Predicting Fatalities in Air Accidents using CHAID XGBoost Generalized Linear Model Neural Network and Ensemble Models of Machine Learning

Nikita Pande, Devyani Gupta, Jitendra Shreemali, Prasun Chakrabarti

Abstract—The study examines the historical data of about 4700 air crashes all over the world since the first recorded air crash of 1908. Given the immense impact on human beings as well as companies, the study aimed at utilizing Machine Learning principles for predicting fatalities. The train-test partition used was 75-25. Employing the IBM SPSS Modeler, the machine learning models used included CHAID model, Neural Network, Generalized Linear Model, XGBoost, Random Trees and the Ensemble model to predict fatalities in air crashes. The best results (90.6% accuracy) were achieved through Neural Network with one hidden layer. The results presented also include comparison of the predicted versus observed results for the test data.

Keywords— Neural Network, CHAID, Ensemble Model, XGBoost Model, Air accidents, Fatalities

I. INTRODUCTION

Transportation or travelling plays an important role in human's life. As compared to other forms of transportation, flying in an aeroplane saves time and covers large amounts of distances. It is a safe mode of transportation. On average a person is 65 times more likely to die in a car travelling the same distance as an airliner. The safety of commercial passenger accidents service remains a worldwide concern. Aviation accidents always result in great damage to human life and also to the economy of the country. Plane crashes are mostly fatal and deadly because of its size and weight.

Air accidents are caused due to pilot error, mechanical error, bad weather, sabotages or human mistakes. These errors or mistakes can happen any time. Flying mischance cases have a great degree of complexity.

Between 1990 and 2006, an average of over 1,000 passengers and 130 crew members died in commercial passenger service accidents every year.

Common Causes of Aviation accident

1. Pilot error: The general error that happened in an aviation accident is done by pilot, which accounts for approximately half of all plane crashes. Flying a plane is considered as a very complex and difficult job in today's world, despite modern technologies of features of air travel. For instance a pilot needs to monitor multiple maps and readouts the proper direction during the entire duration of the flight, something that could look extremely daunting to non-pilots or non-technical people. Any miscalculation or even a minor misinterpretation can result in a deadly crash.
2. Mechanical defects: If there are any defects or failure in any one system then the situation leads to a lethal outcome. The possibility of repairing or replacing faulty parts mid air is almost close to zero.
3. Weather problems: If in any situation the weather suddenly changes or bad weather will be there that situation is also dangerous for flying as well as hard to handle the flight for the pilot. Weather conditions such as heavy rainstorms, fog and snow makes it more difficult and dangerous and can lead to severe accidents.
4. Human error: When a pilot works for long hours some fatal errors may be committed by him while operating the aircraft which can lead to fatal results . Some other jobs which require a great amount of precision are the jobs of air traffic controllers,dispatchers and loaders and even a minor error on their end can lead to life-threatening situations.
5. Other: During wars or attacks by other countries through aeroplane the large amount of disturbance in environment or natural resources. In addition to this, while transporting highly inflammable substances a great deal of precaution has to be taken otherwise if there's any leakage or even a little spark, the result could be major system failures and eventual loss of life and property.

The paper is further structured as follows literature survey, data and methodology and conclusion and finding. The study helps to understand the real-time crashes, who are being affected and how to prevent this problem.These studies help to get the solution for the general population. An analysis has been made using crash-deceleration pulse data from a crash-dynamics program on general aviation airplanes and from transport crash data available in the literature.

Revised Manuscript Received on March 15, 2020.

Nikita Pande*, Techno India NJR Institute of Technology, Udaipur, India. E-mail: nikipande00@gmail.com

Devyani Gupta, Techno India NJR Institute of Technology, Udaipur, India.

Jitendra Shreemali, Techno India NJR Institute of Technology, Udaipur, India.

Prasun Chakrabarti, Techno India NJR Institute of Technology, Udaipur, India.

II. LITERATURE SURVEY

According to the ACRP Report 62 () ICAO, the aero must be monitored "regular, mandatory, systematic, and harmonized safety audits" these rules audits by the United State in 2007. According to Clinton V. ,Oster, Jr., the accidents categorized into the type of service being conducted when the accident occured. The type of service related to the passenger facilities and fatal accidents. The passenger facilities occured in the scheduled service for domestic as well as international where large aircraft are more commonly used.

Vane, R. (2016).Analysis of flight delay and how this system can be improvised is possible through big data. International Research Journal of Engineering and Technology (IRJET), 03 (06), pp 778-780.

DeAngelis, said that if in any case weather issues or some human error will be there then airlines should train pilots in such a way that they can solve the problem easily.If the important point to maintain and improve flight safety as airline and government must invest in human factors research.

III. DATA AND ANALYSIS METHODOLOGY

The data set used included over 5000 data points after cleaning the data set 4700 data points starting from the first reported air crash of the US Army flyer flown by Orville Wright. The data attributes include data on time, location of the air crash, the Operator, number of passengers abroad and details of fatalities.The study examines the possibility of predicting the number of fatalities based on the available data through using multiple machine learning algorithms. Data was taken from www.kaggle.com/cgurkan/airplane-crash-data-since-1908.

The models used included relatively simpler ones like multiple regression and more advanced ones like generalized linear model, random trees, XGBoost tree model, CHAID, the neural network and the ensemble model. The IBM SPSS Modeler tool was used to run the models with a training set comprising 75% of the data and test set comprising 25%. Findings/output of these models is presented below:

Neural Network: The neural network algorithm is inspired by the neural network present from different human brains. It consists of an input layer which comprises of all the features or properties observed in the input data, an output layer which identifies the cluster to which the input data belongs.Between the input and output layers, there will be one or more hidden layers present. Number of hidden layers depends on the requirements. In these layers neurons compute the weighted inputs to produce output for the next layer with the help of some activation function and biases. Neural network creates a layered directed weighted graph.

Ensemble Model: The ensemble model is a type of supervised learning algorithm.It is a hybrid model, in other words, two or more models are combined together to make predictions and to produce desired output from the given input data. This approach is used to minimize the prediction error in the final output. Instead of running various models on the same data set, it is significant to use a combination of those models to reduce time and complexity. Here classification can be done in a group.

Chi-squared Automatic Interaction Detection model: It applies a decision tree technique, based on adjusted significance testing. These can be used for prediction as well as classification, and find interaction between variables. It is a tool used to discover different relationships between variables and analysis builds a predictive model, or tree, to help determine how variables best merge to explain the outcome in the given dependent variable.

XGBoost Tree Model: XG Boost follows the decision tree model of Machine Learning algorithm which uses boosting framework. In prediction problems involving unstructured data. Artificial neural networks is the top all other algorithms or frameworks. This model performs well for small-to-medium structured/tabular data, decision tree based algorithms are considered best-in-class right now.

Random Forest Algorithm: Random forest is a supervised learning. It is an ensemble classifier obtained by bagging of decision trees that have been trained on randomly selected r dimensions out of d dimensions of input x, where r dimension selected randomly of tree learned in this way.

Generalized linear model: The generalized linear model refers to linear regression. In linear regression, there is a linear relationship between variables. In general form, y=mx+c, where the value of y and x may differ from different points.

Multiple linear regression: Multiple linear regression models refer to having two or more linear relationships between more variables.

It is used to predict the values from the given data.

The figure below gives a schematic of the IBM SPSS Modeler screen for the seven models used in several cases, namely, Neural Networks, Ensemble model, Chi-squared automatic interaction detection model, XGBoost tree model, Multiple linear regression, Random Forest Algorithm and Generalized linear model.

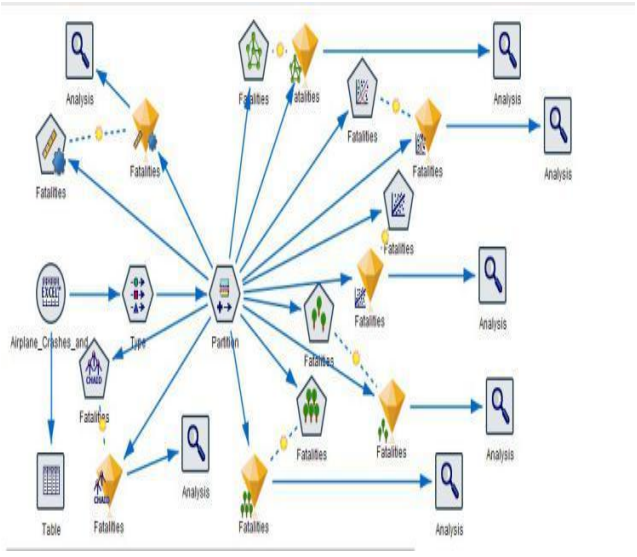

Fig 1: Models: IBM SPSS Modeler

IV. CONCLUSION AND FINDINGS

The findings are briefly presented below:

1. Classification using the Neural network shows over 90.6% correctly classified cases of airplane crashes from the test data.

Model Summary

Target	Fatalities
Model	Multilayer Perceptron
Stopping Rule Used	Minimum error ratio achieved
Hidden Layer 1 Neurons	15

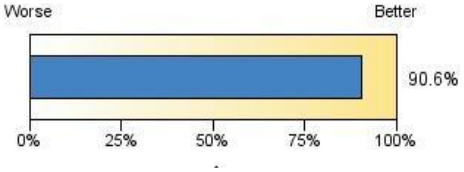

Fig 1: Neural Network Schematic

The neural network used is presented below:

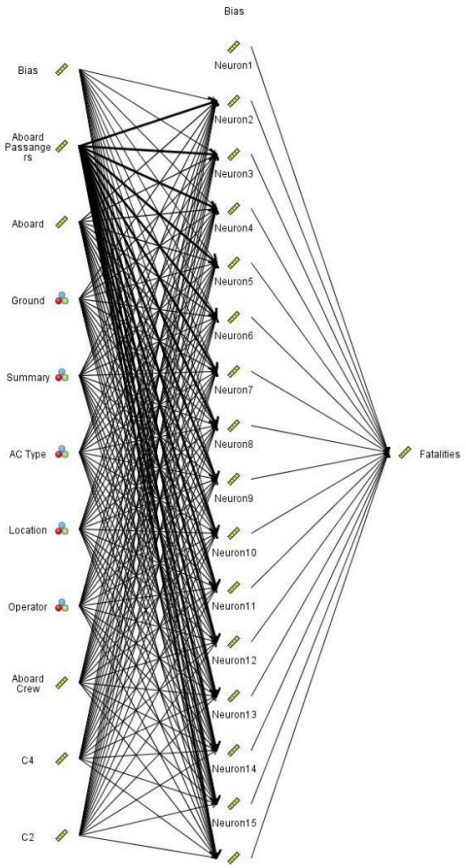

Fig 2: Neural Network Layers

2. The plot of actual versus predicted data shows a high level of resemblance (closeness) between the two, thus suggesting that the neural network model has the potential of being a useful predictive model for air casualties.

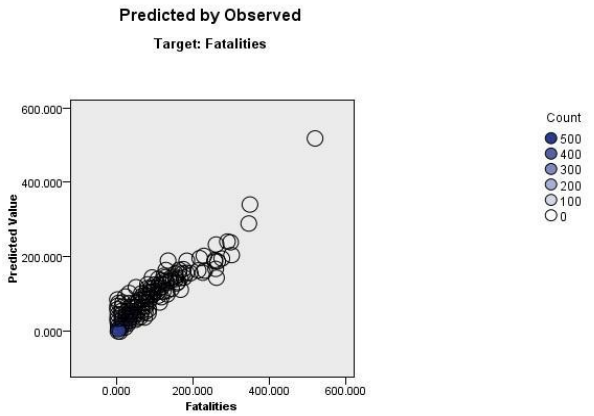

Fig 3: Predicted Observed

3. The CHAID model provides over 67.8% correct classification of the flight casualties data.

Table 1: CHAID Classification

'Partition'	1_Training	2_Testing
Minimum Error	-102.321	-102.321
Maximum Error	417.679	480.679
Mean Error	-0.0	-0.329
Mean Absolute Error	10.082	9.879
Standard Deviation	24.645	25.445
Linear Correlation	0.732	0.678
Occurrences	3,518	1,201

4. Ensemble classification of flight casualties data provides the correlation of 78.1% correctly classified cases of flight for the test data.

Table2: Ensemble Classification

'Partition'	1_Training	2_Testing
Minimum Error	-94.606	-161.689
Maximum Error	225.207	204.309
Mean Error	0.183	-0.43
Mean Absolute Error	7.363	9.527
Standard Deviation	15.716	21.539
Linear Correlation	0.91	0.781
Occurrences	3,518	1,201

Table 3: Ensemble Model

Sort by: Use | Ascending | Descending | Delete Unused Models | View:

Use?	Graph	Model	Build Time (mins)	Correlation	No. Fields Used	Relative Error
☑		Random ...	< 1	0.749	6	0.472
☑		Linear 1	< 1	0.747	4	0.454
☑		XGBoost ...	< 1	0.745	6	0.453

5. Using a generalized linear model, the airplane crashes showing the 74% correlation from the data.

Table 4: Generalized linear model

'Partition'	1_Training	2_Testing
Minimum Error	-302.863	-207.342
Maximum Error	211.633	199.329
Mean Error	-0.0	-0.089
Mean Absolute Error	13.328	12.884
Standard Deviation	27.197	25.928
Linear Correlation	0.751	0.74
Occurrences	3,518	1,201

6. XGBoost model gives the classification of 74.5% correlation.

Table 5: XGBoost Model

'Partition'	1_Training	2_Testing
Minimum Error	-114.22	-191.494
Maximum Error	102.02	248.018
Mean Error	0.68	0.214
Mean Absolute Error	7.301	9.968
Standard Deviation	13.9	23.135
Linear Correlation	0.929	0.745
Occurrences	3,518	1,201

7. Multiple regression works on different parameters show the relationship of 74%..

Table 6: Multiple linear regression model

'Partition'	1_Training	2_Testing
Minimum Error	-303.678	-216.971
Maximum Error	210.735	198.408
Mean Error	-0.0	-0.049
Mean Absolute Error	13.343	12.799
Standard Deviation	27.321	25.983
Linear Correlation	0.748	0.74
Occurrences	3,518	1,201

8. Random Trees model provides a correlation of 67.1% from the test data.

Table 7: Random Trees Model

'Partition'	1_Training	2_Testing
Minimum Error	-128.658	-164.04
Maximum Error	373.661	462.645
Mean Error	-0.054	-0.453
Mean Absolute Error	8.988	10.018
Standard Deviation	22.028	25.815
Linear Correlation	0.795	0.671
Occurrences	3,518	1,201

V. LIMITATION OF STUDY

The primary limitation of this study is clubbing of data from 1908 till the recent past (2019). Considering the immense changes in aircraft technology since 1908, it may be more appropriate to categorize the data based on period of the accident and analyse data in recent years. This also presents an area of further study wherein the data could be segregated differently and conclusions drawn from recent data.

REFERENCES

1. Walton T. "What Are The Most Common Causes Of Aviation Accidents" *Walton Telken injury Attorneys*. Retrieved from waltontelken.com/common-causes-aviation-accidents/
2. Rao S., Shruthi, Shruti Vinaya M, Rao P. and Naik S.R. "Airplane Crash Analysis Using LDA " International *Research Journal of Engineering and Technology (IRJET)* Volume: 05 Issue: 04. Apr-2018 PP 4929 . Retrieved from www.irjet.net/archives/V5/i4/IRJET-V5I41086.pdf
3. Carden H.D. (1982) "Correlation and Assessment of Structural Airplane Crash Data with Flight Parameters as Impact" *NASA Technical Paper 2083*. NASA TP-2083. Nov 1982 Retrieved from ntr.nasa.gov/archive/nasa/casi.ntrs.nasa.gov/19830006250.pdf
4. Oster C.V, Strong J.S. and Zorn C.K. "Why Airplanes Crash: Causes of Accidents Worldwide". Retrieved from pdfs.semanticscholar.org/9e7a/e442f0d799e53e88470a038322df608cfe33.pdf%20/
5. "Airport ApronManagement and Control Programs" *The National Academies Press* chapter 2. Retrieved from www.nap.edu/read/22794/chapter/3#3
6. Broach D. (2004) "Methodological Issues in the Study of Airplane Accident Rates by Pilot Age: Effects of Accident and Pilot Inclusion Criteria and Analytic Strategy" *Final Report U.S. Department of Transportation Federal Aviation Administration* DOT/FAA/AM-04/8 May 2004. Retrieved from libraryonline.erau.edu/online-full-text/faa-aviation-medicine-reports/AM04-08.pdf
7. Vane R. (2016) "Flight delay analysis and possible enhancements with big data" *International Research Journal of Engineering and Technology (IRJET)*. Volume: 03 Issue: 06. June-2016. Retrieved from www.irjet.net/archives/V3/i6/IRJET-V3I6144.pdf
8. DeAngelis T. (2008) "Why airplanes crash". *American Psychological Association*. Vol 39, No. 3, PP 32. March, 2008. Retrieved from www.apa.org/monitor/2008/03/airplanes-crash

AUTHORS PROFILE

Nikita Pande is graduating in B.TECH (CSE) from Techno India NJR Institute of Technology Udaipur. Her areas of interest are cloud computing, machine learning, python and the internet of things. She has good programming skills in C, C++, Python and R. She has been an absolute all rounder in her college time. She is responsible, punctual and creative. She has good leadership qualities and has been a responsible colleague. She has participated in SIH software and hardware for two years - SIH software'20 and SIH hardware'18. Worked on city air pollution tracking project as an analyst and developer.

Devyani Gupta is graduating in B.TECH (CSE) from Techno India NJR Institute of Technology, Udaipur. Her areas of interest are data analytics, machine learning, programming and python. She has good programming skills in C, C++, Python and Java. She has been a team worker, responsible and confident person and a sincere colleague. She has been an extraordinary student in academics and has volunteered in many events. She is responsible, punctual and creative. She has participated in SIH software and hardware for two consecutive years - SIH software'18 and SIH hardware'19. Worked on city air pollution tracking project as an analyst.

Prof Jitendra Shreemali is a graduate from IIT Madras with post graduate from IIM Bangalore. He is working as Professor of the Department of Computer Science and Engineering of Techno India NJR Institute of Technology Udaipur. He has worked in reputed companies in India & abroad for about a decade and half followed by about two decades of academic/research/training experience. He has taught a very wide variety of subjects/courses including operations management, research methodology, and data science besides others. His areas of work include data science, optimization, mathematical modeling and machine learning.

Prasun Chakrabarti received his PhD (Engg) from Jadavpur University in 2009. He is working as Executive Dean (Research and International Linkage) and Institute Distinguished Senior Chair Professor, Techno India NJR Institute of Technology. He has several publications, books and 31 filed Indian patents in his credit. He has supervised ten PhD candidates successfully. On various research assignments, he has visited Waseda University Japan (2012 availing prestigious INSA-CICS travel grant), University of Mauritius (2015), Nanyang Technological University Singapore (2015,2016,2019), Lincoln University College Malaysia (2018), National University of Singapore (2019), Asian Institute of Technology Bangkok Thailand (2019) and ISI Delhi (2019). He is a Fellow of IETE, ISRD(UK), IAER(London), AE(I), CET(I) and Senior member of the IEEE(USA). FIET