Airplane Crash Analysis

Project Members:

- Adam Goldstein
- Ratika Bhuwalka
- Ryan Cummings
- Sai Srinivas Lakkakula

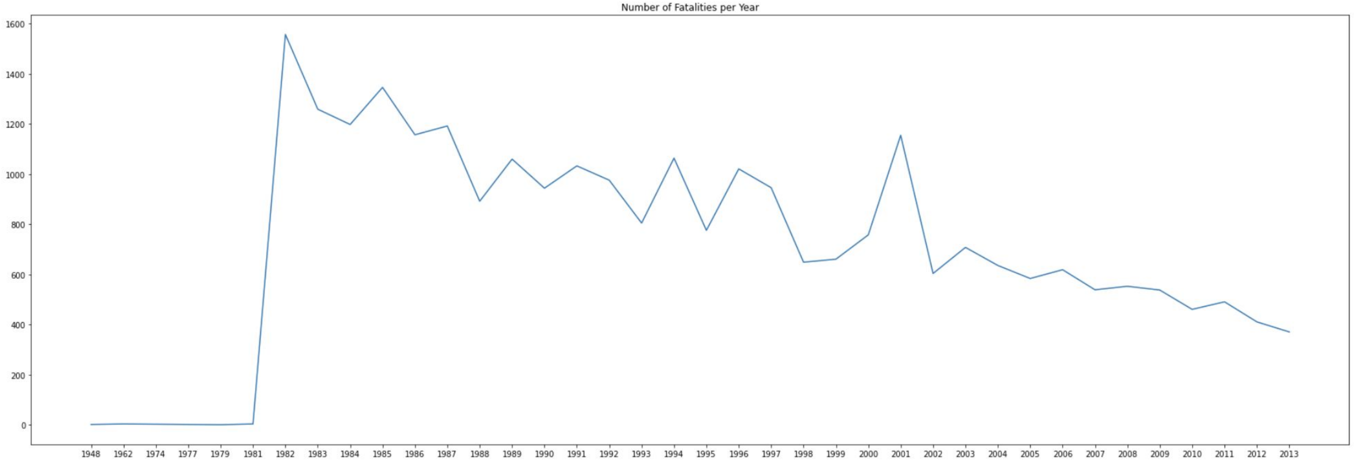
Introduction

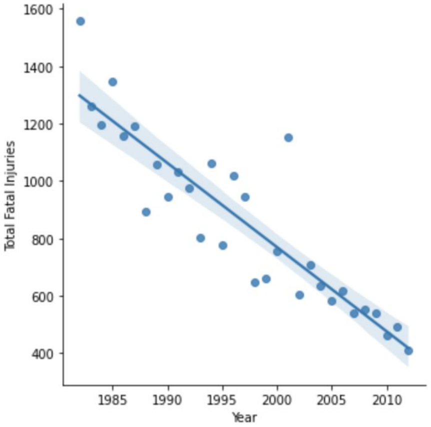
- National Transportation and Safety Board, consists of over 74 thousand accidents and incidents from 1948 to 2013 (https://data.ntsb.gov/avdata).
    - 23% (13089 rows) of the data does have fatalities (1 or more deaths), while the remaining 77% (57668 rows) is non-fatal airplane crashes.
    - Columns like: Date, Latitude, Longitude, Weather type, Amateur Built, Make, Model, Number of Engines, etc

Methods

Answering 2 questions:

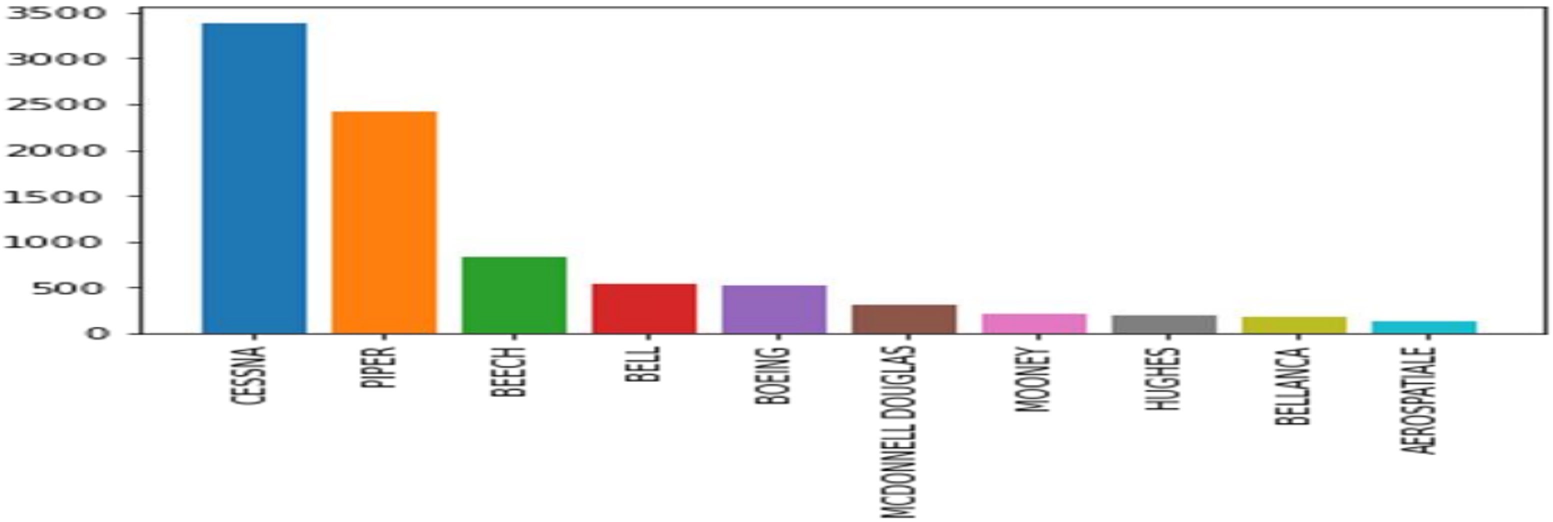
1) Are flights becoming safer over time?
    - Used Basic Data Analysis techniques to answer this question
        - Used graphs and linear models to get slope and then determine the trends based on the slope value.





Further Research

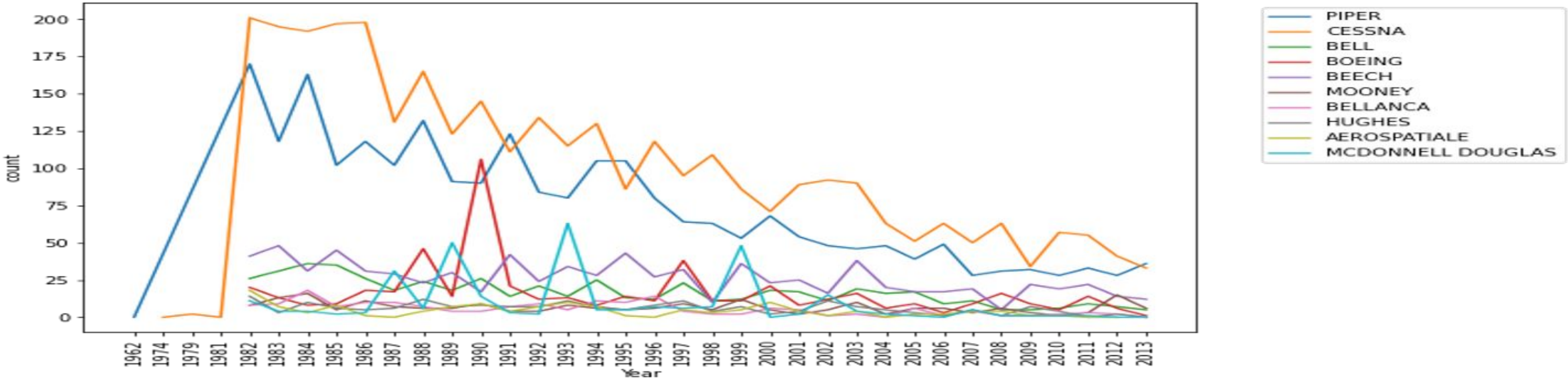
Top 10 manufacturers with most fatality rate.



Things to know

-Small flights are more prone to fatalities than the large commercial ones

- Improved design and safety measures over the years following FAA regulations.



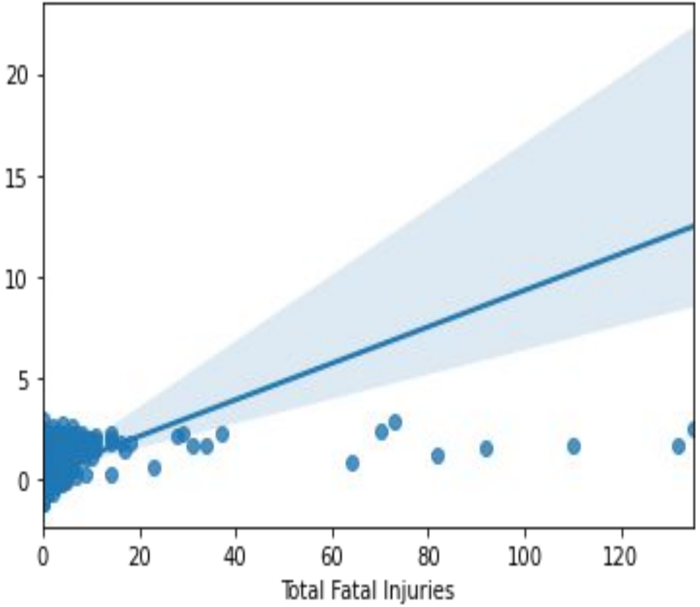
Methods (2)

2) Can we predict the number of fatalities or if a flight will contain any amount of fatalities?

- Predicting number of fatalities:
  - Multiple Linear Regression

- Predicting whether or not there will be fatalities on a flight (Binary Classify)
  - Multiple Logistic Regression
  - Decision Tree(s) and Random Forest(s)

Linear Regression



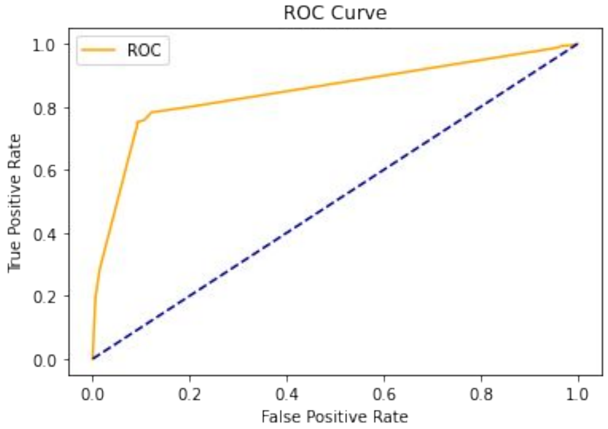
- Tried predicting no of fatalities with parameter like flight weather condition, no.of engines, location, Built, Phanse of flight etc.
- RMSE is found to be 3.4
- Most of the fatalities fall under 0-4 band which the model seems to have predicted correctly.

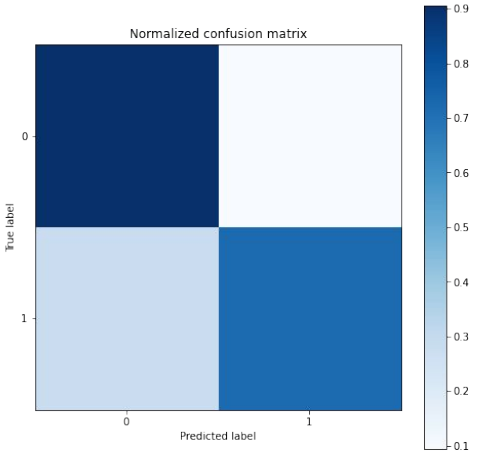
Logistic Regression

- Model designed to predict whether an airplane crash is fatal or not
- Predictions are based on
  - Aircraft Damage
  - Number of Engines
  - Weather Condition

Logistic Regression

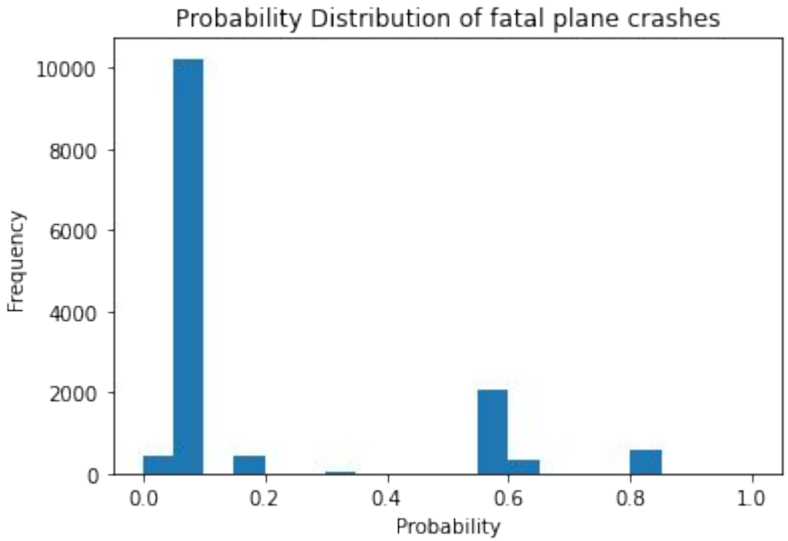
- Model resulted in 87% prediction accuracy with an AUC value of 83%
- Accuracy 10% better than baseline assuming all crashes are fatal
- Hyperparameter tuning using gridsearch marginally improved prediction accuracy
- Tuned model showed marginal improvement in accuracy





Logistic Regression

- Probability distribution is unbalanced as 77% of crashes had no fatalities
- Utilized specific flight characteristics to predict likelihood of fatal plane crash
- Certain characteristics such as ‘high aircraft damage’ or ‘bad weather’ will lead to model predicting around 0.6 likelihood of fatal crash

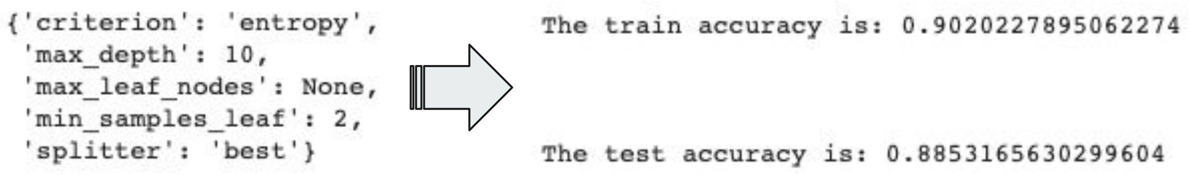


Decision Tree

- Used Decision Tree to Binary Classify whether a plane will have any fatalities.

- Did compute DT with default parameters and with Hyperparameter tuning through Grid Search.

- Ultimately used the accuracy scores of these models to compare against Random Forest(s) since we know RF would outperform DTs.



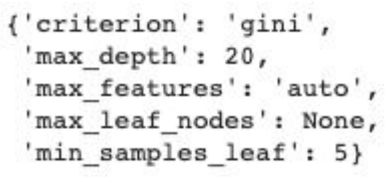
```
{'criterion': 'entropy',
 'max_depth': 10,
 'max_leaf_nodes': None,
 'min_samples_leaf': 2,
 'splitter': 'best'}
```

```
The train accuracy is: 0.9020227895062274


The test accuracy is: 0.8853165630299604
```

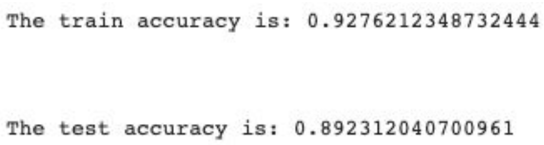
Random Forest

Created 2 RFs:

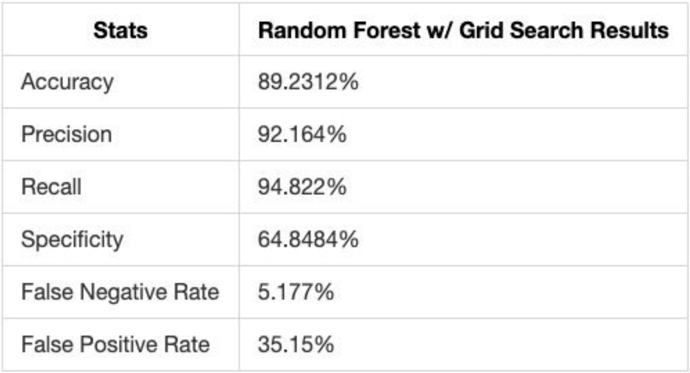
1) RF with Default Parameters (used as baseline)
2) RF with Grid Search



```
{'criterion': 'gini',
 'max_depth': 20,
 'max_features': 'auto',
 'max_leaf_nodes': None,
 'min_samples_leaf': 5}
```



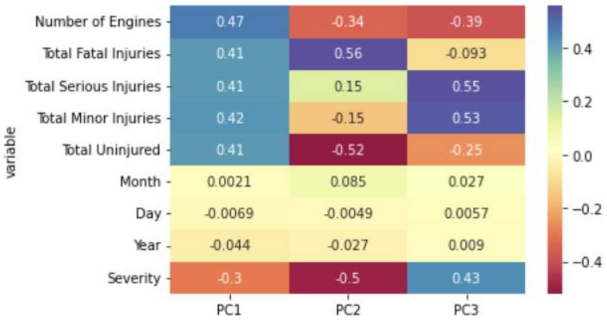
```
The train accuracy is: 0.9276212348732444


The test accuracy is: 0.892312040700961
```



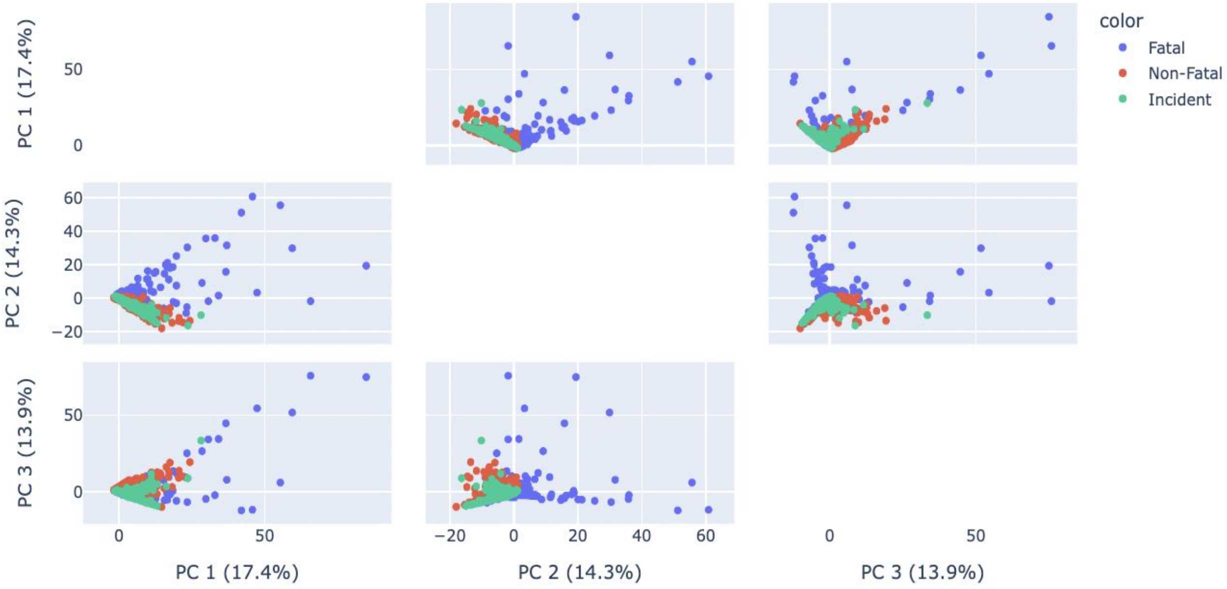
Stats	Random Forest w/ Grid Search Results
Accuracy	89.2312%
Precision	92.164%
Recall	94.822%
Specificity	64.8484%
False Negative Rate	5.177%
False Positive Rate	35.15%

Random Forest gave us the best results for Binary Classification!

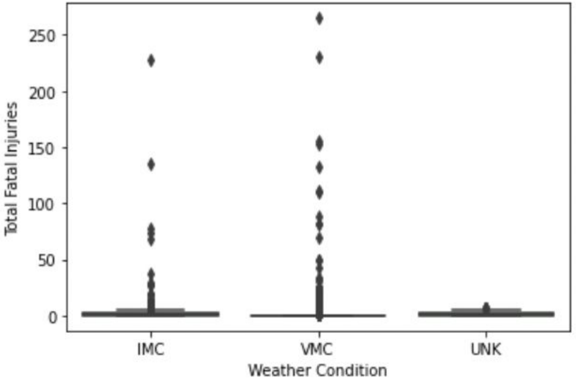
Principal Component Analysis

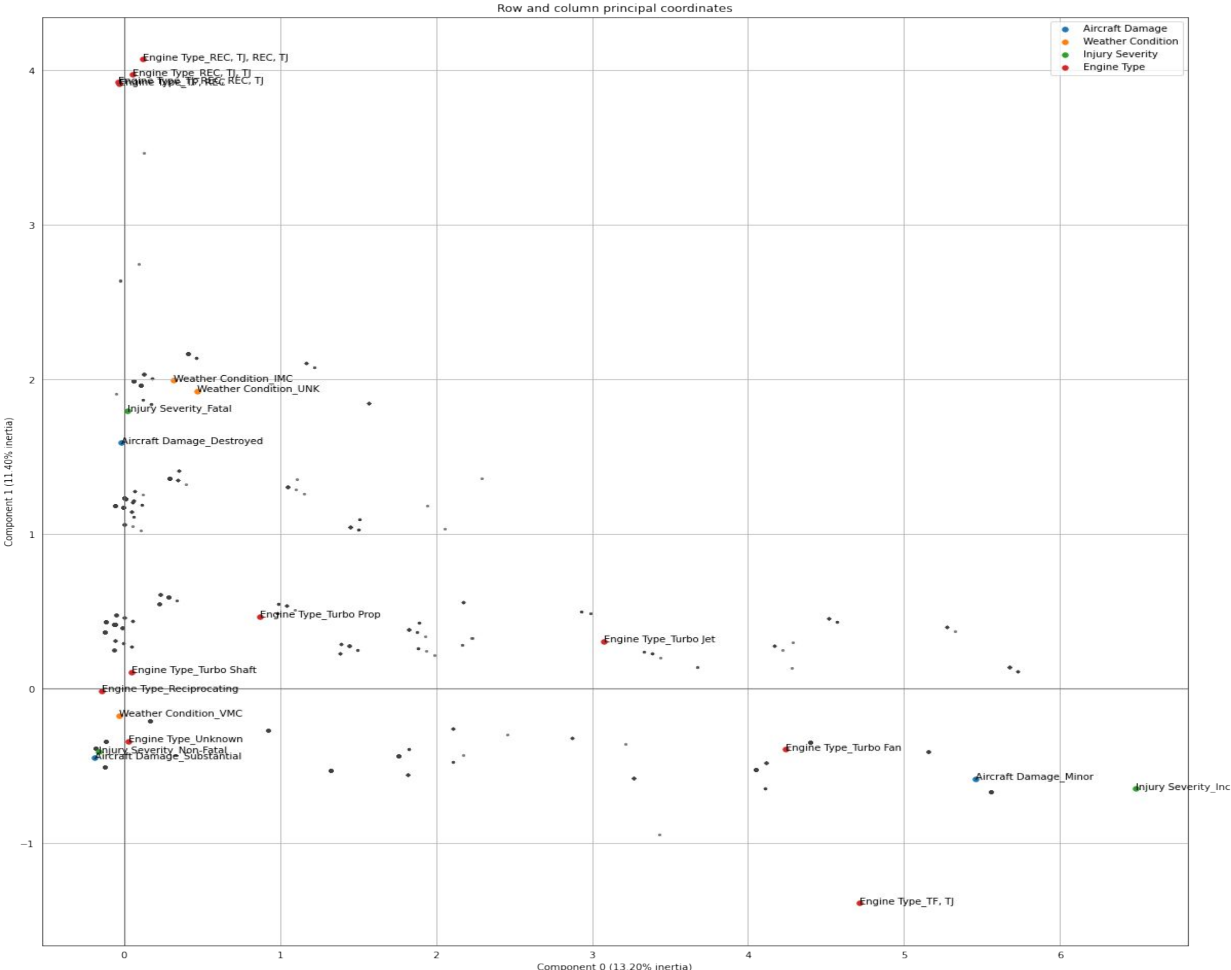


variable	PC1	PC2	PC3
Number of Engines	0.47	-0.34	-0.39
Total Fatal Injuries	0.41	0.56	-0.093
Total Serious Injuries	0.41	0.15	0.55
Total Minor Injuries	0.42	-0.15	0.53
Total Uninjured	0.41	-0.52	-0.25
Month	0.0021	0.085	0.027
Day	-0.0069	-0.0049	0.0057
Year	-0.044	-0.027	0.009
Severity	-0.3	-0.5	0.43



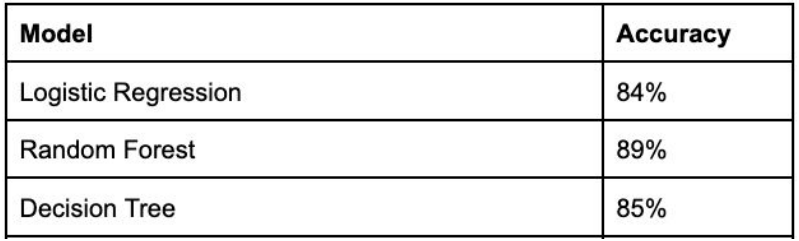
Multiple Correspondence Analysis





Comparisons

When comparing the different binary classification methods such as Logistic regression, Random Forest and Decision Tree, **it is completely reasonable that Random Forest outperformed both Decision Tree and Logistic Regression.**



Model	Accuracy
Logistic Regression	84%
Random Forest	89%
Decision Tree	85%

The performance of Random Forest is largely due to its emphasis on feature selection, ability to ignore linear relationships within predictors and utilization of ensemble learning.

Conclusion

- Analysis focused on flight safety over time
- When predicting whether a flight is fatal, Random Forest drastically outperformed decision trees and logistic regression
- Aviation clearly becoming safer overtime
- When crashes do occur, it is feasible to predict if a crash will be fatal based on flight characteristics

Questions?