SYNOPSIS OF MINI PROJECT (KCS 554)

EXPLORATORY DATA ANALYSIS ON TITANIC DATASET

Submitted by

Nishit Chaudhary (1901920130119)

Pankaj Sharma (1901920130120)

Piyush Sharma (1901920130121)

Submitted to

DR. ARUN KUMAR SINGH
(Associate Professor)

Department of Information Technology

G. L. Bajaj Institute of Technology and Management
Greater Noida, Uttar Pradesh.
(2021-22)

ACKNOWLEDGMENT

I would like to express my special thanks of gratitude to my Mini Project Supervisor **Dr. Arun Kumar Singh** as well as our HoD **Prof P.C Vashist** who gave me the golden opportunity to do this wonderful project on "**EXPLORATORY DATA ANALYSIS ON TITANIC DATASET"** and for their intellectual, valuable guidance which helped me a lot in doing my work. This project would have been their enormous help and worthy experience. Whenever I was in need, they were there behind me. It makes me to do a lot of research work by which I came to learn about so many new technologies.

Although, this project has been prepared with utmost care and deep routed interest. Even then it was not possible to complete it without any support.

During this period I‟ve come a long way in using statistics and data visualization to understand relationships among variables in the data. I‟m glad to admit that I feel comfortable using Statistics to understand patterns in data. I‟m looking forward to using more advanced statistical methods to derive insights from more complex data with entangled variables.

Last but not the least, I would also extend my gratitude to my parents and friends who helped me a lot in finalizing this project within the limited time frame.

SYNOPSIS

In this project, We'll be trying to do data analysis on titanic dataset and predict a classification - survival or deceased using Logistic Regression. We'll use a "semi-cleaned" version of the titanic data set present on largest machine learning learning platform named kaggle.

The most infamous disaster which occurred over a century ago on April 15, 1912, that is well known as sinking of “The Titanic”. The collision with the iceberg ripped off many parts of the Titanic. Many classes of people of all ages and gender where present on that fateful night, but the bad luck was that there were only few life boats to rescue. The dead included a large number of men whose place was given to the many women and children on board. The men travelling in second class were dead on the vine.

Machine learning algorithms are applied to make a prediction which passengers survived at the time of sinking of the Titanic. Features like ticket fare, age, sex, class will be used to make the predictions. Predictive analysis is a procedure that incorporates the use of computational methods to determine important and useful patterns in large data. Using the machine learning algorithms, survival is predicted on different combinations of features.

The objective is to perform exploratory data analytics to mine various information in the dataset available and to know effect of each field on survival of passengers by applying analytics between every field of dataset with “Survival” field. The predictions are done for newer data sets by applying machine learning algorithm. The data analysis will be done on applied algorithms and accuracy will be checked. Different algorithms are compared on the basis of accuracy and the best performing model is suggested for predictions .

Steps involved in Exploratory data analysis (EDA) :

- Import Dataset & Headers
- Replace Missing Data
- Evaluate Missing Data
- Dealing with Missing Data
- Correct Data Formats
- Data standardization
- Data Normalization
- Binning.
- Indicator variable.

Key Words Used: Logistic Regression, Data Analysis , Kaggle Titanic Dataset, Data pre-processing . Cross validation, Confusion Matrix

1) DATA ANALYTICS AND ITS CATEGORIES :

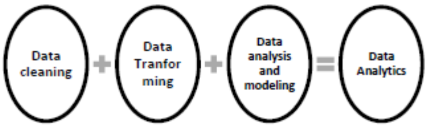

Fig 1: Data Analytics

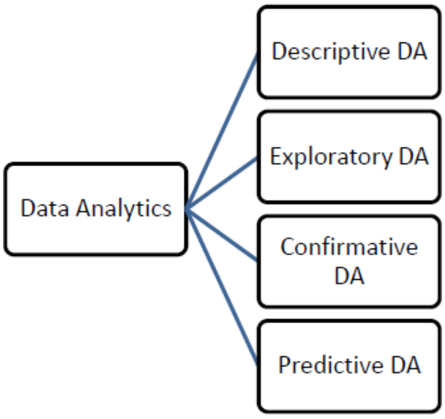

Fig 2: Categories of Data analysis

2) PROCESS FLOW :

There is a step by step approach to choose a particular model for the current problem. We need to decide whether a particular machine learning model is suitable for our problem or not. Here we can see process flow being followed

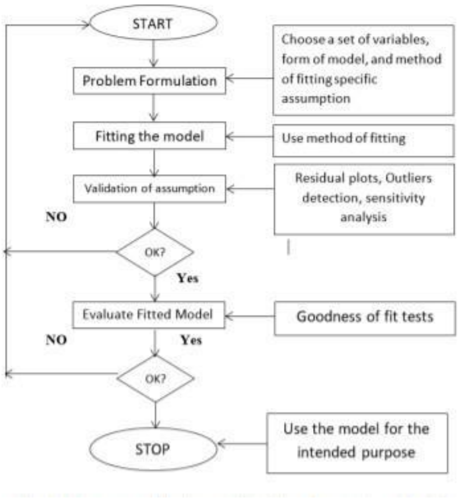
START
Problem Formulation
Choose a set of variables, form of model, and method of fitting specific assumption
Fitting the model
Use method of fitting
Validation of assumption
Residual plots, Outliers detection, sensitivity analysis
NO
OK?
Yes
Evaluate Fitted Model
Goodness of fit tests
NO
Yes
OK?
STOP
Use the model for the intended purpose

Fig 3 : Process fitting

3. ALGORITHM

3) (a) DATA PREPROCESSING

In the dataset available for the prediction some of the data values are missing or unknown. This missing data was resulting in reducing the accuracy of the overall prediction model and also reduces the size of pure training data which in turn reduces accuracy. Data preprocessing is a technique that involves transforming raw data into an understandable format.

Real-world data is often incomplete, inconsistent, and/or lacking in certain behaviors or trends, and is likely to contain many errors. Data preprocessing is a proven method of resolving such issues. Data preprocessing prepares raw data for further processing. Missing values are replaced by average of that column. So, the missing and unknown data of the passengers which is easily predictable is filled up by this step.

3) (b) CLASSIFICATION ALGORITHM

(i) LINEAR REGRESSION

Second step of the algorithm is using a classifier to classify the available information. Logistic Regression is the appropriate regression analysis to conduct when the dependent variable is dichotomous (binary). Like all regression analyses, the logistic regression is a predictive analysis.

(ii) LOGISTIC REGRESSION

Logistic regression is used to describe data and to explain the relationship between one dependent binary variable and one or more nominal, ordinal, interval or ratio-level independent variables. It uses a method of using he regression line between dependent and independent variable to predict the value of the dependent variable

3) (c) CROSS VALIDATION

Dataset is divided into two main parts namely Train and Test data. Training data will be considered for the training of the machine. Test data will be used for validating the machine. Cross validation technique used here is K-Fold.

The method has only one parameter called k that refers to the number of groups into which a given data sample is to be split. As such, the method is also called k-fold cross validation. When a particular value for k is chosen, it may be used in place of k in the reference to the model, such as k=10 becoming 10-fold cross-validation.

Analysis of confusion matrix Confusion matrix is used to show the performance of the algorithm. Accuracy of the model can be predicted using the confusion matrix. It is a plotting of relation between real and predicted outputs. It allows us to check the accuracy and performance of the algorithm. In this case we are using two attributes at a time for the confusion matrix plotting. Test case data is used to build the confusion matrix. The values shown in the confusion matrix are the probability of survival of the individual considering only those parameters.(d) RESULTS

The logistic regression gives the accuracy of 95% which is based on the confusion matrix. The parameters used here are accuracy and false discovery rate. Accuracy is a measure of the correctness of the prediction of the model. Higher accuracy is always better and is calculated by :

(TN + TP)/Total number of rows *100

False discovery rate are the false positive measures of confusion matrix where the model predicts that the passenger would survive but in reality, it doesn‟t. This would prove dangerous as the prediction may go wrong and hampers the accuracy of the results.

The attempts are being made to increase the accuracy rate and reduce the false discovery rates.

4) DATASET INFORMATION

The sinking of the Titanic is one of the most infamous shipwrecks in history. On April 15, 1912, during her maiden voyage, the Titanic sank after colliding with an iceberg, killing 1502 out of 2224 passengers and crew. This sensational tragedy shocked the international community and led to better safety regulations for ships.

One of the reasons that the shipwreck led to such loss of life was that there were not enough lifeboats for the passengers and crew. Although there was some element of luck involved in surviving the sinking, some groups of people were more likely to survive than others, such as women, children, and the upper-class.

Titanic data : Contains demographics and passenger information from 891 of the 2224 passengers and crew on board the Titanic.

TABLE 1: DATASET INFORMATION

Variable	**Definition**	**Key**
Survived	Survival	0 = No, 1 = Yes
Pclass	Ticket class	1 = 1st, 2 = 2nd, 3 = 3rd
Sex	Sex	Male or female
Age	Age in years	Integer values acceptable, drop fractional values
Sibsp	# of siblings / spouses aboard the Titanic	Integer value

Parch	# of parents / children aboard the Titanic	Integer value Unique number Unique value Unique value
Ticket	Ticket number	
Fare	Passenger fare	
Cabin	Cabin number	

Pclass: A proxy for socio-economic status (SES)
1st = Upper
2nd = Middle
3rd = Lower

Age: Age is fractional if less than 1.

Sibsp: The dataset defines family relations in this way...
Sibling = brother, sister, stepbrother, stepsister
Spouse = husband, wife (mistresses and fiancés were ignored)