Assignment 2: Logistic Regression

1st Rohith D
Electrical Engg.
IIT Madras
Chennai, India
ee18b148@smail.iitm.ac.in

***Abstract*—This is a study on the survival rates on the titanic and the factors affecting it. Logistic regression and is used to analyze a training data-set that lists the passengers of the titanic and various information about them, and we build a model to predict survivors on a test data-set. Using the model, we analyze which factors affected the survival rates.**

***Index Terms*—logistic regression, titanic survivor, accuracy, visualization**

I. Introduction

The sinking of the titanic is perhaps the most well known disaster from modern history. Due to the lack of adequate life boats on the ship, there was a large loss of life on the ships maiden, and only voyage. This assignment is an application of logistic regression to predict survivors of the titanic using data about their age, social class, sex, etc. We then analyze the model parameters to get an idea of how the factors change survival rate.

We do an exploratory analysis of the given data using plots to get a preliminary idea of the relationship between data and the ground truth. For the continuous data, we use logistic regression to build a model between them. The model establishes a non-linear relationship between the input and output variables.

The aim is to predict whether a person survives based on information like their age, number of siblings on board, what their ticket class was, etc. It could help us better understand the sinking of the titanic and the evacuation and rescue process during the same.

We start with an overview of logistic regression. Then, we do an exploratory analysis based on some plots. We present our observations based on the data both quantitatively and visually. We then analyze the fitted logistic regression models. From these plots we arrive at a conclusion.

II. Logistic Regression

Logistic Regression is a supervised machine learning algorithm. We use the model to find the probability of a certain class or event. It can be used when the data is linearly separable and the output is dichotomous (binary) in nature. Since our output is binary (Survived/ Not-survived), we can use logistic regression.

Simple Logistic regression models a relationship between two variables. We have a predicted probability p and an input X.

$$y = wX + b \tag{1}$$

where b is the bias and w is the weight. y is the linear regression output. We pass this to a sigmoid function to get a probability. It maps any predicted values into a range of 0 to 1.

$$p = \sigma(y) = \frac{1}{1+e^{-y}} \tag{2}$$

Here p is the predicted probability.

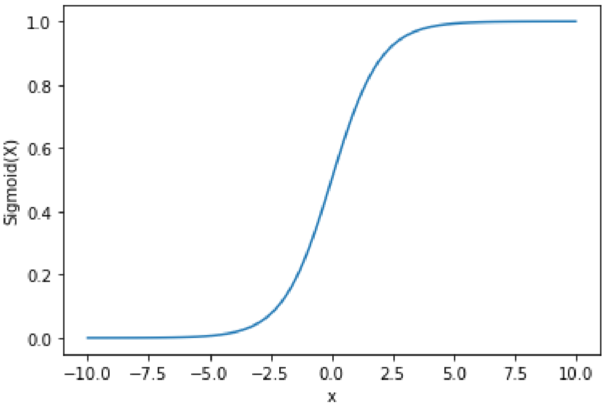

Fig. 1. Sigmoid Function

When we have more than one predictor, it is called multiple logistic regression. The equation for it is similar to simple logistic regression.

$$p = \sigma(y) = \frac{1}{1+e^{-y}} = \frac{1}{1+e^{-(b+w_1X_1+w_2X_2+...w_kX_k)}} \tag{3}$$

where b is the bias and w_i is the weight for the ith input variable.

To learn the model parameters, we minimize cross-entropy loss function which is used to measure performance of a classification model whose output is a probability value. For N samples, the loss function is thus:

$$CE = \sum_{i=1}^{N}[y_i logp_i + (1-y_i)log(1-p_i))] \tag{4}$$

where y_i is the true output (either 0 or 1) and p is the predicted probability of y being 1.

III. The Problem

The survivors of the titanic are of a diverse variety, in terms of age, sex, class, etc. We look at their data and see if we can find something common among them that increased their survival rates.

We first clean the data before working with it:

- Columns such as those of names and ticket number have been removed from the data frame as that data is not used. Since the cabin number is missing for a lot of rows, we drop it.
- As we expect age to have a high effect on surviving, we drop rows with missing age.
- We convert gender (Female - 0, Male - 1) and departure data into numbers so that it can be input to logistic regression.

To start, we try to draw some preliminary relations from the data using graphs. We check for the percentage of people of each gender surviving using a pie chart.

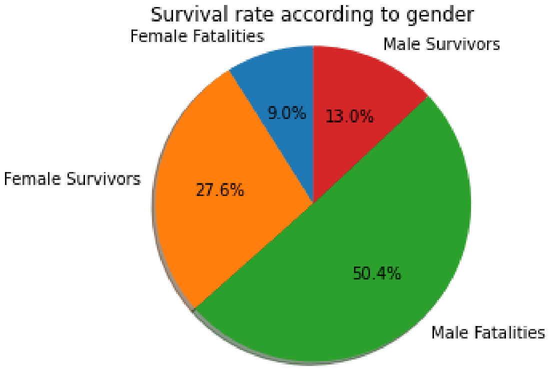

Fig. 2. Survival according to Sex

We see that the percentage of survivors in the female passengers is much higher that that of the male passengers. Female passengers may have been given priority in rescue operations.

Next we analyze how class affected survival.

Class	Sex	Survival
1	Female Male	0.96 0.40
2	Female Male	0.92 0.15
3	Female Male	0.46 0.15

We see that the higher classes had a much higher chance of surviving compared to lower classes, and there is gender disparity among all classes.

We look at age vs survival.

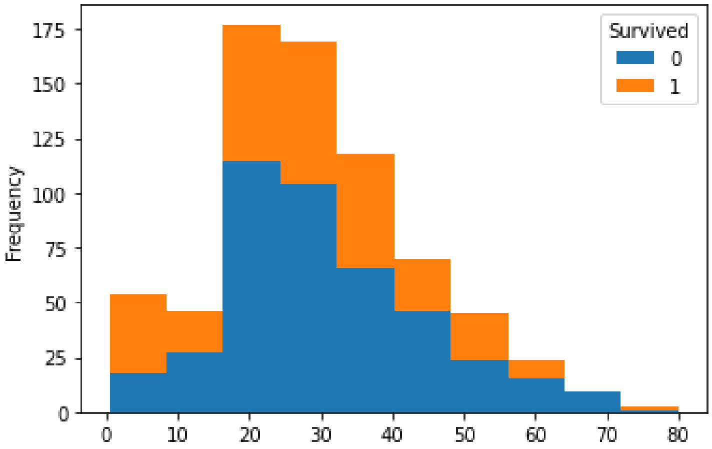

Fig. 3. Survival according to Age

We see that younger passengers passenger were more likely to survive and senior citizens were least likely to survive.

So our inferences are that the classes that are most likely to survive are female, upper class, and young.

We start constructing logistic regression models for the input variables to classify the passengers as survivors and non-survivors. We first construct a multivariate regression model with the following inputs: Class, Sex, Age, Siblings/ Spouses, Parents/ Children, Fare, Embarked ,and survived as output. We get a high accuracy on the training data of **80.2%**. We analyze the coefficients.

We see a negative coefficient corresponding to passenger class and sex, this implies that lower value of class, i.e. , upper class, is preferred, and lower value of sex (female - 0/ male - 1), i.e. , female is preferred. Similarly, lower value of age has higher survival prediction since the coefficient is negative. The number of siblings/ spouses and parents/ children on board does not matter as much to survival as the coefficient is relatively lower (-0.344 and -0.048) in comparison to the data values. Similarly, Fare and Embarking Port don't have much of an effect though the survival chances go up when the Travel Fare goes up.

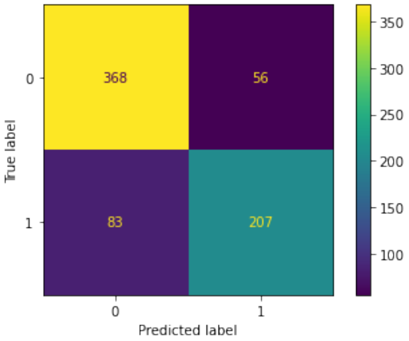

Fig. 4. Confusion Matrix

Now we use the test data. In the test data, for the missing age and fare values, we use an imputer to fill the missing cells with the mean of the column. We run a prediction on the test data, attach it to the data frame, and write it to a csv file. We get a **38.8%** survival rate.

IV. Conclusions

We have seen which features affected the survival rate during the sinking of the titanic the most and how it is affected. We see that the lower income/ low class population was much more likely to die during the accident compared to the higher class population. The female population was much more likely to survive due to being evacuated with higher priority. The same goes for children/ younger part of the population compared to the senior citizens. Further study is possible on the correlation between fare and class, as well as the title (Mr, Ms, Sir, etc.) from the passenger's name and survival rate.

References

[1] https://towardsdatascience.com/building-a-logistic-regression-in-python-step-by-step-becd4d56c9c8

[2] https://www.analyticsvidhya.com/blog/2021/07/an-introduction-to-logistic-regression/

[3] https://www.geeksforgeeks.org/understanding-logistic-regression/