Implementation of fraud detection in advanced databases

Nedeljko Šikanjic
PhD Student at Pan-European University APEIRON
Banja Luka, Bosnia and Herzegovina
nedeljko.sikanjic@hotmail.com

Zoran Ž. Avramović
Professor at Pan-European University APEIRON
Banja Luka, Bosnia and Herzegovina

Abstract— **In today's word amount of data is increasing in rapid way that it is almost impossible to store, track and find data. In these large volumes of data sets, there are data that is being abused deliberately or accidentally. We will use the new implementations of machine learning in advanced databases to find easy and reliable way to inspect data for fraud.**

Keywords 1; Advanced Databases 2; Benford's Law 3; SQL

I. INTRODUCTION

Large volume of data increases inability to process and slows down analysis required to find fraudulent data. Every business area of data transactions is exposed and vulnerable for being manipulated by fraudulent activities. Implementation of automatic or generic approach is best way to fast and secure discover fraudulent tendencies in our data. In this paper, we will show how to implement in advanced databases fraud detection and what are possibilities of combining different technologies.

II. FREUD DETECTION

In today's world fraud represents a business worth in billion dollars. Survey done by PricewaterhouseCoopers in 2018 [1], a multinational company who is well known for its data analysis and financial services, shows that from 7000 companies almost 50% had been affected with some sort of fraud. Comparing with survey from 2016 done by the same company, it shows an increase of 14% of fraud done in economics.

Fraud increases with growth of information technologies, along with reorganization and business changes where with having less control opportunities for fraud increases.

For detecting fraud there data analysis methods which requires time consuming and complicated investigations. These methods are combined from multi-disciplinary sciences like economics and law. Problem with fraud that is not easily detected and require recognizing behavior in content and appearance of fraud instances.

In order to prevent fraud, data analysis was first used by companies in insurance, banking sector and telephone companies. This later has led to incorporating artificial intelligence in fraud detecting systems. One of first software created with AI is FICO Falcon from FICO (Fair Isaac Corporation), a company famous for their credit scoring system.

Most exposed are internet transactions due its accessibility to everyone with computer and internet connection. Internet fraud was responsible for more than 1.4 billion dollars lost in 2017 [2]. There are more types of internet fraud ranging from online scams, stealing online credentials and passwords, to email spam.

Fraud is also affecting POS and retail systems. One of most known POS fraud [3] was in USA where hackers have remotely installed "Trojan" software with which they have stolen credit and debit cards data.

Fraud represents form of adaptive crime, where traditional forms of data analysis is not able to cope effectively with it. That is the reason why intelligent forms of data analysis is being develop and continue to improve. Some of areas where data analysis is growing is Data Mining, Expert Systems and Knowledge Databases, Statistics and Machine Learning.

Fraud detection have main classification such as artificial intelligence and statistics.

For data analysis techniques in statistic we can list them in following order:

- Preprocessing of data in order to detect and validate data including correction of errors.
- Formula calculation based on different parameters like average values, baselines and thresholds.
- Clustering data in order to recognize certain pattern in behavior of our data
- Algorithm creation for detecting anomalies in data, elimination of false alarms and predicting results.

For artificial intelligence we can list the following techniques:

- Data mining that will segment and classify data in order of finding links and patterns in possible fraudulent behavior
- Export systems and Knowledge database that will recognize rules of possible fraud instances

- Neural networks and machine learning models that will learn the behavior of creation of fraudulent data and how to recognize them

Best results in detecting fraud is combination of these techniques and depending of domain where we want to implement fraud detection and prevention.

III. BENFORD'S LAW THEORY

Benford's Law is helpful in finding discrepancy in data and its anomalies. It identifies characteristics of data in order to predict following patterns in data.

Benford has found out that certain numbers occur more often than others [4]. This finding was discovered before Benford back in 1881, Professor Simon Newcomb by looking at Logarithm book, where he observed that digits (like 1, 2 and 3) were more worn out then other pages with higher numbers [5]. Difference why this is not being called Newcomb's Law lays in fact that he did not provide any statistical observation and has provided no practical implementation. Later in 1938 Frank Benford without knowing of Professor Newcomb findings, have observed similar behavior by his Logarithm Book. However, Frank Benford did test his theory and made more research for it, with different set of data such as electricity bills, town population, stock market etc.

Mathematic formula he created follows like this:

$$Pr(d)=\log10(d+1)-\log10(d) \quad (1)$$

Or simplified

$$Pr(d)=\log10(1 + 1/d)$$

In formula (1) "d" presents a number value (1,2,3..9) and "Pr" stands for probability.

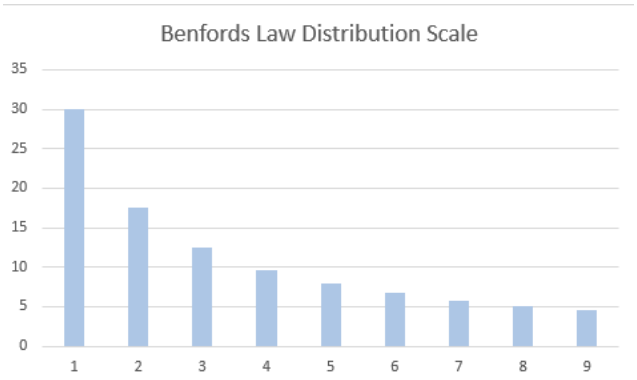

Image 1. Benfords Law diagram

As we can see on the Benford's Law diagram (image 1.), it is plotted histogram of first digits. First digit is the number 1, which occurs more often than number 2, following less occurrence of number 3, following less occurrence of number 4, then 5, going to the number 9 which occurs at least [6].

Table 1. Expected Frequencies

Digit	Expected Frequencies based on Benford's Law		
	1st place	*2nd place*	*3rd place*
0		.11968	.10178
1	.30103	.11389	.10138
2	.17609	.19882	.10097
3	.12494	.10433	.10057
4	.09691	.10031	.10018
5	.07918	.09668	.09979
6	.06695	.09337	.09940
7	.05799	.09035	.09902
8	.05115	.08757	.09864
9	.04576	.08500	.09827

In the table 1 we can see calculated values of leading digits where difference between first number and second number are 50% and then decremented based on formula (1) calculation.

IV. IMPLEMENTATION OF BENFORD'S LAW WITH SQL SERVER R SERVICES

To implement statistical computation Microsoft as a leader in Databases has brought together R language as statistical tool with databases. Combination of statistical language within database present huge step for data analysis in terms of quickly and safely handle the sensitive data.

Existing infrastructure of database server including roles, security and database, will help us operationalize the results. These results will be processed with R language using SQL Server enterprise capability to cope with large set of data [7].

R is widely being used with machine learning. Machine learning models are being used as learning patterns with having data sets provided, with little or no help from humans. This is extremely helpful with automatic detection of fraudulent data, while performance of transaction storing is not being affected. Also it is important to know domain that we are preparing the algorithm for, in order not get bad results. As people can learn bad habits, also can badly architected machine learning model.

We have used data set from one bank with more one million transactions in them. In matter of data sensitivity, this data has been replaced with real account names with numeric ids. Other data like amount and sums are real.

In order to have easily readable data, we will use diagrams where we compare the values. This is being done with Benford's Law diagram (image 2.), where we can easily spot anomalies in data.

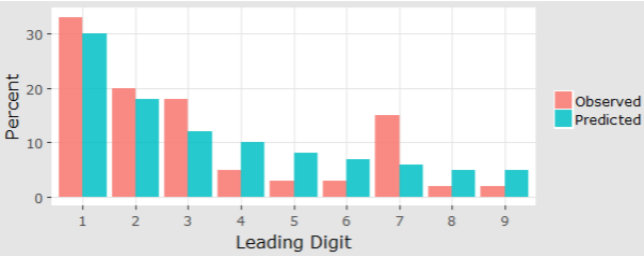

Image 2. Observed vs Predicted values diagram

Here we will present example framework how data is being processed in order of fraud detection (image 3.). First, we collect the data or receive transaction coming to our system.

In our example, data is coming from transaction database (RDBMS) as primary source. Sample data was imported as CSV file into SQL Server table.

This data will be first filtered through set of objects that will have implemented formulas and algorithms. Set of objects that we are using in this example are stored procedures and user defined functions. Reason is that these objects are already within secure environment, where the observed data already resides.

```
Transactions
AS
( SELECT AccountId, Amount
            ,
            cast(round(
            CASE
                WHEN Amount >=1000000000 THEN Amount / 1000000000
                WHEN Amount >=100000000 THEN Amount / 100000000
                WHEN Amount >=10000000 THEN Amount / 10000000
                WHEN Amount >=1000000 THEN Amount / 1000000
                WHEN Amount >=100000 THEN Amount / 100000
                WHEN Amount >=10000 THEN Amount / 10000
                WHEN Amount >=1000 THEN Amount / 1000
                WHEN Amount >=100 THEN Amount / 100
                WHEN Amount >=10 THEN Amount / 10
                WHEN Amount <10 THEN Amount
             END,0,1 ) as int) AS Numbers
            ,Count(1) OVER (partition by AccountId) as Transaction
    FROM [dbo].[Transaction])
```

Results | Messages

AccountId	Numbers	FreqeuncyOfOccurence
1	1	66
1	2	51
1	3	51
1	4	9
1	5	18
1	6	18
1	7	15
1	8	7
1	9	4
2	1	159
2	2	94

Image 3. Preparing data for checking

We will use user defined function where previously source table will be processed in the way of sorting frequency occurrence of leading digits. For example, if we have amount paid on account, we will prepare data for further inspection by finding number of occurrence for each account with specific leading digit. As we can see on image 3., account 1 for leading digit 1 has 66 occurrences, for digit 2 account 1 has 51 occurrences, etc.

Then data is going through verification and check without being stored in final destination. Final destination in our example is table in data storage. This verification will be processed by fast implementation, in our case it will be done with R language. This implementation is being done in staging faze.

```
EXEC sp_execute_external_script
@language=N'R',
@script='N
library(reshape2);
dataset=dcast(InputDataSet, AccountId ~ Numbers, value.var ="FreqeuncyOfOccurence");
OutputDataSet=cbind(
dataset,
## Benfords Law formula that will process the data with given baseline
apply(dataset[,-1],1,function(xx) chisq.test(xx,p=(log(1+1/1:9))/log(10)))$p.value));
colnames(OutputDataSet) <- subset(OutputDataSet,PRvalue<Baseline);
',
@input_data_1=N'
SELECT AccountId, Numbers, FreqeuncyOfOccurence
FROM TransactionStage;
',
@params=N'@Baseline float',
@Baseline=@Baseline
WITH RESULT SETS
((AccountId VARCHAR(50), Digit1 INT, Digit2 INT, Digit3 INT,
Digit4 INT, Digit5 INT, Digit5 INT, Digit6 INT, Digit7 INT, Digit8 INT, Digit9 INT,
PRValue FLOAT));
```

Image 4. Passing data to R language for processing

On image 4. is T-SQL code written together with R language, where R code is being invoked as external script [8] by calling system stored procedure *sp_execute_external_script* and passing required parameters. This code is being used as source data for reporting purposes and alerting purposes. In R code we have defined input or source data (*input_data_1*) from our user function (*TransactionStage*), then we have Benford's Law formula equation [9] and we have output values as result from equation and input data. We have also *Baseline* parameter for filtering results for probability value *PRValue*. *PRValue* is value calculated from chi-square [10] test library in R language. *Baseline* parameter is used for adjusting the level of accuracy in numeric irregularities. In this case we have set the value to 0.1 (*PRValue*<*Baseline*). We will show on final report how data will be trough graph easily spotted the difference between observed and predicted values (image 2.).

In the case of possible fraud, system is being alerted. These data will be "flagged" and provided in reports that are easy readable for further investigation. We must keep the all data processed including flagged ones, because there are data that can be false positive or commonly called "false alarm". This means anomaly has occurred but it is not necessary a fraud. These transactions will be inspected by additional checks and verifications. Other data is being stored to designating tables in database storage.

Here we have a diagram of data flow on global perspective.

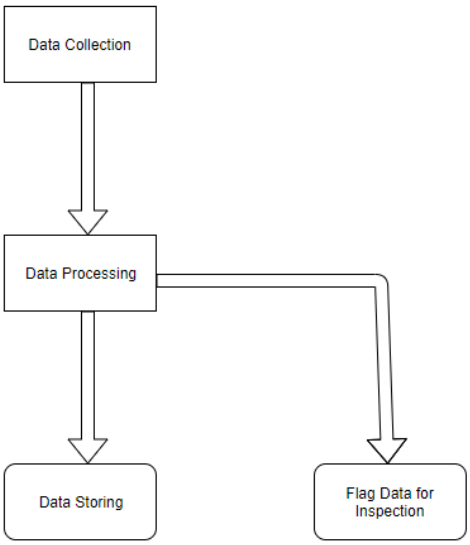

Image 5. Data inspection diagram

Data that are suspicious is being processed for reporting purposes [11]. When data, which was previously flagged by the system, is audited then it will be through the system approved or disapproved. Disapproved data will be stored in special tables for history analytics.

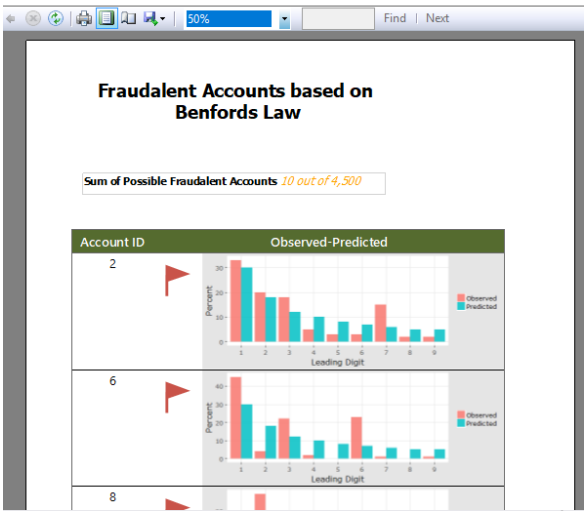

Image 6. Reporting based on finding of fraudulent data

As we can see on the image 6. account with ID 2 has anomaly easily spotted in Benford's diagram on leading digit number 7, where observed value is different than predicted. On the report is also added indicator as red flag and green flag, in the case of call of parametrized report, so we can check all data with simple overview without going in depth of chart values.

Testing data set have million transactions with 4500 accounts where we have identified 10 accounts with possible fraudulent behavior. This has saved us hours of work, if not even days.

V. CONCLUSION

Here, with implementation of Benford's Law with SQL Server Database and R language, we have shown small partition of what advanced databases can do in area of data analysis. We have verified the benefits of combining technologies such as RDBMS and statistical programing language R, that are proven in efficiently storing and retrieving data, performing cleaning, exploration and manipulation of data, and biggest advantage is being done at single place.

Data is being processed in fast way using powerful database engine within secure environment, with no need for external resources and additional authentications.

What is certain that now with possibilities of advanced databases, forensic analytics is really fast-paced advancing in providing solutions for future challenges and Benford's Law is one important part of it.

REFERENCES

[1] Lavion, Didier; et al. "PwC's Global Economic Crime and Fraud Survey 2018" (PDF). PwC.com. Retrieved 18 December 2018.

[2] FBI.gov. Federal Bureau of Investigation. "2017 INTERNET CRIME REPORT" 27,November 2018.

[3] U.S. Department of Justice, "Four Romanian Nationals Charged with Allegedly Participating in Multimillion Dollar Scheme to Hack into and Steal Credit Card Data from U.S. Merchants", Office of Public Affairs, Press Release Number:11-1598

[4] F. Benford, The law of anomalous numbers, Proc. Amer. Philos. Soc. 78 (1938) 551-572.

[5] S. Newcomb, Note on the frequency of use of the different digits in natural numbers, Amer. J. Math. 4 (1881) 39-40.

[6] M. J. Nigrini, Benford's Law: Applications for forensic accounting, auditing, and fraud detections (John Wiley & Sons, 2012).

[7] Tomaz Kastrun, Julie Koesmarno, SQL Server 2017 Machine Learning Services with R: Data exploration, modeling, and advanced analytics, ISBN-13: 978-1787283572,

[8] Denis Rothman, Artiőicial Intelligence By Example: Develop machine intelligence from scratch using real artificial intelligence use cases, ISBN-13: 978-1788990547

[9] Raghav Bali, R Machine Learning By Example, ISBN-13:978-1784390846

[10] https://www.statisticssolutions.com/using-chi-square-statistic-in-research/ accesed on 10.12.2018

[11] https://docs.microsoft.com/en-us/sql/advanced-analytics/r/sql-server-r-services?view=sql-server-ver15 accessed on 10.12.2018.