ISSN 2347 - 3983

Volume 8. No. 9, September 2020

International Journal of Emerging Trends in Engineering Research

Available Online at http://www.warse.org/IJETER/static/pdf/file/ijeter152892020.pdf

https://doi.org/10.30534/ijeter/2020/152892020

Design and Development of Financial Fraud Detection using Machine Learning

J.Rajendra, Prasad[1], S.SaiKumar[2], B.V.SubbaRao[3]

[1]Professor,Dept.ofIT, P.V.P.Siddhartha Institute Of Technology,India, rp.rajendra@rediffmail.com

[2]Assistant Professor,Dept.ofIT, P.V.P.Siddhartha Institute Of Technology, India, saikumar.senagavarapu@gmail.com

[3]Professor& HOD,Dept.ofIT, P.V.P.Siddhartha Institute Of Technology, India, bvsrau@gmail.com

ABSTRACT

Financial fraud is a growing concern with far reachingconsequences in the government, corporate organizations,finance industry, In Today's world high dependency oninternet technology has enjoyed increased credit cardtransactions but credit card fraud had also acceleratedasonline and offline transaction. As credit cardtransactions become a widespread mode of payment, focus has been given to recent computational methodologies to handle the credit card fraud problem. Finance fraud is a growing problem with far consequences in the financial industry and while many techniques have been discovered. Machine learning has been successfully applied to finance databases to automate analysis of huge volumes of complex data. Datamining has also played a salient role in the detectionof credit card fraud in online transactions. Frauddetection in credit card is a data mining problem, Itbecomes challenging due to two major reasons–first,theprofiles of normal and fraudulent behaviors changefrequently and secondly due to reason that credit cardfraud data sets are highly skewed. This paper uses ofDecision tree, Random Forest, SVM and logisticregression on highly skewed credit card fraud data.Recent research has shown that machine learningtechniques have been applied very effectively to theproblem of payments related fraud detection. Such MLbased techniques have the potential to evolve anddetectpreviously unseen patterns of fraud. In thispaper, weapply multiple ML techniques based onLogisticregression and Support Vector Machine to theproblemof payments fraud detection using a labeleddatasetcontaining payment transactions. We show thatourproposed approaches are able to detect fraudtransactions with high accuracy and reasonably lownumber of false positives.

Key words: Fraud in credit card, data mining, decisiontree, SVM, random forest.

1. INTRODUCTION

In the existing real world there is no real fraud detection mechanism on the working planet. Fraud works and falsified records are going on in the world rapidly and many people are committing crimes and falsifying records of the documents .This leads to the unethical prosperity of the world and also financially it leads to huge losses to the government and private sector. According to Global Payments Report 2015, credit card is the highest used payment method globally in 2014 compared to other methods such as e-wallet and Bank Transfer. The huge transactional services are often eyed by cyber criminals to conduct fraudulent activities using the credit card services. Financial fraud is defined as the unauthorized usage of card, unusual transaction behavior, or transactions on an inactive Card. Therefore, it is necessary to develop credit card fraud detection techniques as the counter measure to combat illegal activities. Credit card detection is one of the fraud detection techniques .We can also go with the insurance detection and also different comparative studies and feature analysis[1][2][3].

The existing system results in the following drawbacks:

- The classification rules cannot differentiate the fraud and genuine users of different corporations of the world.
- The present mechanism we are using affects so much the financial status of different bodies of different streams as the fraud detection is impossible considering the present technology.
- The existing system also affects the moral and common users of the credit cards, insurance etc as the company fails to accommodate the insurance to all the users even the fraud.

2. PROPOSED SYSTEM

The goal of this paper is to develop an efficientmethod to perform Financial Fraud Detection (FFD)from the analysis of data. These are required to befiled annually with the SEC, and are made public viathe database. According to the SEC, in 2006 more than245 large companies with market capitalization of $75million submitted a financial restatement, which is amodification of a previously filed statement, and oftenrequired when fraud is detected. However, due torapidly increasing number of documents, traditionalaudittechniques relying on human judgment have beomeprohibitively expensive. Natural Language Processingtechniques, which utilize the power of machine learningto do auditing instead of human, has jumped to thestage to fill the gap. Natural Language Processing(NLP) is an area of research and application that usescomputers to sift through large numbers of textdocuments to extract patterns that can be correlated with humanly discernable text content. Combined with Supervised Machine Learning (SML) techniques, such as Support Vector Machine, Neural Networks, Binomial Logistic Regression, and Ensemble techniques, we attempt to develop a methodology for financial fraud detection that is cost effective relative to human efforts. Our approach differs from existing text mining approaches to FFD, which consider the underlying semantic structure of the documents to identify the fraud [1]. Instead of relying on the semantic intricacies of the document, which is notoriously hard to capture, we use the probability distributions of the words across documents asa heuristic for classification[4],[5],[6],[7],[8],[9].

The proposed system results in the following advantages:

- The advantage of this technique is it can able to work on commercial databases without capacity limitations.
- It has a sophisticated graphical user interface.
- It is easily extensible to integrate with other intelligent techniques for credit card fraud detection.
- The performance is satisfactory.

Modules

1.Data collection and pre-processing
2.Data Analysing
3.Applying supervision machine learning methods.

Data collection and Preprocessing

Generally, the data will be split into three different segments – training, testing, and cross-validation. The algorithm will be trained on a partial set of data and parameters tweaked on a testing set. The performance of the data is measured using cross-validation set. The high performing models will be then tested for various random splits of data to ensure consistency in results

Data Analyzation

The main application of machine learning used in fraud detection is the prediction. We want to predict the value of some output (in this case, a boolean value that is true if the payment is fraudulent and false otherwise) given some input values (for example, the country the card was issued in and the number of distinct countries the card was used in the past day). The data that is used to train the ML models consists of records with both the output values for various input values. The records are often obtained from historical data.

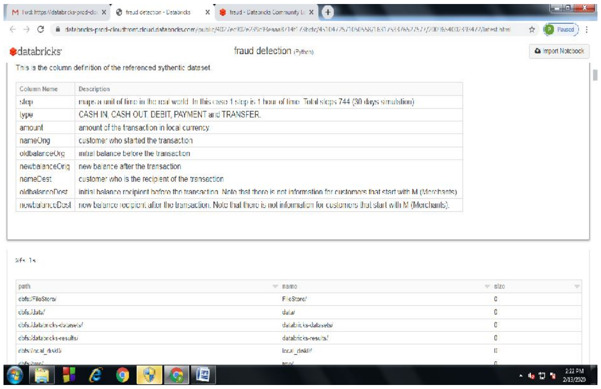
Applying Supervised Machine Learning methods

Building models is an essential step in predicting the fraud or anomaly in the data sets. We determine how to make that prediction based on previous examples of input and output data. We can further divide the prediction problem into two types of tasks:
1.Classification
2. Regression

1. Logistic Regression Regression analysis is a popular, longstanding statistical technique that measures the strength of cause-and-effect relationships in structured data sets. Regression analysis tends to become more sophisticated when applied to fraud detection due to the number of variables and size of the data sets. It can provide value by assessing the predictive power of individual variables or combinations of variables as part of a larger fraud strategy. In this techniques, the authentic transactions are compared with the fraud ones to create an algorithm. This model (algorithm) will predict whether a new transaction is fraudulent or not. For very large merchants these models are specific to their customer base, but usually, general models will apply.

2. Decision Tree This is a mature machine learning algorithm family used to automate the creation of rules for classification tasks. Decision Tree algorithms can use for classification or regression predictive modeling problems. They are essentially a set of rules which are trained using examples of fraud that clients are facing. The creation of a tree ignores irrelevant features and does not require extensive normalization of the data. A tree can be inspected and we can understand why a decision was made by following the list of rules triggered by a certain customer. The output of the machine learning algorithm might be a model like the following decision tree. This gives a probability score of fraud based on earlier scenarios.

3. RESULTS


Fwd: https://databricks-prod-clo
fraud detection - Databricks
fraud - Databricks Community E
databricks-prod-cloudfront.cloud.databricks.com/public/4027ec902e239c93eaaa8714f173bcfc/4510472571050558/1637253376652757/2009654002393477/latest.html
Paused
databricks
Import Notebook
2:22 PM
2/13/2020

fraud detection (Python)

This is the column definition of the referenced sythentic dataset

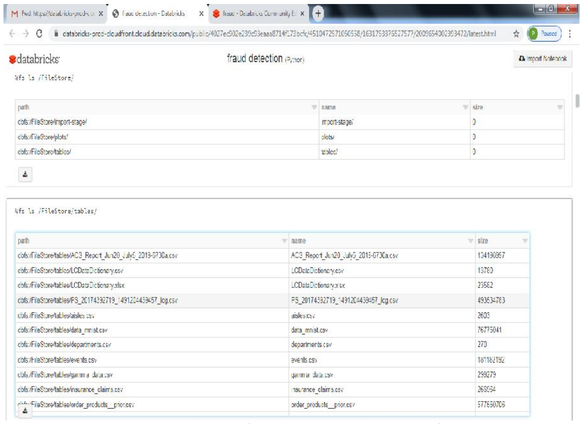
Column Name	Description
step	maps a unit of time in the real world. In this case 1 step is 1 hour of time. Total steps 744 (30 days simulation)
type	CASH IN, CASH OUT, DEBIT, PAYMENT and TRANSFER.
amount	amount of the transaction in local currency.
nameOrig	customer who started the transaction
oldbalanceOrg	initial balance before the transaction
newbalanceOrig	new balance after the transaction
nameDest	customer who is the recipient of the transaction
oldbalanceDest	initial balance recipient before the transaction. Note that there is not information for customers that start with M (Merchants)
newbalanceDest	new balance recipient after the transaction. Note that there is not information for customers that start with M (Merchants).

```
%fs ls
```

path	name	size
dbfs:/FileStore/	FileStore/	0
dbfs:/data/	data/	0
dbfs:/databricks-datasets/	databricks-datasets/	0
dbfs:/databricks-results/	databricks-results/	0
dbfs:/local_disk0/	local_disk0/	0
dbfs:/tmp/	tmp/	0

Description: The dataset we are working on

We have taken the data set of a company containing transactions of credit card that contains different transactions like transfer, credit, debit, cash-in, cash-out etc.There are nearly 90000 transactions in the datasets containing different persons dealing with money giving money and receiving money. There are attributes in this dataset like oldbalance, newbalance, oldbalancedest, newbalancedest.We are recording the people who have transferred the money to someone and the people who have received the money. We are also considering the amount of money before transaction and after the transaction.



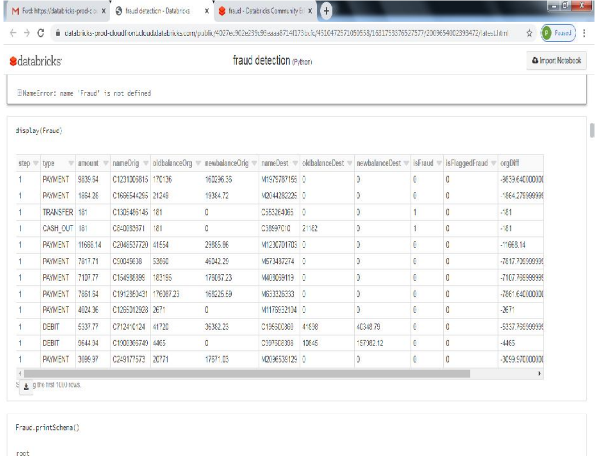
```
%fs ls /FileStore/
```

path	name	size
dbfs:/FileStore/import-stage/	import-stage/	0
dbfs:/FileStore/plots/	plots/	0
dbfs:/FileStore/tables/	tables/	0

```
%fs ls /FileStore/tables/
```

path	name	size
dbfs:/FileStore/tables/ACS_Report_Jun20_July5_2019-0730a.csv	ACS_Report_Jun20_July5_2019-0730a.csv	134190967
dbfs:/FileStore/tables/LCDataDictionary.csv	LCDataDictionary.csv	13780
dbfs:/FileStore/tables/LCDataDictionary.xlsx	LCDataDictionary.xlsx	23562
dbfs:/FileStore/tables/PS_20174392719_1491204439457_log.csv	PS_20174392719_1491204439457_log.csv	493534783
dbfs:/FileStore/tables/aisles.csv	aisles.csv	2603
dbfs:/FileStore/tables/data_mnist.csv	data_mnist.csv	76775041
dbfs:/FileStore/tables/departments.csv	departments.csv	270
dbfs:/FileStore/tables/events.csv	events.csv	181182192
dbfs:/FileStore/tables/gamma_data.csv	gamma_data.csv	299279
dbfs:/FileStore/tables/insurance_claims.csv	insurance_claims.csv	268064
dbfs:/FileStore/tables/order_products__prior.csv	order_products__prior.csv	577550706

Description: We are finding the path of the dataset



```
NameError: name 'Fraud' is not defined
```

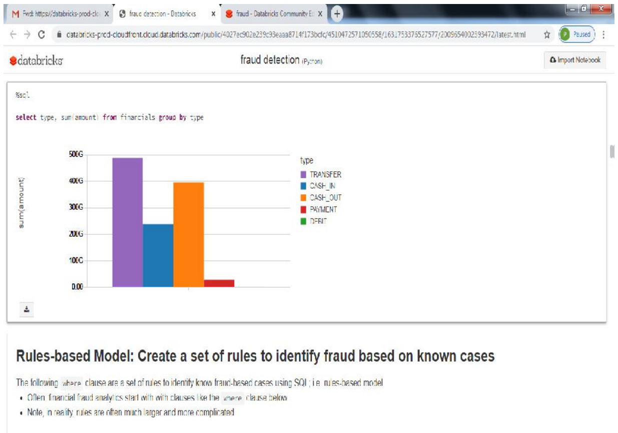
```
display(Fraud)
```

step	type	amount	nameOrig	oldbalanceOrg	newbalanceOrig	nameDest	oldbalanceDest	newbalanceDest	isFraud	isFlaggedFraud	orgDiff
1	PAYMENT	9839.64	C1231006815	170136	160296.36	M1979787155	0	0	0	0	-9839.64000000
1	PAYMENT	1864.28	C1666544295	21249	19384.72	M2044282225	0	0	0	0	-1864.27999999
1	TRANSFER	181	C1305486145	181	0	C553264065	0	0	1	0	-181
1	CASH_OUT	181	C840083671	181	0	C38997010	21182	0	1	0	-181
1	PAYMENT	11668.14	C2048537720	41554	29885.86	M1230701703	0	0	0	0	-11668.14
1	PAYMENT	7817.71	C90045638	53860	46042.29	M573487274	0	0	0	0	-7817.70999999
1	PAYMENT	7107.77	C154988899	183195	176087.23	M408069119	0	0	0	0	-7107.76999999
1	PAYMENT	7861.64	C1912850431	176087.23	168225.59	M633326333	0	0	0	0	-7861.64000000
1	PAYMENT	4024.36	C1265012928	2671	0	M1176932104	0	0	0	0	-2671
1	DEBIT	5337.77	C712410124	41720	36382.23	C195600860	41898	40348.79	0	0	-5337.76999999
1	DEBIT	9644.94	C1900366749	4465	0	C997608398	10845	157982.12	0	0	-4465
1	PAYMENT	3099.97	C249177573	20771	17671.03	M2096539129	0	0	0	0	-3099.97000000

S g the first 1000 rows.

```
Fraud.printSchema()

root
```

Description: The dataset we are about to work on



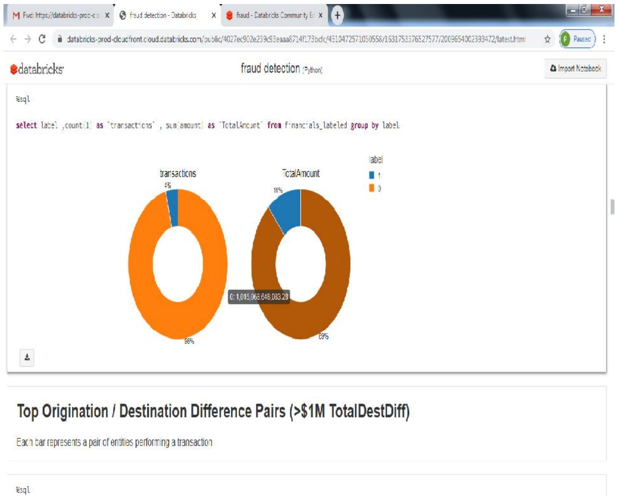
```
%sql

select type, sum(amount) from financials group by type
```

Rules-based Model: Create a set of rules to identify fraud based on known cases

The following `where` clause are a set of rules to identify know fraud-based cases using SQL; i.e. rules-based model

- Often, financial fraud analytics start with with clauses like the `where` clause below
- Note, in reality, rules are often much larger and more complicated

Description: The percentage of each transaction of all transactions.



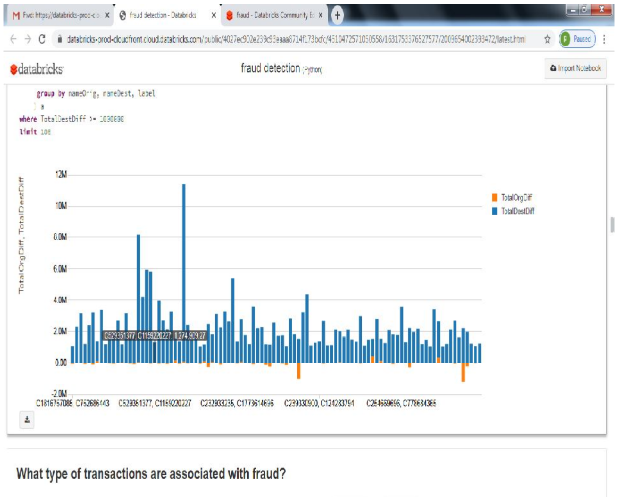
```
%sql

select label ,count(1) as `transactions` , sum(amount) as `TotalAmount` from financials_labeled group by label
```

Top Origination / Destination Difference Pairs (>$1M TotalDestDiff)

Each bar represents a pair of entities performing a transaction

```
%sql
```

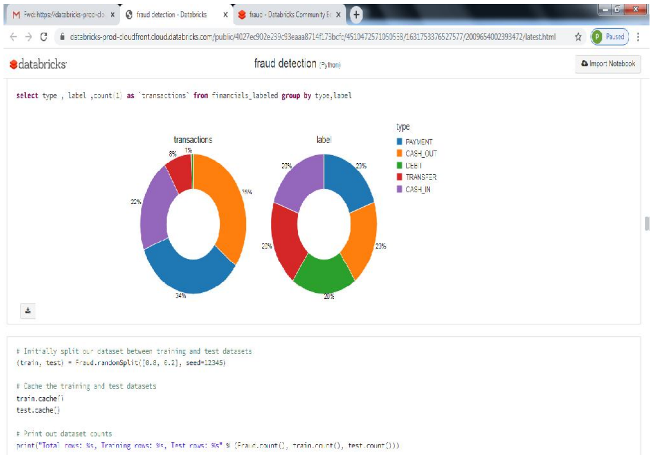
Description: The total amount of transactions and amount that are fraudlent.



```
    group by nameOrig, nameDest, label
    ) a
where TotalDestDiff >= 1000000
limit 100
```

What type of transactions are associated with fraud?

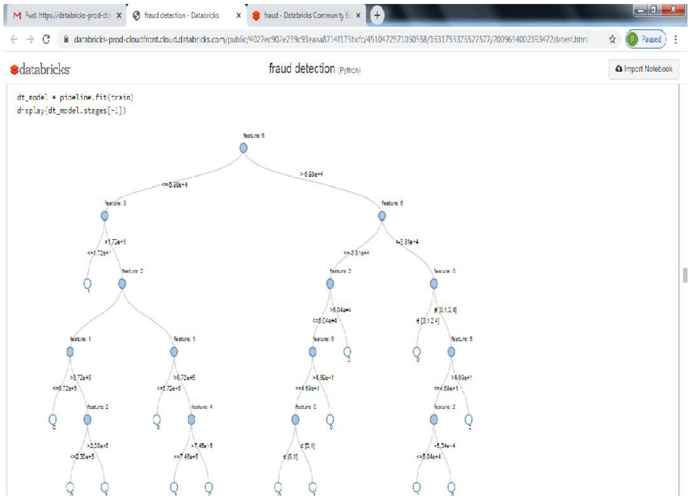
Description: The transactions in the form of bar graph form.



```
select type , label ,count(1) as `transactions` from financials_labeled group by type,label
```

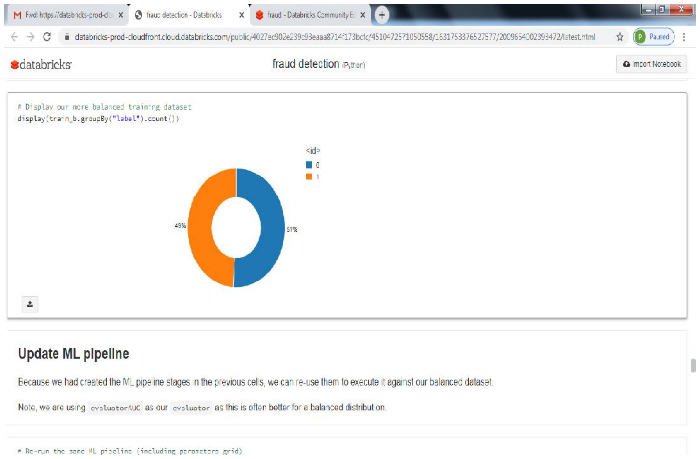
```
# Initially split our dataset between training and test datasets
(train, test) = Fraud.randomSplit([0.8, 0.2], seed=12345)

# Cache the training and test datasets
train.cache()
test.cache()

# Print out dataset counts
print("Total rows: %s, Training rows: %s, Test rows: %s" % (Fraud.count(), train.count(), test.count()))
```

Description: The percentage of the each type of transaction of all transactions and for the label data.



```
dt_model = pipeline.fit(train)
display(dt_model.stages[-1])
```

Description: The initial fitting of the decision tree model.



```
# Display our more balanced training dataset
display(train_b.groupBy("label").count())
```

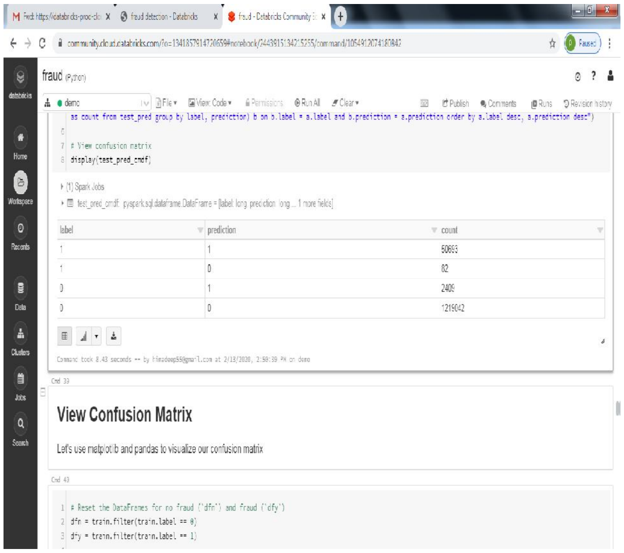
Update ML pipeline

Because we had created the ML pipeline stages in the previous cells, we can re-use them to execute it against our balanced dataset.

Note, we are using `evaluatorAUC` as our `evaluator` as this is often better for a balanced distribution.

```
# Re-run the same ML pipeline (including parameters grid)
```

Description: The training dataset for the test of accuracy.



```
as count from test_pred group by label, prediction) b on b.label = a.label and b.prediction = a.prediction order by a.label desc, a.prediction desc")

# View confusion matrix
display(test_pred_cmdf)
```

(1) Spark Jobs

test_pred_cmdf: pyspark.sql.dataframe.DataFrame = [label: long, prediction: long ... 1 more fields]

label	prediction	count
1	1	50693
1	0	82
0	1	2409
0	0	1219042

Command took 8.43 seconds -- by himadeep55@gmail.com at 2/13/2020, 2:50:39 PM on demo

Cmd 39

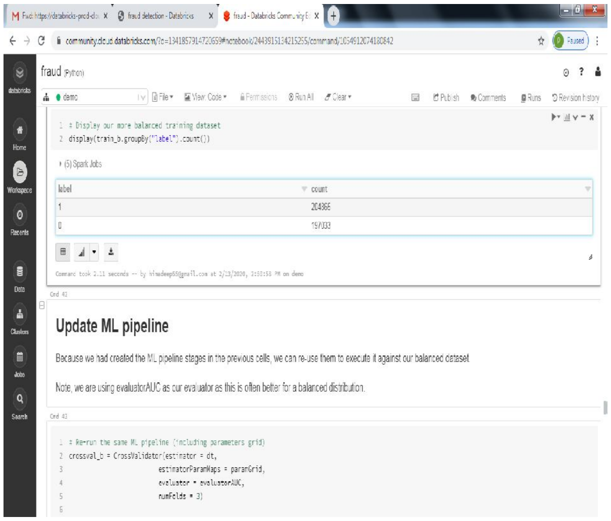
View Confusion Matrix

Let's use matplotlib and pandas to visualize our confusion matrix

Cmd 40

```
# Reset the DataFrames for no fraud (`dfn`) and fraud (`dfy`)
dfn = train.filter(train.label == 0)
dfy = train.filter(train.label == 1)
```

Description: The confusion matrix predictions



```
# Display our more balanced training dataset
display(train_b.groupBy("label").count())
```

(5) Spark Jobs

label	count
1	204366
0	197033

Command took 2.11 seconds -- by himadeep55@gmail.com at 2/13/2020, 2:53:58 PM on demo

Cmd 42

Update ML pipeline

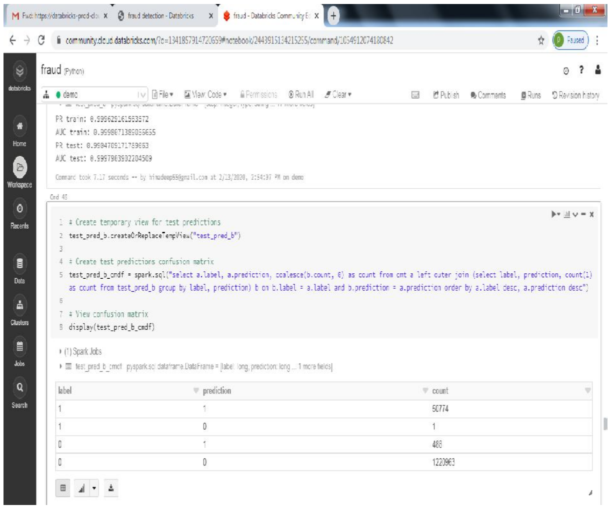
Because we had created the ML pipeline stages in the previous cells, we can re-use them to execute it against our balanced dataset

Note, we are using evaluatorAUC as our evaluator as this is often better for a balanced distribution.

Cmd 43

```
# Re-run the same ML pipeline (including parameters grid)
crossval_b = CrossValidator(estimator = dt,
                            estimatorParamMaps = paramGrid,
                            evaluator = evaluatorAUC,
                            numFolds = 3)
```

Description: the training dataset fraudulent cases



```
PR train: 0.999629161563572
AUC train: 0.9998071389055655
PR test: 0.9984709171789663
AUC test: 0.9997903902204509
```

Command took 7.17 seconds -- by himadeep55@gmail.com at 2/13/2020, 2:54:37 PM on demo

Cmd 45

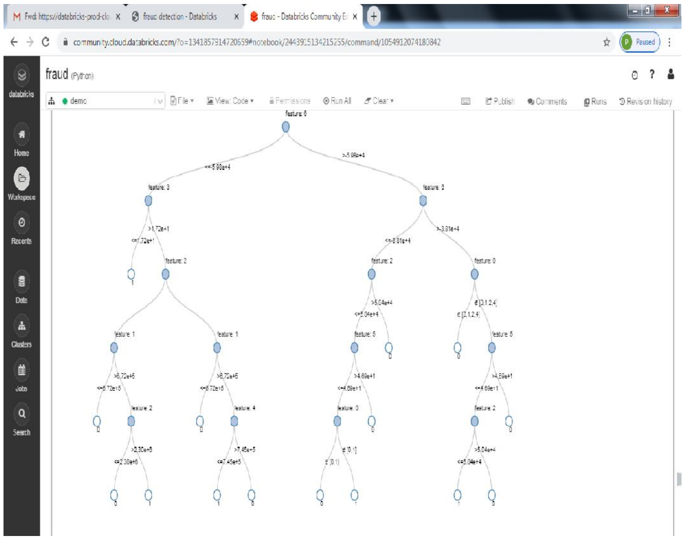
```
# Create temporary view for test predictions
test_pred_b.createOrReplaceTempView("test_pred_b")

# Create test predictions confusion matrix
test_pred_b_cmdf = spark.sql("select a.label, a.prediction, coalesce(b.count, 0) as count from cmt a left outer join (select label, prediction, count(1) as count from test_pred_b group by label, prediction) b on b.label = a.label and b.prediction = a.prediction order by a.label desc, a.prediction desc")

# View confusion matrix
display(test_pred_b_cmdf)
```

(1) Spark Jobs

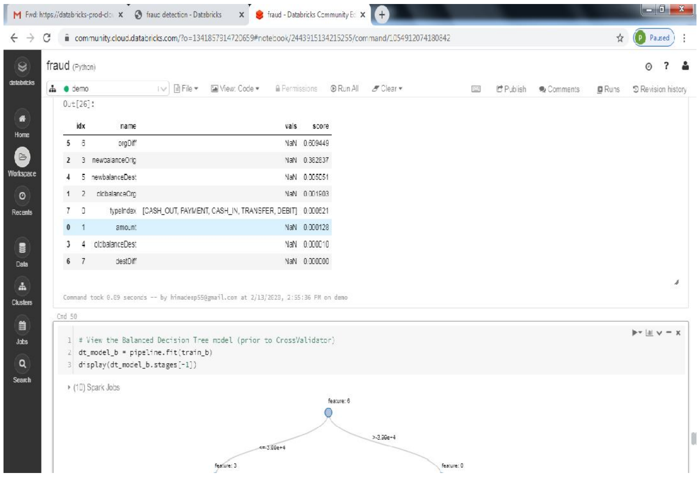
test_pred_b_cmdf: pyspark.sql.dataframe.DataFrame = [label: long, prediction: long ... 1 more fields]

label	prediction	count
1	1	50774
1	0	1
0	1	488
0	0	1220963

Description: The predictions of fraud on the test dataset



Description: The fitting of the decision tree model on the training dataset


Fwd: https://databricks-prod-clo
fraud detection - Databricks
fraud - Databricks Community Ed
community.cloud.databricks.com/?o=1341857914720659#notebook/2443915134215255/command/1054912074180842
Paused
fraud (Python)
demo
File
View: Code
Permissions
Run All
Clear
Publish
Comments
Runs
Revision history
databricks
Home
Workspace
Recents
Data
Clusters
Jobs
Search

Out[26]:

	idx	name	vals	score
5	6	orgDiff	NaN	0.609449
2	3	newbalanceOrig	NaN	0.382837
4	5	newbalanceDest	NaN	0.005051
1	2	oldbalanceOrg	NaN	0.001903
7	0	typeIndex	[CASH_OUT, PAYMENT, CASH_IN, TRANSFER, DEBIT]	0.000621
0	1	amount	NaN	0.000128
3	4	oldbalanceDest	NaN	0.000010
6	7	destDiff	NaN	0.000000

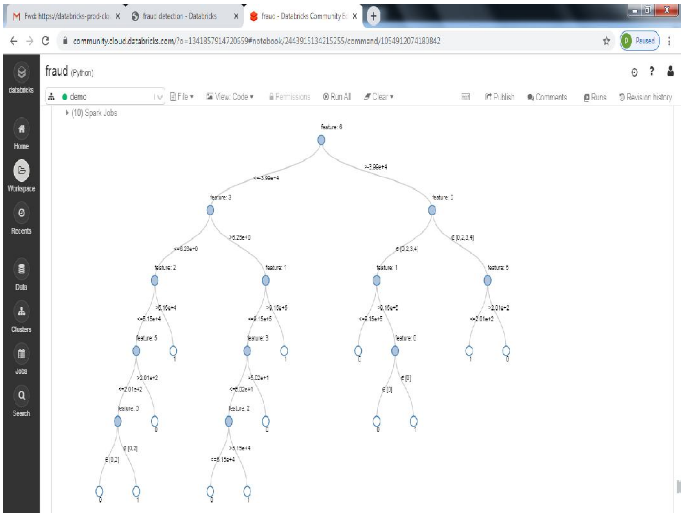
```
Command took 0.89 seconds -- by himadeep55@gmail.com at 2/13/2020, 2:55:36 PM on demo
```

Cmd 50

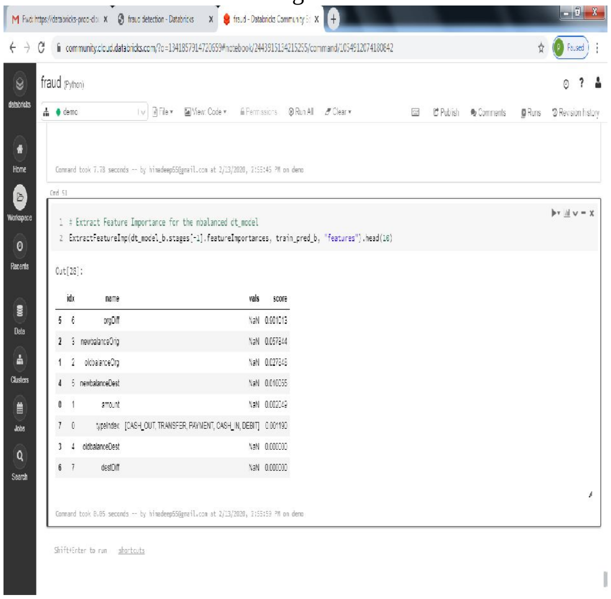
```
# View the Balanced Decision Tree model (prior to CrossValidator)
dt_model_b = pipeline.fit(train_b)
display(dt_model_b.stages[-1])
```

(10) Spark Jobs

Description: The result of the dataset after fitting the decision tree model.



Description: The fitting of the decision tree model on the training dataset


Fwd: https://databricks-prod-clo
fraud detection - Databricks
fraud - Databricks Community Ed
community.cloud.databricks.com/?o=1341857914720659#notebook/2443915134215255/command/1054912074180842
Paused
fraud (Python)
demo
File
View: Code
Permissions
Run All
Clear
Publish
Comments
Runs
Revision history
databricks
Home
Workspace
Recents
Data
Clusters
Jobs
Search

```
Command took 7.78 seconds -- by himadeep55@gmail.com at 2/13/2020, 2:55:45 PM on demo
```

Cmd 51

```
# Extract Feature Importance for the nbalanced dt_model
ExtractFeatureImp(dt_model_b.stages[-1].featureImportances, train_pred_b, "features").head(10)
```

Out[28]:

	idx	name	vals	score
5	6	orgDiff	NaN	0.901013
2	3	newbalanceOrig	NaN	0.057844
1	2	oldbalanceOrg	NaN	0.027845
4	5	newbalanceDest	NaN	0.010055
0	1	amount	NaN	0.002049
7	0	typeIndex	[CASH_OUT, TRANSFER, PAYMENT, CASH_IN, DEBIT]	0.001190
3	4	oldbalanceDest	NaN	0.000000
6	7	destDiff	NaN	0.000000

```
Command took 0.05 seconds -- by himadeep55@gmail.com at 2/13/2020, 2:55:59 PM on demo
```

Shift+Enter to run shortcuts

Description: The result of the dataset after fitting the decision tree model.

4. SCOPE FOR FURTHER DEVELOPMENT

Since machine learning is a very popular field amongacademicians as well as industry experts, there is ahuge scope of innovation. Experimentation withdifferentalgorithms and models can help your business indetecting fraud. Machine learning techniques areobviously reliable than human review and transactionrules. The machine learning solutions are efficient,scalable and process a large number of transactions inreal time. But extracting data and training data sets forcorrect prediction is a tough task

Here in our paper we only determined the noof fraud cases out of the cases of the dataset and thento prove the accuracy of the model we used we diduse decision tree supervised algorithm .For that weclassified the data into training and test data .So wecan take a lot of other attributes such as location sowe can implement the model based on the location andgive us the output where the fraud is a lot .And wecan extend it in a lot of ways in a lot of directions.As already we took a lot of data ,data doesn’t concernus .More the data ,more the accuracy .So our paper isas flexible as it can get and as reliable as it can get.

5. CONCLUSION

Machine learning has been instrumental in solving someof the important business problems such as detectingemail spam, focused product recommendation, accuratemedical diagnosis etc. The adoption of machine learning(ML) has been accelerated with increasing processingpower, availability of big data and advancements instatistical modeling. Fraud management has been painfulfor banking and commerce industry. The number oftransactions has increased due to a plethora of payment channels – credit/debit cards, smartphones, kiosks. Atthe same time, criminals have become adept at findingloopholes. As a result, it’s getting tough for businessesto authenticate transactionsDatascientists have beensuccessful in solving this problem with machinelearning and predictive analytics. Automated fraud screening systems powered by machine learning can helpbusinesses in reducing fraud.

Such ML based techniques have the potential to evolve and detect previously unseen patterns of fraud. In this paper, we apply multiple ML techniques based on Decision Tree and Support Vector Machine to the problem of payments fraud detection using a labeled dataset containing payment transactions. We show that our proposed approaches are able to detect fraud transactions with high accuracy and reasonably low number of false positives.

REFERENCES

1 Raj S.B.E., Portia A.A., **Analysis on credit cardfraud detection methods**, Computer, Communicationand Electrical Technology

International Conferenceon (ICCCET) (2011), 152-156.

2 Jain R., Gour B., Dubey S., **A hybrid approach forcredit card fraud detection using rough set anddecision tree technique**, International Journal ofComputer Applications 139(10) (2016).

3 Dermala N., Agrawal A.N., **Credit card frauddetection using SVM and Reduction of falsealarms**,International Journal of Innovations in Engineeringand Technology (IJIET) 7(2) (2016).

4 Phua C., Lee V., Smith, Gayler K.R., Acomprehensive survey of data mining-based frauddetection research. arXiv preprint arXiv:1009.6119(2010).

5 Bahnsen A.C., Stojanovic A., Aouada D., OtterstenB., **Cost sensitive credit card fraud detection usingBayes minimum risk**. 12th International Conferenceon Machine Learning and Applications (ICMLA)(2013), 333-338.

6 Carneiro E.M., Dias L.A.V., Da Cunha A.M.,Mialaret L.F.S., **Cluster analysis and artificial neuralnetworks:** A case study in credit card frauddetection, 12th International Conference onInformation Technology-New Generations (2015),122-126.

7 Hafiz K.T., Aghili S., Zavarsky P., **The use ofpredictive analytics technology to detect credit cardfraud in Canada**, 11th Iberian Conference onInformation Systems and Technologies (CISTI)

8 Dr.A.M.Mahaboob Basha, Dr.M.Rajaiah, Dr.P.Penchalaiah, Dr.CH.Raja Kamal, B.Niranjana Rao **Machine Learning-Structural Equation Modeling Algorithm: The Moderating role of Loyalty on Customer Retention towards Online Shopping,** International Journal of Emerging Trends in Engineering Research, Volume 8, Issue No.5, May 2020, ISSN 2347-3983,PP.1578-1585.

9 Cho Do Xuan, Tisenko Victor Nikolaevich, Nguyen Quang Dam, Nguyen Quoc Hoang, Do Hoang Long **Malicious domain detection based on DNS query using Machine Learning,**International Journal of Emerging Trends in Engineering Research, Volume 8, Issue No.5, May 2020, ISSN 2347-3983, PP.1809-1814.