Prediction of Aviation Accidents using Logistic Regression Model

Aswathy Benny[1], Maria Johny[2], and Linda Sara Mathew[3]
[1]*Aswathy Benny, Mar Athanasius College of Engineering, Kothamangalam*
[2]*Maria Johny, Mar Athanasius College of Engineering, Kothamangalam*
[3]*Linda Sara Mathew, Mar Athanasius College of Engineering, Kothamangalam*

Abstract **- When it comes to long distance movement the only easiest and fastest option, we got is aircraft. Plane crashes have always been a big tragedy. Even though we are able to create machines that can carry 850 plus passengers, the safety of this aircraft comes with some questions. No mode of transport is safe. Even a child riding a bicycle is not. But we cannot turn our back on the growing world, speaking of which aircraft play a major role in the development of the society. Just because it is not safe, or a few does not reach their destination humanity can't refuse airplanes. The study about recent aircraft accidents proves there is a strong chance of an unlikely end. A flight crash is caused due to multiple factors. If we can save the lives of people, delay the undeniable death, we are making the world great again. Here we are trying to build Machine Learning models to anticipate and classify the severity of any airplane accident based on past incidents. With this method, the entire aviation industry can predict the airplane accidents caused due to various factors. Then they can make a plan of action to minimize the risk of accident. We have used logistic regression to identify whether a particular feature is important or not and then we adapted random forest technique for classification. Finally, we used XGBoost, which provides a gradient boosting framework for python to produce the model. The final result of the method will give the aviation accident prediction based on severity of the accidents.**

Index Terms **- Aviation accidents, Logistic regression model, XGboost, Random Forest**

I.INTRODUCTION

Aviation is a key factor of our modern life. Popularity of aviation is increasing day by day. Aircrafts are complex machines, and it requires a lot of management. Field of aviation is acquiring technological advances fastly. Since the technology is advancing, the safety of aircrafts is questioned. When it is in the air, the responsibility lies on the pilot's shoulder. If we are able to predict the accidents before the takeoff, chances of saving lives are high. Here we are introducing a machine learning technique to predict the severity of airplane accidents based on the past incidents, which consists of 40 years of data. Airplane accidents can occur due to different causes such as human error, mechanical failure, weather etc. When we analyze the previous year's data, we can get different aspects of plane crashes, which directly leads us to the root causes of the very same plane crash. We can use this information to predict accidents. Analyzing 40 years of data is a difficult task, probably impossible. With the help of capable data mining tools, we can acquire information from data. Mining is the process of finding hidden materials from the raw materials. Similarly, data mining will find out information from large data. The data consists of many rows and columns. The columns will be given the features whereas rows will give the records. In our method we are using a logistic regression model to identify the importance of each feature. Feature selection is an important procedure to get the proper information. When we select the right feature, we can build the model more precisely. Logistic regression model is a supervised classification algorithm. In supervised classification algorithms, we have to know the output initially. In our method we have used random forest technique for the classification. Classification is a big part of machine learning. We want to know one particular object belongs to which class. The precise prediction of classes is extremely valuable. In a random forest classification, we are checking the classes of each decision tree. The output of the random forest will be the most number of classes predicted by different decision trees in it. Finally, we used XGBoost, which provides a gradient boosting

framework for python to produce the model. The output of the project will give the aviation accident prediction based on severity of the accidents.

II. RELATED WORKS

As of now numerous studies are done for recognizing the main reasons related to aircraft accidents. Shyur [1] proposed a model for examining the impacts of safety in aviation and the evaluating risks related to aviation. Xia Feng et.al [2] has done a differentiation set mining procedure to contemplate mishaps related to pilot and occurrences within the previous years and analyse the behaviour of the pilot. The examples of components that were distinguished were all more altogether related with mishaps than occurrences and they were more likely brought about mishaps really than anticipated. Zubair et al. [3] proposed CBR strategy to foresee and dissect the air occurrences and mishaps. They demonstrated up to 87 rate exactness in their model. Dishank Kaj [4] is identifying whether a flight is ready for takeoff by performing a comparative analysis of different algorithms such as Naive Bayes and Decision tree. Christopher et.al.[5] performed an experimental investigation, furthermore, talked about different bunching and characterization techniques for foreseeing the admonition level of airplane accidents. This paper mainly discussed different techniques like Neural Network, Decision Tree, Naive Bayes, K-closest neighbor for the prediction of warning level in flight accidents. Li et.al. [6] is considering various techniques like DBSCAN, Filter Cluster, etc for making the evaluation model of cabin crew fatigue risk. The paper is mainly organized into different sections. Section 2 is about the methodology used for the aircraft data analysis and accident prediction. Section 3 contains the results. Section 4 is the discussion about the logistic regression model, Random forest and XGBoost that is used in the proposed prediction model. Section 5 is the conclusion of this research.

III. METHODOLOGY

The workflow of the proposed system is shown in figure 3.1. First, we collect the data set and clean it. Data cleaning includes removal of unwanted and irrelevant data from the data set. Then we split the data set into train and test sets. After analyzing the training data, we applied the logistic regression to identify the importance of each feature. Then a random forest classifier is applied for the classification process and last applied XGBoost classifier for the aviation accident prediction.

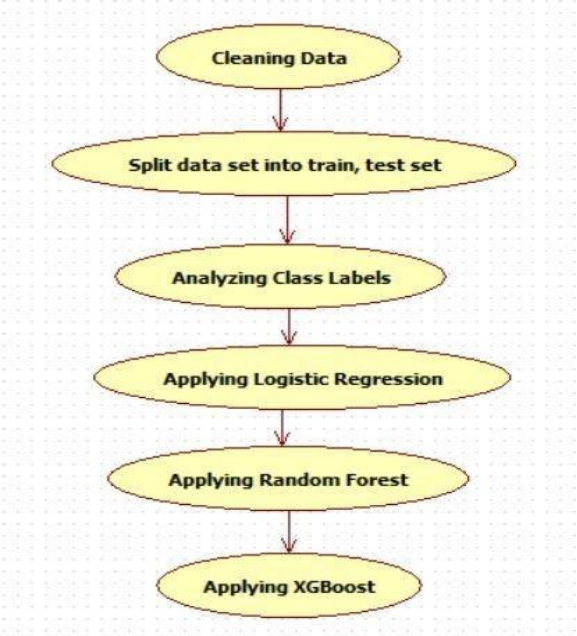

Figure 3.1: System Workflow

Our proposed method consists of following four modules such as: analyzing training data, applying Logistic Regression, applying Random Forest, applying XGBoost Classifier.

A. Analyzing Training Data

The proposed model is for the prediction of aviation accidents severity based on past aviation accident data. For that we first collected aviation accident data set and then trained the model with the past accident-causing attributes. The training data set consists of attributes such as cabin temperature, adverse weather metric, turbulence, safety score, control metric, days since inspection, total safety complaints, accident type code, max elevation, severity, and violations [9]. By analyzing and understanding all the features, we trained the model more efficiently. Based on the analysis on the training data, we understood that in our training dataset contains less quantity of class labels called 'Significant damage and fatalities'. And except this particular class label, the values of safety score follow a gaussian distribution. After analyzing the features such as turbulence and control metric, we

understood that they are negatively correlated. i.e., If there is an increase in the turbulence, then the control of the pilot on the plane will decrease.

B. Applying Logistic Regression

Logistic Regression is a Machine Learning algorithm which is used for binary classification. It is a simple algorithm that could be used as a performance baseline and is simple to enforce. Like many other techniques of Machine learning, it is taken from the sector of data statistics. And regardless of its name it is not a set of rules for regression problems, in which we want to expect a non-stop outcome. Instead, Logistic Regression is the go-to approach for binary classification. It gives a discrete binary final result between 0 and 1. Logistic Regression measures the relationship among the variables that are dependent (our label, what we need to predict) and the one or more variables that are independent (our features), by estimating probabilities using its underlying logistic characteristic. The logistic feature is defined as:

$\log(\text{odds}) = \text{logit}(P) = \ln (P/1 - P)$ (Equ:3.1)

These probabilities are then transformed into binary values so as to absolutely make a prediction. This is the challenge of the logistic function, also referred to as the sigmoid function. The Sigmoid-Function is an S-shaped curve that could take any real-valued number and map it into the range of zero and 1, but never precisely at those limits. These values between 0 and 1 will then be transformed into either 0 or 1 the use of a threshold classifier. While considering logistic regression, there are some assumptions such as: binary logistic regression calls for the dependent variable to be binary, for a binary regression, the aspect degree 1 of the dependent variable should constitute the preferred outcome, only include the meaningful variables, the independent variables need to be independent of each other. That is, the model has to have very little multicollinearity, the independent variables are related to the log odds linearly, logistic regression needs quite large sample sizes.

These probabilities are then transformed into binary values so as to absolutely make a prediction. This is the challenge of the logistic function, also referred to as the sigmoid function. The Sigmoid-Function is an S-shaped curve that could take any real-valued number and map it into the range of zero and 1, but never precisely at those limits. These values between 0 and 1 will then be transformed into either 0 or 1 the use of a threshold classifier. While considering logistic regression, there are some assumptions such as: binary logistic regression calls for the dependent variable to be binary, for a binary regression, the aspect degree 1 of the dependent variable should constitute the preferred outcome, only include the meaningful variables, the independent variables need to be independent of each other. That is, the model has to have very little multicollinearity, the independent variables are related to the log odds linearly, logistic regression needs quite large sample sizes.

C. Applying Random Forest

The random forests could be a supervised machine learning algorithm used for classification and regression. This algorithm is user friendly and scalable. A random forest is composed of many trees. The random forests build decision trees on randomly selected data samples and get predictions from every tree and pick the simplest and best answer by voting. A random forest is more durable as there are more trees in it. This offers a relatively smart predictor of the value of the function. Random forests have a variety of uses, along with the image classification and selection of functions. It can be used to classify loyal mortgage applicants, pick out fraudulent activity and prediction of accidents. It lies at the bottom of the Boruta algorithm, which selects features in a dataset that are important. Alternatively, the random Forest package includes two extra data snippets: a proportion of the importance of the indicator variables, and a measure of the information's internal structure [7]. Algorithm 1 displays the overall procedure included in the Random Forest.

ALGORITHM 1: RANDOM FOREST ()

1. Select random samples from a given dataset.
2. Construct a selection tree for every sample and get a prediction end result from each decision tree.
3. Perform a vote for each expected result.
4. Select the prediction end result with the most votes as the final prediction.

In the proposed method, we are classifying the severity of aviation accidents by the usage of random forest classifiers. Scikit-learn includes another model variable showing the relative significance or influence of each function within the prediction. In the training process, it mechanically calculates the relevant score for each attribute. So, it scales down the relevance

until the sum of all the scores is 1. This score will assist to select the features that are most important and drop the least crucial ones for building the model. Random Forest makes use of gini significance or mean decrease in impurity (MDI) to sum the significance of each characteristic. Gini significance Is likewise classified as the whole lower in node impurity. That is how the model will match or the accuracy will diminish when you reduce a variable. If the decrease is more than the variable is more significant. Here, the significant parameter for variable selection is the mean decrease. The overall explanatory strength of the variables can be described by the Gini index.

D. Applying XGBoost Classifier

XGBoost is an enhanced distributed gradient boosting library intended to be fantastically economical and bendy. It uses the Gradient Boosting framework to implement machine learning algorithms. The Extreme Gradient Boosting (XGBoost) strategy is the advancement of a gradient boosting model that has prevalent outcomes and procedure speeds. XGBoost has the well-liked position of getting ready missing value data while not ascription of knowledge toward the beginning of learning [8]. XGBoost presents a parallel tree boosting (also known as GBDT, GBM) that solves totally different data science issues in a very quick and accurate way. XGBoost is often used for the training purpose of decision trees which are gradient-boosted and other gradient boosted models. Random forests are using the same inference and model representation, as gradient boosted decision trees, but a different training algorithm. XGBoost may be used for training a standalone random forest or use random forest as a base model for gradient boosting. Here we are using random forest as a base model for gradient boosting. Random forest functionality is offered by XGBRFClassifier and XGBRFRegressor which are SKL-like classes. They are versions of XGBRegressor and XGBClassifier that train random forest instead of gradient boosting and having feature default values and that means some of the parameters adjusted accordingly. XGBoost classifier uses the Gradient Boosting framework for the prediction of aviation accidents based on their severity. In this method XGBoost provides a parallel tree boosting for solving the prediction in a fast and accurate way.

IV. RESULTS

We have developed a model for the aeronautics mishaps to determine whether or not the flights are susceptible to accidents. The model usually forecasts the circumstances under which mishap occurs. Aviation accident prediction based on historical data analysis can aid in the provision of previous accident information. It is seen that by using a logistic regression model and random forest classifier, we can effectively discover the correct and usable data from the dataset. The findings reported in this research are easy for mediators to understand. With the aid of the results obtained here, predictions can be made, and the safety measures can be taken against the aircraft malfunctions.

V. DISCUSSIONS

In this method we are able to predict severity of aviation accidents based on the past incidents which consists of 40 years of data. We are living in a world where the field of aviation influences our modern life, because aviation is a major means of transport when it comes to long distance movement. The results obtained in our method help to save lives and money. We used machine learning to predict the extent of aircraft accidents, so that we can make the best decisions before the aircraft starts. In order to take this decision, we were analyzing the features of previous aviation accidents data. The limitation of our system is that we are only able to predict the probability of an accident happening, but not a 100% decision. In the future using technological advances in both aviation and computer science, we will be able to arrive at a definitive decision.

VI. CONCLUSION

In this paper, we analyze historical data of long-term aviation accidents of about 40 years and predict the aviation accident severity. Logistic Regression is used to analyze each of attributes and to identify the importance of these attributes. Random forest classifier is used for the classification of accidents into different classes based on the intensity of damages and injuries occurring during aviation accidents. Finally, the XGBoost classifier is used for the prediction of accident severity. This analysis and prediction are very helpful to reduce the future accidents. So that the impact of accidents can be foreseen and take the necessary precautions to rescue the passengers. In

future we can extend this method for the prediction of flight delay based on flight features.

REFERENCES

[1] H.J. Shyur, A quantitative model for aviation safety risk assessment, Computers, and Industrial Engineering (2007). Pp. 34 - 44.

[2] Xia Feng and Juanjuan Li, Analyzing Pilot – related accidents and Incidents by Data Mining, International Conference on Computer Application and System Modelling (ICCASM 2010), pp. 325 – 327, 2010.

[3] Maria Zubair, Malik Jahan Khan and Mian Muhammad Awais, Prediction and Analysis of Air Incidents and Accidents Using Case Based Reasoning, Third Global Congress on Intelligent Systems, pp. 315-318, 2012.

[4] Dishank Kaji, Dhrumil Mehta, Divyesh Sanghani, Harsh Shah and Rashmi Malvankar, Study of Prediction Algorithms on Aviation Accident Dataset using Rapid Miner, International Research Journal of Engineering and Technology (IRJET), pp. 1670-1672, 2019.

[5] A.B. Arockia Christopher and S. Appavu alias Balamurugan, Performance of Different Clustering Methods and Classification Algorithms for Prediction of Warning level in Aircraft Accidents: An Empirical Study, Fourth International Conference on Computing, Communications and Networking Technologies (ICCCNT), pp. 1-7, 2013.

[6] Wei Li, A Cabin Crew Fatigue Risk Comprehensive Evaluation Model, 12th International Conference on Fuzzy Systems and Knowledge Discovery (FSKD), pp. 1596 –1600, 2015.

[7] A. Liaw and M. Wiener, "Classification and Regression by randomForest", R News, vol. 2, no. 3, pp. 18-22, 2002.

[8] Widya Fajar Mustika, Hendri Murfi and Yekti Widyaningsih, Analysis Accuracy of XGBoost Model for Multiclass Classification - A Case Study of Applicant Level Risk Prediction for Life Insurance, 2019 5th International Conference on Science in Information Technology (ICSITech), pp. 71-77, 2019.

[9] Flight safety foundation Last updated: 31 July 2020 AviationSafety network Accessed: April. 25, 2020 https://aviation-safety.net