Titanic Cross Validation

*Ankush Laxmeshwar*

## **Predict survival of passengers on the Titanic ship.**

## Methodology:

## The analysis was carried out on *Logistic Regression*. The reasons for this are:

* A statistical method was required which could provide a binomial output and predict survivors and non survivors.
* This dataset has been worked on famously on the internet using Decision Trees and Random Forests in addition to other prediction techniques such as KNN and Naïve Bayes. I wanted to try another technique which wasn’t used much for this dataset and whose output is not as direct and easy to interpret as the other techniques in which everything works more like a black box and has lesser user involvement.
* There is a dire need to first understand the initial statistical analysis methods before moving on to the higher machine learning methods. Even from the industrial requirement standpoint this regression testing knowledge is considered crucial.
* Adequate cleaning has been carried out on the dataset without focus on intensive fine tuning.

## Steps Performed:

1. The *xlsx* package is pulled in to import the dataset in and the working directory for the session is set.
2. The data is imported to the data frame *mydata*. It is checked for data included, data types and NAs using *str* and *summary* functions.
3. *mydata$Name, mydata$Ticket and mydata$Cabin* are coerced to character strings as they don’t need to have levels. They are supposed to be unique values. The NAs in *mydata$Age* are replaced by the mean value of the same column and the NA value in *mydata$Embarked* is replaced by *‘S’* which is the most popular location to board the ship.
4. Set the seed for reproducibility in the case of upcoming sampling.
5. Divide the data into 70% training dataset and 30% testing dataset.
6. Fit the logistic regression model using *glm* function on the training dataset and check the summary.
7. Since some coefficients were statistically insignificant, as in they do not significantly contribute and discriminate in predicting the response, we remove them and once again fit the model for which we check the model summary.
8. Now that we have a model with significant predictors, we use the *predict* function with the fitted model and testing dataset.
9. We check the summary of the predicted vector of values and hence see the range which falls between 0 and 1 as we want it.
10. Since we have probabilities and not concrete 0 and 1 values, we assign a cut off which bifurcates the results in 1 and 0. This bifurcation is set to 80% (the reason will be discussed in the result section).
11. A confusion matrix is created to compare the actual and predicted values.
12. Using the confusion matrix, the accuracy of the model is assessed.
13. The ROC curve is plotted and the Area under the Curve is assessed.
14. The PassengerId, Actual Values and Predicted Values are put into data frame and this data frame is written to a CSV file in the current working directory.

## Result and Interpretation:

Following the above steps, we get a vector output with probability of success for each set of inputs.

* The probability needs to be bifurcated into 1 and 0 to indicate survival and death of the Titanic passenger. I have chosen 80% percentile as the separation from trial and error. The probability having more than 0.8 will be set as 1 and 0 otherwise. I have mainly chosen a higher tolerance for 0 and lesser for 1 since most of the passengers died, as per earlier analysis.
* As per this predicted result based on the model fitted and then assigned to 1s and 0s, we tabulate the predicted values against the actual values of survival and get the below table. We can see that only 4 Type 1 errors are made where we predict the survival of the passenger but actually there was no survival. This value is very less and is a tolerable amount. The Type 2 error accounts for 50 cases where we predicted the passenger dies but is actually alive. Again, we should keep it in mind that it’s better to incorrectly predict death of passenger than incorrectly predict that the passenger is alive when in fact he/she died.

**predicted**

**actual 0 1**

**0 161 4**

**1 50 53**

* The above table is evaluated for accuracy by adding the sensitivity and specificity and diving by the total number of cases. For this model we get an accuracy of **79.85%** which is within acceptable standards.
* Going further, we study the ROC curve for the result by comparing the predicted and actual survival values. For this we get the chart as below. For the Area under the curve (AUC) we see that the model tends to the upper left area and this showcases the independent variables have good discriminatory ability and this can justify the 79.85% accuracy the model gave us in the previous point.

