



Clear Skies:

Data-Driven Safety Recommendations From Historical Aircraft Crash Reports

Will Griffin, Hilary Lawlor, and Emily Qiu



Problem Statement

The Federal Aviation Administration (FAA) uses data from the National Transportation Safety Board (NTSB) to evaluate their regulations for civil aviation. The FAA hired our team of data scientists to analyze historic aircraft crash data from the NTSB and make predictions that will help them determine regulations for the future of aviation safety.

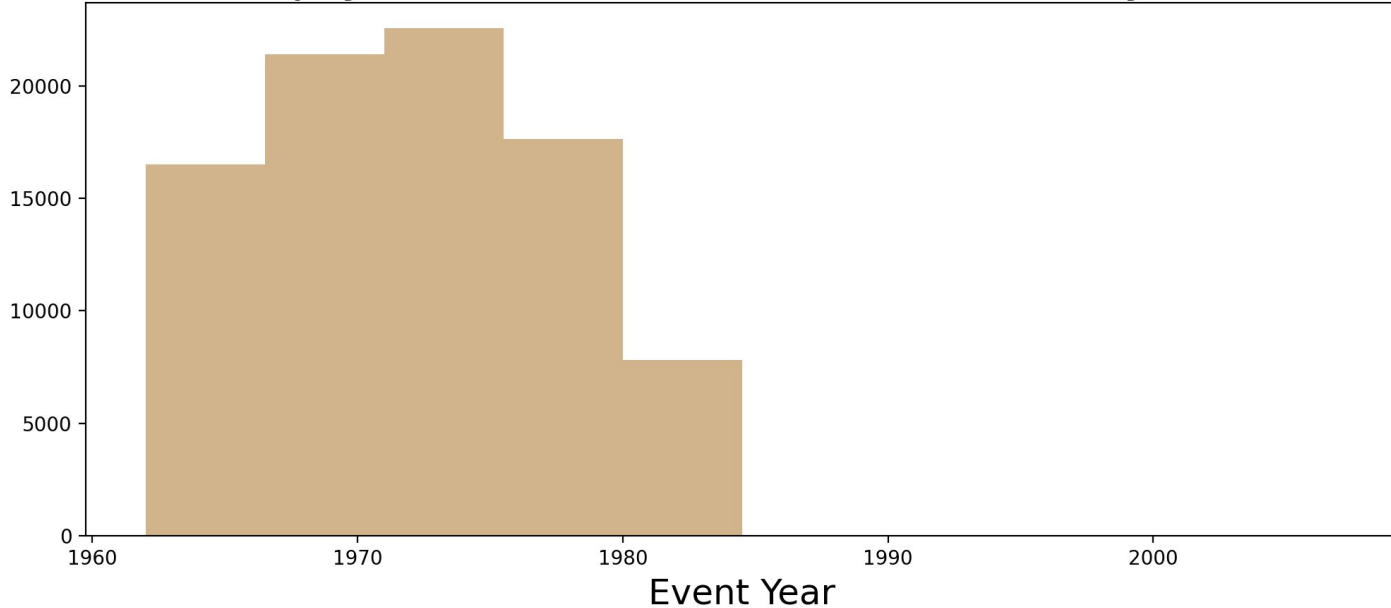


Cleaning/EDA

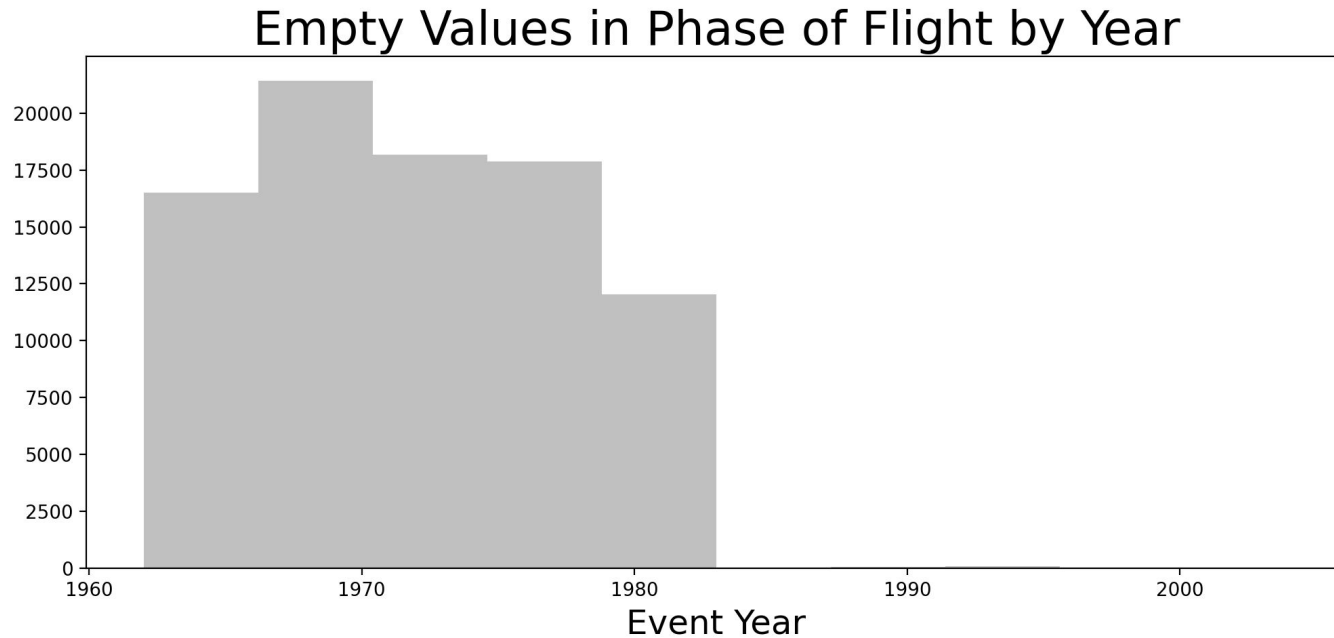
- Initially a raw .txt file from NTSB
 - Dropped all events outside of USA
 - Reformat to remove needless spaces
 - Size: +- 150,000 rows
- Create event_month and event_year columns from event_date
- Cross check injury_severity nulls against total_fatal_injuries
 - If injury_severity was null but total_fatal_injuries was 0, can make injury_severity "Non-Fatal"
 - If total_fatal_injuries was null but injury_severity is "Non-Fatal", can make total_fatal_injuries "0"

Handling Null Values: Weather

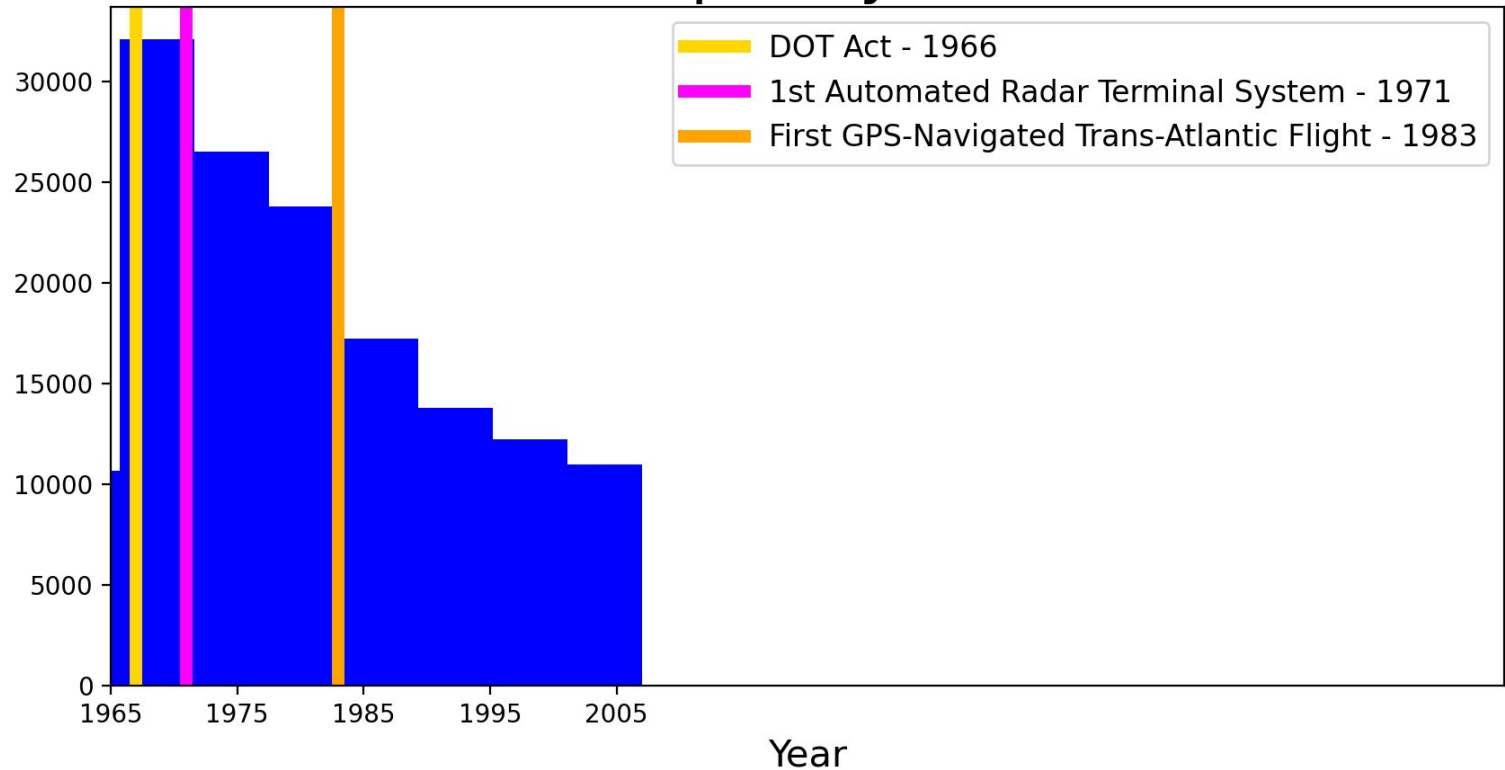
Empty Values in Weather Condition by Year



Handling Null Values: Phase of Flight



Event Frequency Over Time



Chapter 1: Aircraft Damage Evaluation and Repair

Outcome Evaluation (Worth Repair or not??)

Is it possible to develop a repair plan for each damaged aircraft?

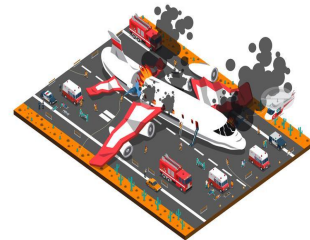
- Complexity of aircraft designs
- Widely differing conditions

It is still a subjective evaluation because of the lack of specific regulations (dated 2018).

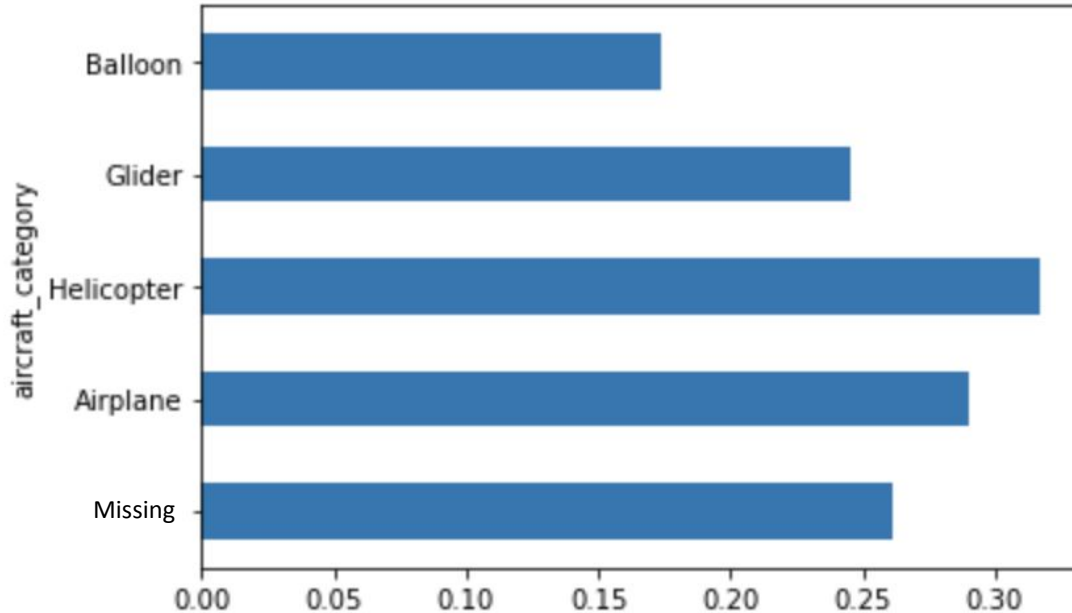
Minor: 'is restorable to airworthy conditions by minor repairs'

Substantially damaged: 'damage adversely affects the structural strength, performance, or flight characteristics of the aircraft'

Destroyed: unable to repair

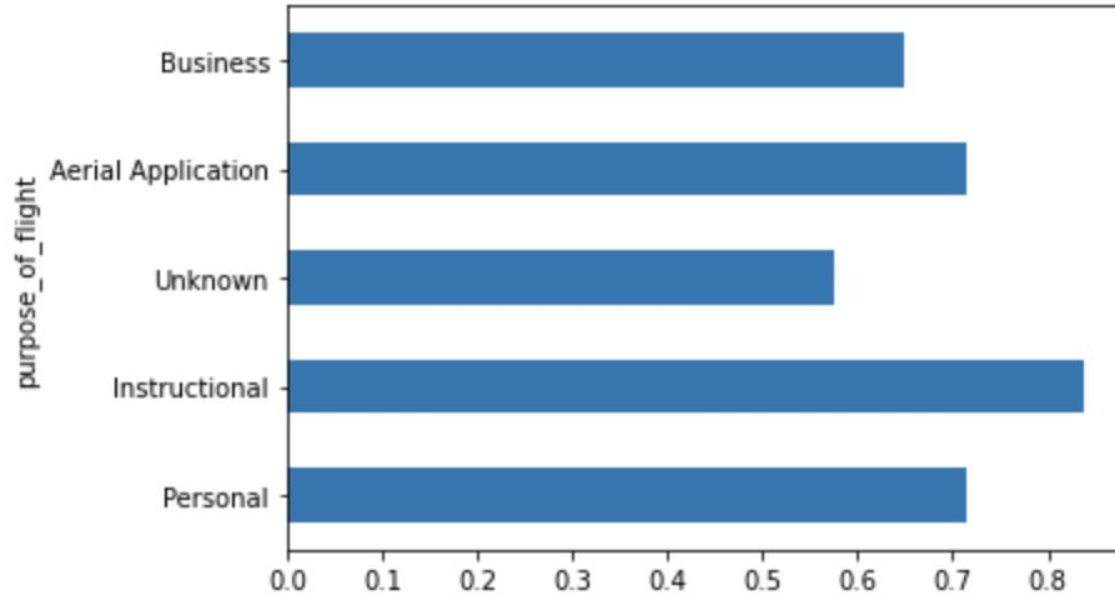


~ 30% of Helicopters and Airplanes are Classified as Destroyed



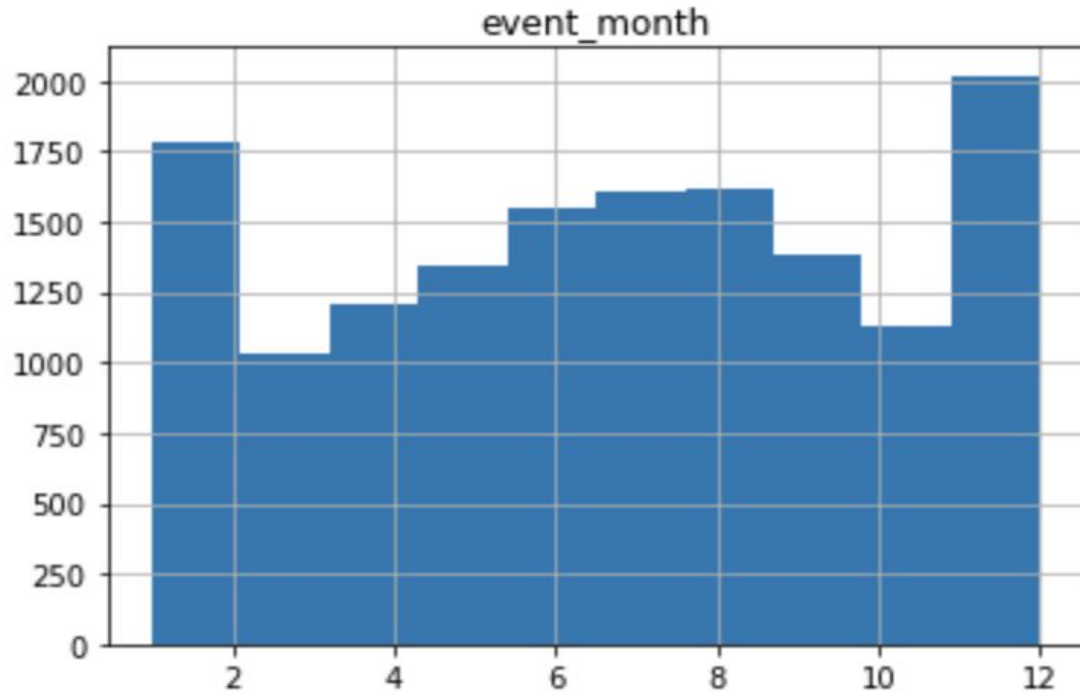
Helicopter accidents can often result in serious injuries because of the size, weight, height, and the speed.

~ 80% aircraft damages happened during instructional session



FAA could possibly consider to enforce stricter regulations and/or recommend better flight models for instructional sessions.

Destruction by Month



Depending on specific geographic area, FAA could consider to enforce stricter policies during winter.

Random Forest & Support Vector Machine

Random Forest Model

- X variables: weather conditions, phases of flights (label encoding)
- Y variable: minor, substantially damaged, and destroyed as [0, 1, 2]
- Training score: 74%
- Testing score: 74%
 - Substantially damaged [1]
 - Precision: 0.75, recall: 0.96, F1-score: 0.84 (weather condition)
 - Precision: 0.71, recall: 1, F1-score: 0.83 (phase of flights)

Support Vector Machine Model

- X variables: weather conditions, phases of flights (label encoding)
- Y variable: minor, substantially damaged, and destroyed as [0, 1, 2]
- Training score: 85%
- Testing score: 85%
 - Substantially damaged [1]
 - Precision: 0.85, recall: 0.97, F1-score: 0.90 (phase of flights)

Logistic Regression Model

Logistic Regression Model

- X variables: total_minor_injuries, total_major_injuries, total_fatal_injuries, event_month
- Y variable: minor, substantially damaged, and destroyed as [0, 1, 2]
- Training score: 85%
- Testing score: 85%
 - Substantially damaged [1]
 - 85% precision
 - 96% recall
 - 90% F1 score
 - Destroyed [2]
 - 85% precision
 - 64% recall
 - 73% F1 score

Predicting Aircraft Damage Level

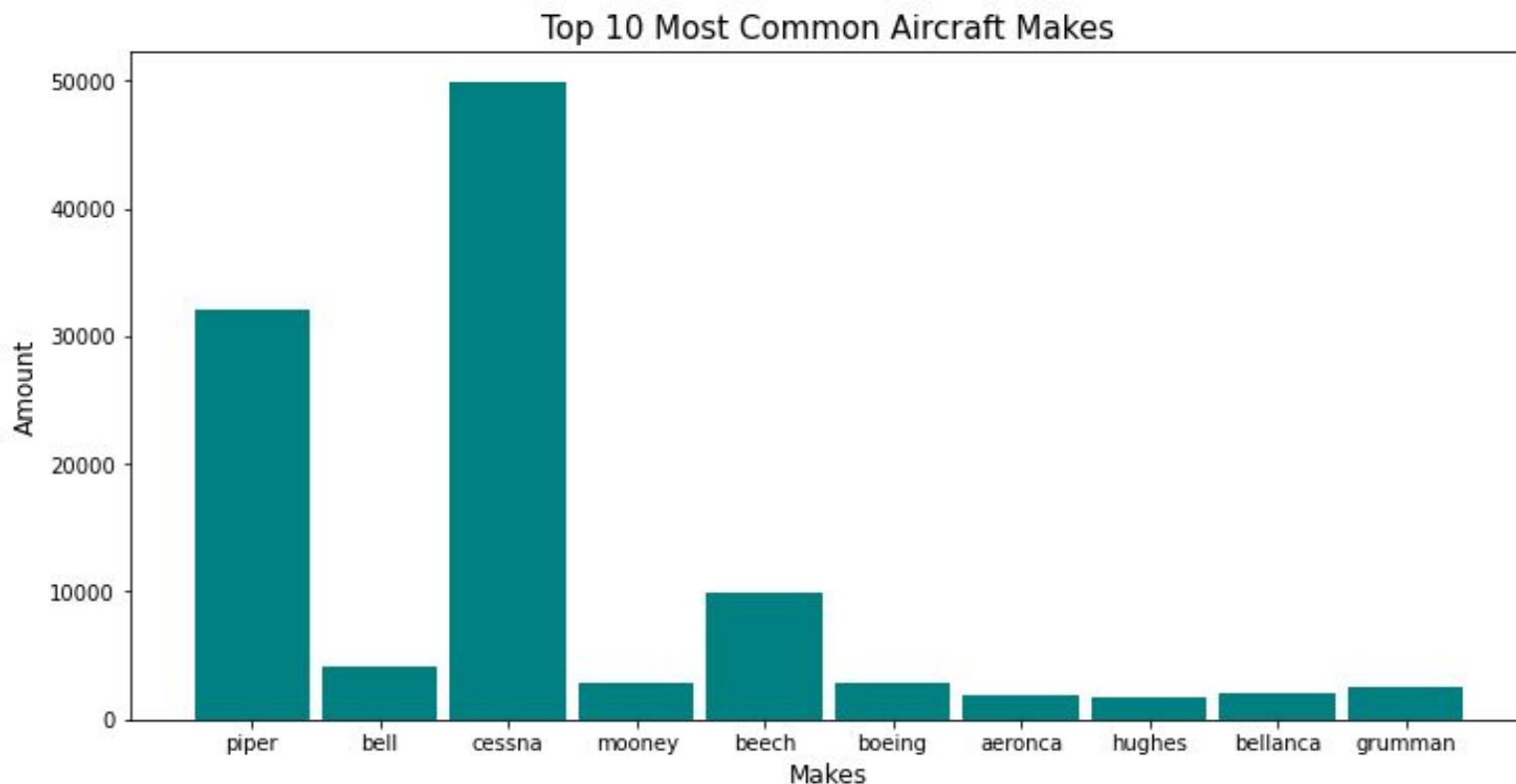
Current aircraft damage percentage breakdown:

- 24% of aircrafts are destroyed
- 73% of aircrafts are substantially damaged
- Only 3% of aircrafts have minor damages

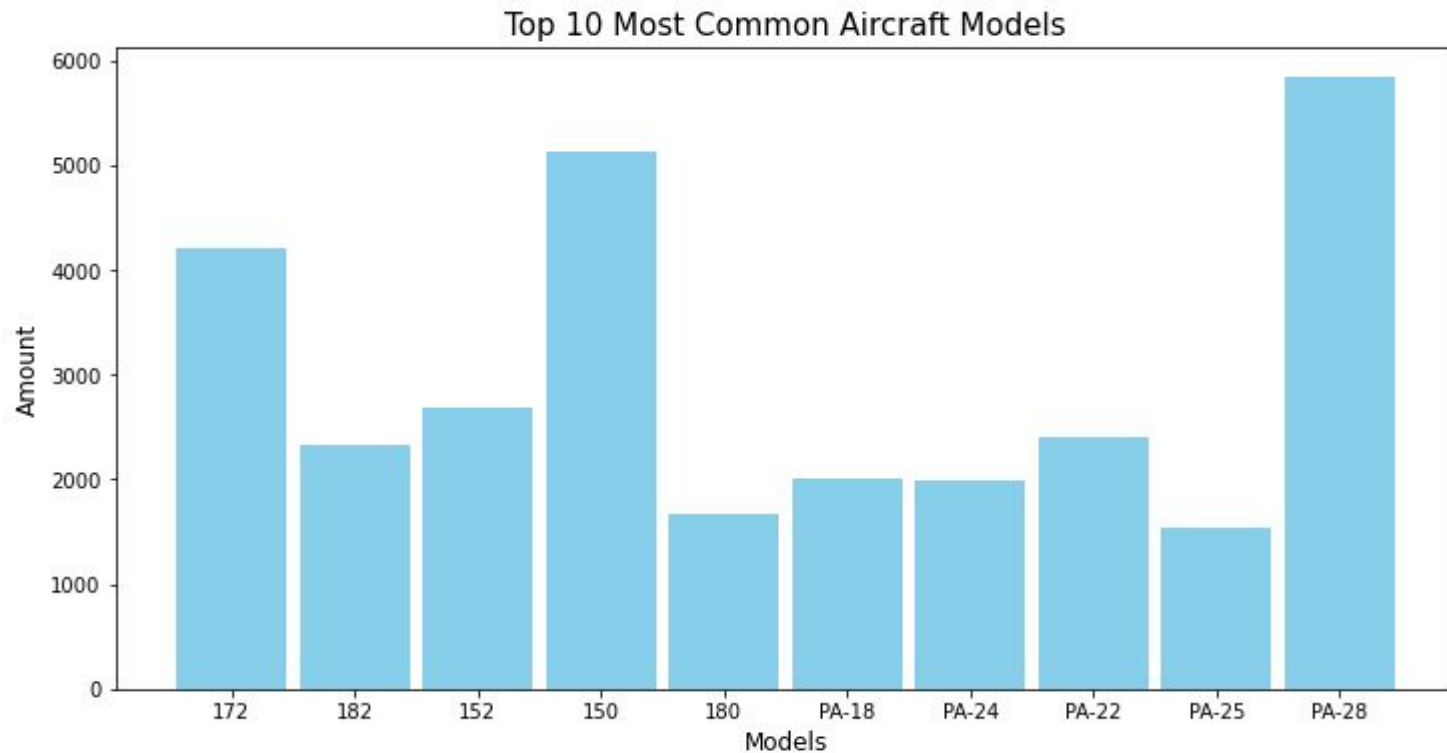
Key takeaways:

1. Logistic regression model (total injuries & event month)
 - a. It is crucial to prevent aircrafts from damaging in the first place.
 - b. Also, it might be more cost effective to focus on prevention than damage repairs in the long term.
2. Random Forest & SVM (weather condition & phases of flights)
 - a. FAA could consider to refine some of the existing policies to minimize damage levels.

Chapter 2: Top 10 Most Common Makes



Chapter 2: Top 10 Most Common Models



Chapter 2: A Little Info

- The Cessna presence on the Top 10 makes/models can be accounted for by the enormous presence of Cessnas in the sky: the [Cessna 172 Skyhawk](#) is the most widely-produced light aircraft of all time.
- The [Piper PA-28 Cherokee](#) is the model most commonly seen in NTSB crash data. This can be accounted for primarily due to the aircraft's popularity, but also due to the fact that the PA-28 is frequently used as a [training aircraft](#) in flight schools.



Cessna 172



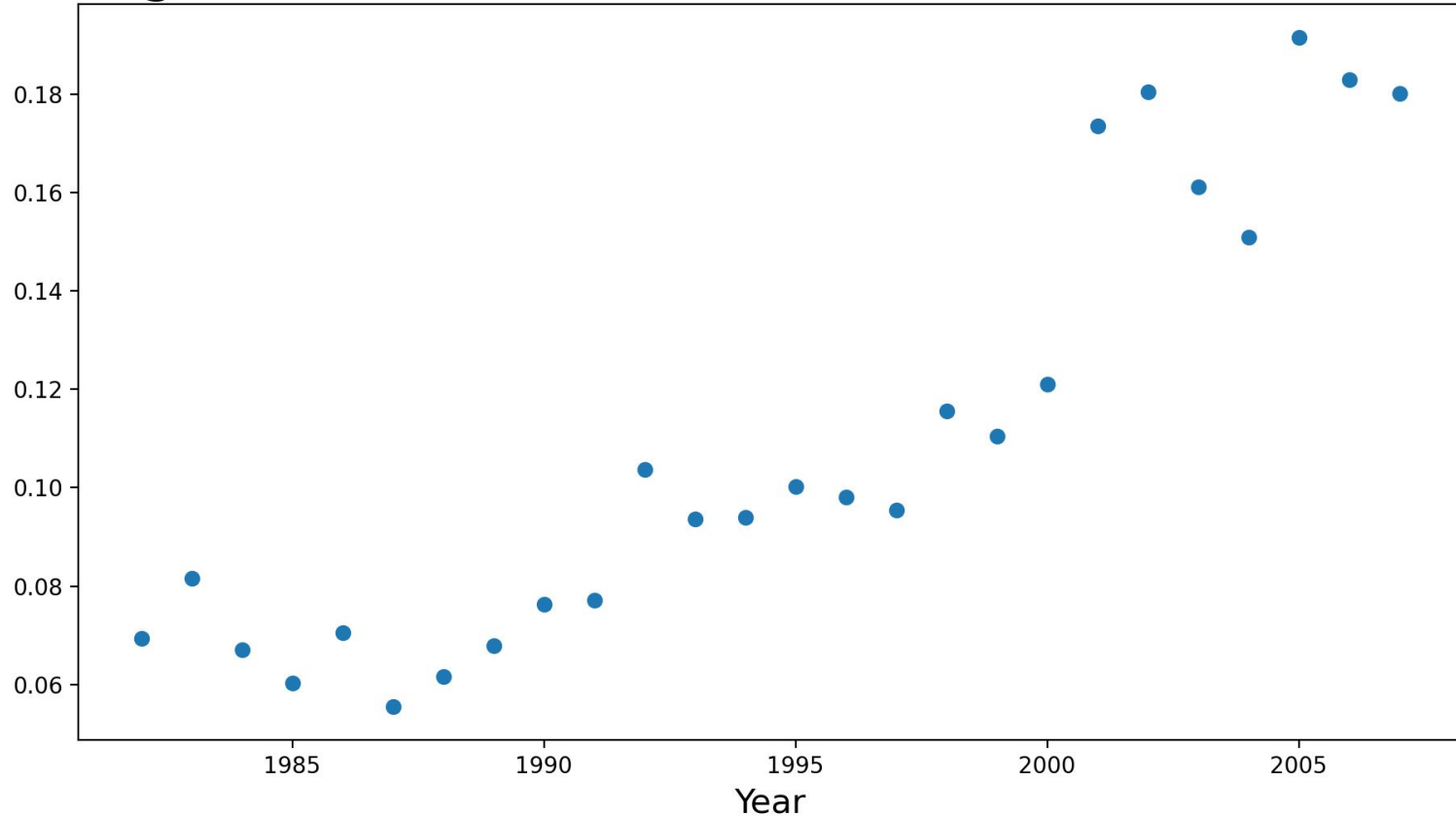
Piper PA-28

Chapter 2: Modeling

Given our results, the FAA has some concerns about the aircraft models that show up most often in this NTSB crash data set. They've asked us to run some models to see if we can predict which models will be troubling in the future, based on the data we have. So, these models were designed to optimize for accuracy while using multiclass classification to predict aircraft models.

Model	Accuracy Score
Baseline	19.58%
Neural Network (1 hidden layer)	42.87%
Random Forest	43.79%

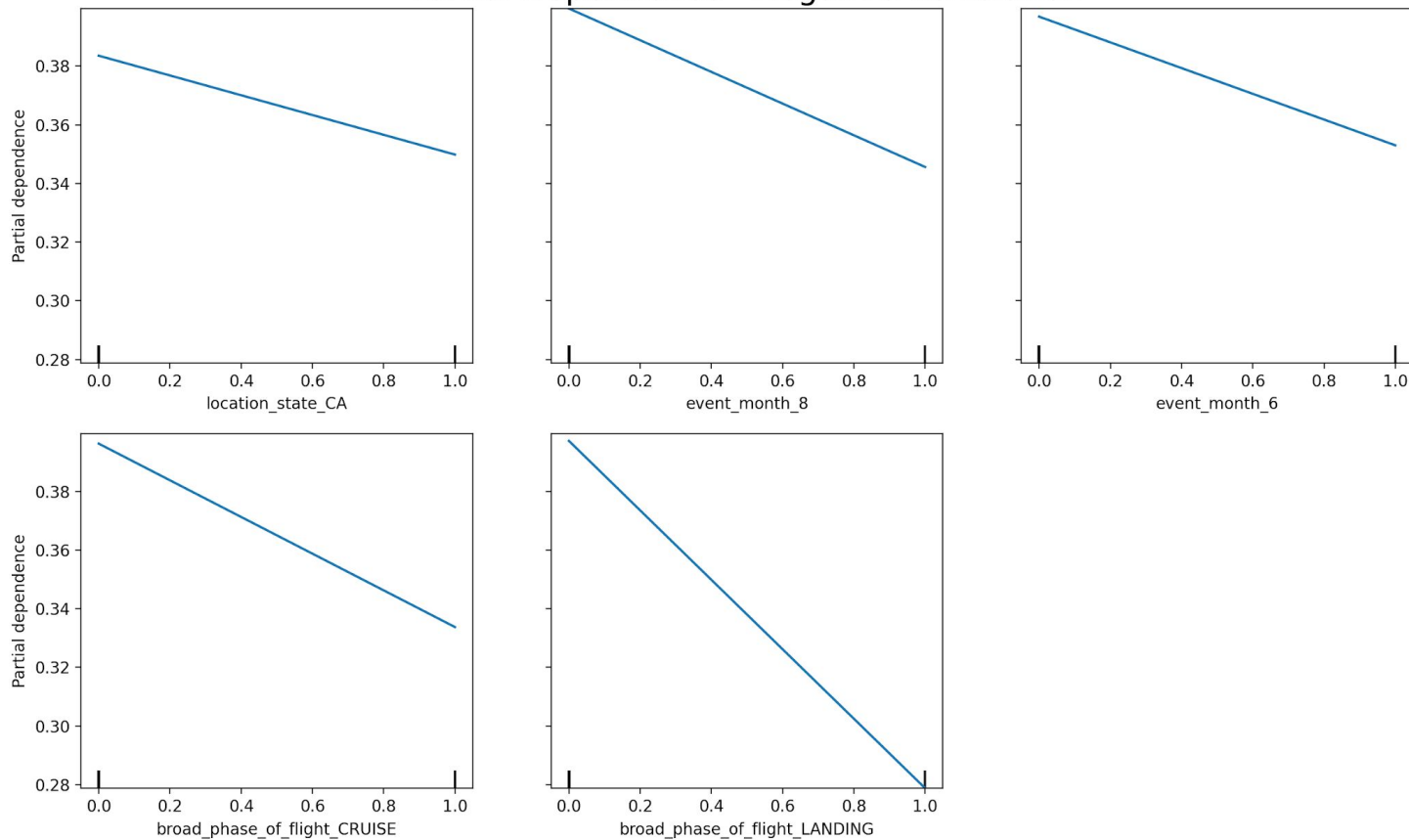
Percentage of Crashed Aircraft that are Amateur Built, by Year



Chapter 3: Amateur Classification

- Is it possible to predict if a craft was `amateur_built` based on the crash information?
- The Problem - baseline accuracy:
 - Non-amateur-built aircraft are over 91% of my data!
 - Initial models struggle to exceed this high bar
- The Solution - Bootstrapping!
 - Added 26,000 random samples from the `amateur_built = 1` rows (with replacement)
 - Brought the value counts to a 60/40 split (non amateur built to amateur built)
- The Best Model: Random Forest
 - Training Score: 98.6%
 - Testing Score: 95.6%

Partial Dependence - Negative Correlation



Recommendations

- **Based on Chapter 1:** Damage prevention might be more cost effective than post damage response. FAA might consider to implement stricter safety policies so that we can see fewer flight accidents. By doing so, we believe we can also drastically reduce the financial loss caused by damaged aircrafts, especially those destroyed ones.
- **Based on Chapter 2:** Given that the results of every model shows that the Piper PA-28 aircraft is likely to remain the model most commonly seen in any crash data, the FAA would do well to make sure that, if the PA-28 is going to be used as a training aircraft, it is as safe as possible - perhaps stricter regulations for training aircraft or pre-training before actual flight.
- **Based on Chapter 3:**
 - Designate April as “Amateur Builder Month” (~10% of crashes in April are Amateur Built, one of the highest months of the year)
 - Partner with companies that produce builder kits meeting FAA minimum safety guidelines to promote and share best practices

Next Steps

- Updating FAA regulations is an annual process - analyze new data from NTSB
- Allow our team to scrape NTSB event records to create more features in our data and further refine and improve predictive models
- Refine models and increase accuracy and F1 scores for all production models

THE END