Assignment 2: a mathematical essay on logistic regression

Krishna Sah Teli
Cyber-Physical Systems (I2MP)
Indian Institute of Technology (IITM)
Madras - 600036, India
ge22m018@smail.iitm.ac.in

***Abstract*—In this paper, logistic regression has been studied extensively with an application exploring titanic data set. It explains about the different parameters which play crucial role in predicting survival of passenger in titanic. With the help of data visualization, correlation between the data features has been explored. Even though number of male passenger was comparatively high, only 19% of them were able to survived during the titanic tragedy. Data analysis shows that female population was more likely to be survived because they might have been given more priority over male while evacuating. It also allows us to visualize from point of view to extract insight such as why only 38% of them only were able to survive and moreover female and child got more preference. It also makes us to think over the survival rate of passenger travelling with different classes of ticket. This paper has explored the ground truth behind the survival of passengers having different socioeconomic background. With the supervised logistic regression machine learning predictive model, the titanic data has been implemented to build a predictive model and has been explored its accuracy, precision and recall properties. In this revised paper, I found overall flow of the paper is well systematic with self explainable visualization plots. However, I have improved introduction as well as conclusion sections.**

***Index Terms*—Logistic regression, Socioeconomic, Data Visualization, Correlation, Titanic**

I. Introduction

Machine learning and data science has brought the revolutionary changes in the field of technology across the world. It has not contributed in the technical field but we can see its impact on almost every field from agriculture to astronomy. Thus, it has become the most trending topic in current time period. As it has great responsibility, it is more likely to vanish a world if it is miss-understood or handed over to wrong parties such Naxal, terrorists, etc. Thus It should come up with restriction and strict punishment for mishandling. It totally depends on the data which is used to train the ml-model. If the data is noisy, ml model would be of less accuracy in the performance. That's why data preprocessing followed by data augmentation and proper formatting is necessary. It is also necessary to understand the different data features and their correlation to make the model robust. Ml-model can predict, estimate, recognize, detect, etc. based on the data and type of training model. In this paper, titanic data has been used to predict the survival of passenger. It explains the correlation between the data features such as socioeconomic background of passenger, age, sex and family group with respect to their survival. With the help of data analysis and visualization, ground truth has been explored followed by development of predictive model. [9]

To build a predictive model, logistic regression, a supervised machine learning model has been implemented. The dependent variable (survived) has two classes i.e binary class and is predicted based on the correlation between other independent variables such as age, sex, fare, sibling/spouse, parent/child, boarding station, etc. Logistic regression is slightly different from linear regression such as logistic regression is used when target variable is categorical(it may be binary or multi-class) [1] while linear regression is used when target variable is numerical. Logistic regression has several application in industries [2], [3] as well as in small scale company or in developing smart gadgets. It is used when decision making model has to be built such as pass/fail, yes/no, sunny/cloud/rainy day, good/bad, wrong/right, etc. It has wide range of application such as weather forecasting, developing AI-based devices, e-commerce stack fall down prediction, health status check up, disease confirmation, banking [3], scientific studies [9], etc.

The main focus in this paper is to demonstrate what proportion of passenger were able to survived and what group of passenger got more preference during the titanic tragedy. It correlates the passenger survival with their socioeconomic status i.e ticket class. With the help of visualization techniques, passenger survival during sinking of titanic with respect to family group, point of boarding, sex, age and fare has been illustrated. It also explains why only 19 % of male passenger got survived though they were 69% of total passenger and 78% of female passenger got survived.

The rest of the paper has been organized as follows: section II outlines the data set related to our problem, section III explains the mathematics behind the logistic regression and tools used for developing our model. Section IV Modelling and section V concludes the paper with key contributions made and directions for future works.

II. Data set

The source of data is kaggle which is famous for the data science and storage of data. This data has totally 12 columns namely PassengerId, Survived, Pclass, Name, Sex, Age, SibSp, Parch, Ticket, Fare, Cabin, and Embarked. Out of 12 data columns, only 8 columns are selected for further preprocessing and building predictive model. The 8 important data columns are Survived, Pclass, Sex, Age, SibSp, Parch, Fare, Embarked. Pclass shows the ticket class and it has three different values namely 1st class, 2nd class and 3rd class where as SibSp explains about the sibling and spouse travelling together and Parch explains the number of child with parent travelling together in the titanic. Similarly, embarked tells about the boarding station which includes totally three different stations namely; S: Southampton, Q: Queenstown, and C: Chernberg.

With the data analysis and data visualization, the correlation among the data features has been studied and correlated with the survival of passengers. The fig.1 shows about 70% of passenger died in the titanic tragedy and about 30% of them were able to survive.

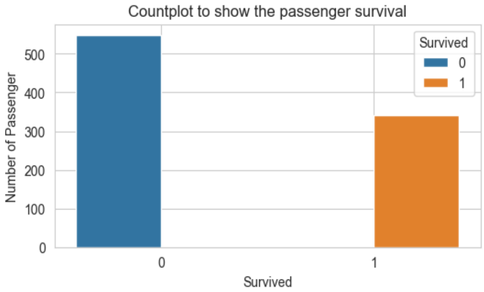

Fig. 1. Showing passenger survival

Similarly, the fig.2 shows out of the passenger who died in titanic tragedy, about 80% of them were male and 20% of them were female. Similarly, approximately 74% of female passengers were able to survived and 20 % of male were able to survive. It may be because female might have been given more priority while evacuating the titanic.

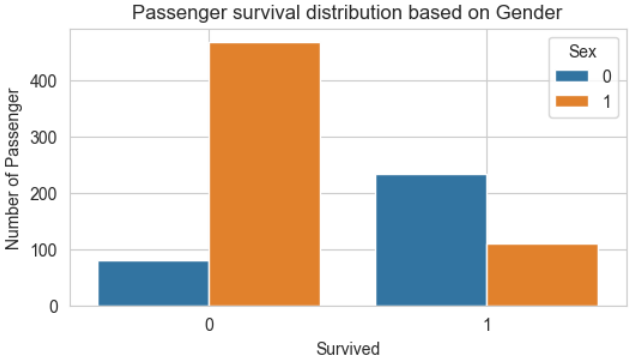

Fig. 2. showing the passenger survival based on gender

The fig. 3 illustrates that the number of passenger who died significantly more were from the 3rd class of passengers. Similarly, 63% of 1st class passengers and 24% of 3rd class passengers survived during that tragedy. It shows that titanic manager gave more priority to first class of passengers.

Likewise fig.3, figure 4 also describes that the passenger who paid more got more preference while evacuating. It also shows there were less high paid passengers. The passenger who died were mainly from mediocre family. This illustrates that first priority highly paid passenger and then others.

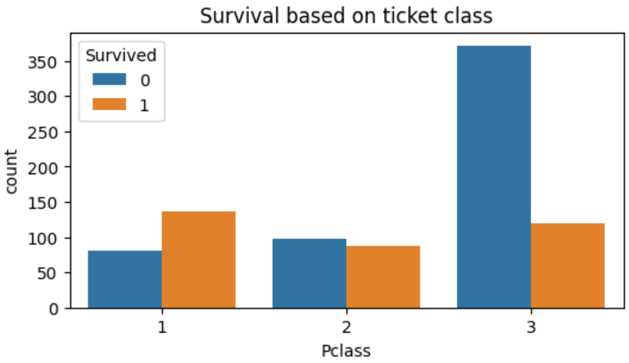

Fig. 3. Survival based on ticket class

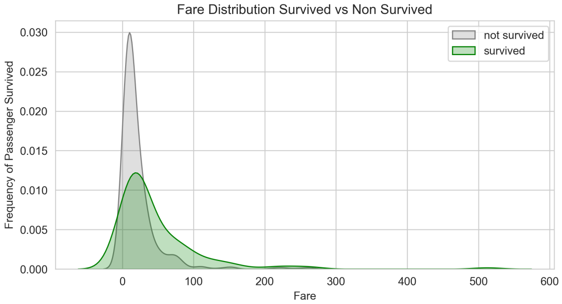

Fig. 4. Survival based on fare

The figure 5 depicts that the teenager passenger had survived more than that of child and old citizen. It shows 12% of them were child passenger and 53% of child passenger got survived where as the 32% of them were old citizen and about 30% of them survived. Similarly, 54% of them were adult group of passenger and 38% of them survived. It shows that teenager were busy in saving child and old group of people and most of them sacrificed themselves for others.

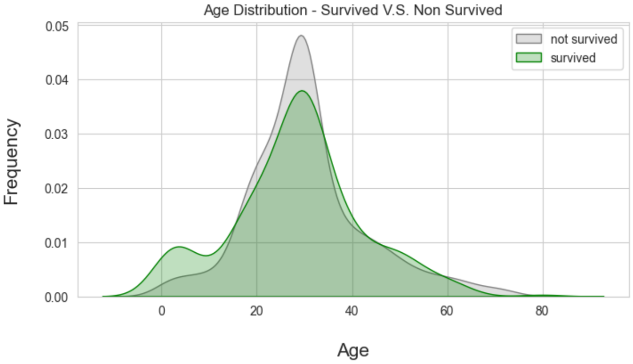

Fig. 5. Survival with respect to different age group of people

The below figure 6 illustrates that the passenger who were travelling with family or any group of people had not better chance of survival. This figure clearly shows that as the number of family member increases, more responsibility keeps on adding. Because of this, most of them becomes nervous while tragedy and loses life.

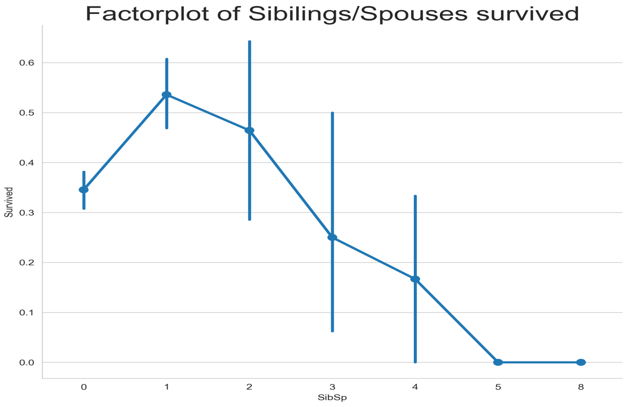

Fig. 6. Survival of passenger travelling with family

Here the figure 7 depicts the relation among the different parameters such as sex, age, emabarked and survival rate. It shows that the passenger who boarded at the station Southamptole died comparatively more than others and most of them were male passengers. It also explains that very less passenger had boarded at Queenstown station.

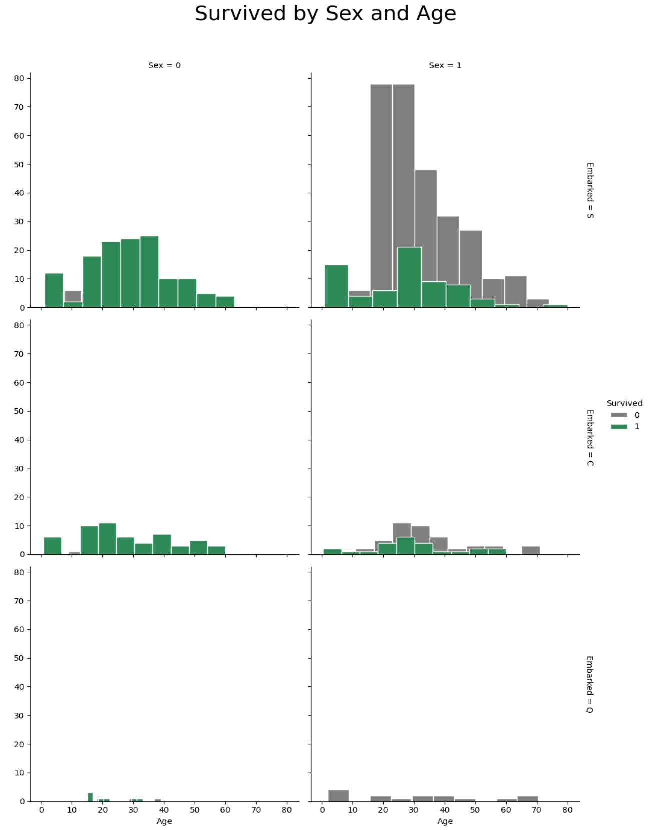

Fig. 7. Survival among the passenger of different age, sex and boarding station

III. Logistic Regression

Regression Analysis is a statistical technique of constructing mathematical models to estimate relationships existing between the response variable Y and the predictor variables X. [4] Logistic regression is a method of modeling the probability of a discrete outcome based on given input variable. It is used when one has to build a model to make a decision. The most frequently used logistic regression models is a binary outcome. Its outcomes are in a pair such as true/false, right/wrong yes/no, and so on. There can be more than two discrete outcomes and this kind of logistic regression is called Multinomial logistic regression [7]. It can be implemented in different aspects [2], [3], [8] such in cyber security for classifying the problems, such as attack detection, pattern recognition, spam detection, fraud detection, etc. Hence, it is a useful analytic technique.

Logistic regression is another useful as well as powerful supervised machine learning algorithm [10]. we can say that logistic regression is a kind of linear regression but for classification problems. It mainly uses a logistic function defined below to model a binary output variable. [6]The main difference between logistic regression and linear regression is that linear regression's range has numerical outcome while logistic regression outcome is bounded between 0 and 1. Moreover, it is not necessary to have linear relationship among inputs and output variables. In order to convert the outcome in range of 0 to 1, a logistic function is used.

$$Logisticfunction = 1/(1 + (exp)^{-x})$$

With the help of logistic function (sigmoid function), a linear regression is transformed into logistic regression. Here x represents the linear relation among the data features such x= $\beta 0$ + $\beta 1$ $*$ x. The loss function used in logistic regression is referred as MLE i.e "maximum likelihood estimation". It is a conditional probability function where prediction is classed as class 0 If the probability is greater than 0.5 else class 1. Before going through logistic regression derivation, let's first define the logit function. Logit function is defined as the natural log of the odds. A probability of 0.5 corresponds to a logit of 0, probabilities smaller than 0.5 correspond to negative logit values, and probabilities greater than 0.5 correspond to positive logit values. Logistic function ranges between 0 and 1 (P ϵ $[-\infty, \infty]$) while logit function can be any real number from minus infinity to positive infinity (P ϵ $[-\infty, \infty]$).

$$odds = P/(1 - P)$$

Logistic regression is a parametric form for the distribution $P(X/Y)$ where Y is a discrete value and X= $x1$,$x2$,$x3$,...., x_n is a vector containing discrete or continuous values. The parametric model of LR can be written as

$$P(Y=1|X) = \frac{1}{1+exp(w_0 + \sum_{n=1}^{n}(w_i - x_{)}}$$

and,

$$P(Y=0|X) = \frac{exp(w_0 + \sum_{n=1}^{n}(w_i - x_{)}}{1+exp(w_0 + \sum_{n=1}^{n}(w_i - x_{)}}$$

This is the mathematics behind the classification or binary classification predictive model. After the model is developed, its accuracy, precision and recall is checked against the testing data. Confusion matrix is used or accuracy score method is used to evaluate the model.

IV. Modeling

Logistic regression is based on conditional probability. It predicts the output based on the given input. It uses a sigmoid function to convert the numerical outcome into discrete i.e 0 and 1. In our model, it relates the data features with some weight and finds the probability of survival. If the probability given input is more than 0.5 then survival =1 otherwise it is 0. Since sex has more correlation with survival features, hence sex has more weight while finding probability. To build a logistic regression model, from sklearn regression model is imported and liblinear as solver is used with l1 penality. For evaluating the model, accuracy_error and mean_absolute_error is used. To develop a perform report, classification_report is used. For this model which has been trained using titanic data set, its accuracy is 81.38 % followed by recall_score 77.27% and precision 71.8%.

Acknowledgement

It is my pleasure to express with deep sense of gratitude to Professor Gaurav Raina, for his guidance, continual encouragement, understanding.

I would like to extend my gratitude to IIT Madras, for providing with an environment to work in and for his inspiration during the tenure of the course.

Conclusion

In this paper, Logistic regression has been used to find out the ground truth behind the biased death of passengers from the given titanic data set. Data analysis and visualization has been performed in order to find out the correlation among the variables and cause-effect. The ground truth is that the death of passenger is socioeconomic oriented as more number of 3rd class passenger i.e 63 % were died compared to 1st class (23%). Out of passenger, about 38% total passenger had survived and 74% out of female passenger and 19% were out of male passenger. From this analysis, it can be said that titanic manager had given more priority to female passenger from 1st class followed by child passenger. This paper also guides future researcher to grow in data science and learn more about logistic regression.

References

[1] X. Liu, "Logistic Regression for Binary Data," in Applied Ordinal Logistic Regression Using Stata: From Single-Level to Multilevel Modeling, 2022. doi: 10.4135/9781071878972.n3.

[2] S. Mehrolia, S. Alagarsamy, and V. M. Solaikutty, "Customers response to online food delivery services during COVID-19 outbreak using binary logistic regression," Int J Consum Stud, vol. 45, no. 3, 2021, doi: 10.1111/ijcs.12630.

[3] R. P. Hauser and D. Booth, "Predicting Bankruptcy with Robust Logistic Regression," Journal of Data Science, vol. 9, no. 4, 2021, doi: 10.6339/jds.201110_09(4).0006.

[4] C. Y. J. Peng, K. L. Lee, and G. M. Ingersoll, "An introduction to logistic regression analysis and reporting," Journal of Educational Research, vol. 96, no. 1, 2002, doi: 10.1080/00220670209598786.

[5] S. Reza, B. Sarwar, R. R. Nawaz, and S. M. N. Ul Haq, "Application of Logistic Regression on Passenger Survival Data of the Titanic Liner," Journal of Accounting and Finance in Emerging Economies, vol. 7, no. 4, 2022, doi: 10.26710/jafee.v7i4.1994. [6] N. Torosyan, "Application of binary logistic regression in credit scoring," University of Tartu, 2017.

[6] A. Borucka and M. Grzelak, "Application of logistic regression for production machinery effciency evaluation," Applied Sciences (Switzerland), vol. 9, no. 22, 2019, doi: 10.3390/app9224770.

[7] A. M. El-Habil, "An application on multinomial logistic regression model," Pakistan Journal of Statistics and Operation Research, vol. 8, no. 2, 2012, doi: 10.18187/pjsor.v8i2.234.

[8] L. Niu, "A review of the application of logistic regression in educational research: common issues, implications, and suggestions," Educational Review, vol. 72, no. 1. 2020. doi: 10.1080/00131911.2018.1483892.

[9] S. Sabouri, A. Hajrasouliha, Y. Song, and W. H. Greene, "Logistic regression," in Basic Quantitative Research Methods for Urban Planners, 2020. doi: 10.4324/9780429325021-13.

[10] S. Sperandei, "Understanding logistic regression analysis," Biochem Med (Zagreb), vol. 24, no. 1, 2014, doi: 10.11613/BM.2014.003.

[11] Connelly, "Logistic regression," MEDSURG Nursing, vol. 29, no. 5, 2020, doi: 10.46692/9781847423399.014.

Source code link: https://drive.google.com/drive/folders/1TldIU49IYyRuwFL_VbGFWaaXFA_Debk0?usp=share_link