Is air travel becoming safer? – Learning From Data

James Thomas

1 Introduction

This project aims to address the safety of air travel over time. The question is rooted in widespread worry associated with planes; stemming from catastrophic incidents over time. With millions of passenger flights embarking worldwide each year, it begs the question of safety. Addressing whether flying is safe is somewhat ambiguous; understanding if flying is safer now than was in the past is what this project aims to address. Going further, Machine Learning (ML) models will be developed to understand the causes of incidents, and the factors indicative of their severity.

The paper by Mehta et al. [2] investigates the use of ML for crash severity prediction. The dataset contained crash coordinates, data on the engine type, and aircraft build quality, amongst others. Interestingly, severity prediction was treated as a categorical classification problem. Varying levels of aircraft destruction and fatality rates were combined to make 9 possible predicted classes. Random Forest (RF) performed best for this, closely followed by Artificial Neural Network (ANN). Gupta et al. [1] conducted similar research, aiming to classify the main causes of aircraft accidents. From classification models built, best accuracy came from XGBoost, followed by ANN.

A dataset on Kaggle.com contains all of the aircraft crash incidents from 1918 to the year 2022 [3]. The dataset was web-scraped from the reputable Bureau of Aircraft Accident Archives. The dataset comprises of 28,536 records containing the conditions of each incident. The accuracy of the records is limited by the knowledge documented of the circumstances. Data may be missing or incomplete, especially for older incidents or crashes in rural areas (MAR).

2 Regression

To establish how occurrences of aircraft crashes have changed over time, records per year can be counted to train a regression model. No preprocessing is required since no NaN values exist in the 'Date' column. A simple Linear Regression (LR) shows a slope gradient of -0.641 (3sf), proving a negative correlation between time and the number of plane crashes a year. Since annual flights have increased over time, this implies flying has become safer. The simple LR shown in Figure 1 yields a R^2 score of 0.0188, meaning an extremely poor model fit. A peak in crashes can be seen in the years of World War Two, due to the uptake of air battles during the war. These years are exceptionally high and prevent reasonable training of a model, as such data before 1960 can be omitted. To improve on simple LR, Polynomial Regression (PR) can be used to capture the complexity of the trend. To find an appropriate degree for the polynomial function, the Bayes Information Criterion (BIC) can be used. Figure 2 demonstrates that a polynomial with degree 3 yields the lowest BIC, therefore is the best choice for a regression model on this data. Using a train-test split on the dataset allows a model to be evaluated on unseen data, reducing the risk of overfitting. Over a 5-Fold Cross Validation (CV), there was an average Mean Absolute Error (MAE) of 23.1, with a relatively low standard deviation (SD) (4.37), meaning MAE is reasonably consistent. An average R^2 of 0.878 shows a great fitness score with a 0.0055 SD supporting consistency of the strong fitness. The PR model (shown in Figure 3) is limited by its degree 3 nature. Over time the gradient begins to increase and thus after 2030 the model predicts rising numbers of crashes, shown in Figure 4. In reality a model would be retrained regularly and predictions ten years ahead would not be trusted. However, this shows a PR model may perform inaccurately for future predictions. To address this, ANN architecture can be used for regression. A sequential model was built using Keras, with Dense layers intertwined with Dropout layers for preventing model overfitting. Sigmoid was chosen as the activation function of the layers, due to its ability to capture non-linear relationships. After some manual hyper-parameter tuning, 100 epochs and a batch size of 4 gave best results during testing. The optimiser "Adam" was chosen due to fast performance and the MAE error metric was

chosen to allow direct comparison to PR. The model can be seen in Figure 5. 5-Fold CV shows an average MAE of 11.1, with a SD of 3.86. The ANN predicts far more accurately, and consistently than PR. Future predictions using this model will perform better than PR, due to the flattening of the curve: indicating continuous low numbers of crashes in future. Figure 6 shows this.

3 Clustering

Understanding the main causes of crashes presents a clustering problem. The 'Circumstances' field of the dataset contains a description of crashes, including notes on weather condition, aircraft faults, and pilot error. Words can be stemmed and, using TF-IDF values, entries can be vectorised to a vector of length 61,645. This is extremely sparse so Principal Component Analysis (PCA) was performed to reduce the dimensionality to two components before clustering. Due to the size of the dataset, PCA was extremely inefficient and results seen in Figure 7 took over an hour to render. This is infeasible for testing. As an alternative, Sklearn's TruncatedSVD was used. This performs dimensionality reduction optimised for sparse NLP vectors. Changing the number of iterations of TruncatedSVD made little change to the outcome. TruncatedSVD results are shown in Figure 8, colour coded by the fatality proportion of the data point. Using DBSCAN to cluster results, just three clusters can be extracted when using a distance of 0.05, chosen due to being very small. Two of the three clusters contain very few data points (2 and 4 points each), highlighting outlier clusters rather than a valuable subset. Figure 9 shows these findings. Using a smaller link distance yields no better results as reducing the metric further creates too many clusters. For meaningful clustering, it is likely that samples of another class (i.e. planes that did not crash) would be required. This would highlight some of the differences in flight circumstances and produce hopefully more meaningful clusters. This data is not handily available and as such it is difficult to use this dataset to find the main thematic causes of aircraft incidents.

4 Classification

To understand factors changing the severity of aircraft crashes, a classification model was trained to predict severity. Severity quantifiers of 'No', 'Low', 'Medium', 'High', or 'Very High', were assigned to data based on death rate and fatality total. To prepare the data, NaN values for year of manufacture were imputed to be 15 years prior to the accident to reflect average aircraft age (MAR since likely missing data for registration). Scalers were applied to numerical features and categorical features were One Hot Encoded. 'Unknown' was imputed for NaNs in any categorical field – MAR since most likely due to poorly documented crash circumstances. Any records with missing or inconsistent onboard/fatality counts were classed MAR, due to age of records, and average onboard/fatality for that aircraft across the dataset was imputed. SMOTE was applied to 'Severity', addressing class imbalance. Random Forest (RF) classification was chosen due to recommendations in literature [2] and resistance to noise. Performing a CV Grid Search on hyper-parameters found an optimal RF architecture yielding a strong 0.780 accuracy and 0.79 & 0.78 for Precision & Recall. Feature importance can be extracted from this model, with the top 20 features shown in Figure 10. The representations of natural language from SVD rank high, making sense due to carrying sentiment regarding the severity. The number of people onboard strongly indicates the crash severity, as does the age of the plane and the year of the crash. Crash site is next most important, showing that crashing near the airport or in mountains changes severity most.

5 Conclusion

To summarise, data indicates that air travel has become safer over the decades and, using Neural Network based regression models, frequency of crashes is likely to continue to decrease year on year. Models indicate crash frequency of passenger flights is likely to plateau, but not reach zero, which is unfortunate but understandable given the volume of flights each year. Random Forest performed strongly at classifying the severity of incidents, with an accuracy of 0.780. Inspecting feature importances shows, plane age, being mid flight, crashing over mountains, or near the airport, are the main factors changing crash severity. Attempts to cluster similar aircraft crashes based on circumstances produced no clear insight due to a deficit of data relating to other flights.

Figures

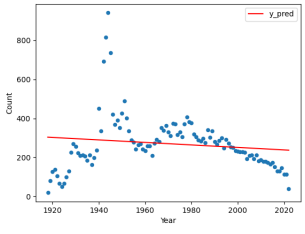

Figure 1: Simple Linear Regression

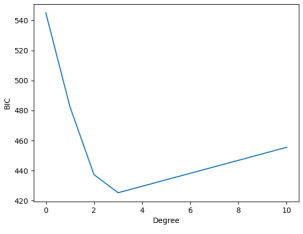

Figure 2: BIC Plot for Polynomial Degrees

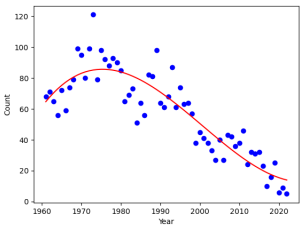

Figure 3: Polynomial Regression

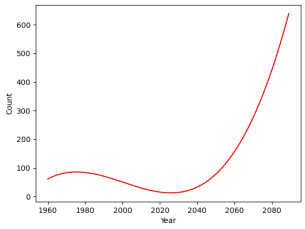

Figure 4: Polynomial Regression Future Prediction

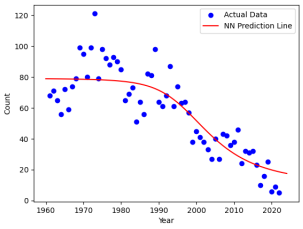

Figure 5: Neural Network Regression

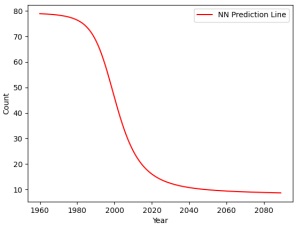

Figure 6: NN Regression Future Prediction

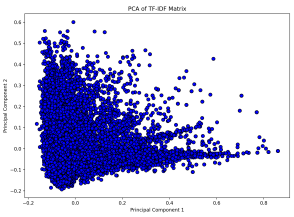

Figure 7: PCA Results on Tf-IDF

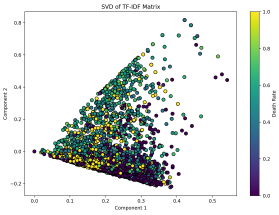

Figure 8: TruncatedSVD Results on TF-IDF

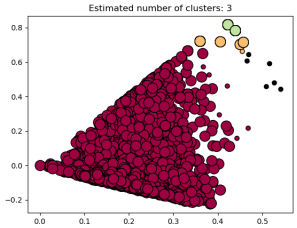

Figure 9: DBSCAN Clusters on SVD Results

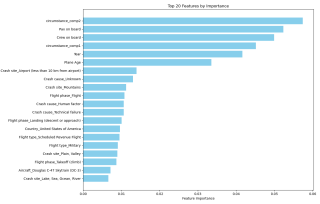

Figure 10: Feature Importance from RF Model

References

[1] Ved Prakash Gupta, M Sajid Mansoori, Jitendra Shreemali, and Payal Paliwal. Predicting causes of airplane crashes using machine learning algorithms. *International Journal of Recent Technology and Engineering (IJRTE)*, 2020.

[2] Jay Mehta, Vaidehi Vatsaraj, Jinal Shah, and Anand Godbole. Airplane crash severity prediction using machine learning. In *2021 12th International Conference on Computing Communication and Networking Technologies (ICCCNT)*, pages 1–6. IEEE, 2021.

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

Generative AI Statement

AI-supported/AI-integrated use is permitted in this assessment. I acknowledge the following uses of GenAI tools in this assessment:

- (NO) I have used GenAI tools for developing ideas.
- (NO) I have used GenAI tools to assist with research or gathering information.
- (NO) I have used GenAI tools to help me understand key theories and concepts.
- (NO) I have used GenAI tools to identify trends and themes as part of my data
- analysis.
- (NO) I have used GenAI tools to suggest a plan or structure for my assessment.
- (NO) I have used GenAI tools to give me feedback on a draft.
- (NO) I have used GenAI tool to generate image, figures or diagrams.
- (NO) I have used GenAI tools to proofread and correct grammar or spelling errors.
- (NO) I have used GenAI tools to generate citations or references.
- (YES) Other: I have used GenAI tools to resolve bugs in my code when the cause is not obvious.

I declare that I have referenced use of GenAI outputs within my assessment in line with the University referencing guidelines.

Appendices

Field Name	Description	Datatype
Date	Date of the incident (YYYY-MM-DD)	String
Time	Time of the incident (hhH mmM ssS)	String
Aircraft	Aircraft type, incl. manufacturer	String
Operator	Aircraft operator company	String
Registration	Tail No./Registration of Aircraft	String
Flight phase	Flight/Takeoff/Landing/etc.	String
Flight type	Military/Training/Cargo/etc.	String
Survivors	Whether there were survivors	Boolean
Crash site	Lake/Forest/Airport/etc.	String
Schedule	Planned route: Takeoff - Land	String
MSN	Manufacturer Serial Number	String
YOM	Year of Manufacture	Integer
Flight no.	The flight number assigned	String
Crash location	Local name of crash location	String
Country	Country name of crash	String
Region	Continent Name	String
Crew on board	Number of crew on board	Integer
Crew fatalities	Number of crew fatalities	Integer
Pax on board	Number of passengers on board	Integer
PAX fatalities	Number of passenger fatalities	Integer
Other fatalities	Other fatalities including ground fatalities	Integer
Total fatalities	Total of the three classes of fatality	Integer
Circumstances	Summary of crash circumstances and occurrence	String
Crash cause	Technical Failure/Human Error/etc.	String

Table 1: Dataset Field Descriptions

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 2: CV Results on Polynomial Regression (3sf)

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

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

Table 4: Severity Rating Quantifier Table

Severity Class	Quantity	Percentage
No	8435	33.7
Low	1333	5.33
Medium	2225	8.90
High	4726	18.9
Very High	8283	33.1

Table 5: Severity Class Imbalance before SMOTE

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 6: Random Forest Cross Validation Scores

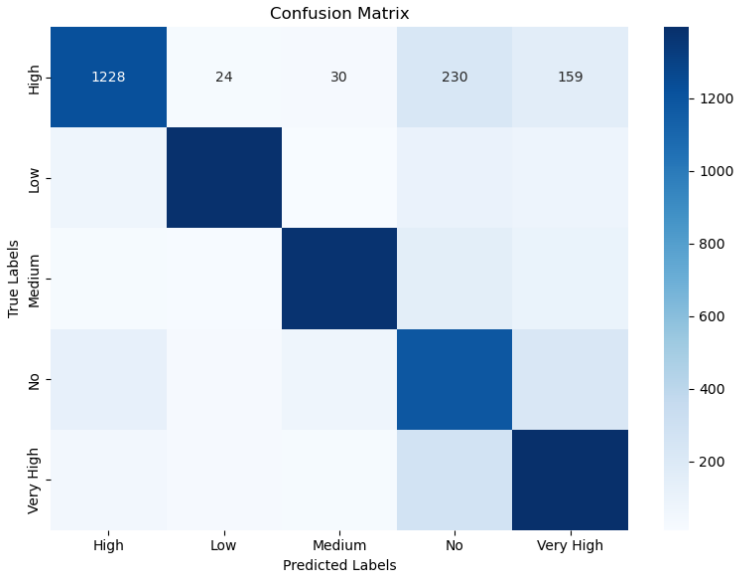

Figure 11: Confusion Matrix for Random Forest Classifier