THE ANNALS OF "DUNĂREA DE JOS" UNIVERSITY OF GALATI
FASCICLE III, 2020, VOL. 43, NO.3, ISSN 2344-4738
ELECTROTECHNICS, ELECTRONICS, AUTOMATIC CONTROL, INFORMATICS
Article DOI: https://doi.org/10.35219/eeaci.2020.3.03

USING MACHINE LEARNING ALGORITHMS TO DETECT FRAUDS IN TELEPHONE NETWORKS

Apostu Sergiu

Department of Computer of Information Technology
"Dunărea de Jos" University of Galati, Romania
Corresponding author e-mail: apostusergiu4995@gmail.com

Abstract: This paper presents an analysis of voice traffic in telephone networks, based on machine learning algorithms to detect frauds made by callers. Starting from the raw data set that includes information about the call date, destination number, duration and caller's number, in our approach we were able to identify fraudulent calls in early stages. For balance, the data set was split in 2 parts: one for training and one for testing. To obtain mean's values from dataset, a standardization technique was applied in order to scale the data before the dimensionality reduction using Principal Component Analysis. Then, the first two components were used as inputs for Logistic Regression and Random Forest models, having the caller as target. Finally, the target was moved on the destination file so as to identify the caller and the moment when the call has started based on a vector representation of words.

Keywords: principal component analysis, logistic regression, random forest, word2vec.

1. INTRODUCTION

Many telephony operators have problems and financial losses created by their subscribers, skilled clients, who are exploiting their special offers with many international minutes, making calls to special high cost international numbers and getting paid to make those phone calls.

(Merve Sahin and Aurélien Francillon, 2018) presented such a scheme, IRSF (International Revenue Share Fraud), in which the malicious parties are teaming to abuse premium phone numbers, to generate unpaid call traffic in order to collect and share revenue paid by the telephony operators. But finally, the operators will share their losses with their clients, inflating the bill.

By 2019 it is estimated that the losses of the operators and their subscribers are over 38 billions $ (Cătălin Cimpanu, 2019).

The main question is *How do callers cheat, if they only make calls on international destinations?* The answer is quite simple: those callers search of apparently no-cost phone numbers from the outside (located in other country) telephone network, numbers for which they are not charged by their own operator because they profit from offers with international minutes included. However, those numbers are in fact premium with a payout each time they are called. So, a client having a subscription with lots of minutes for calls to international destinations will gain some money, in the end of this process.

The main problem for the telephony operator is the financial loss. There are interconnection agreements between phone network providers. For example, if a caller, an X operator subscriber, makes calls to one or more international numbers, the destination network provider will charge the X operator for those calls with a certain amount. And if these destination numbers are premium, with huge costs for operator

X, but no costs for the caller, then the loss will be borne by operator X. It must pay for the traffic generated by its customers to the recipient telephone network which has provided some premium numbers for callers. Then the outside network will split certain amounts of money for callers following calls made for the generated traffic (Fig. 1).

Fortunately, the good part in the effort to spot this kind of situation is the fact that someone who knows about this method to make money will try to consume all free international minutes available as soon as possible.

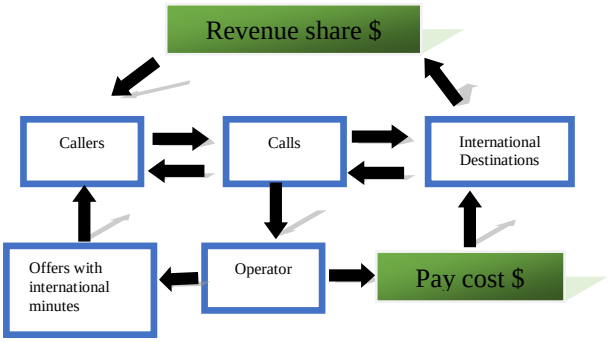

Fig.1. Cashflow diagram

Besides, due to the fraudulent calls, other problems such as equipment degradation or the network cloud overload could appear.

2. IMPLEMENTATION

This analysis was performed using an open-source machine learning library in Python, scikit-learn, a simple and efficient tool for data mining, data scaling and graphic representation (Scikit-learn library, 2020). The dataset used to study the possibility to recognize fraudulent calls was created in May 2020 having the acceptance to use the dataset recorded into a database. The main attributes considered for each call are (Fig. 2):

- Date of the call (start date);
- Destination's number;
- Caller's number;
- Duration of the call.

date	destination	caller
2020-05-04 14:19:03	44736801	40722642
2020-05-04 14:28:36	44736801	40721602
2020-05-04 14:43:08	44736801	40722643
2020-05-04 14:45:29	44736801	40746955
2020-05-04 14:53:15	44736801	40722409

Fig.2. Initial dataset.

The type of the first attribute is the "datetime" so, in order to increase the granularity of the data, this attribute was split into three new attributes: year, month and day (Fig. 3).

year	month	day	destination	duration	caller
2020	5	4	4473680	16.54	4072240
2020	5	4	4473680	16.28	4072264
2020	5	4	4473680	16.23	4072240
2020	5	4	4473680	19.33	4077158
2020	5	4	4473680	16.36	4072825

Fig.3. Dataset used for analysis

Because the attributes present big differences between their value ranges, the data was scaled by standardization (Aniruddha Bhandari, 2020).

	year	month	day	destination	duration
0	0.0	0.0	-1.709787	0.060410	0.682438
1	0.0	0.0	-1.709787	0.060410	-0.074945
2	0.0	0.0	-0.462451	-1.062696	-0.383605
3	0.0	0.0	-1.709787	0.060410	-0.030851
4	0.0	0.0	1.141266	1.803373	0.742095
...	...	...	...	...	...
1176	0.0	0.0	-0.106070	-0.958364	-3.916330
1177	0.0	0.0	0.963076	-1.062696	-0.030851
1178	0.0	0.0	-0.462451	-0.258724	-0.106070
1179	0.0	0.0	-0.462451	0.201566	-0.310979
1180	0.0	0.0	0.784885	-0.258724	0.939222

Fig.4. Standardized training dataset

2.1. Logistic Regression

Principal Component Analysis was used to reduce the dimension of the training data (Ashutosh Tripathi, 2019). The first two principal components were selected and used as input variables for Logistic Regression (Andreas C. Müller and Sarah Guido, 2016). The target (output) variable was the caller.

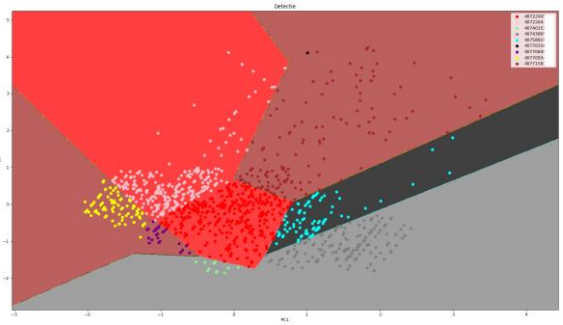

Fig.5. Detected callers who defrauded the calls

Fig.5 shows the 2-dimensional representation of the calls inside the space defined by the selected

principal components. Each color represents a caller. In each area callers with similar behavior can be found. For example, in the red zone, two callers and their calls: red and pink are classified. But the red dots, corresponding to this zone are invisible, having the same color as the area, while the pink caller is highlighted. The pink cells in the red zone correspond to the calls made in same days, to the same destinations, and, most important, with similar long duration. Those subscribers have consumed all their minutes and the algorithm detected those two callers as generating unusual traffic, different than for their personal use.

Another approach uses the *destination* and *duration* features as input for logistic regression. Fig.6. shows the calls in the space of duration (vertical) and the destination (horizontally). The callers were detected, which defrauded? Calls are visible as dots. They have same destination phone numbers and the similar average duration.

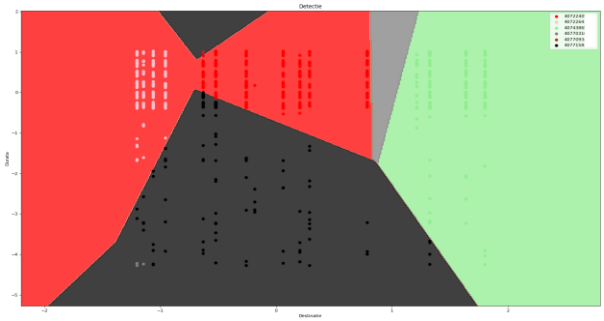

Fig.6. Detected frauds (destination vs duration).

In Fig. 7, the situation where the selected features used for this analysis was the following: day (horizontally) and destination (vertical) is depicted. It can be seen that inside the black zone are brown dots (detected calls identified are being fraudulent). The same situation is inside the red zone, where two types of colors are presented, meaning that the pink caller made nearly the same calls in the same days and to the same destination as the red caller, basically both being fraudulent customers.

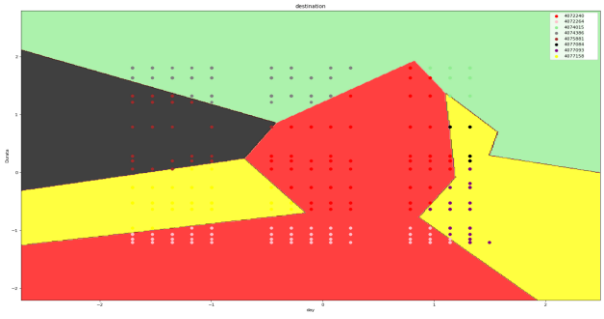

Fig.7. Detected frauds (day vs destination)

2.2. Random Forest

In the second part of this study, Logistic Regression was replaced by Random Forest, an ensemble machine learning algorithm, based on decision trees and majority vote as final forecast of the calls value (Aurélien Géron, 2017). Principal Component Analysis was again used together with Random Forest to benefit from data size reduction and to get a stable prediction. But there is a drawback of the Random Forest approach due to the difficulty in the interpretation of the resulted model (Aurélien Géron, 2017).

In Fig. 8, many detected callers, were tagged after majority vote was applied. Overall, it is difficult to separate the fraud-causing callers from the regular callers (for personal use) - the blue dots.

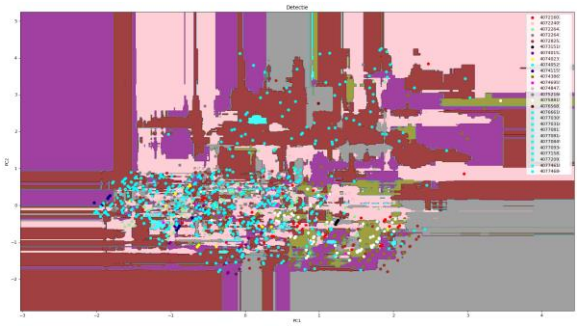

Fig.8. Detected frauds using Random Forest

For this reason other useful metrics are used to quantify the classification performance (Lahiru Liyanapathirana, 2018). Table 1 shows the accuracy, precision, recall and *F*-score obtained for all the above presented approaches.

Table 1 Performance measurement

Algorithm	Accuracy	Precision	Recall	F-score
PCA Logistic Regression	17,76%	8,7%	17,76%	11,27%
Logistic Regression (destination vs duration)	21,23%	6,95%	21,23%	10,15%
Logistic Regression (day vs destination)	19,96%	9,1%	19,96%	12,15%
PCA Random Forest	23,01%	23,19%	23,01%	22,89%

For example, using Random Forest the best results from table were obtained, but this algorithm does not

detect the frauds very efficiently and is difficult to separate callers who cheated and who not cheated. At the same time, Logistic Regression (day vs. destination) does not have good results, but it successfully detects the frauds.

2.3. Vector representation of words

Last study was based on *Word2Vec*, a group of two layer neural network to produce a vector representation of words. This method was developed and patented by Tomas Mikolov and his team (Tomas Mikolov et al., 2013a).

Prateek Joshi developed a recommendation system using *word2vec* with the method *Skip-gram* (Prateek Joshi, 2019). *Skip-gram* is a useful method because it uses the word target to predict the words from the context (Tomas Mikolov, Quoc V. LE and Ilya Sutskever 2013b).

First of all, in this case the data was converted as string and the destinations were set as target word, in order to find the frauds from calls (context). *Gensim* models are an open-source library that is used together with the functions from *word2vec*, applying *Skip-gram* in order to create the model (Radim Řehůřek, 2019).

Secondly, the training data was used to train the model, where the size of the vectors for each word (destination) from vocabulary, was set to 100 using the library *gensim models* together with the functions from *word2vec* and in particular using *Skip-gram*.

```
Word2Vec vocab=15, size=100
```

Fig.9. Model of vocabulary having 15 unique words (destinations).

Fig.10 shows in detail each phone number (destination) having a length of 100 vectors based on the created model. Each destination is the target word because the dataset was transformed into strings. After the training has been successfully completed the vectors will be used as a representation in fraud detection by selecting the destination.

```
4473680[redacted]0 0.44631377 -0.34997067 -0.30058482 0.26477212
4473680[redacted]1 -0.031369355 -0.22439265 -0.22338389 0.1842618
4473680[redacted]2 0.5906196 -0.20653142 -0.094594225 0.05232011
4473680[redacted]8 0.05530334 -0.39908406 -0.24837793 -0.25886658
4473680[redacted]9 0.108516164 -0.18664448 -0.1923563 0.28101787
4473680[redacted]8 0.11233531 -0.36647877 -0.2885639 0.073710404
4473680[redacted]1 0.13769484 -0.38846642 -0.20893128 -0.27745008
4473680[redacted]4 0.17602868 -0.26326588 -0.17346786 0.07508563
4473680[redacted]1 0.34641948 -0.28492838 -0.15876086 -0.1566633
4473680[redacted]7 0.16418687 -0.24215007 -0.15130793 -0.01421344
4473680[redacted]1 0.13281138 -0.37572435 -0.23648503 -0.16639662
4473680[redacted]8 0.37228796 -0.0625688 -0.18022941 0.5799752 -0
4473680[redacted]7 0.29628497 -0.3417781 -0.187262 -0.20271319 0.
4473680[redacted]2 0.31568837 -0.10052972 -0.18849526 0.52767974
4473680[redacted]4 0.2920924 -0.11912105 -0.17070085 0.40819097 -
```

Fig.10. Vectors of size 100 each word's (destination).

After the selection of destinations from the vocabulary we succeeded to identify the callers who generated fraudulent voice traffic and the moment when these calls have started (Fig. 11).

```
[(['4074695    ', '2020-05-06'], 0.9812464714050293),
 (['4074695    ', '2020-05-21'], 0.9806598424911499),
 (['4074386    ', '2020-05-04'], 0.8862554430961609),
 (['4074386    ', '2020-05-04'], 0.732524931430816 7),
 (['4074695    ', '2020-05-04'], 0.7236882448196411),
 (['4074695    ', '2020-05-05'], 0.6497043371200562),
 (['4075881    ', '2020-05-12'], 0.45743077993392944),
 (['4074386    ', '2020-05-13'], 0.44368767738342285)]
```

Fig.11. Fraud detection

3. CONCLUSIONS

To use machine learning algorithms in order to detect fraud in voice traffic in telephone network is a promising approach mainly because these algorithms can detect the fraud attempts in the early stages. If the fraud attempt is promptly recognized, then the financial loss is reduced.

In the sequel is a good way to use others machine learning algorithms to detect the frauds efficiently in order to improve the detection and also to develop a software able to identify early fraudulent calls.

4. REFERENCES

Andreas C. Müller and Sarah Guido (2016), *Introduction to Machine Learning with Python: A Guide for Data Scientists* (O'Reilly Media, Inc.), pp. 238-251.

Aniruddha Bhandari (2020), https://www.analyticsvidhya.com/blog/2020/04/feature-scaling-machine-learning-normalization-standardization/ (Accessed June, 2020).

Ashutosh Tripathi (2019), https://www.analyticsvidhya.com/blog/2020/04/feature-scaling-machine-learning-normalization-standardization/ (Accessed June, 2020).

Ashutosh Tripathi (2019), https://ashutoshtripathi.com/2019/07/11/a-complete-guide-to-principal-component-analysis-pca-in-machine-learning/(Accessed June, 2020).

Aurélien Géron (2017), *Hands-On Machine Learning with Scikit-Learn & TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems* (O'Reilly Media,Inc.), pp. 181-202.

Cătălin Cimpanu (2019), https://www.zdnet.com/article/phone-fraudsters-are-stealing-billions-each-year-through-a-scheme-known-as-irsf/ (Accessed June, 2020).

Lahiru Liyanapathirana (2018), https://heartbeat.fritz.ai/classification-model-evaluation-90d743883106(Accessed June, 2020)

Merve Sahin and Aurélien Francillon (2018). *IRSF: a Billion $ Fraud Abusing International Premium Rate Numbers* (EURECOM).

Prateek Joshi (2019), *Building a Recommendation System using Word2vec: A Unique Tutorial with case Study in Python* https://www.analyticsvidhya.com/blog/2019/07/how-to-build-recommendation-system-word2vec-python/ (Accessed June, 2020).

Radim Řehůřek (2019), https://radimrehurek.com/gensim/models/word2vec.html (Accessed June, 2020).

Tomas Mikolov et al. (2013a), *Efficient Estimation of Word Representations in Vector Space* (Google Inc.), arXiv:1301.3781v3.

Tomas Mikolov, Quoc V. LE and Ilya Sutskever (2013b), *Exploiting Similarities among Languages for Machine Translation* (Google Inc.), arXiv:1309.4168v1.