Is air travel becoming safer?

Learning From Data

James Thomas

December 2024

Research Aims

This research aims to answer:

- Are planes crashing more or less now?
  - Regression
- Can we capture this trend in a model?
  - Regression
- What factors cause aircraft crashes the most?
  - Clustering
- What changes the severity of a crash?
  - Classification

Motivation

- Air travel is considered a high risk transport, despite being one of the safest [1]
- Widespread fears related to flying
- Models may help us predict frequency and danger of future plane crashes
- Raising awareness of the risk associated
- Potential to better predict circumstances of crashes in order to mitigate them

The Data

- The Dataset was found on Kaggle.com, published by Abe Ceasar Perez [2]
  - `https://www.kaggle.com/datasets/abeperez/historical-plane-crash-data`
- All crashes from 1918 to 2022
- This is a reputable website for dataset finding and sharing, trusted by many data scientists [3]
- The data is sourced through webscraping the Bureau of Aircraft Accident Archives [4]
  - `https://www.baaa-acro.com/index.php/crash-archives`

Dataset Overview

	Date	Time	Aircraft	Operator	Registration	...	PAX fatalities	Other fatalities	Total fatalities	Circumstances	Crash cause
0	1918-05-02	NaN	De Havilland DH.4	United States Signal Corps - USSC	AS-32084	...	0.0	0.0	2	The single engine airplane departed Dayton-McC...	Technical failure
1	1918-06-08	NaN	Handley Page V/1500	Handley Page Aircraft Company Ltd	E4104	...	0.0	0.0	5	Assembled at Cricklewood Airfield in May 1918,...	Technical failure
...	...	...	...	...	...	...	...	...	...	...	...
28534	2022-05-29	10H 7M 0S	De Havilland DHC-6 Twin Otter	Tara Air	9N-AET	...	19.0	0.0	22	The twin engine airplane departed Pokhara City...	Human factor
28535	2022-06-03	13H 46M 0S	Cessna 208B Grand Caravan	GoJump Oceanside	N7581F	...	1.0	0.0	1	The single engine was completing local skydivi...	Unknown

Figure: Dataframe Head

- Has shape: (28536, 24)
- Has fields: Date, Time, Aircraft, Operator, Registration, Flight phase, Flight type, Survivors, Crash site, Schedule, MSN, YOM, Flight no., Crash location, Country, Region, Crew on board, Crew fatalities, Pax on board, PAX fatalities, Other fatalities, Total fatalities, Circumstances, Crash cause

NaN Analysis

- Date: 0
- Time: 14587
- Aircraft: 1
- Operator: 0
- Registration: 815
- Flight phase: 638
- Flight type: 57
- Survivors: 1297
- Crash site: 383
- Schedule: 8946
- MSN: 4182
- YOM: 5311
- Flight no.: 28536
- Crash location: 12
- Country: 1
- Region: 1
- Crew on board: 24
- Crew fatalities: 1
- Pax on board: 54
- PAX fatalities: 1
- Other fatalities: 10
- Total fatalities: 0
- Circumstances: 25
- Crash cause: 0

Regression

Are planes crashing more or less now?
Can we capture this trend in a model?

Linear Regression

- Running a simple Linear Regression on the number of crashes per year

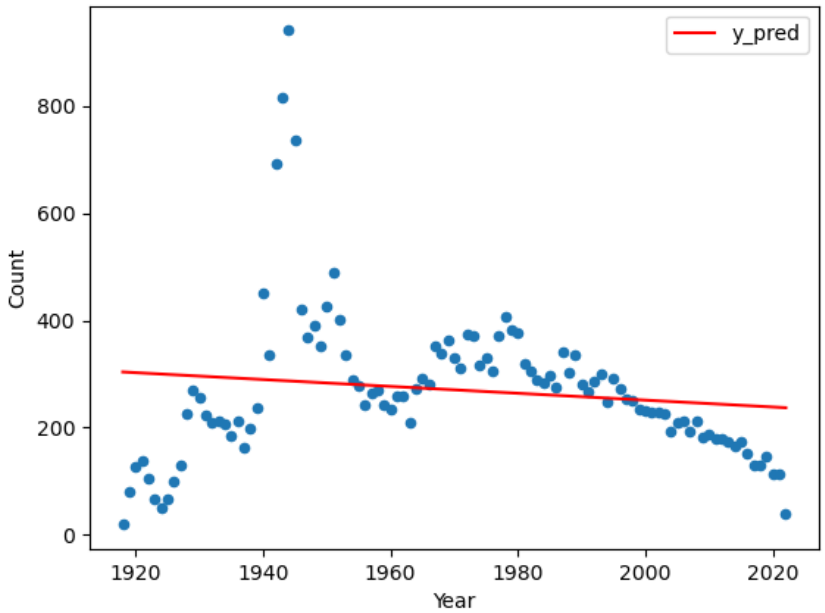

Figure: Simple Linear Regression

- Slope gradient of -0.641 (3sf)

Linear Regression - The Problems

- The years between 1940 and 1950 are very high
- The shape of the scatter is not linear
- R^2 is 0.0188 (3sf)
- Just plain bad!

Linear Regression

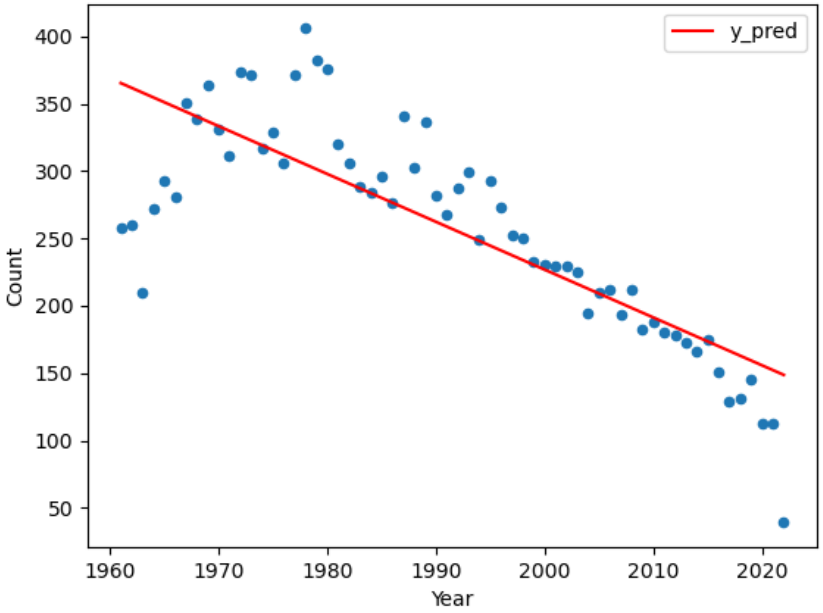

Figure: Linear Regression Post 1960

- R^2 is now 0.658 (3sf)
- Can we do better?

Polynomial Regression - BIC

- Analysing the Bayesian Information Criterion values for polynomials

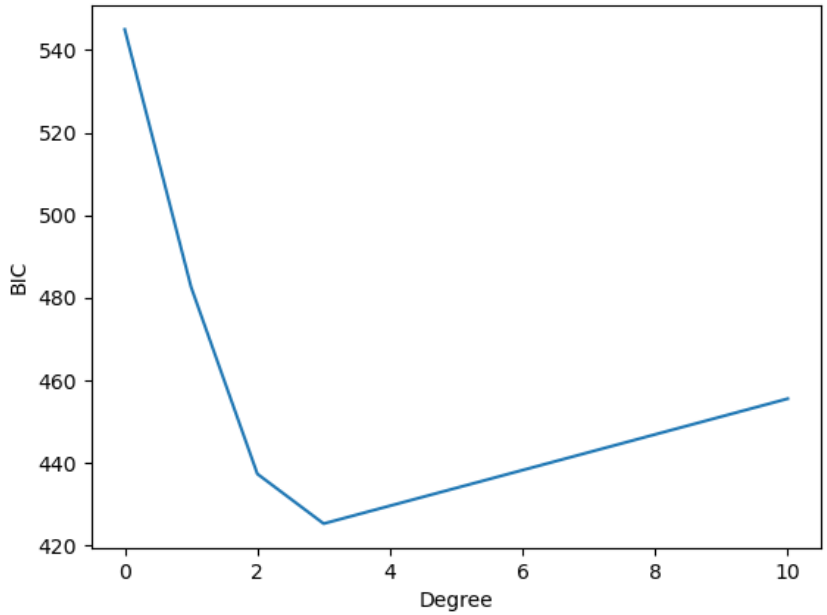

Figure: BIC Analysis for Polynomial Regression

- Lets select a Degree of 3

Polynomial Regression

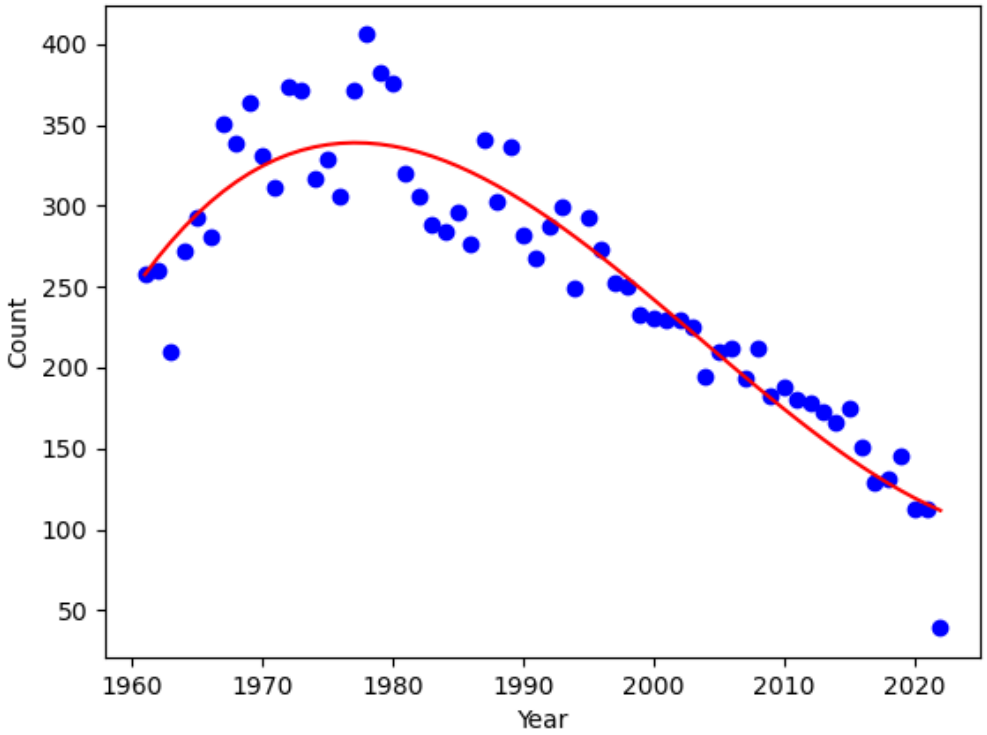

Figure: Polynomial Regression of Degree 3

- R^2 of 0.881 (3sf)
- Can we predict future numbers of crashes?

Polynomial Regression - Cross Validation

- K-Fold CV with k=5

Fold	MAE	R^2
1	17.7	0.880
2	26.9	0.868
3	26.9	0.881
4	19.2	0.881
5	24.9	0.878

Table: CV Results on Polynomial Regression (3sf)

- Mean MAE: 23.1 (σ: 4.37)
- Mean R^2: 0.878 (σ: 0.00550)

Limitations

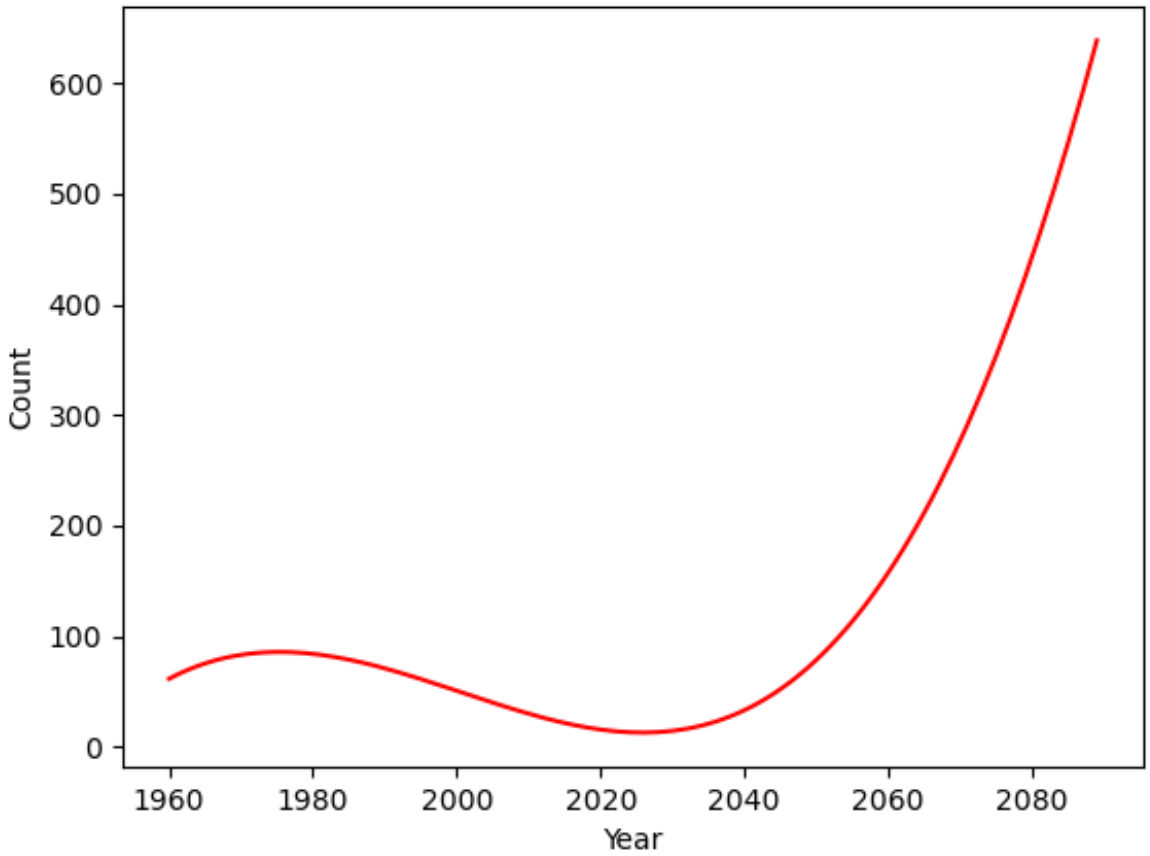

Figure: Future Prediction from Polynomial Regression

Neural Network Regression

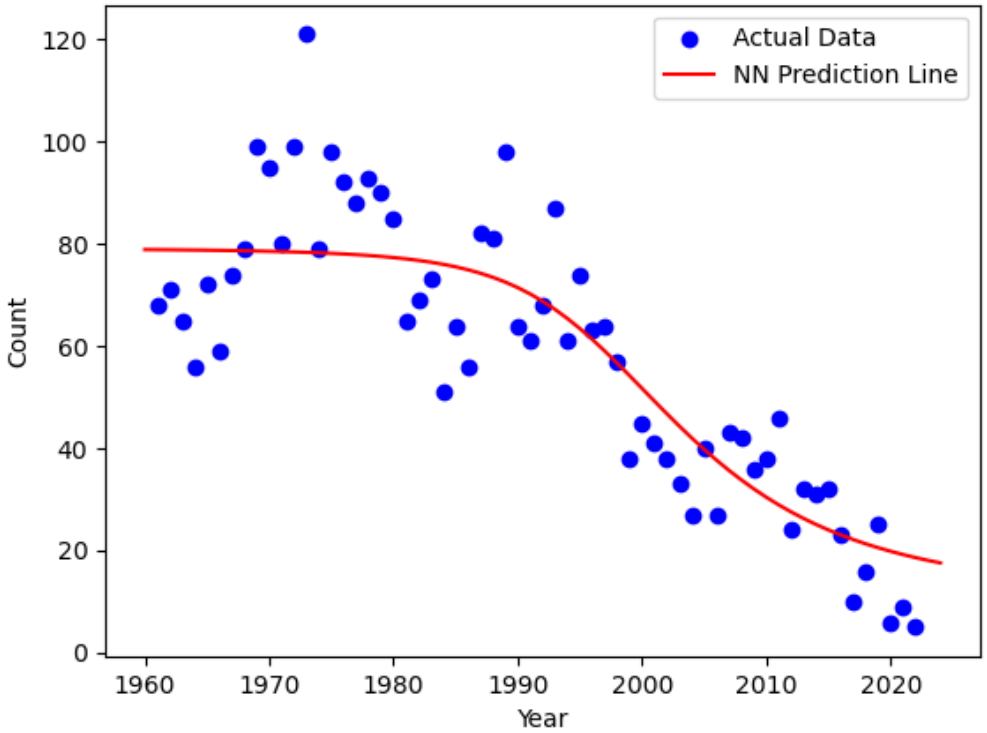

Figure: Neural Network Regression

Neural Network Regression - Cross Validation

- K-Fold CV with k=5

Fold	MAE	Loss
1	8.90	96.8
2	12.9	266
3	17.1	344
4	8.74	110
5	7.96	80.5

Table: CV Results on Neural Network Regression (3sf)

- Mean MAE: 11.1 (σ: 3.86)
- Mean Loss: 180 (σ: 118)

Predictions

- Next year there "will" be 16 crashes
- Around a 0.00003% chance a plane you're on crashes
- Good odds?

Future Outlook

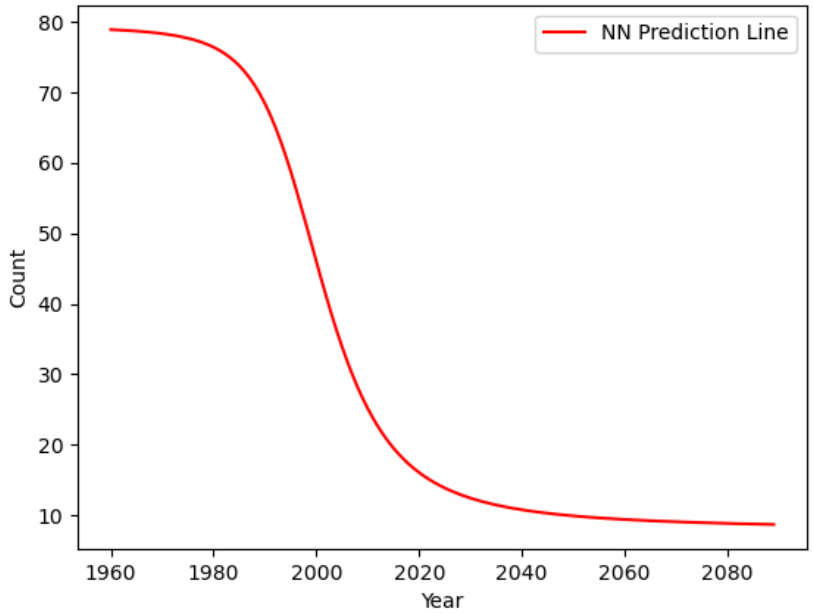

Figure: Future predictions from the Neural Network

Clustering

What factors cause aircraft crashes the most?

TF-IDF – Dimensionality Reduction

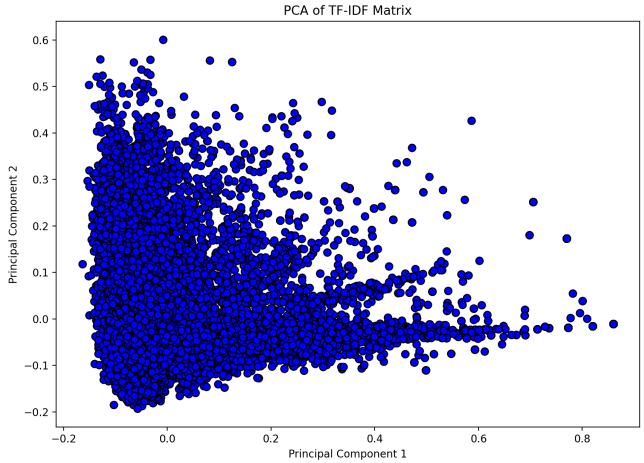

Figure: PCA on 'Circumstances' TF-IDF

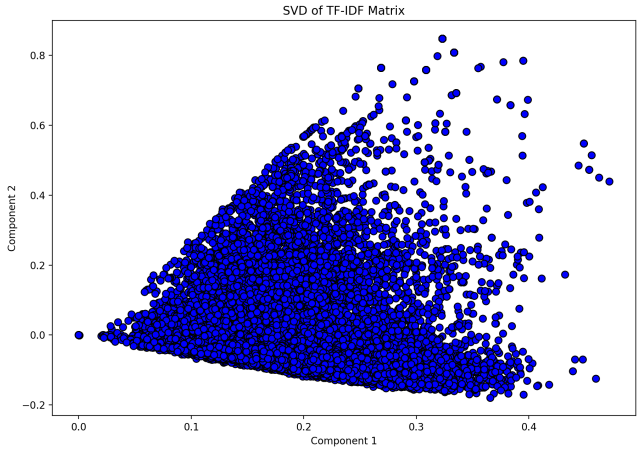

Figure: SVD on 'Circumstances' TF-IDF

DBSCAN

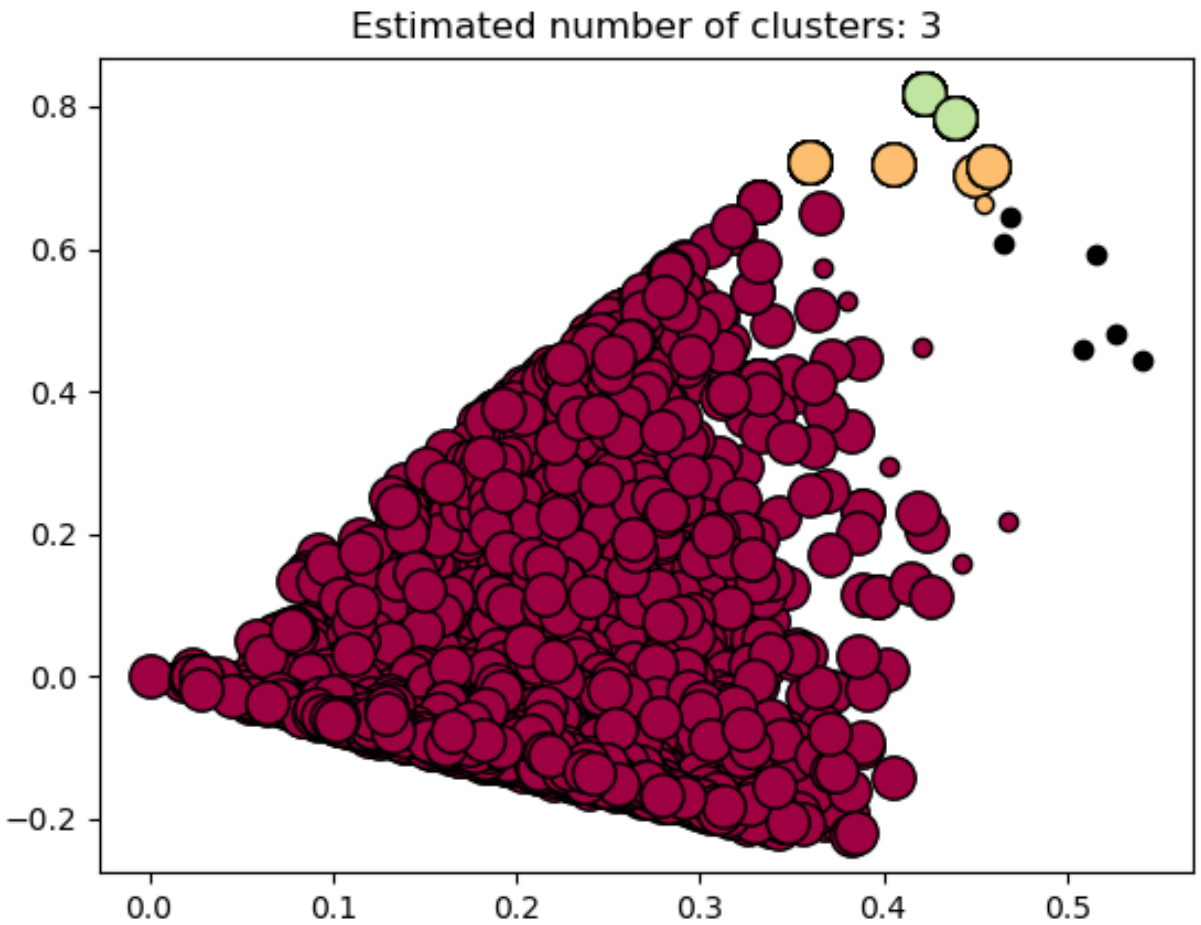

Figure: DBSCAN On TruncatedSVD Results (d=0.05)

Clustering Results

- Inconclusive
- Perhaps works better on whole dataset, not just text field
- Outliers and noise can be picked out
- More data is needed
- Circumstances of planes that did not crash would be useful

Classification

What changes the severity of a crash?

Severity Labelling

		Total Fatalities (F)			
		$F=0$	$0<F\leq 1$	$1<F\leq 3$	$3<F$
Death Rate (D)	$D=0$	No	NA	NA	NA
	$0<D\leq 0.3$	NA	Low	Medium	Medium
	$0.3<D\leq 0.6$	NA	Low	High	High
	$0.6<D\leq 1$	NA	Medium	High	Very High

Table: Severity Rating Quantifier Table

Preprocessing for Classification

- Remove rows not needed
  - Region, Time, YOM, Crash location, fatality columns, date, etc.
- Vectorise 'Circumstances' using TF-IDF, then reduce dimensionality
- Impute an aircraft age of 15 years if year of manufacture missing
- Impute values for onboard/fatalities based on aircraft average
- 'Unknown' replacing missing categorical data
- Remove remaining NaNs since so few
- One Hot Encode categorical data
- Z-Score normalise the X data
- Label scale the 'Severity' target
- SMOTE to balance the dataset
- Test-Train split

Random Forest

- First tried to treat this as a regression problem on 'Death rate' using an ANN
- Random Forest classifier was chosen instead
- Cross Validated Grid search for hyper-parameters

Random Forest Training

Fold	Precision	Recall	F1 Score	Support
0	0.81	0.73	0.77	1671
1	0.93	0.83	0.88	1662
2	0.91	0.82	0.86	1678
3	0.61	0.72	0.66	1646
4	0.71	0.79	0.75	1778
Accuracy			0.78	8435
Macro Avg	0.79	0.78	0.78	8435
Weighted Avg	0.79	0.78	0.78	8435

Table: Random Forest Cross Validation Scores

Random Forest Confusion Matrix

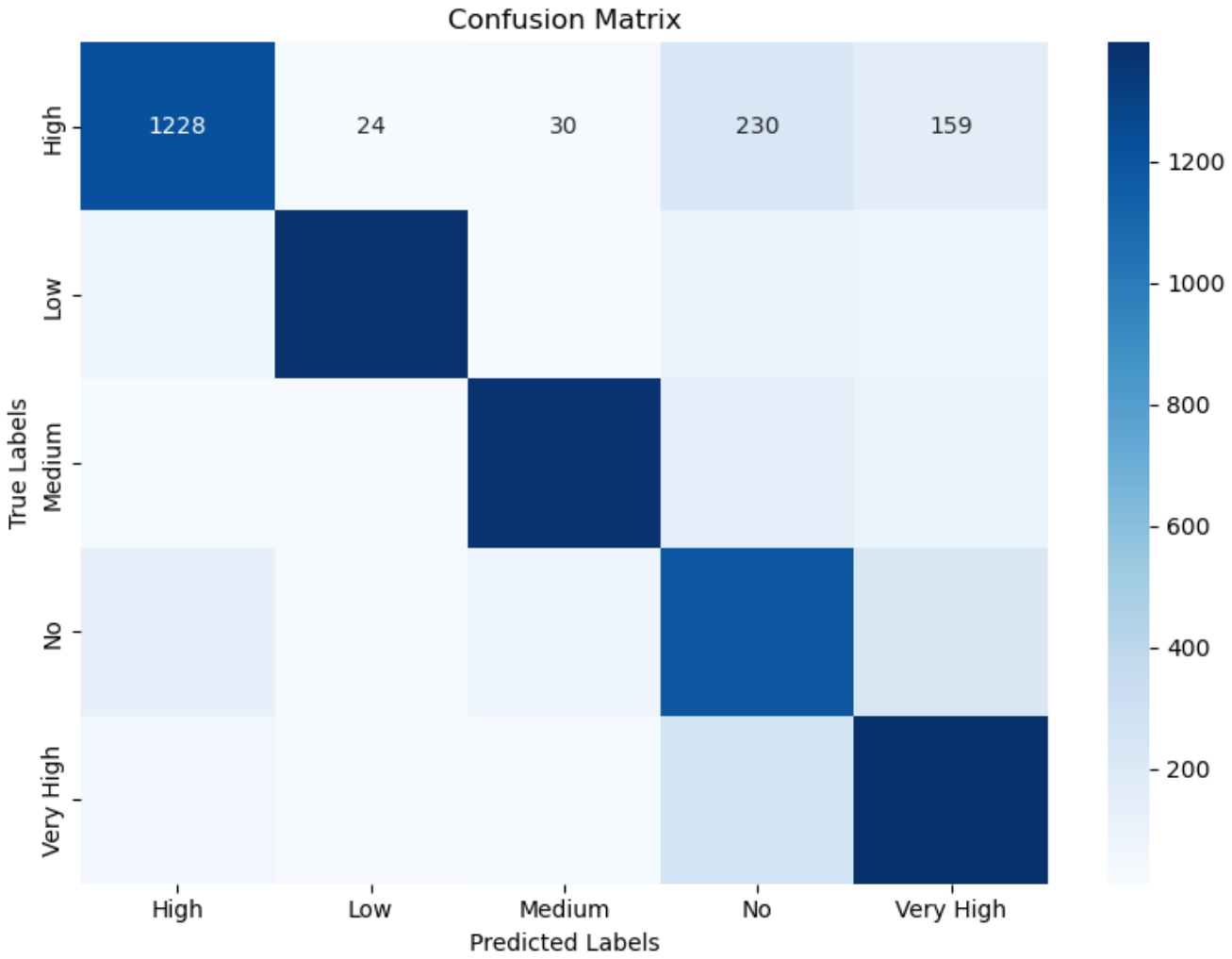

Figure: Confusion Matrix for Random Forest Classifier

Random Forest Feature Importance

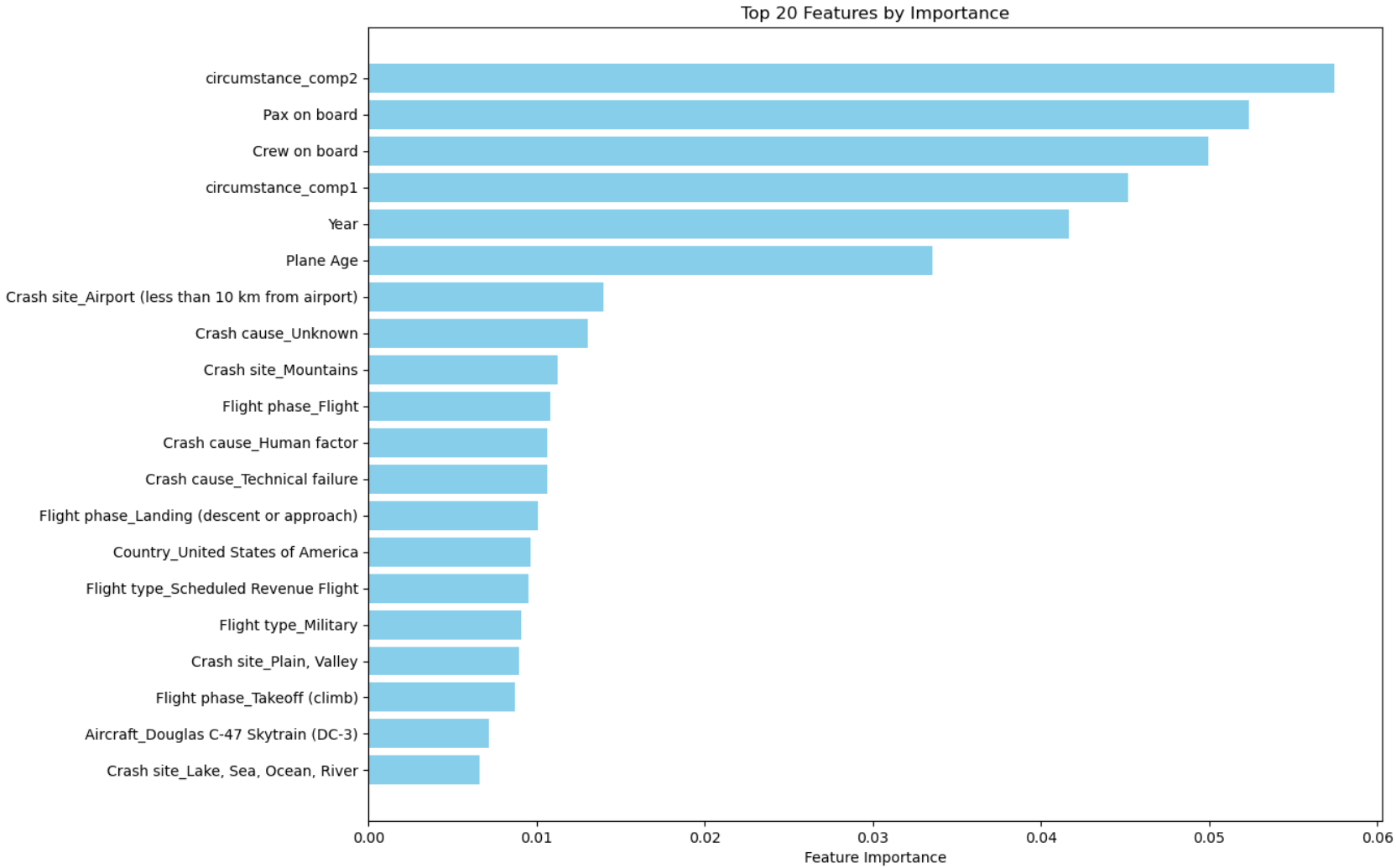

Figure: Feature Importance from Random Forest

Overall Results

My research aims to answer:

- Are planes crashing more or less now?
  - Less
- Can we capture this trend in a model?
  - Yes – Neural Network Predictor
- What factors cause aircraft crashes the most?
  - More data needed
- What changes the severity of a crash?
  - Age of the plane, year of the crash, number of people onboard, being near an airport, being over mountains, being mid-flight, technical failures, human factors

Conclusions

- Flying is remarkably safe
- Flying is safer than it ever has been
- Trends show that flying is likely to continue to become safer
- Somewhere between 12 and 16 commercial flights are estimated to crash in 2025
- Factors like people onboard, crash location, plane age, and crash cause are most indicative of severity

Limitations

- More data is needed
  - Dataset only contains crashed planes
  - Cannot completely understand factors of a crash without having data from planes that did not crash
  - Other data just is not there currently
  - Even if it was it would be enormous
- The dataset was not clean
  - Lots of missing data
  - Various reasons for this
  - Attempts made to mitigate this
  - Will impact model performance

Bonus Question – Does Red Bull give you wings?

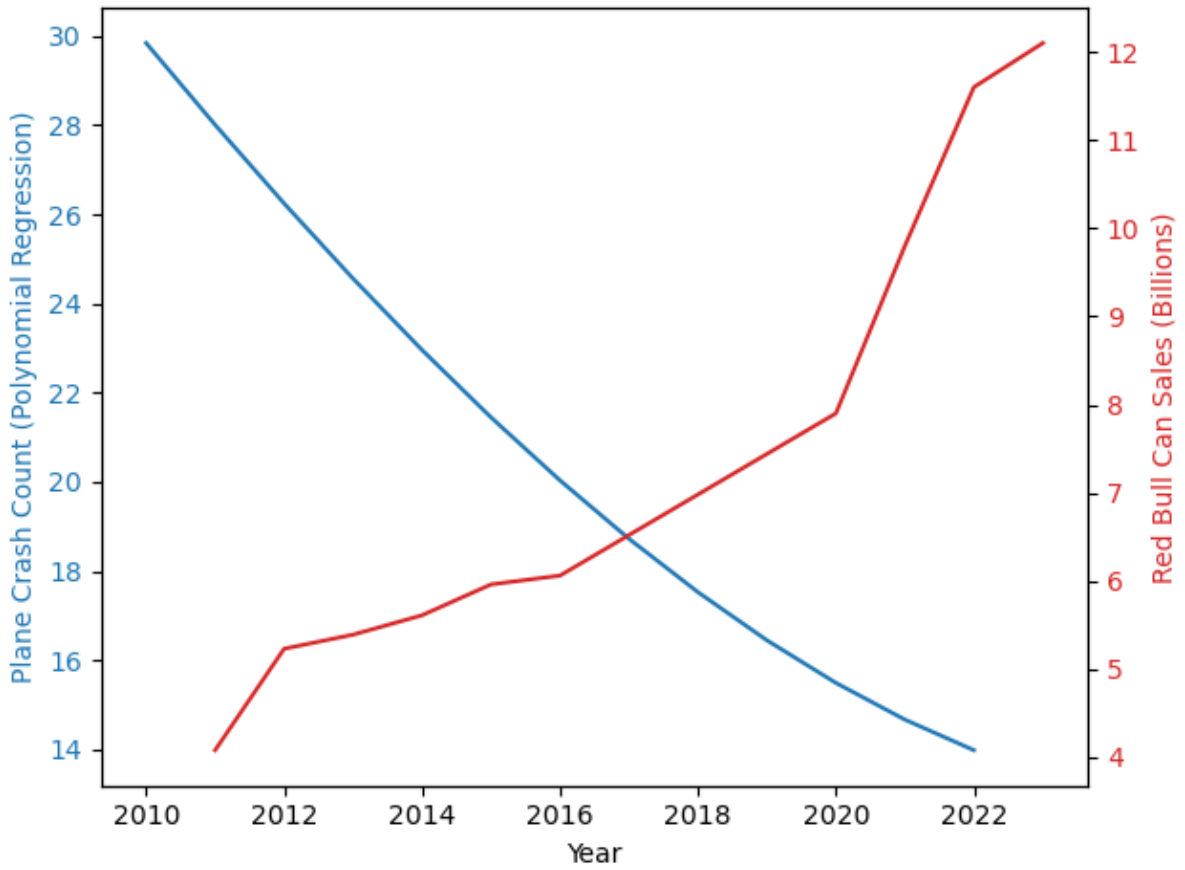

Figure: Red Bull Sales Plotted Against Commercial Plane Crashes

Correlation Coefficient: -0.882 (3sf)

Is air travel becoming safer?

Learning From Data

James Thomas

December 2024

References I

C. Ingraham, "The safest — and deadliest — ways to travel." https://www.washingtonpost.com/news/wonk/wp/2015/05/14/the-safest-and-deadliest-ways-to-travel/, 2015.
[Accessed 21-10-2024].

A. C. Perez, "Historical Plane Crash Data — kaggle.com." https://www.kaggle.com/datasets/abeperez/historical-plane-crash-data, 2022.
[Accessed 22-10-2024].

D. Editor, "Kaggle : All you need to know about this platform — datascientest.com." https://datascientest.com/en/kaggle-all-about-this-platform, 2023.
[Accessed 22-10-2024].

BAAA, "Accident Archives — Bureau of Aircraft Accidents Archives — baaa-acro.com." https://www.baaa-acro.com/index.php/crash-archives, 2024.
[Accessed 22-10-2024].