International Journal of
Emerging Research in Management &Technology
ISSN: 2278-9359 (Volume-6, Issue-6)

Research Article **June 2017**

Prediction of Survivors in Titanic Dataset: A Comparative Study using Machine Learning Algorithms

Tryambak Chatterjee*
Department of Management Studies, NIT Trichy, Tiruchirappalli,
Tamilnadu, India

Abstract—

The Titanic disaster which occurred in 1912 remains as one of the biggest tragedies that occurred in human endeavours. The objective of this paper is to apply different algorithms to check whether a passenger survived the Titanic disaster based on different attributes a passenger possess which is included in the dataset for testing. The results from the application of the different algorithms are compared and analysed.

Keywords— Titanic, Survivor, Logistic Regression, Multiple Linear Regression, dataset, Kaggle, Data Analysis

I. INTRODUCTION

The Titanic was a ship disaster that on its maiden voyage sunk in the northern Atlantic on April 15, 1912, killing 1502 out of 2224 passengers and crew [2]. While there exists conclusions regarding the cause of the sinking, the analysis of the data on what impacted the survival of passengers continues to this date [2], [3]. The approach taken is utilize a publically available data set from a web site known as Kaggle [4]. Kaggle offers businesses and other entities crowd-sourcing of data mining, machine learning, and analysis. Sometimes offering prizes (for example there had been a $200,000 prize being offered from GE through Kaggle in a competition) [1].

II. OBJECTIVE

The dataset found on the Kaggle website has two perspectives. One is a training data and the other is a testing data. The objective of the training data is to create a model which will help in predicting the outcomes of the test data. For the purpose of this research paper, the training data from the Kaggle website will be divided into two parts using three different ratios for training and creating the model and then predicting and the testing data from the Kaggle website will not be used. There will be application of different techniques for predicting whether a person survived the Titanic disaster or not. The data analysis will then be done and the prediction outcomes will be checked for accuracy. The accuracy will then be compared in order to suggest the better performing algorithm with respect to used dataset.

III. METHODOLOGY

The training data from the Kaggle website contains 892 rows and 12 columns. The first row of the dataset describes the different parameters for a passenger. The first column in the dataset gives the PassengerId of a passenger and the second column of the dataset gives whether the person survived or not. The PClass attribute defines the class in which the passenger was travelling in the ill-fated ship.

The following will be the roadmap for the research work:

The 891 data will be divided into different ratios for training and then testing. The different algorithms which will be applied for the purpose of the research are Multiple Linear Regression and Logistic Regression. Apart from applying the above methods, there will be a trivial base case method which assumes all the women and children (whose age is below 18) survived the horrific incident and all the men died. This assumption is based on the records that initially women and children were allowed to come out of the ship.

Different segregation of the data for applying the above three methods are listed as follows:

1. Training Data- PassengerId values from 1-446 Testing Data PassengerId values from 447-891
2. Training Data- PassengerId values from 447-891 Testing Data PassengerId values from 1-446
3. Training Data- PassengerId values from 1-600 Testing Data PassengerId values from 601-891
4. Training Data- PassengerId values from 601-891 Testing Data PassengerId values from 1-600

A. Multiple Linear Regression

Multiple Linear Regression is an extension of simple linear regression [5]. It is used when we want to predict the value of a variable based on the value of two or more other variables. The variable we want to predict is called the dependent variable (or sometimes, the outcome, target or criterion variable). The variables we are using to predict the value of the dependent variable are called the independent variables (or sometimes, the predictor, explanatory or regressor variables).

For example, you could use multiple regression to understand whether exam performance can be predicted based on revision time, test anxiety, lecture attendance and gender. Alternately, you could use multiple regression to understand whether daily cigarette consumption can be predicted based on smoking duration, age when smoking, smoker type, income and gender started.

Multiple regression also allows you to determine the overall fit (variance explained) of the model and the relative contribution of each of the predictors to the total variance explained. For example, you might want to know how much of the variation in exam performance can be explained by revision time, test anxiety, lecture attendance and gender "as a whole", but also the "relative contribution" of each independent variable in explaining the variance.

There are certain assumptions associated with Multiple Linear Regression. They are explained below:

First, multiple linear regression needs the relationship between the independent and dependent variables to be linear. It is also important to check for outliers since multiple linear regression is sensitive to outlier effects. The linearity assumption can best be tested with scatterplots.

Second, the multiple linear regression analysis requires that the error between observed and predicted values (i.e., the residuals of the regression) should be normally distributed. This assumption can best be checked by plotting residual values on a histogram with a fitted normal curve or by reviewing a Q-Q-Plot. Normality can also be checked with a goodness of fit test (e.g., the Kolmogorov-Smirnov test), though this test must be conducted on the residuals themselves. When the data is not normally distributed, a non-linear transformation (e.g., log-transformation) might correct this issue if one or more of the individual predictor variables are to blame, though this does not directly respond to the normality of the residuals.

Third, multiple linear regression assumes that there is little or no multicollinearity in the data. Multicollinearity occurs when the independent variables are not independent from each other.

Fourth, multiple linear regression analysis requires that there is little or no autocorrelation in the data. Autocorrelation occurs when the residuals are not independent of each other. In other words when the value of y(x+1) is not independent of the value of y(x). This for instance typically occurs in stock prices, where today's price is not independent of yesterday's price [8].

The last assumption the multiple linear regression analysis makes is homoscedasticity.

B. Logistic Regression

In statistics, logistic regression, or logit regression, or logit model is a regression model where the dependent variable (DV) is categorical. The binary logistic model is used to estimate the probability of a binary response based on one or more predictor (or independent) variables (features). It allows one to say that the presence of a risk factor increases the probability of a given outcome by a specific percentage. Logistic regression can be binomial, ordinal or multinomial. Binomial or binary logistic regression deals with situations in which the observed outcome for a dependent variable can have only two possible types, "0" and "1" (which may represent, for example, "dead" vs. "alive" or "win" vs. "loss"). Multinomial logistic regression deals with situations where the outcome can have three or more possible types (e.g., "disease A" vs. "disease B" vs. "disease C") that are not ordered. Ordinal logistic regression deals with dependent variables that are ordered. In binary logistic regression, the outcome is usually coded as "0" or "1", as this leads to the most straightforward interpretation. If a particular observed outcome for the dependent variable is the noteworthy possible outcome (referred to as a "success" or a "case") it is usually coded as "1" and the contrary outcome (referred to as a "failure" or a "noncase") as "0". Logistic regression is used to predict the odds of being a case based on the values of the independent variables (predictors). The odds are defined as the probability that a particular outcome is a case divided by the probability that it is a noncase.

Like other forms of regression analysis, logistic regression makes use of one or more predictor variables that may be either continuous or categorical. Unlike ordinary linear regression, however, logistic regression is used for predicting binary dependent variables (treating the dependent variable as the outcome of a Bernoulli trial) rather than a continuous outcome. Given this difference, the assumptions of linear regression are violated. In particular, the residuals cannot be normally distributed. In addition, linear regression may make nonsensical predictions for a binary dependent variable. What is needed is a way to convert a binary variable into a continuous one that can take on any real value (negative or positive). To do that logistic regression first takes the odds of the event happening for different levels of each independent variable, then takes the ratio of those odds (which is continuous but cannot be negative) and then takes the logarithm of that ratio. This is referred to as logit or log-odds) to create a continuous criterion as a transformed version of the dependent variable

We have two possible outcomes 1 or 0 for N observations indicating a success or failure of an event. For N observations, a study of failures in a class is conducted for a particular subject. Following are the outputs:

{ 0, 0,0,1,1,1,1,1,1,1,0,0,0,0,0,…………..}

In which 1 indicates the pass in subject and 0 indicates the failure in subject.

The likelihood principle says that all inference about a parameter should utilize observed data only through how it affects the likelihood function, the probability of observing the observed data given *p*. the likelihood is

$F(Y) = P(Y = (1, 1, 1, 0, 0, 1 \ldots\ldots\ldots\ldots 1, 1, 0) | p)$

$= p * p * p * (1 - p) * (1 - p) * p * \ldots\ldots p * p * (1-p)$

$F(Y) = \sum P^{y} * (1 – P)^{1-y}$ …. Eq 1.1

In our case P takes up the logistic distribution function $K(x) = e^{z} / (1+e^{z})$

$\ln F(Y) = \sum y \ln P + 1-y \ln (1 – P)$

$\ln F(Y) = \sum y \ln P - y \ln (1 – P) + \ln (1 – P)$

$\ln F(Y) = \sum y \ln (P / (1 – P)) + \sum \ln (1 – P)$

$\sum(P / (1 - P)) = e^z / (1+e^z) / 1/(1+ e^z)$
$\sum(P / (1 - P)) = e^z$ where $z = \beta1 + \beta2 * x$
$\sum (1 - P) = 1/(1+ e^z)$
$\ln F(y) = \sum y (\beta1 + \beta2 * x) - \sum \ln(1 + e^{\beta1 + \beta2 * x})$ …. Eq 1.2
Eq 1.2 is called as Log Likelihood function

In maximum likelihood, our aim is to maximize the likelihood function and finding the unknown parameters β1 and β2 in such a way that the probability of observing the value of Y's as high as possible. The resulting expansion will be of nonlinear nature hence we should resort to iterative algorithm like Newton – Raphson method [7]. Logistic regression measures the relationship between the categorical dependent variable and one or more independent variables by estimating probabilities using a logistic function, which is the cumulative logistic distribution. Thus, it treats the same set of problems as probit regression using similar techniques, with the latter using a cumulative normal distribution curve instead. Equivalently, in the latent variable interpretations of these two methods, the first assumes a standard logistic distribution of errors and the second a standard normal distribution of errors.

The definitions of the different variables used for the purpose of the research are listed as follows:

Table I Definitions of Variables in the Titanic Dataset

Variable	**Definition**
Survived	Survival
Pclass	Ticket class
Sex	Sex
Age	Age in years
Sibsp	# of siblings / spouses aboard the Titanic
Parch	# of parents / children aboard the Titanic
Ticket	Ticket number
Fare	Passenger fare
Cabin	Cabin number
Embarked	Port of Embarkation

For the research purpose the dependent variable will be whether the passenger survived or not. The independent variables chosen will be scaled down from the original data. The independent variables chosen for the purpose of the research are Pclass, Sex, Age, Sibsp, Parch, Fare and Embarked.

In the dataset, 0 in the Survived column means the passenger unfortunately did not survive. Pclass value of 1 denotes Upper Class, value of 2 denotes the Middle Class and value of 3 denotes the Lower Class.

In the dataset, 177 rows have missing values for Age column. This field is filled with the mean of all the ages in the dataset (average inclusive of missing data) which comes out to be 29.6 years. So all the missing Ages in the dataset are filled with the value 29.6.

Out of all passengers embarking from location 'C'- 93 survived, from 'Q'- 30 survived and from 'S' – 217 survived. Hence for converting the data into a numeric form, the passengers from location 'C' are assigned value of 2, from 'Q' are assigned value of 3and passengers from 'S" are assigned value of 1 based on the similar property of the Pclass variable in the dataset.

For the missing values of location, the value assigned is C (considering it as unbiased) and hence a value of 2 for computation.

For making the numerical value of gender for taking it into calculation, the women are given a value of 1 and the men are given a value of 0.

In the case of multiple regression, the output obtained from inputting the test data after creating the model from the training data is collected. The average of the output is found out and any value lower than the average is treated as '0' which basically means the passenger died. On the other side, a value of the output of the test data obtained from the model formed from the training data, if higher than the average value of all the outputs obtained from the model, the value is treated as '1' which means the passenger survived.

In the case of logistic regression, the output obtained from the model first acts as the power of the exponential function. This value is added to one and the result acts as the divisor to the older value. The result of this division operation if below '0.5' then the value assigned is '0' implying the passenger is dead and vice versa.

IV. RESULTS

Table 2 Comparative analysis of different methods

Serial No.	**Result Analysis (Qualitative Definition)**			**Metrics**
	Name of Method Used	**Test Data**	**Training Data**	**Accuracy**
1	Assumption Based	Not Applicable	Passengers with PassengerId from 447-891	345/445=77.528%

2	Assumption Based	Not Applicable	Passengers with PassengerId from 1-446	356/446=79.820%
3	Assumption Based	Not Applicable	Passengers with PassengerId from 601-891	226/291=77.663%
4	Assumption Based	Not Applicable	Passengers with PassengerId from 1-600	475/600=79.166%
5	Multiple Regression	Passengers with PassengerId from 1-446	Passengers with PassengerId from 447-891	349/445=78.426%
6	Multiple Regression	Passengers with PassengerId from 447-891	Passengers with PassengerId from 1-446	344/446=77.130%
7	Multiple Regression	Passengers with PassengerId from 1-600	Passengers with PassengerId from 601-891	461/600=76.833%
8	Multiple Regression	Passengers with PassengerId from 601-891	Passengers with PassengerId from 1-600	344/446=77.130%
9	Logistic Regression	Passengers with PassengerId from 1-446	Passengers with PassengerId from 447-891	353/445=79.325%
10	Logistic Regression	Passengers with PassengerId from 447-891	Passengers with PassengerId from 1-446	353/446=79.147%
11	Logistic Regression	Passengers with PassengerId from 1-600	Passengers with PassengerId from 601-891	235/291=80.756%
12	Logistic Regression	Passengers with PassengerId from 601-891	Passengers with PassengerId from 1-600	480/600=80.000%

V. CONCLUSION AND ANALYSIS

It can be seen from Table 2 that all the models which were applied across the different test cases and different algorithms yield almost similar results. In the case of the assumption based methods, there is no test data to create a model. This is because here only gender of the passenger is considered. So basically this model can be represented as y=x where value of x is either ‘0’ or ‘1’. The maximum accuracy obtained from this model is 79.820% for the data of passengers with PassengerId value ranging from 1-446. This indicates that almost a linear model is present.

In the case of Multiple Linear Regression, the maximum accuracy obtained from this model is 78.426% for the training data of passengers with PassengerId value ranging from 1-446 and the testing data of passengers with PassengerId value ranging from 447-891. The model performs better than the Assumption based method only in the first case as listed in the methodology section of this paper.

In the case of Logistic Regression, the maximum accuracy obtained from this model is 80.756% for the training data of passengers with PassengerId value ranging from 1-600 and the testing data of passengers with PassengerId value ranging from 601-891. The model also manages to hit 80% accuracy in last case of the cases listed in the Methodology section of this paper. Though less number of data items were used for training the data model, the data model proved to be significant in case of predicting for a larger number of inputs.

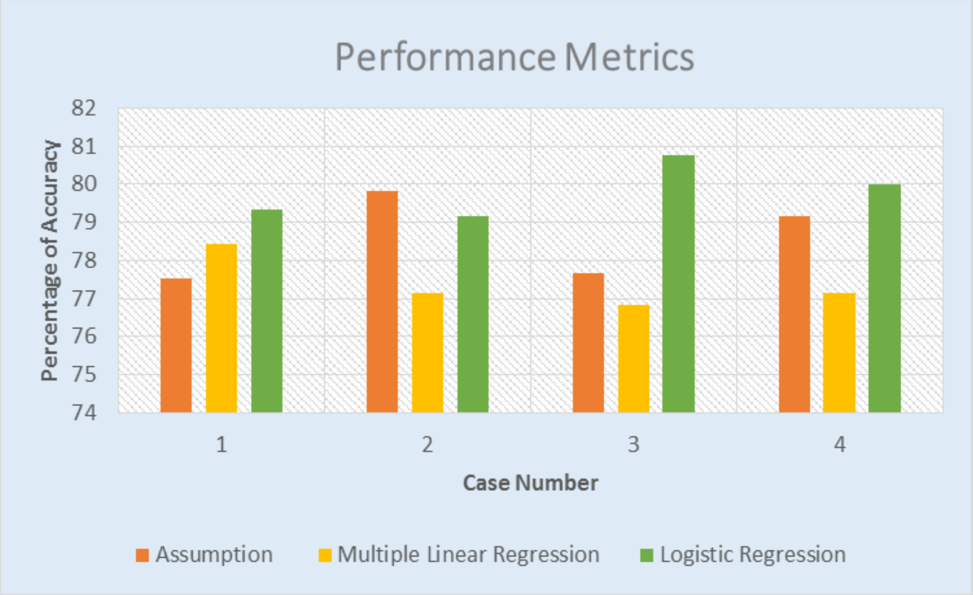

Fig 1. Performance Metrics across different cases comparison

The Logistic regression model performs best and saturates within 10 iterations in all of the above cases. Thus it can be concluded that multiple linear regression performed the worse followed by the assumption based method and then logistic regression the best.

REFERENCES

[1] GE, “Flight Quest Challenge,” Kaggle.com. [Online]. Available: https://www.kaggle.com/c/flight2-final. [Accessed: 2-Jun-2017].

[2] “Titanic: Machine Learning from Disaster,” Kaggle.com. [Online]. Available: https://www.kaggle.com/c/titanic-gettingStarted. [Accessed: 2-Jun-2017].

[3] Wiki, “Titanic.” [Online]. Available: http://en.wikipedia.org/wiki/Titanic. [Accessed: 2-Jun-2017].

[4] Kaggle, Data Science Community, [Online]. Available: http://www.kaggle.com/ [Accessed: 2-Jun-2017].

[5] Multiple Regression, [Online] Available: https://statistics.laerd.com/spss-tutorials/multiple-regression-using-spss-statistics.php [Accessed: 2-Jun-2017].

[6] Logistic Regression, [Online] Available: https://en.wikipedia.org/wiki/Logistic_regression [Accessed: 2-Jun-2017].

[7] Consumer Preferences to Specific Features in Mobile Phones: A Comparative Study [Online] Available: http://ermt.net/docs/papers/Volume_6/5_May2017/V6N5-107.pdf.

[8] Multiple Linear Regression, [Online] Available http://www.statisticssolutions.com/assumptions-of-multiple-linear-regression/ [Accessed: 3-Jun-2017]