Submitted by Suresh Gourigolla

Risk, Fraud, Threat Detection

**INTRODUCTION**

According to Forbes, Fraud losses incurred by banks and merchants on all credit, debit, and pre-paid general purpose and private label payment cards issued globally hit $21.84 billion (bn) in 2015, with the United States (US) accounting for almost two-fifths (38.7%) of the total at $8.45bn. But by 2020 it could surpass $12bn, were the global percentage growth rate of 45% to be mirrored.

**Fraud Detection with Analysis Process**

* Ingest the Data
* Explore the Data
* Futurize Data with SQL
* Train a Model
* Evaluate the Model

**Finding the credit card fraud with 20.9 GB data from Azure Container,**

* we have a series of credit card transactions those credit card transactions have been labelled as fraud or not fraud.
* This data was anonymized and distributed for a fraud detection student competition by for Capital One Financial Corporation.
* Capital One Financial Corporation has authorized Databricks to use of this data.

**STEP 1:**

* Check the size

val fileSize = dbutils.fs.ls("filepath"). map (\_. size). sum / (Math. Pow (1024.0,3))

println(f"Size of file in GB: $fileSize%2.2f") : 20.9 GB

* Register table (auth\_Data) from csv

Read CSV file and write just ¼ amount of total size, delimiter type is PIPELINE from the container and format is parquet while saving as auth\_data table

**STEP 2: Explore Credit card Data**

Print the schema, display the first 1,000 rows, use describe to get some statistics

Run SQL Query on Table or describe statistics of table

display(table("auth\_data").describe())

Examine of source data,

Header Columns:

AUTH\_ID ACCT\_ID\_TOKEN FRD\_IND ACCT\_AVL\_CASH\_BEFORE\_AMT ACCT\_AVL\_MONEY\_BEFORE\_AMT ACCT\_CL\_AMT ACCT\_CURR\_BAL ACCT\_MULTICARD\_IND ACCT\_PROD\_CD ACCT\_TYPE\_CD ADR\_VFCN\_FRMT\_CD ADR\_VFCN\_RESPNS\_CD APPRD\_AUTHZN\_CNT APPRD\_CASH\_AUTHZN\_CNT ARQC\_RSLT\_CD AUTHZN\_ACCT\_STAT\_CD AUTHZN\_AMT AUTHZN\_CATG\_CD AUTHZN\_CHAR\_CD AUTHZN\_OPSET\_ID AUTHZN\_ORIG\_SRC\_ID AUTHZN\_OUTSTD\_AMT AUTHZN\_OUTSTD\_CASH\_AMT AUTHZN\_RQST\_PROC\_CD AUTHZN\_RQST\_PROC\_TM AUTHZN\_RQST\_TYPE\_CD AUTHZN\_TRMNL\_PIN\_CAPBLT\_NUM AVG\_DLY\_AUTHZN\_AMT CARD\_VFCN\_2\_RESPNS\_CD CARD\_VFCN\_2\_VLDTN\_DUR CARD\_VFCN\_MSMT\_REAS\_CD CARD\_VFCN\_PRESNC\_CD CARD\_VFCN\_RESPNS\_CD CARD\_VFCN2\_VLDTN\_CD CDHLDR\_PRES\_CD CRCY\_CNVRSN\_RT ELCTR\_CMRC\_IND\_CD HOME\_PHN\_NUM\_CHNG\_DUR HOTEL\_STAY\_CAR\_RENTL\_DUR LAST\_ADR\_CHNG\_DUR LAST\_PLSTC\_RQST\_REAS\_CD MRCH\_CATG\_CD MRCH\_CNTRY\_CD NEW\_USER\_ADDED\_DUR PHN\_CHNG\_SNC\_APPN\_IND PIN\_BLK\_CD PIN\_VLDTN\_IND PLSTC\_ACTVN\_REQD\_IND PLSTC\_ISU\_DUR PLSTC\_PREV\_CURR\_CD POS\_COND\_CD POS\_ENTRY\_MTHD\_CD RCURG\_AUTHZN\_IND RVRSL\_IND SENDR\_RSIDNL\_CNTRY\_CD SRC\_CRCY\_CD SRC\_CRCY\_DCML\_PSN\_NUM TRMNL\_ATTNDNC\_CD TRMNL\_CAPBLT\_CD TRMNL\_CLASFN\_CD TRMNL\_ID TRMNL\_PIN\_CAPBLT\_CD DISTANCE\_FROM\_HOME

Mock Data :

4.45E+07 4.52E+07 null 2063.32856 2271.134268 3671.488059 1307.746119 null 23.67176051 null 4 null 1.443767445 0.006895252 null null 47.69801525 2.422723283 null 1.005063857 null 137.4062856 1.705829458 1014.877486 null 1.203904373 1.823679279 16.54531475 null 30590.44823 null 0.90934408 null null 1.070499825 682083.055 0.553171278 156.9285116 0.014623591 268.7501087 null 5844.265985 839.390814 68.0155152 5.021719442 null null null 356.4020061 null 5.565679338 77.03955118 null null null 839.9078728 0 0.404635355 3.152213568 2.12527761 Infinity 0.984512904 417.7634912

2.57E+07 2.61E+07 null 2984.023809 4862.731745 5981.70736 2219.582935 null 29.20064837 null 0 null 1.572042791 0.095188483 null null 779.3760684 3.205928201 null 0.071021931 null 404.0773877 31.35552666 246.528401 null 7.423808942 1.424426792 901.6241136 null 32793.35851 null 0.441755688 null null 1.916918544 468701.9029 1.780526671 318.7755609 0.507097293 400.7643651 null 974.5534246 16.83404494 219.5356257 0.254335592 null null null 297.249979 null 12.11158783 29.42622541 null null null 8.90337542 0 0.509437637 2.453466369 2.418803134 NaN 1.386189002 781.0533842

Added sample of 2 records with required columns values.

**STEP 3: Visualization**

**Show the distribution of length**

acct\_id\_token

auth\_id fraud\_reported

account\_ageactivation\_age

time\_since\_first\_use

time\_of\_day

case when FRD\_IND = 'Y' then 1 else 0 end fraud reported

cast (AUTHZN\_RQST\_PROC\_DT as bigint) - cast (ACCT\_OPEN\_DT as bigint) account age

cast (PLSTC\_ACTVN\_DT as bigint) - cast (ACCT\_OPEN\_DT as bigint) activation age

cast (PLSTC\_FRST\_USE\_TS as bigint) - cast (ACCT\_OPEN\_DT as bigint) time\_since\_first\_use

Sample values for the above derived columns,

Auth\_Id Fraud\_Protected Account\_Page Account\_Page Activation\_Age Time\_Since\_First\_Use

1602072 0 35078400 1036800 1073038 11

1602167 0 238723200 235094400 235394812 13

1602264 0 248572800 215740800 217024661 9

1602360 0 248227200 215740800 217024661 12

1602454 0 277689600 205459200 205515912 14

Along with taking required columns and above columns are derived columns

AUTHZN\_AMT ACCT\_AVL\_CASH\_BEFORE\_AMT ACCT\_AVL\_MONEY\_BEFORE\_AMT ACCT\_CL\_AMT ACCT\_CURR\_BAL AUTHZN\_OUTSTD\_AMT AUTHZN\_OUTSTD\_CASH\_AMT APPRD\_AUTHZN\_CNT APPRD\_CASH\_AUTHZN\_CNT ACCT\_PROD\_CD AUTHZN\_CHAR\_CD AUTHZN\_CATG\_CD CARD\_VFCN\_2\_RESPNS\_CD CARD\_VFCN\_2\_VLDTN\_DUR POS\_ENTRY\_MTHD\_CD TRMNL\_ATTNDNC\_CD TRMNL\_CLASFN\_CD DISTANCE\_FROM\_HOME

Create DATAFRAME from the above SELECT query.

Step 4: Model creation

* Use MLLib to train a Logistic Regression Model
* 20 features and 20 million rows
* Split data into Categoricals
* Use StringIndexers to Index the Categoricals
* Use a StandardScaler to scale the numeric features based on a normal distribution

Steps to be followed:

1. **Define Feature Columns & Default Null Values**

Defining the features function to handle the fraud\_reported, acct\_id\_token, auth\_id columns.query.clone filter not to be done.

Defining the categoricals and numerics to filter “\_CD”

Defining the authDataSet which handles query.na.fill(0.0).na.fill("N/A")

And then data is ready to train the model.

**acct\_id\_token auth\_id fraud\_reported account\_age activation age time\_since\_first\_use time\_of\_day AUTHZN\_AMT ACCT\_AVL\_CASH\_BEFORE\_AMT ACCT\_AVL\_MONEY\_BEFORE\_AMT ACCT\_CL\_AMT ACCT\_CURR\_BAL AUTHZN\_OUTSTD\_AMT AUTHZN\_OUTSTD\_CASH\_AMT APPRD\_AUTHZN\_CNT APPRD\_CASH\_AUTHZN\_CNT AUTHZN\_CHAR\_CD AUTHZN\_CATG\_CD CARD\_VFCN\_2\_RESPNS\_CD CARD\_VFCN\_2\_VLDTN\_DUR POS\_ENTRY\_MTHD\_CD TRMNL\_ATTNDNC\_CD TRMNL\_CLASFN\_CD DISTANCE\_FROM\_HOME**

18499888 1602360 0 248227200 215740800 217024661 12 333.6 7650 6657.36 7650 992.64 333.6 0 1 0 V 0 M 96506 1 0 4 226.52313

15073846 1602454 0 277689600 205459200 205515912 14 75 4000 7.79 4000 3992.21 0 0 0 0 W 7 M 90453 81 0 0 1443.822

48267719 1602543 0 57628800 691200 1154983 19 9.99 100 0 400 436.3 0 0 0 0 N/A 0 M 16022 81 0 0 1565.5552

1. **Define ML Pipeline to Engineer Features**

Defining the user defines functions with ML lib,

* stringIndexers
* catColumns
* vaNumerics
* vaCategorical
* scaler
* allFeatures
* stages
* featurePipeline

1. **Engineer Features**

By using the fit function into the featurePipeline just pass the authDataset as parameters

featurePipeline.fit(authDataSet)

By using this model we are labialized probability of statistics of users detections and create featured data set.

61478690 1602072 0 35078400 1036800 1073038 11 25.62 150

21298323 1602167 0 238723200 235094400 235394812 13 11311.18 10000

18499888 1602264 0 248572800 215740800 217024661 9 109.25 7650

1. **Cache Features in Memory**

featurizedDataset.cache().rdd.count

1. **Train Model Against Features**

Train the model, by using the LogisticRegression(), ParamGridBuilder(), BinaryClassificationEvaluator(), TrainValidationSplit() and fit into the feature data set model.

1. **Evaluate the Model**

Check the Area Under the Curve which 0.75 (which is an average score not that great)

Examine the intercept and the coefficients

Lastly we examine the absolute value of the coefficients as a general measure of the importance of the various variables we used in the predictive model