Logistic Regression for Predicting the Survival Rate in the Titanic Sinking

Pérez Mariana, ISC A01206747, Tecnológico de Monterrey Campus Querétaro

***Abstract*— This document presents the implementation and explanation of an analysis of the type of people that were likely to survived in the sinking of the RMS Titanic, one of the most infamous shipwrecks in history, and the model used to predict if the given passenger survived or not the incident.**

I. INTRODUCTION

On April 15, 1912, the RMS Titanic sank after colliding with an iceberg killing 1502 out of 2224 passengers and crew, this means a 32% survival rate. The following report is divided into 6 sections that will give insight and information about the state of the art, the proposal of the model, and the test and validation of said model, to predict the likelihood of survivance of a certain passenger according to their characteristics.

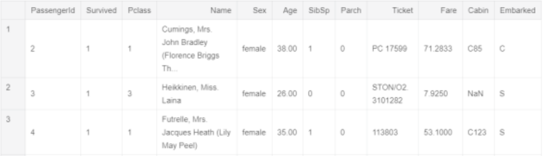
II. STATE OF THE ART

Logistic regression is a statistical model used to analyse and determine an outcome based on a dataset where there are one or more independent variables. The result of a logistic regression is a discrete value meaning that there are only two possible outcomes [1]. Logistic regression us used in a variety of fields such as marketing, human resources or even finance. Logistic regression is a very powerful algorithm, even for very complex problems it may do a good job, in some cases it can even achieve a 95% accuracy. Logistic regression is also used in even more complex algorithms such as neural networks, were each neuron in the network can be viewed as a logistic regression with an input, weights, and the bias as a result of a dot product. [2]

III. DATA SET

The dataset comes from a Kaggle[3] problem which contains records from passengers aboard the famous ship "The Titanic". The given dataset is separated into two different files which one of them is used to train the model and the other is used to test the resulting parameters of the model. The train data has 891 records meanwhile the test data has 418.

Figure 1 shows and example of the data in the given datasets. The twelve columns that are in the dataset are the id of the passenger, if the given passenger survived or not, the class of travel, the name of the passenger, the gender , the age, the number of siblings or spouse aboard, the number of parent and child aboard, the ticket id, the fare, the cabin and the port in which a passenger had embarked.



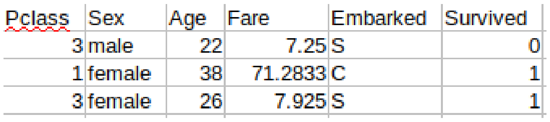
	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.00	1	0	PC 17599	71.2833	C85	C
2	3	1	3	Heikkinen, Miss. Laina	female	26.00	0	0	STON/O2. 3101282	7.9250	NaN	S
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.00	1	0	113803	53.1000	C123	S

Fig 1. Example of the data in the train dataset.

IV. MODEL PROPOSAL

An approach used to get the final model is based on a Kaggle tutorial, were it stated that it is needed to correlate the given data to see in detail which features of the dataset contributed significantly to the dependent variable. When checking the correlation of the data, it was found that some of the fields in the dataset where empty, meaning that the data also needs correcting in order to find a competent model. The fields that were dropped based on the correlation were the Ticket field, because it contained 22% of duplicates and it didn't affect directly the survival rate. The Cabin feature was also dropped as it is highly incomplete. The PassengerId and the Name columns were also dropped from the training dataset as it apparently does not contribute to the survival rate. Finally, the number of related passengers were put on hold to not make the model too complex and to make it understand better the pattern.

Figure 2 shows an example of the resulting data after dropping the fields that didn’t have a high correlation to the survival rate.



Pclass	Sex	Age	Fare	Embarked	Survived
3	male	22	7.25	S	0
1	female	38	71.2833	C	1
3	female	26	7.925	S	1

Fig 2. Example of the remaining fields used in the model

With the resulting dataset it was possible to analyse the data to make observations that would help to determine a model. Figure 3 shows a graph of the relationship between age and survival rate. This graphic representation of the age field and the survival rate helps us observed that the infants from age less than or equal to 4 had a high survival rate, all the passengers that had 80 years survived, and most of the passengers had between 15 and 35 but there is a low surviving rate between the passengers between 16 and 25. The result of this analysis gave an insight of the possible age groups in order to help the model.

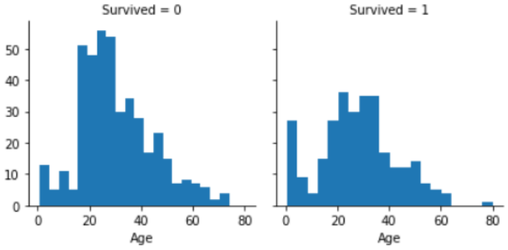


Fig 3. Visual representation of the relationship between age and survival rate

The model divides the age feature into 5 groups, in order to achieve this the cells that have a missing value are replaced with a 1 so there won’t be any null values that could affect the function. The first group is between 0 and 15, because it had the highest survival rate, the second group is between 16 and 32, which is the group with the lowest survival rate, the third group is between 33 and 48, and finally the last group is of the passengers over 64, which is another group with a high survival rate.

The next correlation made was the relationship between the travel class of the passenger. Figure 4 shows a visual representation of the relationship between the class and the survival rate. Thanks to the graph of Pclass we are able to see that the travel class number 3 had the most passengers, however most of them did not survived, most of the infants that travelled in class number 2 and 3 survived, there is a high survival rate in the passengers that were in class 1 and there is a relationship between the age of the passenger and the class they travelled in.

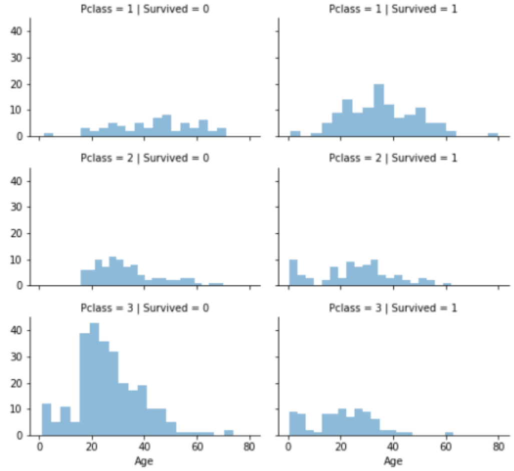


Fig 4. Visual representation of the relationship between travel class and survival

The next correlation analysed was the sex field. Figure 5 shows a visual representation of the relationship between the survival rate and the gender of the passenger. It is possible to observe that the sex of the passenger has a high correlation with the survival rate, this histogram helps us observe that female passengers had a better survival rate than males.

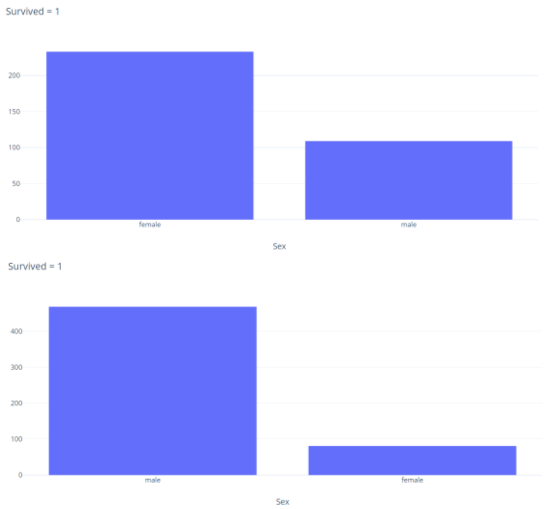


Fig 5. Visual representation of the relationship between gender and survival rate.

The dataset has as input “female” and “male” in the sex field, which are strings that cannot be used in the logistic regression function, so the model converts all the females into a 1 and all the males into a 0. These values can now be used by the algorithm to converge and give as a result the desire goal.

The final correlation that was analysed was the relationship between survival rate, the sex, the fare and the port in which the passenger embarked. Figure 6 shows a histogram of the given relationship.

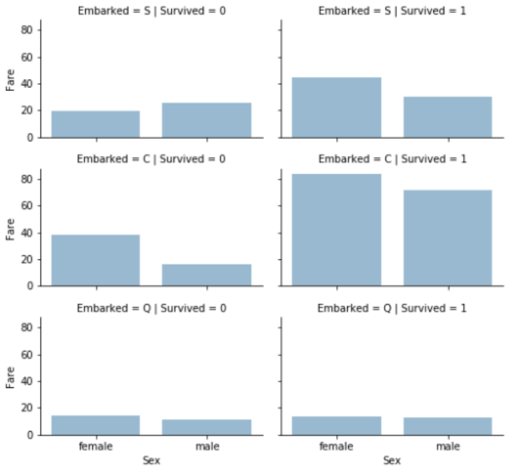
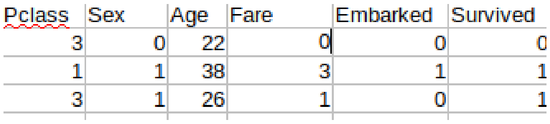


Fig 6. Visual representation of the relationship between fare, sex, port in which the passenger embarked and survival rate.

The histogram shows that even though males did have a low rate of survival in general, they had better probabilities of surviving if they embarked on C and if they were in the traveling number 3, it also shows that a passenger had a better chance at surviving if they pay a higher fare. The dataset gives a string in the embarked field which the model then changes into a numeric value, if the value of embarked is an S then the model assigns a value of 0, if it’s a C a value of 1 is assigned and if it’s a Q then a value of 2 is assigned. Figure 7 shows the resulting data of dropping some fields and having all the values in a numeric form. Finally, the fare was also grouped to help the model converged. The fare is divided into 4 groups which represent the low fare that is when the cost of the fare is less than 7.91 and outputs a value of 0, the low middle fare which is when the value is between 7.92 and 14.454 which outputs a value of 1, the middle high fare which is between 14.455 and 31 which is mapped into a 2 and the high fare which is a 4 and is when the value of the fare is bigger than 31.



Pclass	Sex	Age	Fare	Embarked	Survived
3	0	22	0	0	0
1	1	38	3	1	1
3	1	26	1	0	1

Fig. 7 Resulting data

The resulting data is then given to the logistic regression part of the model. To get the result the model uses the following gradient descent function to minimize the cost until the mean error of the whole iteration is less than 0.0001, the new parameters are the same as the older parameters or the number of epochs or iterations is equal to 10000.

$$\theta j = \theta j - \alpha \sum_{i=1}^{m} \left(h\left(x^{i}\right) - y^{i}\right) xj^{i}$$

Alpha is the learning rate and is set in the model to 0.03 and the hypothesis function is the sigmoid function. This function is used because it maps the given input times the calculated parameters into a resulting value between 0 and 1. [4].

$$\frac{1}{1 + e^{-(\theta_0 b + \theta_1 X_1 + \cdots)}}$$

The prediction made by the sigmoid function returns a continues value between 0 and 1, therefore a threshold value is selected in which in this case is going to be 0.5. This states that everything that is below 0.5 is going to be a 0 and everything that is bigger is going to be mapped into a 1. Figure 8 shows the graph of the sigmoid function used with the chosen decision boundary.

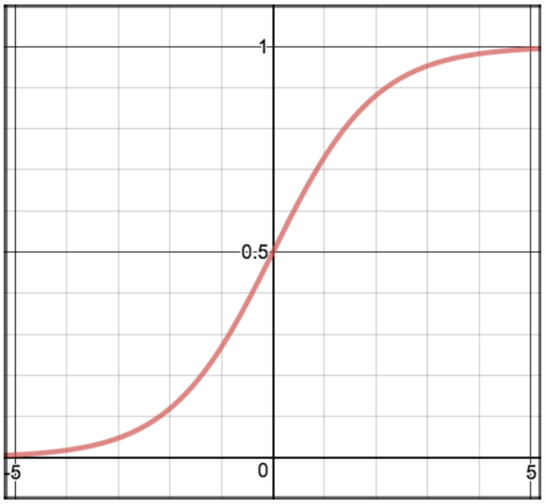


Fig 8. Graph of the sigmoid function with a decision boundary of 0.5

Figure 9 shows the cost function used in linear regression, which is the cross-entropy function, this function measures the performance of the model based on the real dependent value and the calculated from the hypothesis of the model. The cross-entropy function uses logarithms to penalize confident and wrong predictions and reward confident and right predictions.

$$J(\theta) = \frac{1}{m} \sum_{i=1}^{m} \mathrm{Cost}(h_\theta(x^{(i)}), y^{(i)})$$

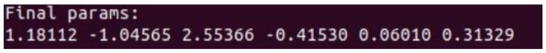
$$\begin{array}{ll} \mathrm{Cost}(h_\theta(x), y) = -\log(h_\theta(x)) & \text{if y = 1} \\ \mathrm{Cost}(h_\theta(x), y) = -\log(1 - h_\theta(x)) & \text{if y = 0} \end{array}$$

Fig 9. Cross entropy function

Therefore, every iteration a copy of the current parameters is made and saved, the parameters are then changed according to the result of minimizing the cost using gradient descent and the hypothesis function, and finally the mean error that the model makes is taken from the cross-entropy function.

V. TEST AND VALIDATION

The results of running the previously presented model [5] using the train dataset outputs the following parameters.



```
Final params:
1.18112 -1.04565 2.55366 -0.41530 0.06010 0.31329
```

The model achieved a mean error of 0.00018 with the training dataset and when using the previous parameters with the test dataset the model achieved a mean error of 0.00032. This means that in general the presented model is learning the general pattern of the data instead of making an overfitting. Fig 10 represents the error of the model using the training dataset over time, this shows that the presented model achieved to reach a local minimum of the function using gradient descent as an optimization algorithm, and that over time it did converged.

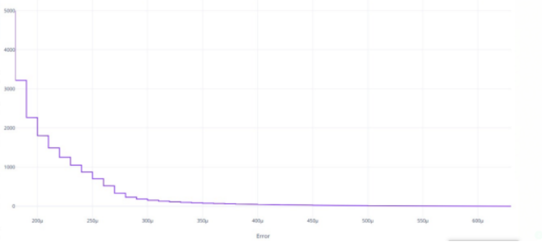


Fig 10. Graph of the mean error vs time

VI. CONCLUSIONS

The presented model performs well with the train and the test datasets, because it is learning the general pattern instead of memorizing the train dataset and trying to make the test data to fit the train data. To improve the model, it is possible to add the family dimension and test the accuracy of the model to see the improvement and performance, other forms improving are changing the way the data is grouped.

VII. REFERENCES

[1]Logistic Regression. (n.d.). Retrieved from https://www.medcalc.org/manual/logistic_reg

ression.php

[2]Shulga D. (2018). 5 Reasons “Logistic Regression” should be the first thing you learn when becoming a Data Scientist. Retrieved from https://towardsdatascience.com/5-reasons-logistic-regression-should-be-the-first-thing-you-learn-when-become-a-data-scientist-fcaae46605c4

[3]Titanic: Machine Learning from Disaster. (n.d.). Retrieved from https://www.kaggle.com/c/titanic

[4]Logistic Regression (n.d.) Retrieved from https://mlcheatsheet.readthedocs.io/en/latest/logistic_regression.html

[5]Perez M. (2019). Logistic Regression. Retrieved from https://github.com/pirty6/Machine-Learning/blob/master/Logistic_Regression/main.java

Sehgal M. (n.d.). Titanic Data Science Retrieved from https://www.kaggle.com/startupsci/titanic-data-science-solutions