Predicting the Cause of Aircraft Accidents

Springboard Capstone 2
Ashley Grindal

Background

40% of US adults have a fear of flying

2.5% have a clinically diagnosed phobia

Flying has gotten safer, but what causes accidents, and how can airlines address it?

Can previous data help investigators determine the cause of accidents faster?

Data

Database of aircraft incidents in the US and international waters

Data used in model: 2008-2021

FAR Part 121 (most passenger airlines)

FAR Part 135 (Small Carrier for both commuter and ‘air taxi’ small aircraft)

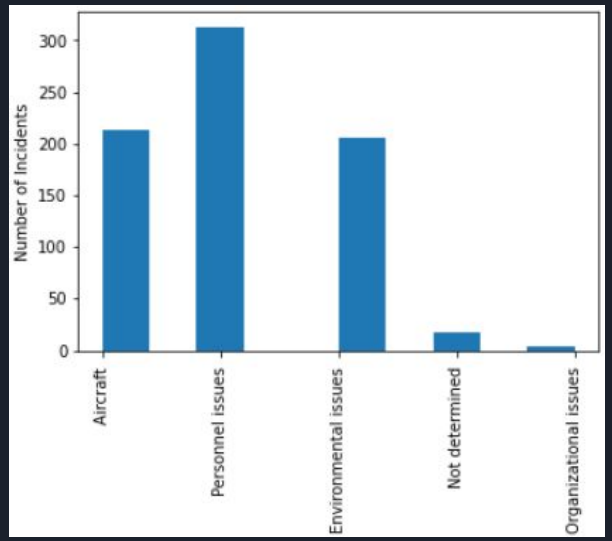
Data Includes:

- Accident date
- Highest Injury Level: none, minor, serious, fatal
- Injury Counts: minor, serious and fatal
- Location
- Accident Findings (main and supporting causes)

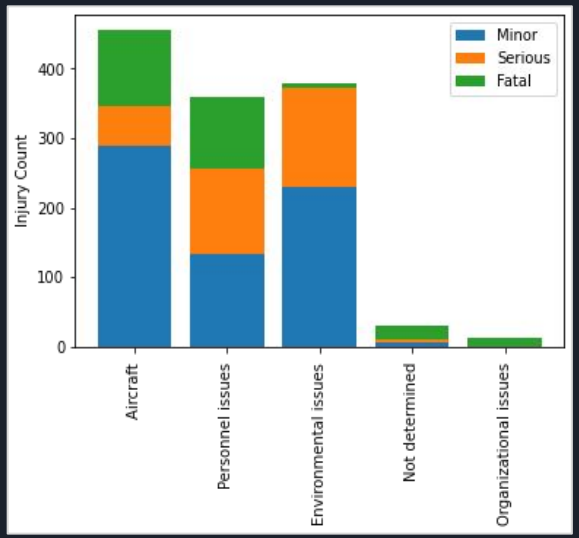
How does accident cause relate to the number of incidents and injuries?

Data Analysis [MAIN FINDING]

Incidents per Accident Type



Injuries per Accident Type



More incidents doesn't mean more injuries.
Aircraft cause the most injuries and the most fatalities.
Environmental has a high number of injuries but very few fatalities.

Data Analysis [SUPPORTING FINDINGS]

Aircraft	**Most incidents: 'Not attained/maintained' and 'Failure'** **Most Injuries: 'Fatigue/wear/corrosion'**	• Related to maintenance issues • Related to improper use by onboard crew
Personnel	**Most incidents: 'Pilot'** **Most injuries: 'Pilot'**	• More training • Possibly more health checks if related to medical issue
Environmental	**Most incidents: 'Turbulence'** **Most injuries: 'Turbulence'**	• Top concern of anxious flyers • No fatalities in the last 13 years, and no crashes in modern era • Injuries from passengers not properly seated and buckled

Pre-Processing

Accident Type	Number of Accidents
Aircraft	214
Personnel	312
Environmental	206
Not Determined	17
Organizational	4

Other	21

- Organizational issues and Not determined were combined into one category
- Location was broken into the nine US Census regions and divisions along with Pacific Ocean, Caribbean and Other
- Feature variables: month of incident, region (location), injury counts, highest injury level, FAR part number
- Target Variable: accident ‘Finding’

Pre-Processing & Modeling

Initial Steps

- One Hot encode categorical features
- Split data into train/test sets with 70/30 split
- Scale injury count numbers

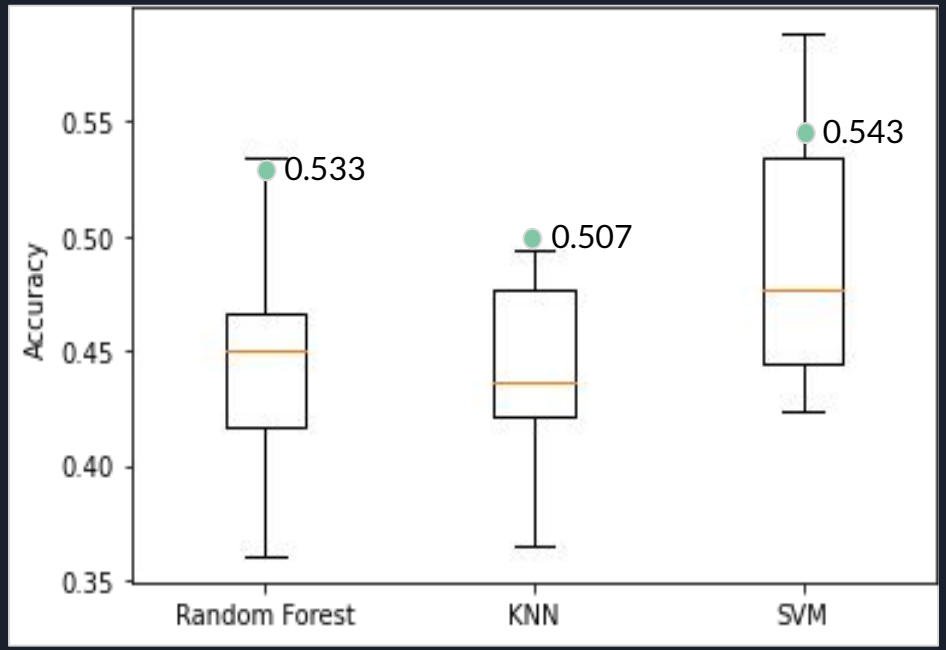
Class Imbalance

- Tune class weights in random search for each model
- Oversample the two smallest classes using SMOTE

Modeling Steps

- Supervised classification models: Random Forest, K-Nearest Neighbor, Support Vector Machine
- Determine optimal parameters and evaluate performance on test data
- Evaluation method: accuracy and confusion matrix

Modelling



Initial Cross Validation Comparison

Accuracy after tuning model parameters and class weights

Results

Model	Training Acc	Testing Acc	Precision	Recall	f1-score
RF	0.568093	0.533937	0.523804	0.533937	0.519987
KNN	0.811284	0.506787	0.496050	0.506787	0.498260
SVM	0.550584	0.542986	0.545119	0.542986	0.535635
Oversampled RF	0.734633	0.520362	0.535497	0.520362	0.523450
Oversampled KNN	0.653673	0.479638	0.538317	0.479638	0.496952
Oversampled SVM	0.646177	0.547511	0.553316	0.547511	0.539050

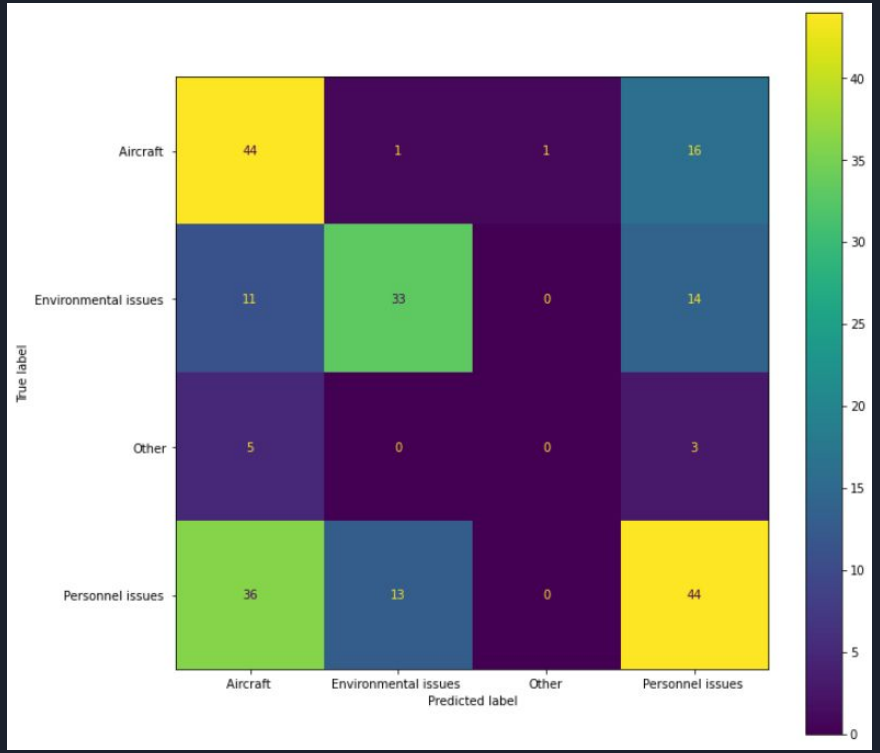
Support Vector Machine with oversampling is the best option

parameters: C = 100, class_weight = {'Aircraft ': 1, 'Environmental issues ': 1, 'Personnel issues ': 1, 'Other': 1}, gamma = 0.01.

Results

Most difficulty distinguishing aircraft and personnel.

Limitations: very little supporting data for the Other class which is made up mostly of ‘Not-Determined’ accidents



Future Work

- Best model only reached a 55.3% accuracy
- More historical data is available, but it is structured in a different way so it would need more processing and input from someone with aircraft investigation knowledge
- Does type of aircraft affect model?