I.J. Mathematical Sciences and Computing, **2019, 1, 42-52**
Published Online January 2019 in MECS (http://www.mecs-press.net)
DOI: 10.5815/ijmsc.2019.01.04

***Available online at** http://www.mecs-press.net/ijmsc*

Deplyoing Advance Data Analytics Techniques with Conversational Analytics Outputs for Fraud Detection

Sunil Kappal

GH5 and 7/775 Second Floor Paschim Vihar,India, New Delhi-87

Received: 11 October 2017; Accepted: 19 July 2018; Published: 08 January 2019

Abstract

This paper outlines the application of various classification methods and analytical techniques to identify a potential fraud. The aim of this document is to showcase the usefulness of such classification and analytical techniques for fraud detection. Considering the fact that there are hundreds of statistical methods and procedures to perform such analysis. In this paper, I would like to present a hybrid fraud detection method by using the Bayesian Classification technique to identify the risk group; followed by Benford's Law (The Law of First Digit) to detect a fraudulent transaction done by the identified risk group. Though this analysis focuses on the healthcare dataset, however, this technique can be replicated in any industry setup. Also, by adding the Voice of the Customer data to these classification and statistical methods, makes this analysis even more powerful and robust with improved accuracy.

Index Terms: Data Mining, Benford's Law, Bayesian Classification Method, Conversational Analytics, Interaction Analytics.

1. Introduction

Conversational Analytics refers to extracting and mining information from the media files (audio files) capturing the voice of the customer with metadata. This technology helps to make the audio files searchable. Whereas data mining refers to extracting information from large-scale datasets. There are numerous techniques like regression, decision trees, neural nets etc. exists. Here, this document will only cover few data mining techniques which would be considered important to detect a potential fraud and they are I) Naïve Bayes

Sunil Kappal. Tel.: +91-9013830386
E-mail address: skappal7@gmail.com / datageek7@gmail.com

Classification method to identify risk group, and ii) Benford's Law for assessing how those groups are not following the Benford's Law curve of the first digit. Further, the output of Benford's Law will be validated using the chi-square test to identify if the variation is significant or not.

As the world is getting more tech savvy and advancements made in the information technology especially in the healthcare industry has opened areas in data mining and machine learning. Within the area of data mining one technique which has gained a lot of popularity as well as skepticism among the auditors and fraud detectives is Benford's Law or "The Law of First digit

In the past some researchers in Canada used the Benford's Law distribution to detect anomalies within the claims amount data for one of the healthcare organization. In this paper we will understand the mechanics of this technique and will also look at its practical usage on some random claims amount data. However, nobody till yet has ever used Benford's Law in conjunction with Naïve Bayes Classification method.

2. Related Work

Benford's Law (1881 Simon Newcomb): The recent work of Mark j Nigrini, PhD, a professor at the College of New Jersey and an author of the book Forensic Analytics (Wiley) describes how this technique can be used to identify spurious patterns and biases in the financial data. Based on the article written by Canadian Capitalist on the 12th April 2010 indicated that the Canada Revenue Agency employs Benford's Law to flag Tax Cheats for further scrutiny.

As per the article written by Robin Wigglesworth on the 21st of April 2016 states that Deutsche Bank's financial data scientists developed a model based on the Benford's Law theory, Deutsche Bank's Javed Jussa wrote that companies not confirming to the Benford's Law may exhibit some sort of irregularities.

What is Benford's Law?

Benford's Law is a probability distribution with strong bearing to financial frauds and anomalies. There has been a lot of research that has undergone related to this area of fraud detection technique (refer to the above source links).

Being a mathematical formula Benford's Law specifies or indicates the probability of leading digit sequences appearing in a set of data. Let's understand what is meant by *leading digit sequence* based on the below data set.

(1) $\boldsymbol{S = \{213, 212, 122, 21, 124, 14, 2154, 129, 12, 128, 63, 1\}}$

There are twelve data entries in the above set of data where set is denoted as S. The digit sequence "21" (referred to as first and second position) appears 4 times. Hence, the probability of the first two digits being "21" is 4/9 ≈ 0.44. This probability is calculated out of 9 as only 9 entries have that position within the data set. The formula for Benford's Law is:

(2) $\boldsymbol{P(D = d) = log10(1 + 1d)}$, where $\boldsymbol{P(D = d)}$ is the probability of observing the digit sequence d in the first 'y' digits and where d is a sequence of 'y' digits.

Requirements for Benford's Law

In order to apply the above stated equation successfully there are certain data requirements that needs to be met before employing this technique which are:

1. Data with values from disparate set of distributions
2. No built in maximum and minimum values or cut off values
3. Skewed data where mean is greater than the median

4. The data should have more small values than large ones
5. Numbers should not consist of assigned numbers like Telephone number, Zip codes, SSNs

Na ïve Bayes Classification Method

It is a classification methodology based on Bayes Theorem assuming independence amongst the predictors.

1. It assumes that the features are not correlated with each other
2. Being easy to use and understand it can be used with large data sets.

Naive Bayes model is known for outperforming other complex and sophisticated classification models.
The Na ïve Bayes Algorithm:

(3) $\boldsymbol{p(c|x) = p(x|c)p(c) \,/\, p(x)}$
(4) $\boldsymbol{p(c|x) = p(x1|c) * p(x2|c) * - - - - * p(xn|c) * p(c)}$

Where p = probability, c= class, x = predictor variables
There are different types of Na ïve Bayes classifiers and they primarily differ based on their assumptions that they make with regards to their distribution of $p(x|c)$.
Na ïve Bayes is considered as a superior algorithm compared to its counterparts like decision tree, neural nets and other sophisticated algorithms.
It is worth mentioning that the decoupling of the class conditional feature distributions allows Na ïve Bayes algorithm to estimate each distribution independently as one dimensional distribution which in turns helps to get rid of dimensionality issue.

3. Applying the Fraud Detection Methodology

The data mining process of Fraud Identification is divided into 4 phases, which are very critical for this piece of analysis Refer to Figure 1. Also, Google has invented a patent that automatically identifies as probable fraud via its voice verification engine.

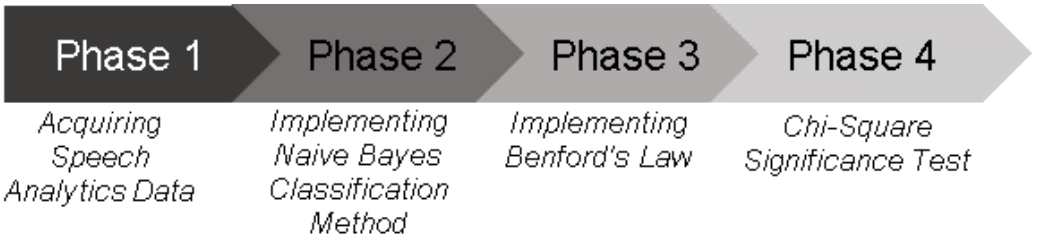


Fig.1

Phase 1: Acquiring Conversational Analytics Data

As we know that Conversational analytics allow its users to query the media files to identify the emerging topics. Therefore, this phase revolves around creating set of queries (structured rules) to identify potential fraud indicating discussions within the Conversational analytics application. Refer to Figure 2 for various fraud indicating scenarios.

Identifying Fraud Scenarios using Conversational analytics tool

Fraud Scenarios	Speech Analytics Data	Call Example	Analytics Insight
Providers billing for services not provided	Voice of the customer complaining that they have bee charged for the "service not rendered"	"I have been charged for Cancer screening but this service has not been rendered to me"	Multiple Service not rendered % by provider to isolate outliers
Providers charging more than peers for the same services	Voice of the Member complaining about overcharging	"I received this service/test last month at a lower cost than this month"	Overcharge complaints by Providers and location
Provider Administering more expensive equipments that are not medically necessary	Member inquiring on the number of procedures on their statement	"I only received a blood test for my cholesterol, I don't know why other tests are listed on my statement."	Quantify Providers with these types of concerns
Provider conducting medically unrelated procedure and service	Voice of the agent this code is not supported	"Please check with your provider as this code is not accepted"	Quanityf Providers with high percentage of "Code Not Supported" issues
Benefits Assignment	Voice of the Member or Agent stating the benefits were assigned without consent	"I am not sure if my benefits are assigned to some one else"	Quantify Members with Benefits Assignment Issues

Fig.2

Once the above scenarios are created in the form of a query *(A query is a rule that helps to make the conversations searchable and reportable. For Example: search for the mentions of "service not rendered", "Outdated CPT code used" etc. and this query will bring all the mentions of such topics).* It further helps the analyst to fetch datasets related to the fraud scenarios with additional metadata to prepare that information for the next phase of Na ïve Bayes classification process.

Phase 2: Implementing Na ïve Bayes Classification Model to the Conversational Analytics Output

Na ïve Bayes Classification model is not only a supervised learning method, but it is also a statistical method for classifying scenarios that may have a high or a low probability of being classified as a fraud outcome. NB Classification model allows to calculate the uncertainty about a particular outcome (in this case it is being a "Fraud" or a "Non-Fraud" outcome). Naïve Bayes is known for its robustness towards the noise in the input data. (Please refer to section B of this document to read more about Na ïve Bayes)

Here is the Bayesian Classification method to predict the probability of a fraud instance. Using the output classification results as "Fraud" or a "Non-Fraud "scenario (as defined by the Conversational analytics structured rules / query(s)). In this phase we aim to pick the scenario with the highest probability of being a fraud Instance.

Once we have that scenario we will pick the specific features related to that scenario to fetch the data and plug that information into the Benford's Law Distribution.

The above table shows the Conversational analytics outputs along with the metadata associated with that conversation in a grid format. This output will be further treated in a spreadsheet program to facilitate the classification process post discretizing the data.

In order to calculate the probabilities based on various conditions (may also be stated as "based on various dimensions of the data") to pick the instance with the highest probability of being a fraud scenario the analyst might have to use statistical software. However, for the purposes of this paper let's look at a couple of conditions as an example that we will pursue throughout this paper to understand the usage of Bayesian theorem's output with the Benford's Law of First Digit distribution curve.

Record	Structured Rule Set	Metadata_Fields			
	Outcome	Association	Age_Bins	Provider Type	Female
1	Fraud	In-Network	30 to 35	Dispensor	Female
2	Fraud	In-Network	56 to 60	Chiropractor	Male
3	Non-Fraud	In-Network	56 to 60	Doctor	Male
4	Non-Fraud	Out-Network	41 to 45	Chiropractor	Male
5	Non-Fraud	Out-Network	30 to 35	Dispensor	Male
6	Fraud	In-Network	30 to 35	Dispensor	Male
7	Fraud	Out-Network	36 to 40	Dispensor	Female
8	Fraud	Out-Network	36 to 40	Doctor	Male
9	Fraud	In-Network	36 to 40	Doctor	Male
10	Fraud	In-Network	36 to 40	Dispensor	Male
11	Non-Fraud	In-Network	46 to 50	Homeopath	Female
12	Non-Fraud	Out-Network	56 to 60	Homeopath	Female
13	Fraud	Out-Network	30 to 35	Doctor	Male
14	Fraud	Out-Network	30 to 35	Dispensor	Female
15	Non-Fraud	In-Network	46 to 50	Dispensor	Female
16	Non-Fraud	Out-Network	30 to 35	Doctor	Female
17	Fraud	In-Network	46 to 50	Homeopath	Male
18	Fraud	In-Network	56 to 60	Dispensor	Female
19	Fraud	In-Network	41 to 45	Doctor	Male

Fig.3

Example:

What is the probability of an **In-Network** provider who is a **Homeopath** a **Female** within the **Age_Bin** of **36** to **40** of being a fraud perpetrator, compared to a provider who is **Out-Network**?

To answer the above question we have to first convert the information *(shown in the above sample table Figure 3)* presented in the grid into a frequency table.

Fraud	Non-Fraud	Total
1000	2775	3775
26%	74%	

Association	Fraud	Non-Fraud
In-Network	375	1777
Out-Network	625	998

Gender	Fraud	Non-Fraud
Female	675	1000
Male	325	1775

Age_Bins	Fraud	Non-Fraud
30 to 35	200	390
36 to 40	150	430
41 to 45	270	398
46 to 50	145	458
51 to 55	135	655
56 to 60	100	444

Provider Type	Fraud	Non-Fraud
Chiropractor	266	655
Dispensor	265	664
Doctor	250	658
Homeopath	219	798

Fig.4

The above tables in figure 4 is the first step to calculate the probability. This table shows the count of various attributes (that may act as various conditions in which a fraud instance might happen) under the "Fraud" and "Non-Fraud" scenarios that we defined using the structured rules in the phase 1 of this methodology.

The second step is to calculate the likelihood %. To calculate these percentages, we have to first calculate % distribution of each condition within the Fraud and Non-Fraud scenarios.

Working Example: In-Network Providers

In-Network Fraud % = $375/1000 = 22\%$
Age_Bin 36 to 40 Fraud % = $150/1000 = 15\%$
Provider Type Homeopath Fraud % = $219/1000 = 22\%$
Gender Female Fraud % = $675/1000 = 68\%$

Similarly calculate the Non-Fraud % for the above attributes:

In-Network Non-Fraud % = $1777/2775 = 64\%$
Age_Bin 36 to 40 Non-Fraud% = $430/2775 = 15\%$
Provider Type Homeopath Non-Fraud % = $798/2775 = 29\%$
Gender Female Non-Fraud% = $1000/2775 = 36\%$

Once we have the above results we will calculate the Likelihood % for a fraud scenario by multiplying each attribute under Fraud scenario with the overall fraud %.

Example Calculation (In-Network):

L Fraud=In-Network % * Age_Bin% * Provider Type % * Gender% * Fraud % = Likelihood % Fraud

L Non-Fraud=*I*n-Network % * Age_Bin% * Provider Type % * Gender% * Non-Fraud % = Likelihood % Non-Fraud

Results based on the above calculations:

L Fraud = $38\% * 15\% * 22\% * 68\% * 26\% = 0.22\%$
L Non-Fraud = $64\% * 15\% * 29\% * 36\% * 74\% = 0.75\%$
Note: L = Likelihood

To calculate the Fraud and Non-Fraud probabilities we just need to divide the likelihood % of "Fraud %" with the sum of Fraud and Non-Fraud %. i.e.
0.22%/sum (0.22%+0.75%) = 23 % (Fraud Probability given the conditions applied)

Similarly, to calculate Non-Fraud Probability:

0.75%/sum (0.22%+0.75%) = 77 % (Non-Fraud Probability given the conditions applied)

Working Example: Out-Network Providers

Out-Network Fraud % = 625/1000 = 63%
Age_Bin 36 to 40 Fraud % = 150/1000 = 15%

Provider Type Homeopath Fraud % = 219/1000 = 22%
Gender Female Fraud % = 675/1000 = 68%

Similarly calculate the Non Fraud % for the above attributes:

Out-Network Non-Fraud % = 998/2775 = 36%
Age_Bin 36 to 40 Non-Fraud% = 430/2775 = 15%
Provider Type Homeopath Non-Fraud % = 798/2775 = 29%
Gender Female Non-Fraud% =1000/2775 = 36%

Once we have the above results we will calculate the Likelihood % or a fraud scenario by multiplying each attribute under Fraud scenario with the overall fraud %.

Example Calculation (Out of Network Providers):

L Fraud=Out-Network % * Age_Bin% * Provider Type % * Gender% * Fraud % = Likelihood % Fraud
L Non-Fraud=Out-Network % * Age_Bin% * Provider Type % * Gender% * Non-Fraud % = Likelihood % Non-Fraud

Results based on the above calculations:

L Fraud $= 63\% * 15\% * 22\% * 68\% * 26\% = 0.36\%$
L Non-Fraud $= 36\% * 15\% * 29\% * 36\% * 74\% = 0.42\%$
Note: L = Likelihood

To calculate the Fraud and Non-Fraud probabilities we just need to divide the likelihood % of "Fraud %" with the sum of Fraud and Non-Fraud %. i.e.

$0.36\%/sum\ (0.36\% + 0.42\%) = 46\ \%$ **(Fraud Probability given the conditions applied)**

Similarly, to calculate Non-Fraud Probability:

$0.42\%/sum\ (0.36\% + 0.42\%) = 54\ \%$ (Non-Fraud Probability given the conditions applied)

The above calculation based on the Naïve Bayes Classification method clearly indicates that an Out-Network provider with above conditions is at a greater risk of being a fraud perpetrator compared to an In-Network. In theory this makes perfect sense. Providers who are associated with a particular payer will be far more cautious while billing a Member compared to an Out-Network service provider.

With this piece of information, we can now move on to our phase 3 where the intent is to further scrutinize the Out-Network Provider's claims amount data based on the lead that we have got from our Phase 1 and Phase 2 investigation.

Phase 3: Implementing the Benford's Law on the Out-Network Provider Claims Data.

In this paper we have already understood some basics about the Benford's Law, The Law of the First Digits and its usage and effectiveness based on its usage in vivid setups *(please refer to the introduction section of this paper to better understand the usage of Benford's Law)*

This phenomenon of "First Digit Law" also gained a lot of popularity and attention when it was used in the television crime dramas like Numbers and Running Man Season 2.

This law can be often used with as an indicator of fraudulent data and can assist with auditing financial data. Benford's distribution is non-uniform, with digits starting with 1 is more likely to appear than the larger digits like 9.

Below is the Benford's distribution table image:

First Digits	*Probability*
1	0.301
2	0.176
3	0.125
4	0.097
5	0.079
6	0.067
7	0.058
8	0.051
9	0.046

Fig.5

Therefore, based on the discovery that we made so far, we will be looking at Out-Network female provider's claims amount data where the age group is in between 36-40 years and where the service type provided is homeopathy.

Now we will look at the distribution to ensure that the data qualifies the requirements stated in the Introduction section of Benford's Law in this paper.

Note: To perform this part of the analysis I used the Minitab statistical tool to understand the Data Distribution. I also used the R Package Rattle (GUI) to perform the Benford's Distribution analysis by parsing 35K claims entries. Excel spreadsheet program was also used to perform Bayesian Classification calculations.

Understanding the Distribution

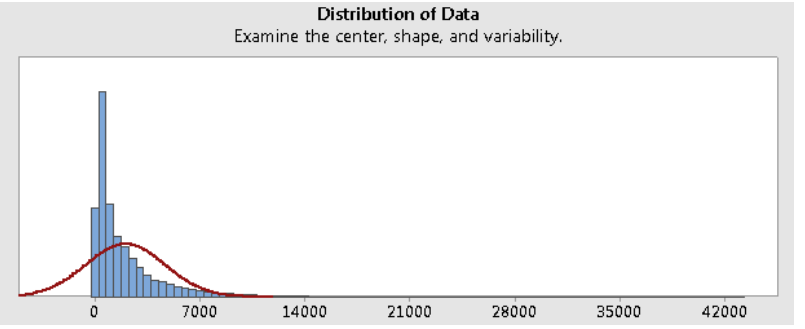
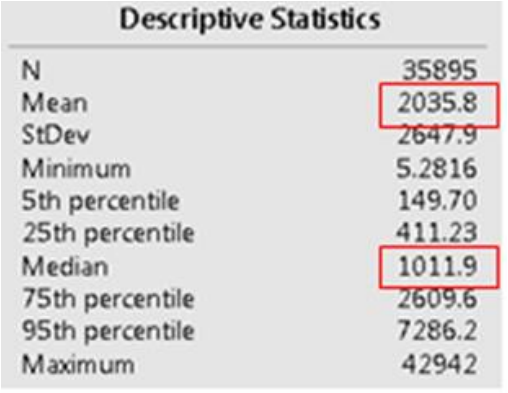


Fig.6

Based on the 35K claims that I parsed *(keep the Bayesian Classification criteria in mind).* It was pretty evident that the data for the Out-Network Homeopathy Female providers with the age group 36-40 is right skewed. *(Refer to Figure 6)*

Also, the mean is greater than the median which is also meeting the Benford's Law analysis criteria. *(Refer to Figure 7)*



Descriptive Statistics	
N	35895
Mean	2035.8
StDev	2647.9
Minimum	5.2816
5th percentile	149.70
25th percentile	411.23
Median	1011.9
75th percentile	2609.6
95th percentile	7286.2
Maximum	42942

Fig.7

Post establishing that this data is apt for the Benford's Law technique. Let's see how this technique can uncover some interesting patterns within this claims submission data set. The goal of applying Benford's Law is to understand how "natural" these claim submissions are.

The Process:

Sample the Data: "The more the merrier" as this expression says the more observations the better. However, for illustration purposes I am using 35K claims submission out of ~100K claim submission data.

Parse the leading digit – As discussed above that Benford's Law focuses on the leading digits in sets of naturally occurring numbers. The actual claims amount, whether it is $100, $200, $300 etc. is unimportant and this can be achieved by using the Excel "Left" formula to get the lead digit for each dollar amount or R based packages like Rattle can perform this analysis in no time.

Create Frequency Distribution – The next step is to create the frequency distribution of the leading digit that have been parsed from the sample data. This can be achieved by either using the "count if" formula or by using the pivot function within MS Excel.

Compute the Distribution – Per the Benford's Law ~30.1% percent of lead digits should be a 1 and 9 should be the least i.e. ~5% keeping this as a standard in mind compute the actual distribution of the leading digits. Once the distribution is computed compare it with Benford's Law distribution and identify any potential outliers. Refer to the image below to see how the end results will look like.

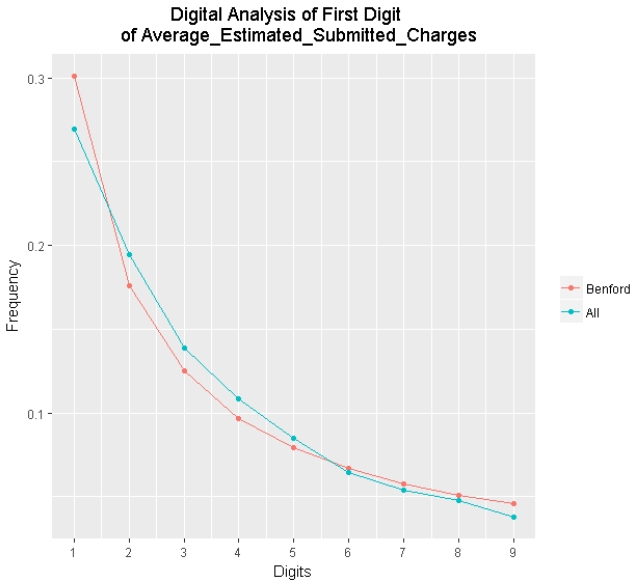


Fig.8

The above graphs clearly indicate that there is an unusual amount of claim submissions with leading digits 2,3,4 & 5. This clearly highlights a potential manipulation, error or even a fraud. Auditors can further apply tests like Chi Square test which acts as a "goodness of fit" statistic that measures how well the data distribution complies with the hypothetical distribution explained in the theory.

Phase 4: Chi Square Test of Significance

It is pretty clear by looking at the observed data that it varies or does not match with the expected values very well, they question that remains unanswered is "how far off these numbers are?".

To answer this question statistically, we implemented the Chi Square Test to provide some guidance to make that decision. Chi Square Test can be easily deployed within Excel spreadsheets by using the CHITEST function. Alternatively, Minitab statistical application can be used to perform this analysis.

This test enables the auditors to test the "Goodness of Fit" i.e. it helps to measure how well the distribution from a sample matches the hypothesized distribution per the Benford's Law theory.

The chi-square statistic from Excel's CHITEST or Minitab's Chi Square Test Goodness of Fit indicates the likelihood that the actual values in the sample follow the prescribed (Benford) distribution. High values such as 93% indicate a good match between actual and expected distributions, while small values such as 3 percent indicate a poor match.

Most business data, such as count of sales, costs, accounts receivables, payments, and even the buyer's street addresses, can be considered as logical or naturally occurring numbers. By connecting the first-digit frequency distribution of naturally occurring data with Benford's probability curve, auditors can easily spot possible data flaws or fraudulent transactions. Hence, when used appropriately, Benford's law can be a valuable and low-costing tool for identifying spurious transactions for advance analysis.

4. Conclusion

In this paper we studied how we can utilize multiple advanced analytical techniques like Naïve Bayes,

Benford's Law and Chi Square Tests to make better decisions. Also, we saw the importance of emerging technologies like Conversational Analytics to provide the most efficient way to parse the audio data and include much talked about "Voice of the Customer' data to performing meaningful analysis

References

[1] http://as.wiley.com/WileyCDA/WileyTitle/productCd-1118152859.html
[2] http://www.canadiancapitalist.com/cheating-on-taxes-and-benfords-law/
[3] https://www.ft.com/content/afeea0be-01b9-11e6-99cb-83242733f755
[4] http://www.techrepublic.com/article/Conversational-analytics-why-the-big-data-source-isnt-music-to-your-competitors-ears/
[5] http://www.cs.unb.ca/~hzhang/publications/FLAIRS04ZhangH.pdf (H. Zhang (2004)
[6] https://www.researchgate.net/publication/241401706_The_Effective_Use_of_Benford's_Law_to_Assist_in_Detecting_Fraud_in_Accounting_Data
[7] https://faculty.uoit.ca/fletcherlu/LuECML05.pdf (Fletcher Lu and J Efrim Boritz)
[8] https://www.ijariit.com/manuscripts/v4i3/V4I3-1165.pdf (Sai Kiran, Jyoti Guru, Rishabh Kumar, Naveen Kumar, Deepak Kataria, Maheshwar Sharma
[9] http://www.iosrjournals.org/iosr-jce/papers/Vol18-issue4/Version-4/F1804042632.pdf (Carolyne Milgo)
[10] https://www.forbes.com/sites/bernardmarr/2016/08/08/the-amazing-potential-of-voice-analytics/
[11] https://www.datasciencecentral.com/profiles/blogs/fraud-analysis-using-speech-analytics-output-with-monte-carlo
[12] https://www.ey.com/Publication/vwLUAssets/ey-audio-analytics/$FILE/ey-audio-analytics.pdf
[13] http://www.speechtechmag.com/Articles/Editorial/FYI/Market-Spotlight-Banking-on-Speech-to-Prevent-Fraud-109084.aspx
[14] https://www.ft.com/content/fd711f44-dc4d-11e3-a33d-00144feabdc0
[15] https://patents.google.com/patent/WO2014107141A1

Authors' Profiles

Sunil Kappal works as an Advanced Analytics Consultant. He has more than 21 years of experience in Data Analytics, Business Intelligence, Statistical Modeling, Predictive Models and implementation of Six Sigma Methodologies in various setups.

Sunil has also earned a certificate of Distinction in the field of Healthcare Innovation and Entrepreneurship from Duke University of North Carolina, Project Management Specialization from University of California Irvine Extension, Business Analytics Specialization from Wharton School of Business. He has delivered multiple lectures at the University of Texas Dallas on the usage of various advanced analytical techniques and machine learning practices. Sunil was also recently invited as a guest speaker at the Symbiosis Institute of Operations Management to talk on Big Data and Machine Learning.

How to cite this paper: Sunil Kappal,"Deplyoing Advance Data Analytics Techniques with Conversational Analytics Outputs for Fraud Detection", International Journal of Mathematical Sciences and Computing (IJMSC), Vol.5, No.1, pp.42-52, 2019.DOI: 10.5815/ijmsc.2019.01.04