Deliverables

**Databricks Notebook**: Nebaditn\_HW4

**URL**: <https://dbc-0fffd743-9b36.cloud.databricks.com/?o=2801300804394420#notebook/1523402910002523/command/2013421380133393>

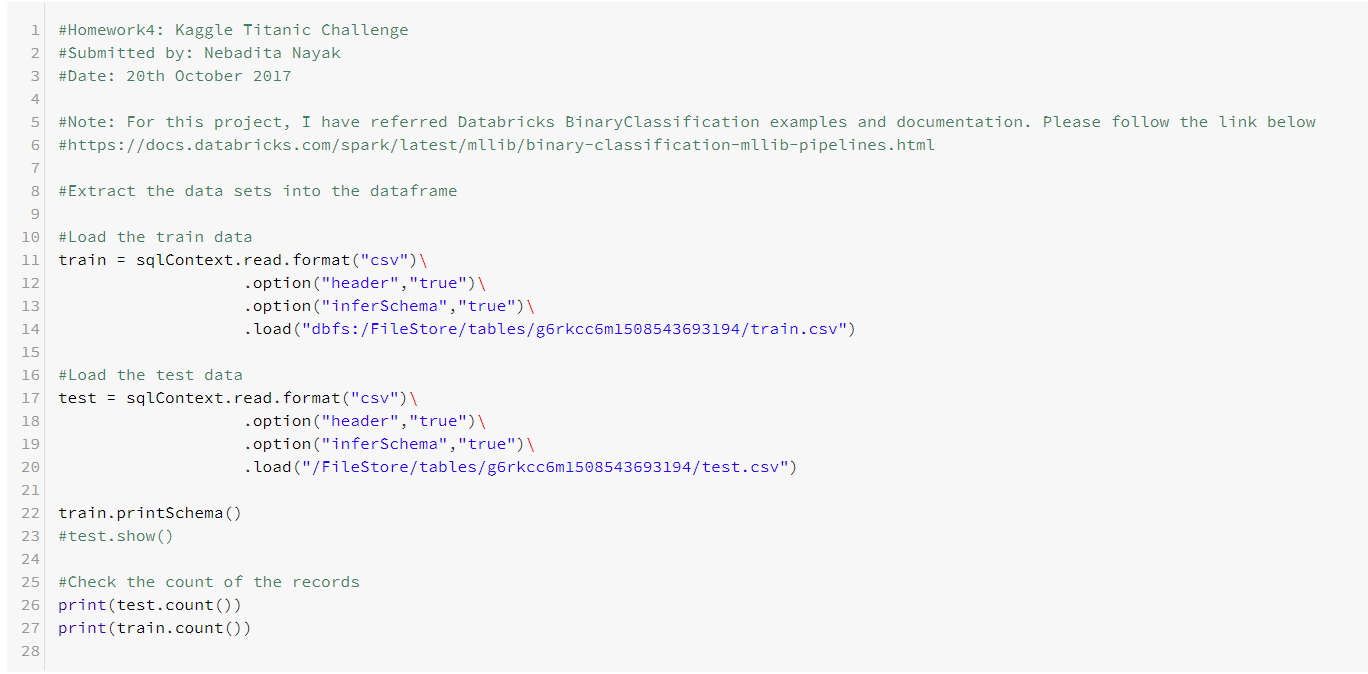
**Problem:** In this challenge, we have been given certain features or parameter of a titanic passenger like Age, Sex, Class of travel, Cost of ticket, Number of family members and we are expected to find the ground truth: Whether a passenger on RMS Titanic will survive or not with some probability.

**Approach**: The problem is categorized as a Binary Classification problem, since we only have two outcomes (Surviced:1 or Not Survived:0).

**Steps to solve the problem:**

*Step 1*: The first step in our model preparation is uploading the data and performing data cleaning.

We will first upload the data in the respective data frames



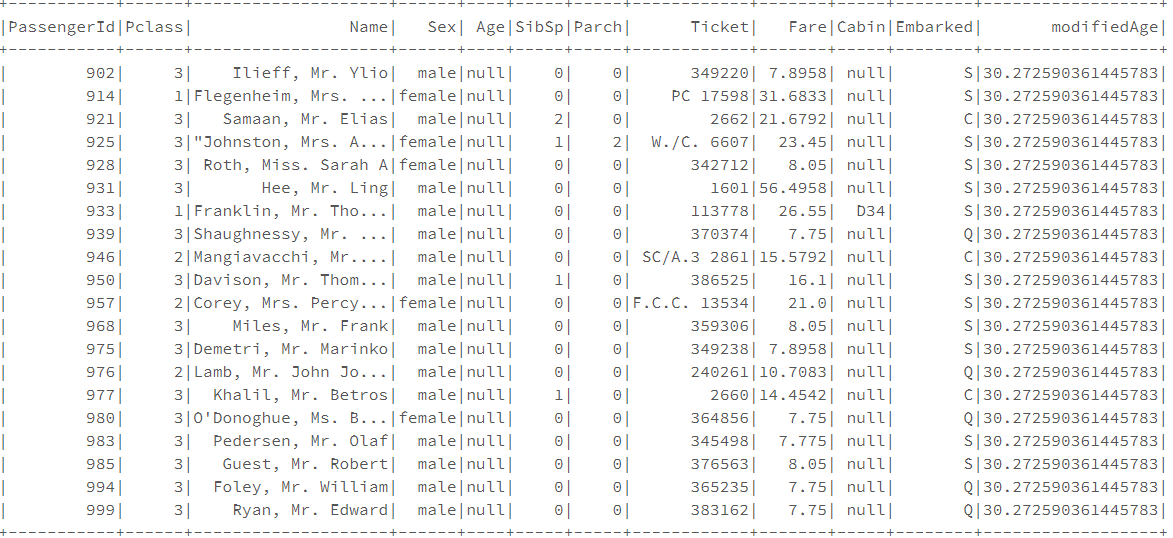
Once, we have uploaded the data, we will explore the columns with the null values. We explored that there following Null values in test and train dataset

1. Train: Age, Embarked
2. Test: Age, Fare

We will then plan to replace this null value using the mean value. We are planning to use a feature transformer Imputer. This transformer will add a new column with the null value replaced by the mean value of the column.

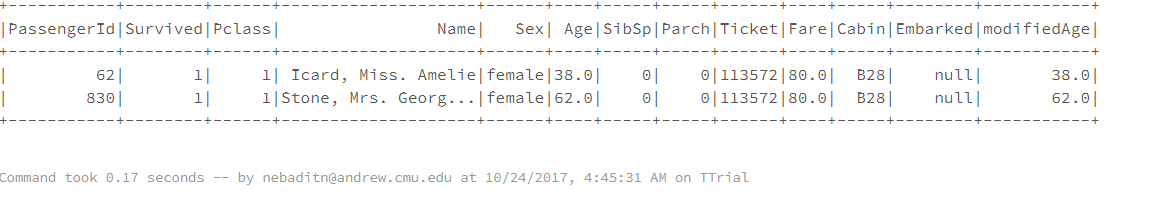


We will add a column Modified Age to our table, which will have the average age of 30.275



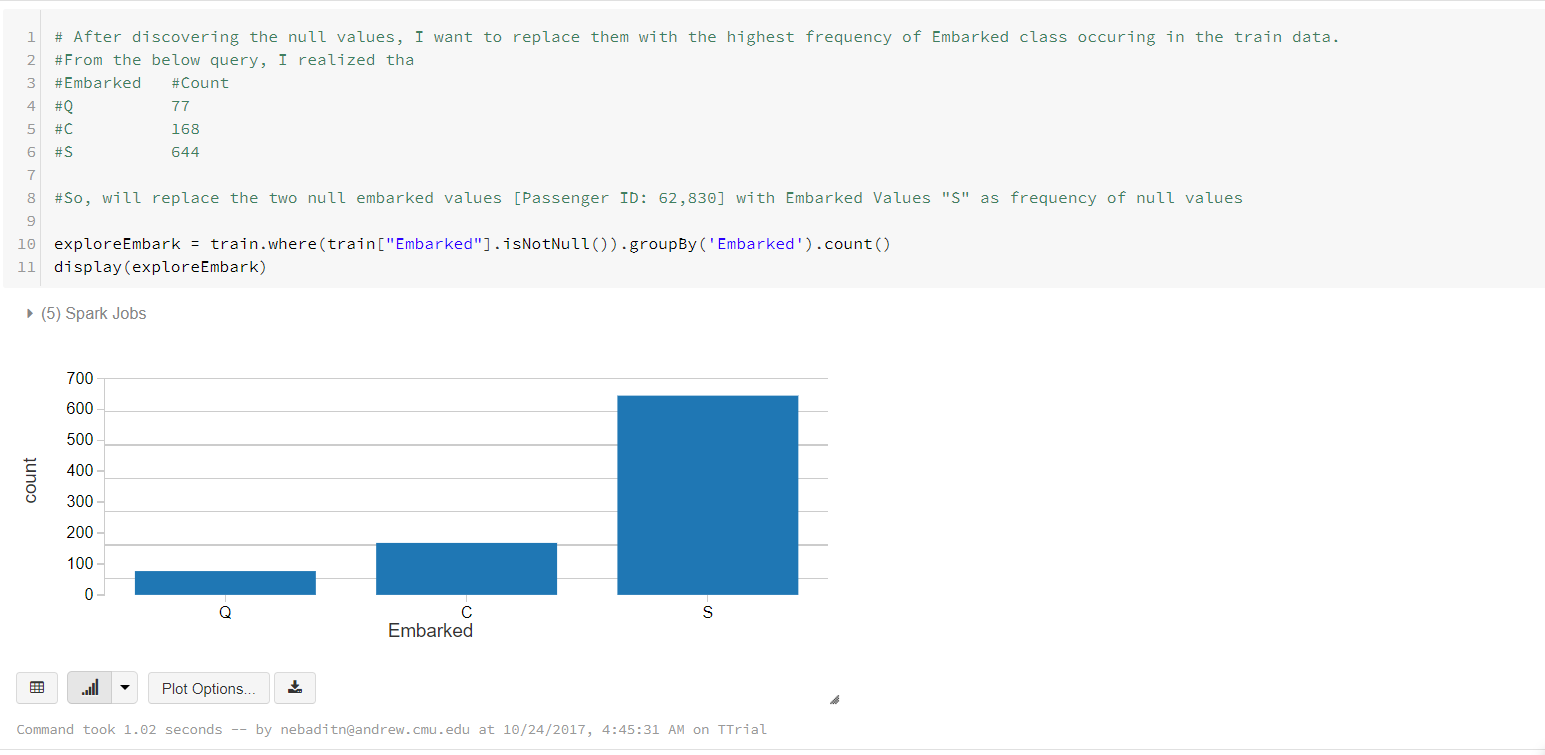
Then we will modify the embarked column to fill the missing values:

There are two records with the values of the embarked column missing.



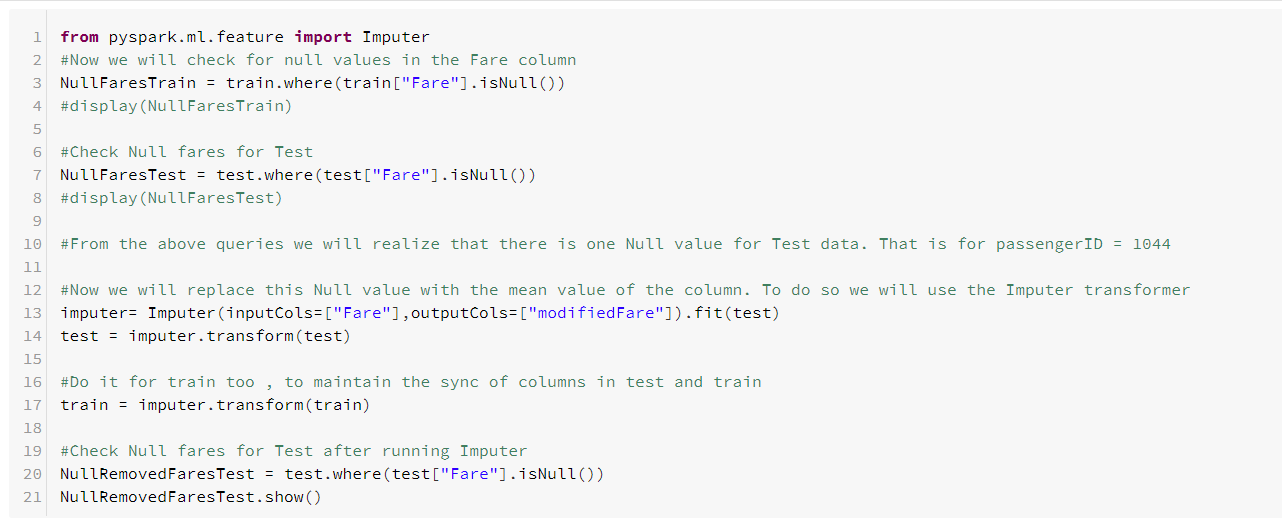
Here will replace the null values with highest frequency of the Embarked category.

From our analysis, we found, that the highest frequency of Embarked category is ‘S’ =644



We will now replace the missing values with the embarked values as ‘S’

Next, we observed that we have null values in the Fare column in the test data. We replace the missing values again with the mean values using the imputer transformation

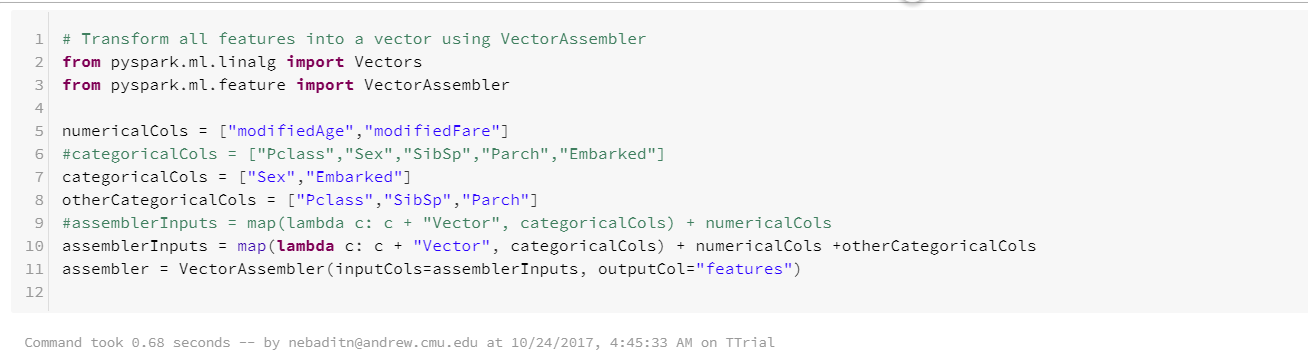


*Step 2*: Feature Engineering

In this step, we will convert the categorical data like Male, Female into numerical indexes. We first use String Indexer to transform String categorical values into numerical values. Then we will use OneHotEncoding to convert them into Binary vectors with at most one non zero value. Here one thing we have observed, we cannot use OneHotEncoding for data sets whose number of categories varies in the test and in the train data set. Hence with categories like Pclass, SibSp we cannot use OneHotEncoding. As the number of categories differ in the test and the train data set.

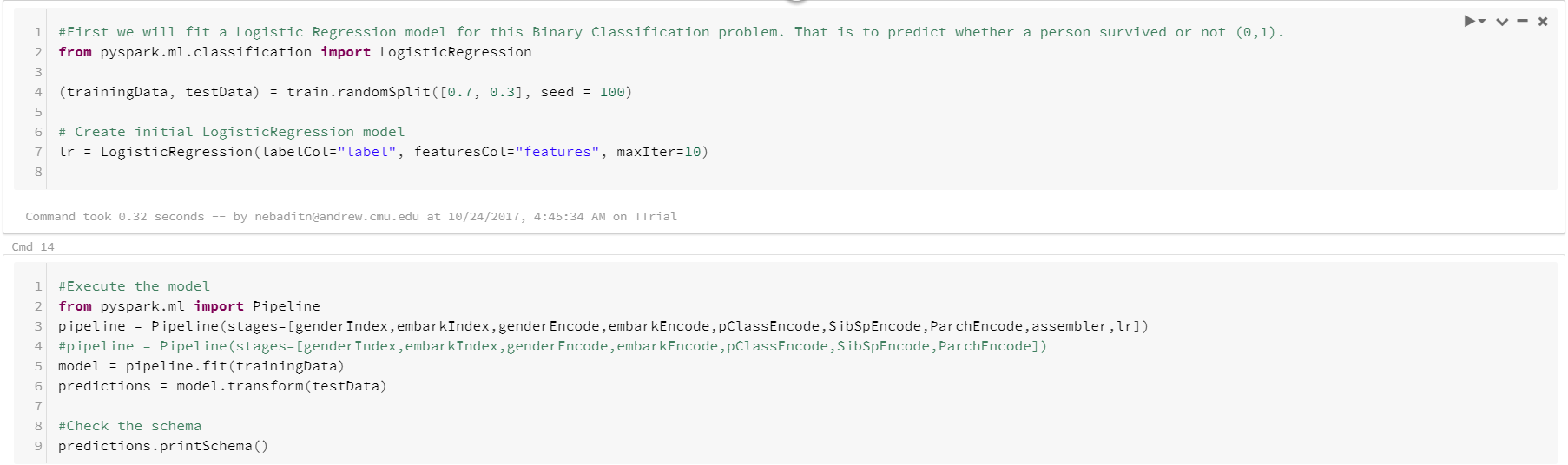


Now, we will assemble all the features in a single vector to produce a feature column against our label column – “Survived” using a Feature Transformer Vector Assembler



Step 3: Create a pipeline

Now, that we are done with the Data cleaning and Feature transformation step. We will create a pipeline to implement different estimators. Here we have used LogisticRegression and RandomForest Classifier

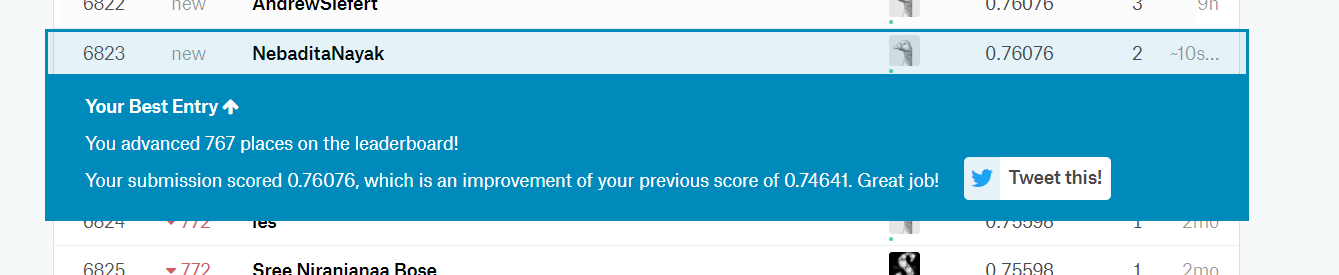


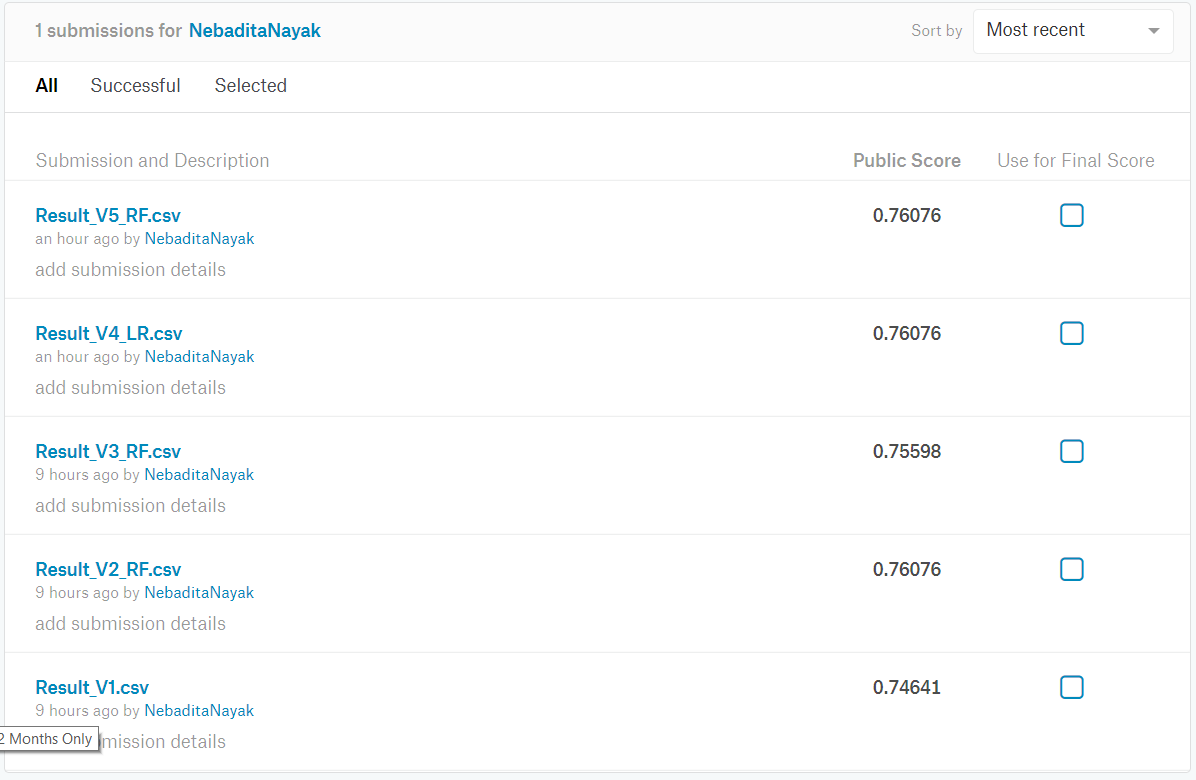
We will use Cross Validation to find the best model and make predictions on the test data set using that best model.

Step 4: In the end we will use the predictions of our model to extract survival prediction for all the passengers in the test data set and upload the result to Kaggle

Result:







We have performed multiple iteration using the Logistic Regression and Random Forest Classifier to find the best prediction model.

References: <https://docs.databricks.com/spark/latest/mllib/binary-classification-mllib-pipelines.html>

Classmates Notebook seen : srossg\_hw4

Python Source Code



Notebook

