0         0           0                     1
  Substantial,Unknown            0                   0                     0     0                 0    0         0           0                     0
  Unknown                        0                   0                     0     0                 0    0         0           0                     0
                        Reference
Prediction               Substantial,Minor Substantial,None Substantial,Substantial Substantial,Unknown Unknown
  Destroyed                              0                0                       0                   0       0
  Destroyed,Destroyed                    0                0                       0                   0       0
  Destroyed,Substantial                  0                0                       0                   0       0
  Minor                                  0                0                       0                   0       0
  Minor,Substantial                      0                0                       0                   0       0
  None                                   0                0                       0                   0       0
  None,None                              0                0                       0                   0       0
  Substantial                            0                0                       0                   0       6
  Substantial,Destroyed                  0                0                       0                   0       0
  Substantial,Minor                      0                0                       0                   0       0
  Substantial,None                       0                0                       0                   0       0
  Substantial,Substantial                1                0                       1                   0       0
  Substantial,Unknown                    0                0                       0                   0       0
  Unknown                                0                0                       0                   0       0

Overall Statistics

               Accuracy : 0.8656
                 95% CI : (0.8383, 0.8899)
    No Information Rate : 0.8628
    P-Value [Acc > NIR] : 0.44

                  Kappa : 0.0693

 Mcnemar's Test P-Value : NA
```

Figure 6. Variable Importance of Random Forest Model