Large Scale NYX Taxi Data Analysis

Yue Wen

Step1:

Download the data from:
https://s3.amazonaws.com/nyc-tlc/trip+data/yellow_tripdata_2016-01.csv

We use 'wget' to download the data and store it into the cluster.

```
wget https://s3.amazonaws.com/nyc-tlc/trip+data/yellow_tripdata_2016-01.csv
gcloud compute copy-files yellow_tripdata_2016-01.csv  cluster-1-m:~/
hadoop fs -put yellow_tripdata_2016-01.csv
```

Step2:

Construct a bounding box for New York City (http://boundingbox.klokantech.com/ and remove any data points that have coordinates outside this box. Use the box: westlimit=-74.2635; southlimit=40.4856; eastlimit=-73.7526; northlimit=40.9596

We simply used a filter function to realize the purpose.

Step3:

Divide the drop off and pickup zones into bins.

Here, we use a trick to create the bucket.

Firstly, perform a linear projection such that westlimit=-74.2635 is mapped to 0 and southlimit=40.4856 is mapped to 20, then all the datapoints' longitude falls between interval (0,20),then we ceil the double to integer. Then all the data points' longitude falls into set{0,1,2,3,..19}.

Similarly, we can perform another linear transformation to map all the data points' latitude to interger{0,1,2,3,4,5,6,7,8,9}.

Then we use the map(longitude)*10+(latitude) to get the 200 bucket from 0 to 199.

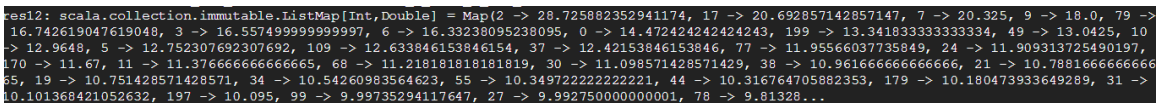
The overall formula is displayed as following:

```
((p.long.toDouble*39.1466+2907.16353).toInt)*10 + (p.lai.toDouble*21.09705-854.1267).toInt
```

Then we apply this function to both pick up coordinate and drop off coordinate to get the zone.

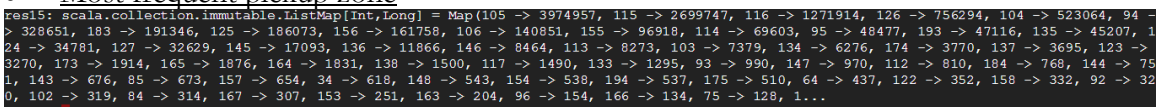
Step4: Data analysis

- Top five drop off zones ordered by high average tip.



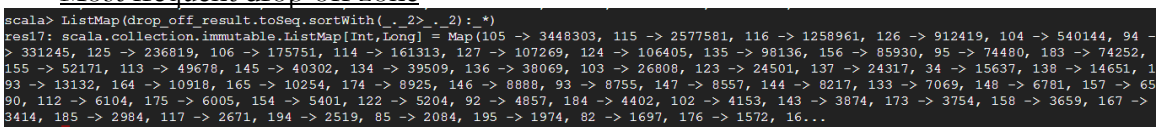
Then the top five zone is:2, 17, 7,9,79

- Most frequent pickup zone



105,115,116,126,104

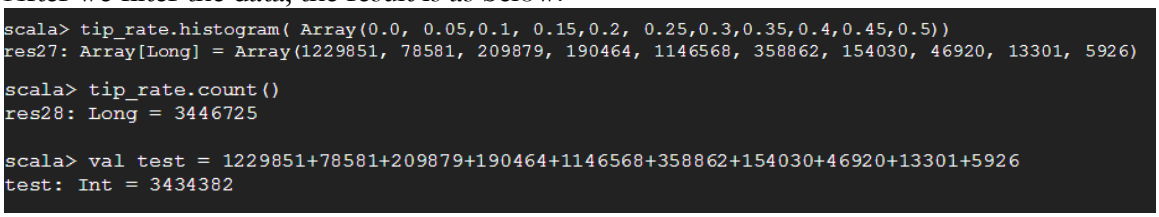
- Most frequent drop-off zone



105,115,116,126,104

- Histogram

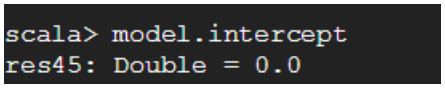
The most frequent one is 105
After we filter the data, the result is as below.



Note here, we only check the tip rate arrange from 0 to 0.5, and the last two results shows that we do not lose too much data points by doing so.

Step5: Logistic Regression

Intercept term: 0.0



The coefficient of (pick up zone, drop off zone, trip distance, passenger count and payment type) is ([0.06369797187134343,

0.053406770692665737,

-1.3580632628969216E-8,

0.08314692711037643,

-9.952407913717822])

```
scala> model.weights
res46: org.apache.spark.mllib.linalg.Vector = [0.0636979718713434 3,0.053406770692665737,-1.3580632628969216E-8,0.0831469271103764 3,-9.952407
913717822]
```

We might also be interested in the other two performance measures.

See the results below.

```
scala> precision.collectAsMap.maxBy(_._2)
res50: (Double, Double) = (0.9139650697297554,0.9652097885856893)
```

```
scala> val auROC = metrics.areaUnderROC
auROC: Double = 0.9661462061881627
```