Shaunak Joshi

1001417919

Big Data Assignment 2

1. Dataset used – Titanic
2. Spark ML algorithms/ models - Logistic Regression, Decision Tree Classifier, Random Forest Classifier, and Naive Bayes Classifier.
3. I have done this project/ assignment in IBM Data Science Experience. Initially, the model is trained over the trainingData and then tested on the testData.

4) The dataset is split in this fashion: 0.7, 0.3.

* 0.7 – trainingData
* 0.3 - testData

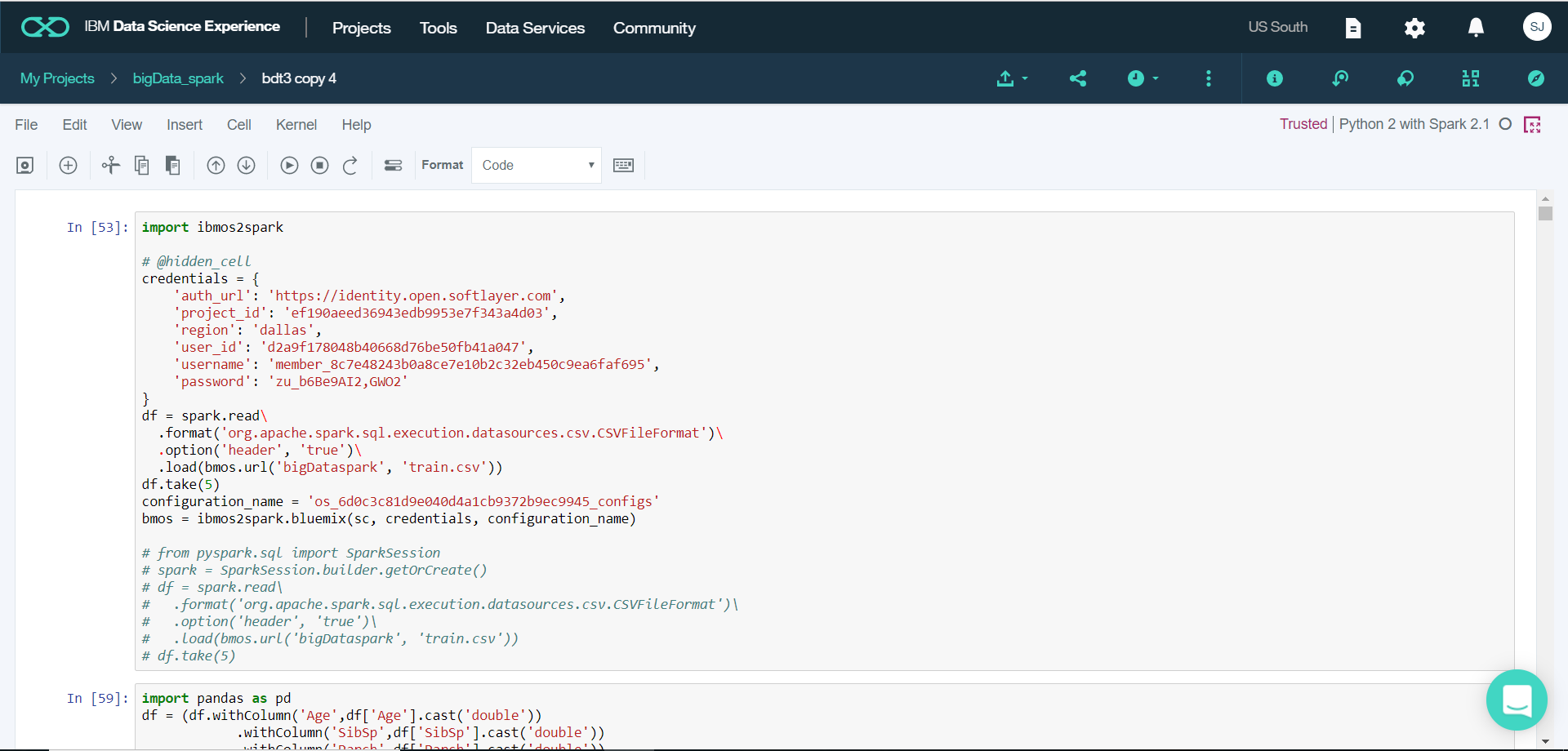
5) Below is the list of algorithms used with its prediction result along with the value after cross validation.

* Logistic Regression - 0.8261634199134199 - 0.8411525974025977
* Decision Tree Classifier - 0.8105248917748916 - 0.8286796536796538
* Random Forest Classifier - 0.8195616883116882 - 0.8270833333333332
* Naive Bayes - 0.46298701298701317

6) References:

* <https://6chaoran.wordpress.com/2016/08/13/__trashed/>
* <http://localhost:8888/notebooks/Downloads/MS%20in%20CS/Fall%202017/Big%20Data%20Analytics/ipython%20Notebook%20on%20Spark/ML_Example_Final.ipynb>

1. The first part of the code was not published on GitHub due to credentials. Hence, I am pasting the screenshot of the same.



Code:

import ibmos2spark

# @hidden\_cell

credentials = {

'auth\_url': 'https://identity.open.softlayer.com',

'project\_id': 'ef190aeed36943edb9953e7f343a4d03',

'region': 'dallas',

'user\_id': 'd2a9f178048b40668d76be50fb41a047',

'username': 'member\_8c7e48243b0a8ce7e10b2c32eb450c9ea6faf695',

'password': 'zu\_b6Be9AI2,GWO2'

}

df = spark.read\

.format('org.apache.spark.sql.execution.datasources.csv.CSVFileFormat')\

.option('header', 'true')\

.load(bmos.url('bigDataspark', 'train.csv'))

df.take(5)

configuration\_name = 'os\_6d0c3c81d9e040d4a1cb9372b9ec9945\_configs'

bmos = ibmos2spark.bluemix(sc, credentials, configuration\_name)