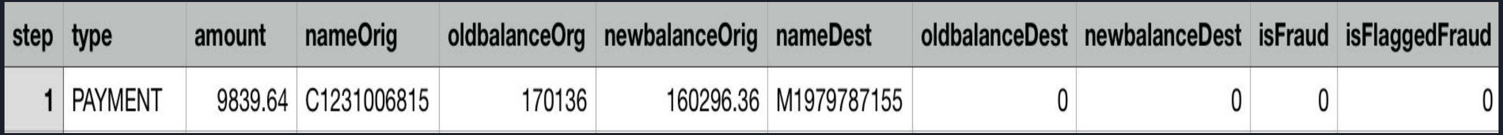
CS6240 WorthlessWithoutCoffee

- Sree Siva Sandeep Palaparthi
- Nikhila Raya
- Velugoti Sai Bhargavi

PaySim Fraud Detection

- Classification and prediction of fraudulent transactions
- Dataset used - PaySim
- Number of records - 6 Million
- Number of fraudulent transactions - 8647
- Training data - 70% Testing data - 30% (both including fraudulent and non-fraudulent records)



step	type	amount	nameOrig	oldbalanceOrg	newbalanceOrig	nameDest	oldbalanceDest	newbalanceDest	isFraud	isFlaggedFraud
1	PAYMENT	9839.64	C1231006815	170136	160296.36	M1979787155	0	0	0	0

Classification and prediction in Spark MLlib

Approaches used

- Ensemble of Random Forest Model
- Ensemble of Gradient Boosted Trees
- Decision Trees

Random Forest Parameters

- Depth - 5
  Number of trees - 2
  Accuracy - 94.9547762376862 %

- Depth - 10
  Number of trees - 5
  Accuracy - 99.48214461720489 %

- Depth - 15
  Number of trees - 5
  Accuracy - 99.6143491049264 %

Increasing the depth and the number of trees , does not improve the accuracy significantly.
After training with 500 trees, the accuracy did not change much.

Gradient Boosted

- Depth - 10
  Number of iterations - 10
  Accuracy - 99.96398497838576 %

Decision Tree

- Depth - 10
  Max bins- 10
  Accuracy - 99.95335845311294 %

Increasing the number of trees or the number of bins, does not improve the accuracy.

Results

Random Forest - 99.6143491049264 %
Gradient Boosted - 99.96398497838576 %
Decision Tree - 99.95335845311294 %

All the models gave good accuracy of 99.6%.

However, we were unable to train models for higher parameters since it was taking too long to run on AWS
number of Iterations - 100, number of trees - 100, depth - 30

K-Nearest Neighbor

Approach:

- Partition + Broadcast

Metrics

- Accuracy - 88%
- F1 measure = 2*(Precision*Recall)/(Precision+Recall)
- Data shuffle
  - From Mappers to Reducers - |TestData|*|TrainData|
  - Reducers to HDFS - |TestData|

K-Nearest Neighbor

Approach:

- Partition + Broadcast with Top-K
- Block Partition

Metrics

- Accuracy - 88%
- F1 measure = 2*(Precision*Recall)/(Precision+Recall)
- Data Shuffle
    - From Mappers to Reducers - |TestData|*k*numberOfMappers
    - From Reducers to HDFS - |TestData|

Challenges

- Data transfer from Mapper to Reducer
- Heap Memory
- Larger Number of actual negatives

Thank You