2021

Predicting Survivors of Titanic Shipwreck

Heath Thompson
Computer Science
CS477 Spring 2021

I. Introduction (*Heading 1*)

The goal of this project is to predict, as accurately as possible, the survivors of the Titanic Shipwreck on April 15, 1912 given three datasets provided by https://www.kaggle.com/c/titanic/data, a community focused on data science and machine learning. There is a training set used to develop the model, a test set used to evaluate the model, and a third gender submission set used to gain extra information on the ship's passengers. The training set contains data pertaining to each of the ship's passengers while the test set only contains the passenger's outcomes, the target to be predicted.

II. METHOD

The model architecture will compose logistic regression as its source of prediction. This apporach was chosen because of it's known practical use in binary classificaton, the problem type to which our project fits best considering we are to determine one of two possible outcomes for each of the passengers. Logistic Regression uses a sigmoid function to that produces a range of values from zero to one where we can attach labels, zero for a passenger who did not survive and one for a paasenger who did. As seen in Figure 1 below,

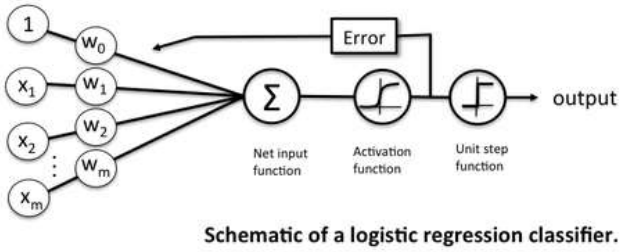


Figure 1: Logistic Regression Model Architecture [1]

parameters are given weights and then passed into the function that then computes the classifcation of either zero or one. These parameters are given to the model by the training set previously discussed. This model, with the help of some tuning, produces accurate predictions based on the parameters fed to it.

III. Experimental Results

The results of this model came slightly under expectations. Overall, the model had a training accuracy of 80.2% and testing accuracy of 80.4%. This information means the model is not overfitted to the training data as the testing accuracy was slightly higher than the training accuracy. The recall of the model was just under 74% while the precision was 75%. Finally, the area under the roc curve as found to be just under 87%. The roc curve is illustrated in Figure 2 below:

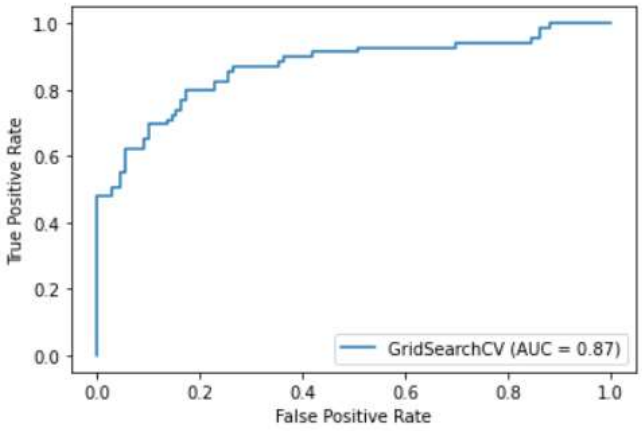
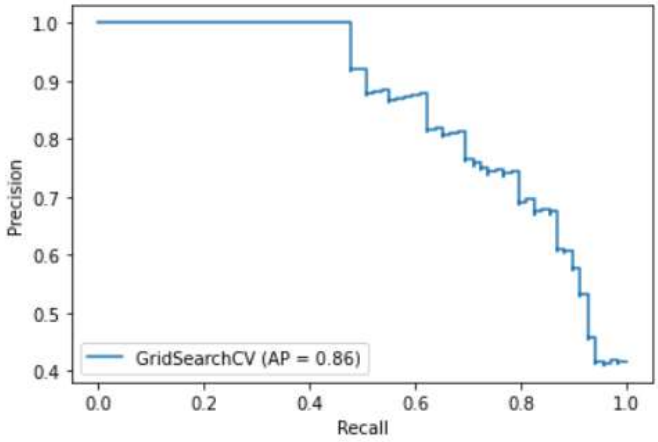


Figure 2: Roc Curve

llustrated in Figure 3 is the precision recall curve.



IV.

Figure 3: Precision Recall Curve

Lastly, we have Figure 4 which illustrated the confusion matrix of the models preodictions:

	Predicted Dead	Predicted Alive
Actually Dead	93	17
Actually Alive	18	51

Figure 4: Model's Confusion Matrix

V. Conclusion

In conclusion, the model developed in this project was able to determine the outcome of eight out of every ten data samples (passengers) correctly. Although this model was not as accurate as expected, it still performed quite well. There were many other tasks that could have been conducted to make the model more accurate. Also, using another type of model like a boosting algorithm or forest classifier may have produced more accurate results.

References

[1] https://sebastianraschka.com/faq/docs/logisticregr-neuralnet.htmlJ. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.