Overview

The sinking of the RMS Titanic is one of the most infamous shipwrecks in history.  On April 15, 1912, during her maiden voyage, the Titanic sank after colliding with an iceberg, killing 1502 out of 2224 passengers and crew. This sensational tragedy shocked the international community and led to better safety regulations for ships.

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. 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.

In this challenge, we ask you to complete the analysis of what sorts of people were likely to survive. In particular, we ask you to apply the tools of machine learning to predict which passengers survived the tragedy.

The data has been split into two groups:

* training set (train.csv)
* test set (test.csv)

The training set should be used to build your machine learning models. For the training set, we provide the outcome (also known as the “ground truth”) for each passenger. Your model will be based on “features” like passengers’ gender and class. You can also use [feature engineering](https://triangleinequality.wordpress.com/2013/09/08/basic-feature-engineering-with-the-titanic-data/)to create new features.

The test set should be used to see how well your model performs on unseen data. For the test set, we do not provide the ground truth for each passenger. It is your job to predict these outcomes. For each passenger in the test set, use the model you trained to predict whether or not they survived the sinking of the Titanic.

We also include gender\_submission.csv, a set of predictions that assume all and only female passengers survive, as an example of what a submission file should look like.

|  |  |  |
| --- | --- | --- |
| Variable | Definition | Key |
| survival | Survival | 0 = No, 1 = Yes |
| pclass | Ticket class | 1 = 1st, 2 = 2nd, 3 = 3rd |
| sex | Sex |  |
| Age | Age in years |  |
| sibsp | # of siblings / spouses aboard the Titanic |  |
| parch | # of parents / children aboard the Titanic |  |
| ticket | Ticket number |  |
| fare | Passenger fare |  |
| cabin | Cabin number |  |
| embarked | Port of Embarkation |  |

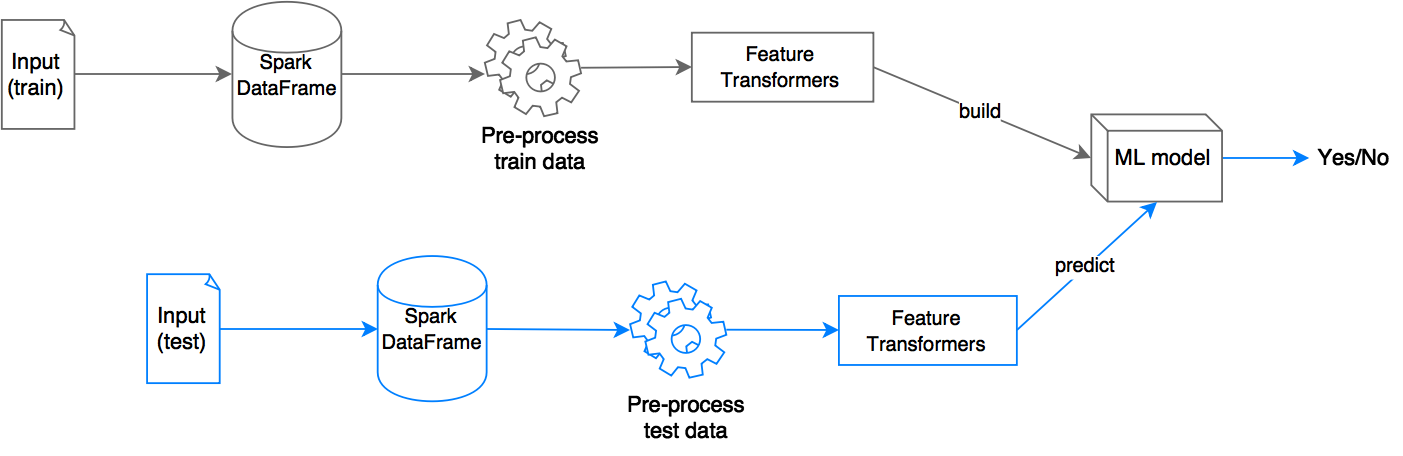
### Variable Notes

pclass: A proxy for socio-economic status (SES)  
1st = Upper  
2nd = Middle  
3rd = Lower  
  
age: Age is fractional if less than 1. If the age is estimated, is it in the form of xx.5  
  
sibsp: The dataset defines family relations in this way...  
Sibling = brother, sister, stepbrother, stepsister  
Spouse = husband, wife (mistresses and fiancés were ignored)  
  
parch: The dataset defines family relations in this way...  
Parent = mother, father  
Child = daughter, son, stepdaughter, stepson  
Some children travelled only with a nanny, therefore parch=0 for them.

In this challenge, we are given a set of details of passengers such as name, gender, fare, cabin etc and if the person survived the Titanic disaster. Based on this we have to build a Model that can predict, given another passenger, if he/she is likely to survive. This is an example of binary classification where there are only two possible classes(1 if passenger survives and 0 if not).

* The first step when trying to build a machine learning model is to analyze and understand the data you have. So that you can decide which all features has to be used for building the model, whether the features are numeric or categorical, what is the mean,max or min of your numerical features and so on.
* Once the data is analyzed, next step is feature selection where we decide which all features are relevant for building the model
* Next is data preprocessing. The input data that you receive for modeling is not going to be good data most of the times. During this stage, for example, we can decide on what to do with the missing values - whether to drop rows having nulls, or fill those with average value of the feature(if feature is numerical), or fill with most occurring value of the feature(if feature is categorical) etc.
* Next comes the Feature engineering and Feature transformation step. In Feature engineering we derive new features from existing ones and during feature transformation we transform existing features so that it can be used for building the model.
* Finally we build the model using the selected features and do prediction on a new set of data.

We will be implementing all of the above steps using Spark and Scala and will be building a machine learning pipeline - the overall flow can be shown by the diagram below. The grey section of the diagram shows the model building flow and the blue section of the diagram shows the flow for making prediction.



### ****Load and Analyze data****

As mentioned earlier, first step is to analyze the data. To do that, we have to first load data into Spark. Download the train.csv file from [here](https://www.kaggle.com/c/titanic/data), and open the file and check the content

As you can see, the file contains a header row which has PassengerId, Survived, Pclass, Name, Sex, Age, SibSp ,Parch ,Ticket ,Fare ,Cabin and Embarked. You can find more information about what each of these fields are from the Kaggle website. Move this file to some folder in HDFS(I have kept mine at /kaggle/titanic/train.csv). The data is in csv format, to load csv files we will use the library [spark-csv](https://github.com/databricks/spark-csv).