Assignment [2]

A Mathematical Essay on Rogistic regression.

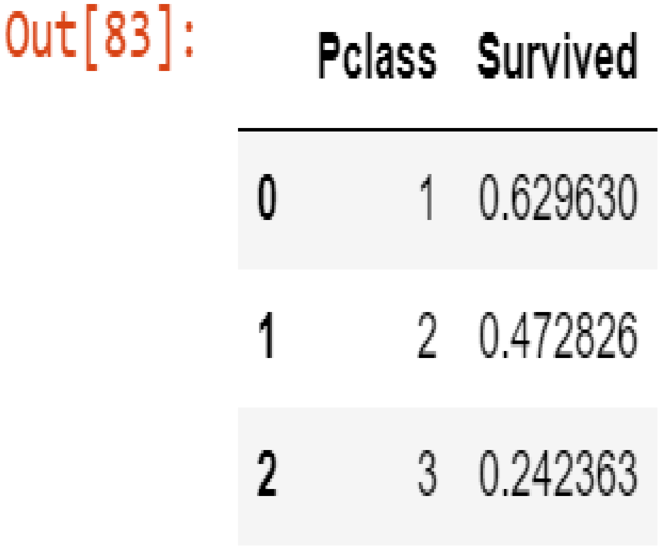
Mustefa Abrahim
Deprt. Data Science
Indian Institute of Technology Madras (IITM)
ge22m014@smail.iitm.ac.in

***Abstract*—The main goal of this algorithm's first to find predictable of survival of people from the sinking of titanic by implementing exploratory data analytics on the available training data, and then to complete the analysis by applying various machine learning models and classifiers. This will indicate which individuals are more likely to survive. [2] We used logistic regression to predict the survivors in terms of class and gender from the Titanic sinking on April 15, 1912, which resulted in the deaths of 1502 out of 2224 passengers and crew. We used re-scaling to correct some outliers, such as in the age and fare column, and in the end, first-class passengers survived than others classes, and females survived than males. [1]**

Key Words: logistic Regression, Exploratory, Modeling

I. Introduction

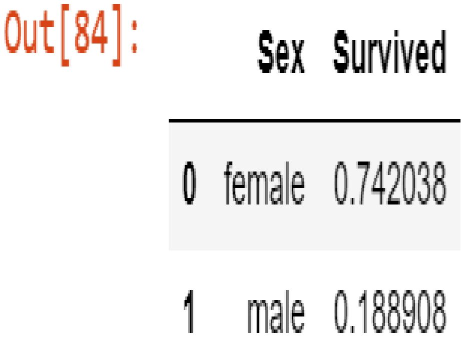
Logistic Regression is a machine learning algorithm that can be used to model the likelihood of a specific class or event. It is used when the data is separable linearly and the outcome is binary or dichotomous. That is, logistic regression is typically applied to binary classification problems. We used logistic regression when we have one dependent variables which depends on the one or more independent variables which make it different from the linear regression. To predict the survival of the people from the sinking of the titanic we used logistic regression as we have more independent variables such as classes, ages, sibsp(sibling and spouses), parent and children this independent variables used to predict the survival of the people,we done also the re-scaling to handle some outlier point and there was some null values which we replace it by median after comparing medial and mean. In this work, we done the prediction of the survived people from the sinking of the titanic and we got first class was survived than other class and female was survived than males as we see in the below table which taken of code.



Out[83]:

	Pclass	Survived
0	1	0.629630
1	2	0.472826
2	3	0.242363

Figure 1: survived vs class
as it show in the above table first class was survived more than others.



Out[84]:

	Sex	Survived
0	female	0.742038
1	male	0.188908

Figure 2: survived vs gender
Survived people in terms of gender was female was survived more than male.

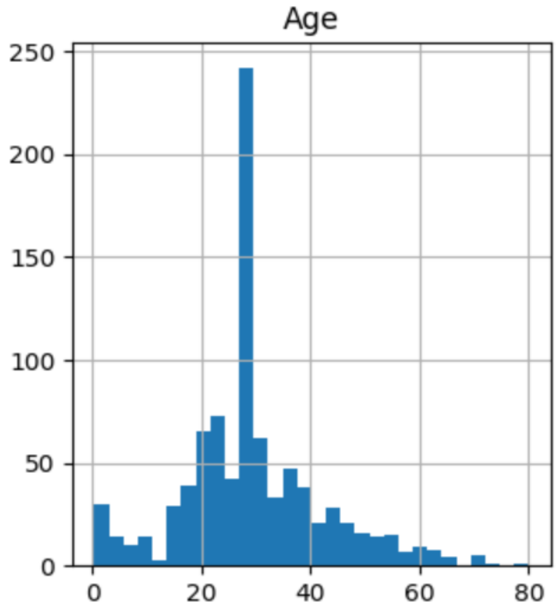
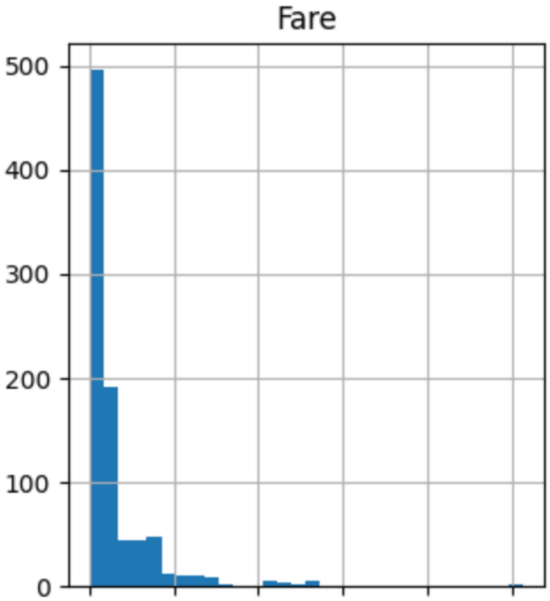


Figure 3: survived vs age

A. MODELING

Logistic Regression is some how similar to Linear Regression. But we map that output from input using logistic function that map this output to some range between [0-1], this is using some function called Sigmoid function, or between [-1-1] like Tanch and other function, but what we are used is the Sigmoid function. This mapping of [0-1] is the estimated probability of the output for some class based on threshold, which map this probability to belong to some class, like if the probability estimation is greater than or equal .5 then, we predict positive class which 1, or less than .5 then, we predict negative class 0 for binary classification problem, as in our case of trying to predict survived and not survived people. In our logistic regression we trying to predict one of the two class either 0 or 1, so in case of survived people which class 1, we should predict 1, so we have to ignore the other class in that case, as opposite is to that when we have actual class is not survived which is 0 and we should predict 0, so we need to ignore the other class which is 1, like Logistic Regression which is of our interest. So overall of instances in our data-set we trying to predict each of these instance, and compare that prediction to the actual output associated with each instance. Which give us the overall of the cost function on all instances.



B. Section 3: Data Set

Nowadays we have massive amounts of data, as well as different ways that the data was generated, such as by machines or your posts on social media or orders on online stores, among other things. Another option is to use surveys and other methods to design your data for a specific purpose. One of the main points of our research is to explore the data, but before that, we need to understand the data and clean it so that we can explore it. For this work we used titanic data, the sinking of the Titanic is one of the most infamous shipwrecks in history. On April 15, 1912, during her maiden voyage, the widely considered unsinkable RMS Titanic sank after colliding with an iceberg. Unfortunately, there weren't enough lifeboats for everyone onboard, resulting in the death of 1502 out of 2224 passengers and crew. While there was some of luck people involved in surviving, it like the first class.

C. Conclusion

The logistic regression has a higher accuracy, It works best with binary dependent variables, which have binary output values such as yes or no, true or false. In our works we got accuracy 80.02 of predicting the survived peoples.

References

[1] Analyzing Titanic disaster using machine learning algorithms-Computing, Communication and Automation (ICCCA), 2017 International Conference on 21 December 2017, IEEE.

[2] Prediction of Survivors in Titanic Dataset: A Comparative Study using Machine Learning Algorithms, Tryambak Chatterlee, IJERMT-2017.

[3] MICHAEL AARON WHITLEY, using statistical learning to predict survival of passengers on the RMS Titanic by Michael Aaron Whitley, 2015.