**Titanic Shipwreck**

**Project Description –**

**Goal**

It is your job to predict if a passenger survived the sinking of the Titanic or not.  
For each in the test set, you must predict a 0 or 1 value for the variable.

**Metric**

Your score is the percentage of passengers you correctly predict. This is known as [accuracy](https://en.wikipedia.org/wiki/Accuracy_and_precision#In_binary_classification).

**Submission File Format**

You should submit a csv file with exactly 418 entries plus a header row. Your submission will show an error if you have extra columns (beyond PassengerId and Survived) or rows.  
  
The file should have exactly 2 columns:

* PassengerId (sorted in any order)
* Survived (contains your binary predictions: 1 for survived, 0 for deceased)

PassengerId,Survived  
892,0  
893,1  
894,0

**Case Description Highlights**

1. On April 15, 1912, during her maiden voyage, the Titanic sank after colliding with an iceberg, killing 1502 out of 2224 passengers and crew. Translated 32% survival rate.
2. One of the reasons that the shipwreck led to such loss of life was that there were not enough lifeboats for the passengers and crew.
3. Although there was some element of luck involved in surviving the sinking, some groups of people were more likely to survive than others, such as women, children, and the upper-class.

**Approach –**

**There are more than 60 predictive modelling algorithms to choose from. The choice of an algorithm is based on factors such as the dataset, results required and performance. These factors help us narrow down to a few models to be implemented. The titanic challenge is a classification and regression type problem, since there is a limited data set which once cleaned, needs to provide a predictive output primarily whether a subject survived the shipwreck or not. The relationship between the output (Survived or dead) is made with the given variables such as age, gender, embarkment port and also the class of ticket they hold. A secondary sect of factors such as number of family members, cabin number etc. are also taken into account. We perform supervised learning since our model is trained with the help of the given data set.**

**Our machine learning model requirement therefore becomes – supervised learning, classification and regression model. Some of these are:**

1. **Logistic Regression**
2. **KNN or k-Nearest Neighbors**
3. **Support Vector Machines**
4. **Naïve Bayes Classifier**
5. **Decision Trees**
6. **Random Forest**

**This report presents the use of logistic regression to provide predicative analysis.**

**Factors Considered –**

**The presented data consisted of different classes i.e., and can be broadly classified as categorical and numerical data. Categorical data consist of factors like survived, sex, embarked pclass etc. while the numerical data consisted age fare, sibsp etc. Cabin number is a unique mixed data type, with a alphanumeric combination.**

**It is difficult to review a large dataset to check for typos and errors. Some of these always require some form of correction. For example. Name feature may contain errors or typos as there are different ways of presenting a name, and its title. Other cases like age and cabin number may miss some integer in their value due to an error. Therefore, such cases must be corrected since we know the actual count of passengers on the ship which was 2224.**

**The data analysis involves completing the data, correcting, creating, and classifying before we can pick the pivoting features.**

To confirm some of our observations and assumptions, we can quickly analyze our feature correlations by pivoting features against each other. We can only do so at this stage for features which do not have any empty values. It also makes sense doing so only for features which are categorical (Sex), ordinal (Pclass) or discrete (SibSp, Parch) type.

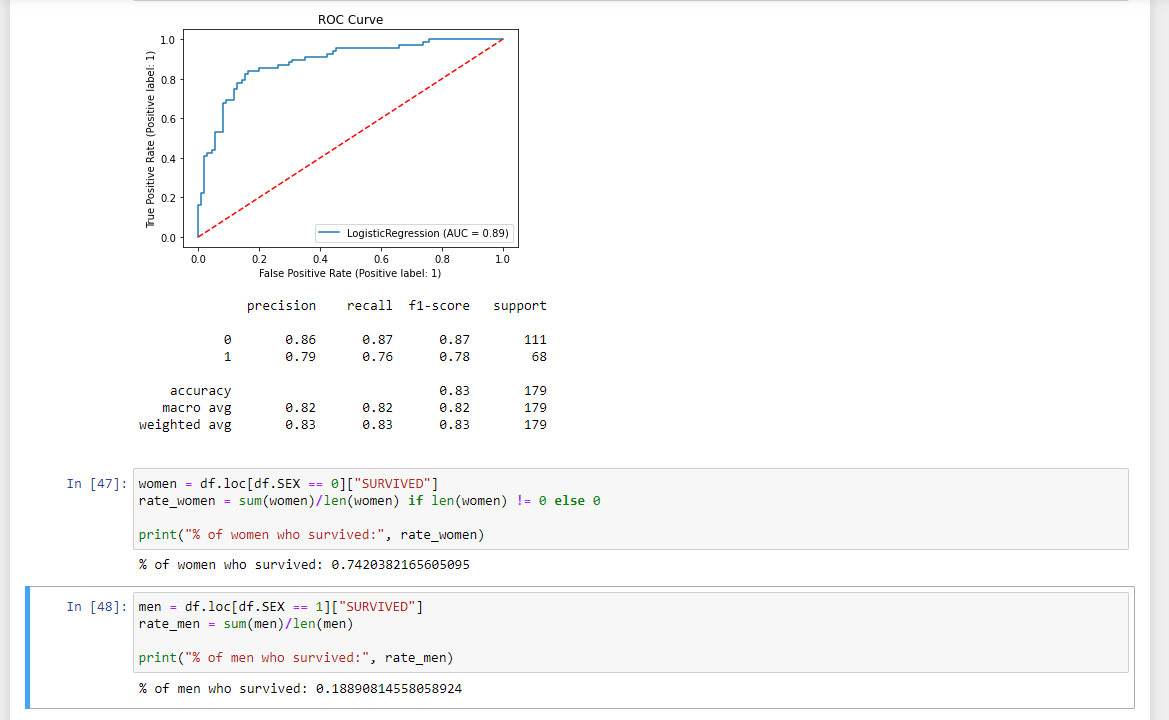
* **Pclass** We observe significant correlation (>0.5) among Pclass=1 and Survived (classifying #3). We decide to include this feature in our model.
* **Sex** We confirm the observation during problem definition that Sex=female had very high survival rate at 74% (classifying #1).
* **SibSp and Parch** These features have zero correlation for certain values. It may be best to derive a feature or a set of features from these individual features (creating #1).

**Flow and Steps –**

**In this section we shall see how the model was developed and implemented. The first aspect involves preparation of the dataset, which is first read and refined. Based on the data dictionary, the data frame in which the csv sheets were added are viewed and sorted to perform exploratory data analysis. Unique features that impact the predicative analysis are identified here. This includes, sex, cabin number and port of embarkment primarily. The data frame is reorganized on the basis of these factors and arranged in the form of decreasing priority of features that impact a passenger’s survival.**

**Once this is done, we then train our machine learning model with the help of scikit learn and initiate logistic regression on our filtered dataset. Here we specify our batch size. A confusion matrix is generated on the basis of the probabilities and simultaneously accuracy score is generated. Other scores like precision, recall and f1 can be visualized and presented from the confusion matrix itself. We then plot our roc curve to see if our model is consistent and also to obtain the scores for each of the categories in the confusion matrix. Finally, a percentage score of the survivors presented on the basis of the gender, which is the primary distinguishing factor for the passengers.**

**Results –**



**Conclusion –**

**There are many things for a greater chance to survive. Being a female or a child will increase you chances. If you have a higher-class ticket, you have the more chance of surviving than a third class ticket. As for a man, you are more likely to survive if embark in Cherbourg compare to Southampton or Queenstown. If you also travel with 1 or 3 people than 0 or more than 3 your survival chances are greater. The younger you are will also make your survival chance. So it comes down to many things to survive on the titanic.**