Aviation Accident Analysis

Madhu Chegondi (m136c192@ku.edu)
Department of EECS, University Of Kansas
Lawrence, KS 66047 USA

Niharika Gandhari (n233g507@ku.edu)
Department of EECS, University Of Kansas
Lawrence, KS 66047 USA

SriMounica Motipalli (s492m183@ku.edu)
Department of EECS, University Of Kansas
Lawrence, KS 66047 USA

Tejaswini Jagarlamudi (t177j342@ku.edu)
Department of EECS, University Of Kansas
Lawrence, KS 66047 USA

Abstract

The project seeks to investigate the civil aviation accidents from the last 54 years starting from 1962 helps us in identifying trends associated with an increase in the total number of injured. We are interested in discovering which environmental or organizational factors, human error and mechanical failures that have caused an increase in fatalities and injuries during air accidents. The project delves to find out the relation between aviation accident factors and the number of injuries.

Keywords: aviation accidents; factors; human error; number of injuries; trends

Background Information

Air travel is one of the most frequently used means to travel by millions of people. Aviation industry has witnessed several accidents despite the sophisticated technology which lead to loss of life and property. Moreover it causes trauma and psychological effects to the family of the deceased and injured. To eliminate these conditions we tried to investigate the aviation accident data taking from National transportation Safety Board website. The data is available in mdb(MS Access file) of size 1 GB. We have decided to use this data source because NTSB is an agency which works under the jurisdiction of Federal Government of United States. The data is updated daily and contains the latest information regarding the aviation accidents. It contains information about all the accidents and incidents within United States and its territories.

Our objective is to analyzing the civil aviation accidents from the past 54 years and to identify the main cause for these accidents causing loss of life. Analyze the density of accidents across different regions which are prone to accidents. We tried to look for scenarios which cause huge loss of property or life.

In this project, we mainly dealt with statistical tabulations and graphs of data compiled from reports of accidents involving US Air carriers, commuter and on demand carriers, and general aviation operations in a particular calendar year. We also did predictive analysis on the rate of fatality or seriously injured in an accident.

Constraints and Limitations

The availability of explicit information on survey data about the causes of the accidents is the main constraint of this project. A Classification scheme is an important factor which aids the study of accidents. The three important classifications are the nature of accident, injury to an individual and damage to material. While the data provides information about the severity of injury an individual met and also whether the aircraft received appreciable or marked damage through the external forces or fire, it doesn’t offer any information about the immediate or underlying cause of the accidents. There is no direct information regarding the root cause as to why an accident occurred. Here is when the target variable comes into picture. We will have to use all the possible predictor variables and compute a target. The target variable will serve as the backbone of our study and help us to extract meaningful information from the aircraft accident analysis.

Even if we come to know the root cause of an accident, it doesn’t necessarily mean that we have the comprehensive knowledge of as to how the accident took place as there might be many other contributing factors which needs to be considered before jumping into conclusions. For example, if the aircraft crashes and the reason cited for the accident is ‘Loss of Control’ which means the inability of pilot to control aircraft. So ‘Loss of Control’ over the aircraft for one pilot may not be for the other one. Thus we need detailed background of the pilots to be 100% sure about the real cause of the accident. So unavailability of the data helpful to identify the characteristics of an important predictor variable is one of the most important limitations which needs to be overcome by using other vast number of given variables finding relationship between them.

Implementing the plan

We implemented our project in below steps.

1. First we reduced the raw data by eliminating rows with null values and filter data according to location.
2. Then, we ignored few variables like Accident Number, Registration Number and Publication date which are no value in analysis.
3. Once the sample data is ready, we shall adopt different software to prepare our analysis and predictive model.

Scope

We considered the survey data after 1962 and ignored data before 1962 to fit in the scope of the project. Initially we got large number of records, we filtered the data based on redundant data, missing values, null values and false values and trimmed it down to a considerable amount. Analyzing the filtered data of last 53 years we are trying to come up with significant information that will help us identify the main causes of these accidents, which can help us to take adequate measures in order to prevent them and save lives and property.

Data Preparation

Data Access

We have taken the data for analysis from National Transportation Safety Board website. Our primary data set consists of 63 variables and 70 thousand records. We decided to use this data source because of following factors.

NTSB is an independent agency which works under the jurisdiction of Federal government of United States. The data is updated daily and contains the latest information regarding the aviation accidents. The latest aviation accident report is that 26th October 2016. The data contains the information about all the accidents within United States and its territories. The factual information about the accidents is updated regularly along with the probable cause and final description.

Data Consolidation

We are mainly using 4 tables – Events (have the data of all the events happened including accidents and incidents), Aircraft (data about the aircrafts) Narratives (contains the summary of the events happened), Main (having the data of all the accidents).

Data Cleaning

There were few missing values for variables like Broad Phase of flight, purpose of flight. But since we have a large volume of data and these predictors are not the most important for our analysis so we are just treating them as nulls. For the variable Schedule there were more than 85% of missing values (as shown in below graph). So we removed the missing values and use a sample set that had only the correct values for this variable.

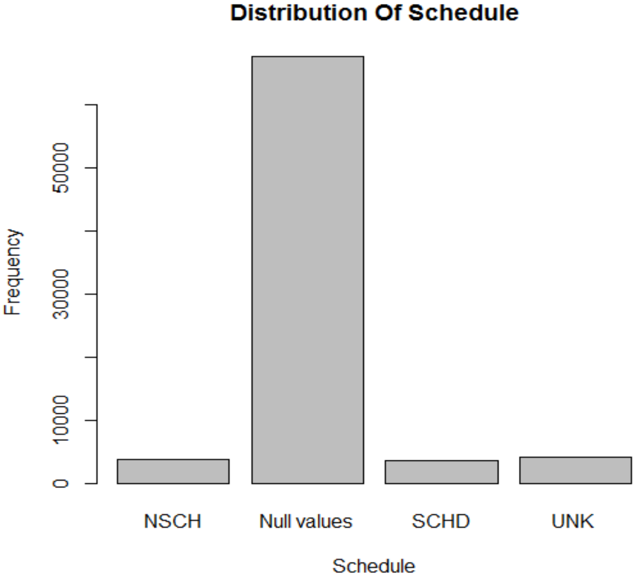


Figure 1: Distribution Of Schedule

For some numerical variables such as *Total Serious Injuries*, *Total Uninjured* we are treating the missing values as '0' which is a logical assumption. There were no outliers present in the data as it talks about the aviation accidents, their impact and severity. Since the data is well maintained, we were easily able to convert some character variables with numeric data such as *Total Serious Injuries*, *Total Uninjured* into numeric variables. We also removed null values for all the variables using R.

Data Analysis

Technological evolution has made Data Visualization more pervasive and powerful than ever before. We have tableau to visualize our data as in our opinion this very fine visualization tool has the ability to visualize data effectively which leads to better understanding.

Accident Analysis by Location

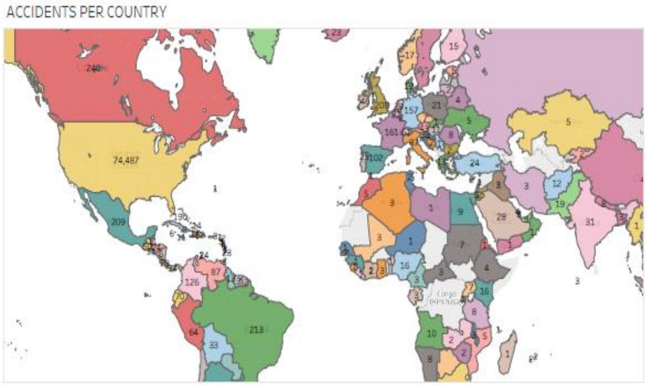


Figure 2: Accidents per Country

We can observe from the above map that most of the accidents are in United States. This is because the data which we took for analysis contains aviation accident data in and around USA.

Aircraft Accidents and their Causes

Because of umpteen people and components involved in flying, there can be any number of errors or mistakes which can lead an aircraft to crash. The different categories under which an aviation accident can fall are:

Human Error This can be further classified into pilot error or other human error. A pilot error represents accidents where the error was on pilot whereas other human error includes air traffic errors, improper aircraft loading error, fuel contamination and proper maintenance methodology.

Mechanical Failure Mechanical failure involves structure component failure, flawed aircraft design and critical system failure. Sometimes it may also occur when outside happenings damage the plane. The causes of such failure can be pretty bizarre. For example, birds striking the aircraft leading to the engine failure, causing the plane to crash.

Weather Condition Heavy winds, storms, thick fog and lightning are examples of inclement weather conditions. They may lead to electrical failure, ignited pipes and fuel tanks, temporary blindness and many other conditions which are deemed hazardous causing aviation accidents.

Organizational failure This is somewhat similar to human error category as it involves error on human part but has been mentioned explicitly as it involves failure of a group as a whole. Flawed management, obscure policies, murky procedures fall under this bracket.

Unknown The aviation accidents for which the reasons are not determined

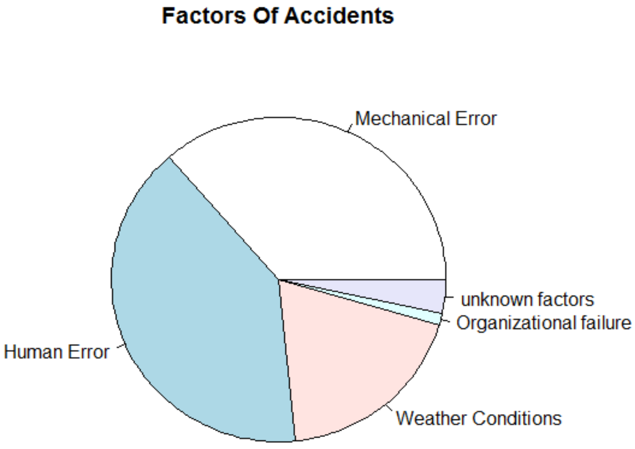


Figure 3: Factors Of Accidents

We can observe from the above pie chart that majority of accidents are caused due to human errors. The second most common cause of aviation accidents is mechanical error, bulk of the remaining error is due to bad weather. Very few accidents are caused due to organizational failure and fewer remains are not known

Total Injuries during Phase of flight

Between the time a person boards an aircraft and the time he debarks, there are different distinct phases

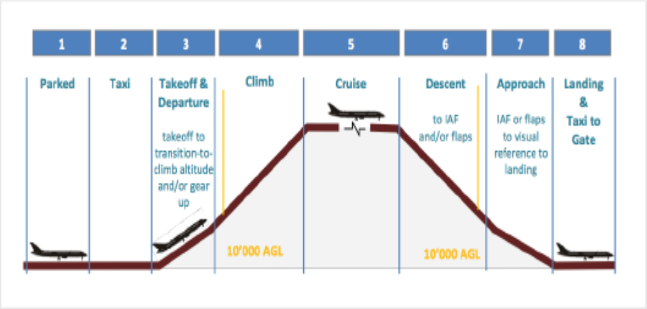


Figure 4: Phases of Flight

Taxi The phase where the aircraft taxis to reach the runway or it taxis after landing to reach the gate.

Climb After the take-off when the aircraft lifts-off and starts to climb until it reaches cruise altitude.

Cruise This is the phase where the aircraft more or less maintains a constant altitude and flies for the longest period.

Descent The aircraft starts moving downwards to get closer to the runway where it can land.

Approach The aircraft before landing aligns itself with the runway axis and approaches the runway entry.

Landing This is the phase of the flight where the aircraft returns to the ground

From the below bubble chart, we can observe that most of the fatal accidents occur during the landing phase of the flight. The second most common phase of flight vulnerable to fatal accidents is the take-off stage. So we can infer that most fatalities take place during the landing and take-off (departure) stages of the flight. During these two phases the aircraft is close to the ground and at susceptible position than during other phases of the flight. We can also observe the least risky phase of the flight is the climb stage where the aircraft is in the position to reach cruise phase.

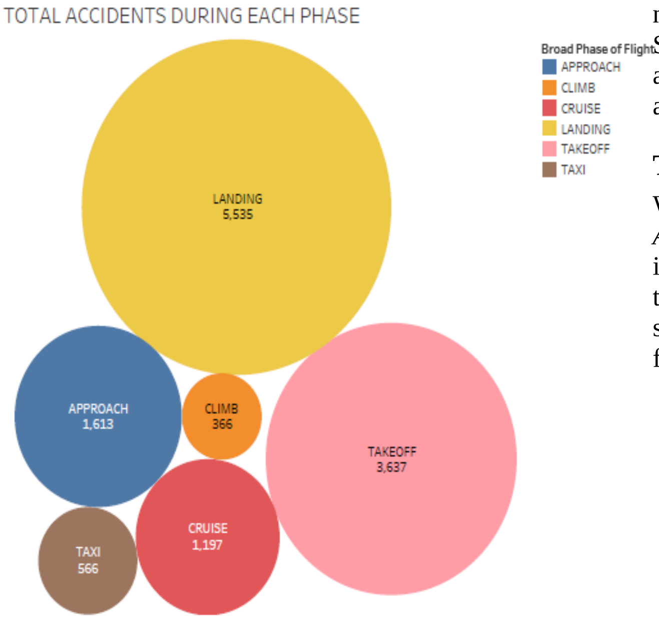


Figure 5:Total Accidents During Each Phase

Total injured by Engine Type

To observe which type of engines are not giving best results we took engine types that met with an accident. We observed a clear trend from the year 2008 to 2015. So we included those results in this paper.

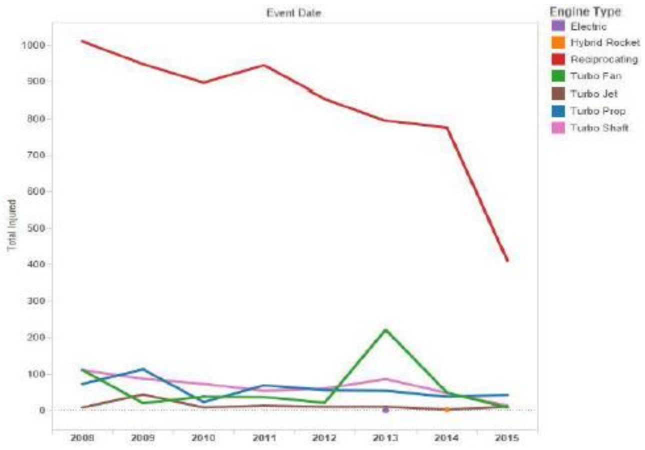


Figure 6: Total Injured By Engine type

From the above plot, we can observe that most number of accidents was reported for the aircrafts using reciprocating engine. Aircrafts equipped with the Turbo Shaft engine had least fatalities, almost equal to none and the figure was constant over all the years. The fatalities caused due to aircraft accidents using Turbo Fan proliferated in 2013 but it would not be wise to conclude that the type of engine used that year was the root cause of the accidents as the fatality figure never increased but went down thereafter. All other engine types used in the aircraft have more or less same number of injuries which is not much, during these years. So we can infer that the aircrafts using reciprocating engines are more vulnerable to fatal accidents as compared to the aircrafts using other Engine Types.

Total Accidents per year

We plotted a bar graphs plotting against *Total Number of Accidents* to the *year*. Here we can clearly observe that there is decrease in the number of accidents when compared to the good olden days. Even though there is a lot much sophisticated technological advancements there is still room for preventing these accidents

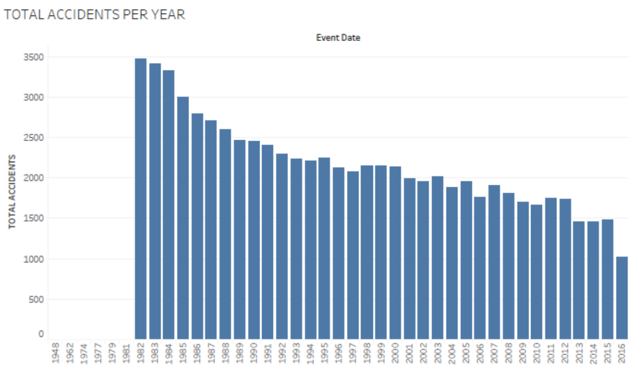


Figure 7: Total Accidents Per Year

Fatalities per Year

Here is the trend how the fatalities are in particular year in all these 54 years.

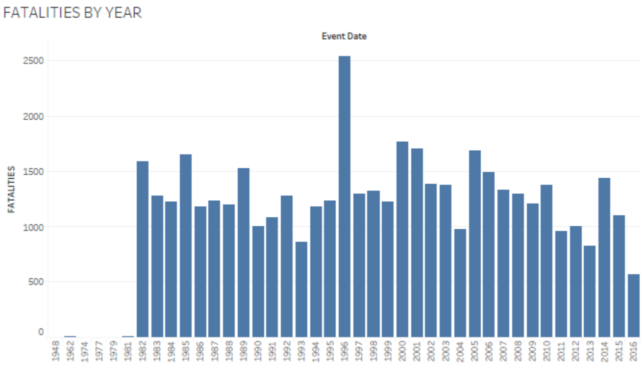


Figure 8: Fatalities By Year

Text Analysis

Textual Analysis is a research method that requires the researcher to closely analyze the content of communication rather than the structure of the content. A textual analysis is most often used to analyze historical documents and narratives[1].

Here main goal is to extract the most common causes of plane crashes. For this we used the table narratives which

[1] Google definition of text analysis

contained a column containing *narr_cause* which describes the cause of the crash of a particular aircraft. We loaded the table into R and did some cleaning on the text, and created a Document-term Matrix, from which we removed most generic terms like flight, crashed, plane, factors etc. We reduced the Document-term Matrix as this is very sparse. To find frequently associated words, we first compute a distance matrix based on our reduced Document-term Matrix, then apply K Means clustering. We built a word cloud to represent frequent words in the causes. Below is the word cloud representing high frequent word.

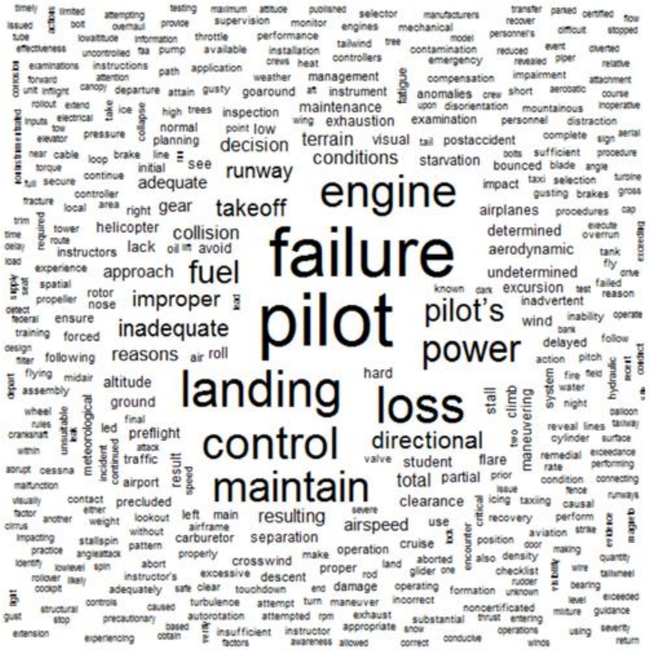


Figure 9: Word Cloud of high frequent words

We considered the 20 most frequent terms and calculated their frequency relative to each other. To add more context to the list, we took most correlated words and plotted the 5 terms that have the higher correlation higher than 0.17

Table 1: Correlation of high frequent words with other words

Highfrequent Word	Correlated Words	Correlation Value
Loss	Partial	0.25
	Starvation	0.24
	Exhaustion	0.23
	Determined	0.22
Control	roll	0.24
	excursion	0.19
	Crosswind	0.18
engine	determined	0.30
	partial	0.30
maintain	adequate	0.33
	airspeed	0.30
	clearance	0.29
power	partial	0.32
	starvation	0.32
	Exhaustion	0.30

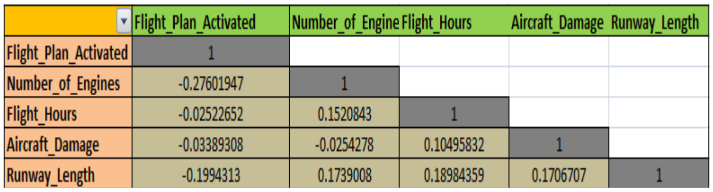
Observations from Text Analysis

The pilots improper fuel management, which resulted in total loss of engine power due to fuel starvation/ Exhaustion. The pilot's inadequate preflight inspection and the flight instructor's supervision led to fuel exhaustion. Failure to maintain adequate airspeed which resulted in improper clearance with the terrain or from the winds. The pilots failure to maintain directional control during the landing roll in crosswinds, which resulted in the runway excursion. The pilots loss of airplane control due to spatial disorientation and instrumental meteorological conditions with gutsy wind.

Predicting Total Fatal Injuries

Multiple Regression

The main assumption of this model is that the independent variable and the dependent variables show a linear relationship. The correlation matrix showing the correlation between independent variables is as shown in figure below



	Flight_Plan_Activated	Number_of_Engine	Flight_Hours	Aircraft_Damage	Runway_Length
Flight_Plan_Activated	1				
Number_of_Engines	-0.27601947	1			
Flight_Hours	-0.02522652	0.1520843	1		
Aircraft_Damage	-0.03389308	-0.0254278	0.10495832	1	
Runway_Length	-0.1994313	0.1739008	0.18984359	0.1706707	1

Figure 10: Correlation among variables

The below scatter plot matrices show that a few independent variables exhibit linear relationship. The variables such as *Number Of Engine, Flight Hours, Aircraft Damage* have a linear relationship as we can see in the below scatter plot.

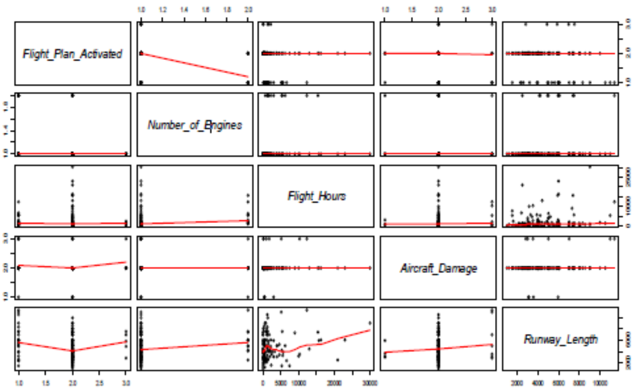
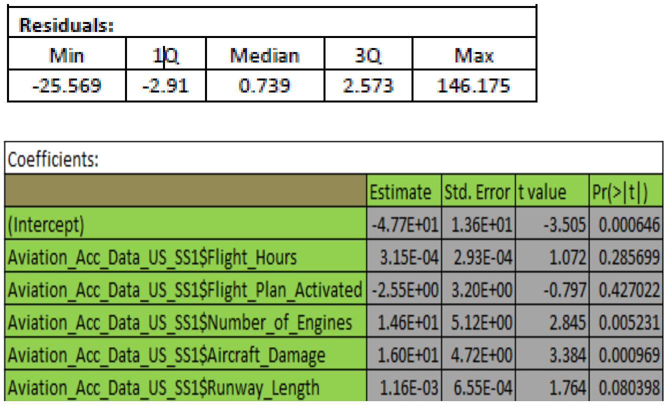


Figure 11: ScatterPlots

Building Model

We built a multiple regression model with considering above independent variables



Residuals:				
Min	1Q	Median	3Q	Max
-25.569	-2.91	0.739	2.573	146.175

Coefficients:				
	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-4.77E+01	1.36E+01	-3.505	0.000646
Aviation_Acc_Data_US_SS1$Flight_Hours	3.15E-04	2.93E-04	1.072	0.285699
Aviation_Acc_Data_US_SS1$Flight_Plan_Activated	-2.55E+00	3.20E+00	-0.797	0.427022
Aviation_Acc_Data_US_SS1$Number_of_Engines	1.46E+01	5.12E+00	2.845	0.005231
Aviation_Acc_Data_US_SS1$Aircraft_Damage	1.60E+01	4.72E+00	3.384	0.000969
Aviation_Acc_Data_US_SS1$Runway_Length	1.16E-03	6.55E-04	1.764	0.080398

Figure 12: Descriptive Analysis

Residual standard error: 15.28 on 118 degrees of freedom
Multiple R-squared: 0.2284, Adjusted R-squared: 0.1957
F-statistic: 6.986 on 5 and 118 DF, p-value: 9.469e-06

The p-value for the regression model is 9.469e-06 which is a statistically a small number. The adjusted R-squared value given by the model is 0.1957 which is considerable. The dependent variables in the model which are statistically significant (p-value greater than alpha=0.05) are aircraft damage and Number of Engines. Runway Length, Flight Plan Activated and Flight Hours are statistically insignificant independent variables. To predict the dependent variable (Total Injured), the model is showing that Aircraft Damage and Number of Engines are helping to predict the total number of injuries in the accidents whereas the dependent variables Runway Length, Flight Hours and Flight Plan Activated are not providing the predictive value of Total Injuries

Predicting Fatalities Given Injuries

We used logistic regression to predict whether there are fatalities given number of injuries for a crash. Initially, we created a separate field in existing data, which indicates whether there are Fatalities or not (i.e., YES or NO). We divided the data into training and validation. When we applied the model on training data we got 97% accuracy in while predicting, when there are no deaths and only 3% accuracy when there are fatalities. When we applied the model on validation data, we got 90% accurate prediction when there are no fatalities and 11% accuracy when there are fatalities. The accuracy of prediction is calculated using confusion matrix. From the results, we can conclude that, one cannot accurately predict the fatalities given injuries. We can predict accurately only when there are no fatalities.

Conclusions

In this project, we tried to find the most frequent causes of accidents. We also found statistically significant effects of several factors on the fatal and serious injury rate. Among them, those which seem practically important in reducing the accident rate are explosion or fire on ground, poor lighting conditions, and existence of weather factors. These findings can be hints to improve passenger and crew survivability in aircraft accidents. But farther research is possible by clarifying definition of variables such as landing surface, adding more variables such as topographical information or equipment, and complementing missing values.

Acknowledgment

We would like to thank our project advisor, Dr. Nicole Beckage, for her precious help all through our project.

References

Gareth James, Daniela Witten, Trevor Hastie & Robert Tibshirani. (2015) *An Introduction to Statistical Elements with Applications in R* New York, USA: Springer.

Joel Grus (2015). *Data Science from Scratch,* California, USA: O'Reilly Media.

Grinstead & Snell. (2006). *Introduction to Probability,* Rhode Island, USA: American Mathematical Society.

Norman S.Matloff (2011). *The art of R programming,* California, USA: O'Reilly Media.