Users with high hit ratio make more purchases

**Attribute selection**

PYSPARK\_DRIVER\_PYTHON=ipython pyspark

import pandas as pd

from pyspark.mllib.clustering import KMeans, KMeansModel

from numpy import array

buyclicksDF = pd.read\_csv('./buy-clicks.csv')

buyclicksDF = buyclicksDF.rename(columns=lambda x: x.strip()) #remove whitespaces from headers

gameclicksDF = pd.read\_csv('./game-clicks.csv')

gameclicksDF = gameclicksDF.rename(columns=lambda x: x.strip()) #remove whitespaces from headers

teamDF = pd.read\_csv('./team.csv')

teamDF = teamDF.rename(columns=lambda x: x.strip()) #remove whitespaces from headers

teamassgnDF = pd.read\_csv('./team-assignments.csv')

teamassgnDF = teamassgnDF.rename(columns=lambda x: x.strip()) #remove whitespaces from headers

userbuy = buyclicksDF[['userId','price']] #select only userid and price

userclicks = gameclicksDF[[‘userId’,’isHit’]]

teamstrength = teamDF[[‘teamId’,’strength’]]

teamassgn = teamassgnDF[[‘team’,’userId’]]

teamassgn = teamassgn.withColumnRenamed(‘team’,’teamId’)

**Training Dataset creation**

totuserbuy = userbuy.groupby('userId').sum()

totuserbuy = totuserbuy.reset\_index()

totuserbuy.columns = [‘userId’,’totalpurchase’]

useravghit = userclicks.groupby(‘userId’).avg()

useravghit = useravghit.reset\_index()

useravghit.columns = [‘userId’,’avgishit’]

userstrength = teamstrength.merge(teamassgn, on=’userId’)