Journal of Engineering Sciences
Vol 11, Issue 7, July /2020
ISSN NO:0377-9254

DETECTION OF FRAUDULENT BEHAVIOUR IN WATER CONSUMPTION USING MACHINE LEARNING ALGORITHMS

[1]GHANTASALA VENU GOPAL, [2]V.BALAJI

[1]Assoc. Professor, Dept. of CSE, PBR VITS, Kavali, A.P, India.

[2]MCA, Dept. of MCA, PBR VITS, Kavali, A.P, India.

Abstract – Data mining is a powerful tool widely used by organizations to enhance their businesses and gain a competitive advantage over their competitors. The data mining process helps in extracting and analysing various data patterns, information or trends from large databases. Various data mining techniques are available to conduct the data mining process. Data mining techniques are used in a variety of applications, one of which is the detection and prevention of different types of frauds. Although there is existing research on data mining and various data mining techniques that can be used to detect and identify different types of frauds, there is little research that synthesizes various facets of fraud that uses the data mining techniques. This research explores the use of two classification techniques (SVM and KNN) to detect suspicious fraud water customers. The SVM based approach uses customer load profile attributes to expose abnormal behaviour that is known to be correlated with non-technical loss activities. The data has been collected from the historical data of the company billing system. To deploy the model, a decision tool has been built using the generated model. The system will help the company to predict suspicious water customers to be inspected on site.

Keywords – Data Mining, KNN (K Nearest Neighbour), SVM (Support Vector Machine) and NTL (Non-technical losses).

I. INTRODUCTION

Water is an essential element for the uses of households, industry, and agriculture. Jordan, as several other countries in the world, suffers from water scarcity, which poses a threat that would affect all sectors that depend on the availability of water for the sustainability of activities for their development and prosperity [3].

The mentioned Irregularities known as non-technical losses (NTLs). NTLs originating from electricity theft and other customer malfeasances are a problem in the electricity supply industry. [4] NTL is a problem in water supply industry too because of the similarity between water and electricity distribution systems in depending on meter technology and load profiling concept.

NTLs include the following activities

1) Losses due to faulty meters and equipment.
2) Tampering with meters so that meters record low rates of consumption.
3) Stealing by bypassing the meter or otherwise making illegal connections.

4) Arranging false readings by bribing meter readers.
5) Arranging billing irregularities with the help of internal employees by means of such subterfuges as making out lower bills, adjusting the decimal point position on the bills, or just ignoring unpaid bills.

Fraud is a serious problem face information system that implemented in various domains. Credit card transactions as a financial system branch had a total loss of 800 million dollars of fraud in U.S.A. and 750 million dollars in U.K. in the year 2004 [1]. In the area of health care according to transparency international [2], the total expenditure exceeds the amount of 3 trillion euro worldwide. That size in the health care industry induces several actors in the field to make a profit by using illegal means, forbidden financial operation committing health care fraud.

A. Problem Statement

This water crisis situation has been aggravated by the rapid population growth and mismanagement. Efforts of the water suppliers to improve water and sanitation services are faced by managerial, technical and financial determinants and the limited amount of renewable freshwater resources.

B. Objective of the project

1) Well-known data mining techniques to build a suitable model to detect suspicious fraudulent customers.
2) Depending on their historical water metered consumptions.
3) Water supplying companies incur significant losses due to fraud operations in water consumption.
4) This model introduces an intelligent tool that can be used to detect fraud customers and reduce their profit losses.

C. Existing System

Earlier research shows that various classification technique in the detection of suspected fraud customers and corrupted measurement meters. The available literature related to detecting the fraudulent activities of Non-Technical Loss in water consumption is limited in comparison to other sectors such as electricity consumption and financial issues. A research was conducted by Humid in the Arabic region related to suspicious water consumption activities. Humid used data mining techniques to discover fraudulent water consumption by customers in Gaza city. The historical data of water consumption was used as a training dataset to build the intelligent model. The author focused on using support vector machine SVM classifier to detect the fraudulent activities.

D. Disadvantages of Existing System

1) Very less knowledge of existing fraud behaviour leads to less accurate result.
2) Imbalanced dataset gives less reliable result.
3) Less reliable result needs more human involvement in verification phase.

E. Proposed system

In this project, we will apply the data mining classification techniques for the purpose of detecting customers' with fraud behaviour in water consumption. We will use decision tree

based classifiers to assemble the best classification models for detecting suspicious fraud customers. The models were built using the customer's historical metered consumption data. The model is trained using the default classifiers parameters. Results inspection will be done using confusion matrix to get the accuracy of the classifiers model. The customers will be classified into frauds or non-frauds category. This model introduces an intelligent tool that can be used to detect fraud customers and reduce their profit losses. The suggested model helps saving time and effort of employees of by identifying billing errors and corrupted meters. With the use of the proposed model, the water utilities can increase cost recovery by reducing administrative Non-Technical Losses (NTL's) and increasing the productivity of inspection staff by onsite inspections of suspicious fraud customers.

F. Advantages of Proposed System

1) Accurate result can increase cost recovery of water suppliers.
2) Higher Performance.
3) Able to get the better insights from the result.
4) Able to increase the productivity of inspection staff.

G. Challenges in Data Mining to detect Fraud:

1) There are millions of transactions each day. To extract large amount of data from a database requires highly efficient techniques.
2) The data or information is noisy.
3) Data labels are not immediately available. Frauds or intrusions usually aware after they have already happened.
4) It is hard to track user's behaviours. All types of users (good users, business, and fraudsters) change their behaviours frequently [5] [6].

II. LITERATURE SURVEY

A. *Real Application On Nontechnical Losses Detection*

The main objective of data mining techniques is the evaluation of data sets to discover relationships in information. These relationships may identify anomalous patterns or patterns of frauds. Fraud detection is a very important problem in telecommunication, financial and utility companies.

B. Artificial Neural Networks And Support Vector Machines For Water Demand Time Series Forecasting

Water plays a pivotal role in many physical processes, and most importantly in sustaining human life, animal life and plant life. Water supply entities therefore have the responsibility to supply clean and safe water at the rate required by the consumer. Water plays a pivotal role in many physical processes, and most importantly in sustaining human life, animal life and plant life. Water supply entities therefore have the responsibility to supply clean and safe water at the rate required by the consumer. The modelling of water resource variables is a very active field of study and definitely there still is a lot ofwork to be done. In the initial stages, modelling of water resource

variables was done using the traditional statistical models. The modelling of water resource variables is a very active field of study and definitely there still is a lot of work to be done. In the initial stages, modelling of water resource variables was done using the traditional statistical models.

C. *Machine Learning Algorithm For Efficient Power Theft Detection Using Smart Meter Data*

Electricity Theft is one of the major problems of electric utilities. The dishonest electric power users produce financial loss to the utility companies. Machine learning algorithm is used for this purpose the trustworthiness of customer is verified and is selected for theft program. This analysis is carried out by tweaking the actual smart meter data to create fraudulent data.

D. *An Approach To Detection Of Tampering In Water Meters*

Meter tampering is nothing but fraudulent manipulation which explains a service that is not billed by a utility company.it is a lack of consumption for the utility company and a main problem because they represent an important loss of income. The algorithms were generated and program after data mining process from the database of the company. They detect three types of consumption patterns.

III. PROPOSED WORK

The Module descriptions of the methodology are as follows:

system architecture is shown below in Figure 1.

A. *Customer Data*

The customers those who are willing to get water through agencies are registered with system. The only ways for user to consume water by customers are through this registration. Customer request for admin to get water and to generate bills.

B. *Verify Feedback*

Bills are generated after checking the limit by on field executives after check the limit. The quantity that they consumed must be equal to noted details by admin. The fraud details can be check through this process. The bills were uploaded after this and find the fraudulent among the customers.

C. *Action Against Fraudulent*

The fraud customers who illegally consumes more water than they used or may be requires can be found by admin and bills also verified by them. Fraud details are set to block by the user and let them not provide any more water to them again and the details handover to cops to punish them with legally.

D. *Graph Analysis*

The graphs are handy to understand the data and based on this analysis admin can find the fraud customers. The business gradually improves as per their understand of where exactly problem arises and to find the place improve and lack. This will gives the clear picture about the current and past picture from the dataset.The

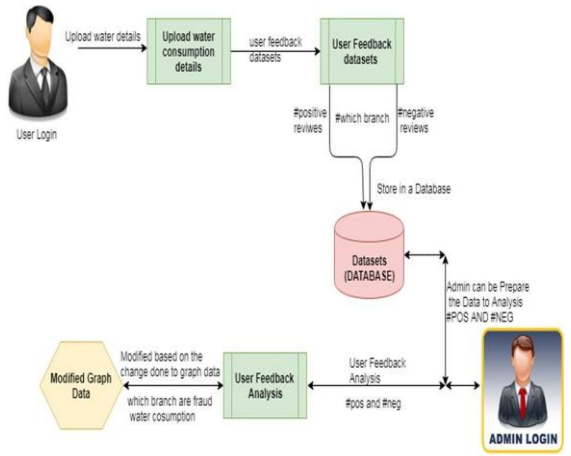

Figure 1: System Architecture

The customers those who are willing to get water through agencies are registered with system. Upload the water consumption details through this registration. On field executives checking the limit according to branch wise and collect the feedback from the customers through this process. Admin can prepare the data to analyse the positive and negative feedbacks. Based on their feedback Fraud details are set to block by the user and let them not provide any more water to them again. Admin can find the fraud customers where exactly problem arises. This will gives the clear picture about the current & past picture from the dataset.

E. Algorithm

In this work the experiments are performed on two important and well known classification algorithms K-Nearest Neighbor (KNN) &SVM are applied to the water customer's dataset which is taken from the Executives. There accuracy is obtained by evaluating the datasets. Each algorithm has been run over the training dataset and their performance in terms of accuracy is evaluated along with the prediction done in the testing dataset. It is one of the world's most popular and most used open source data mining solutions. It has a comfortable user interface, where in a process view analyses are configured. It uses a modular concept, where respective operators are used in the analysis process. These operators have input and output ports through which the operators can communicate with the other operators in order to receive input data or pass the data and generated models over to the following operator. In this way, the entire analysis process creates a data flow. K-Nearest Neighbor makes predictions based on the outcome of the K neighbors closest to that point. Therefore, to make predictions with KNN, we need to define a metric for measuring the distance between the query point and cases from the examples sample.

F. KNN Algorithm

1) Training algorithm:

For each training example $(x,f(x))$, add the example to the list training examples.

2) Classification algorithm:

- Given a query instant x_q to be classified as
  - Let $x_1 \ldots x_k$ denote the k instances from training_examples that are nearest to x_q.
  - Return

$$\sqrt{\sum_{i=1}^{n} (q_i - p_i)^2}$$

G. SVM Algorithm

input :
training dataset D
number of kSVM models T
$rdims$ random attributes used in the kSVM model
k local models in the kSVM model
hyper-parameter of RBF kernel function γ
C for tuning margin and errors of SVMs
output:
T kSVM models

1 begin
2 #pragma omp parallel for
3 for $t \leftarrow 1$ to T do
4 Sampling a bootstrap D_t (train set) from D using $rdims$ random attributes)
5 $kSVM_t = kSVM(D_t, k, \gamma, C)$
6 end
7 return $krSVM - model = \{kSVM_1, kSVM_2, \ldots, kSVM_T\}$
8 end

IV. RESULTS

The results of the system are shown below:

USERNAME	BRANCH	NO OF WATERCANES	NO OF LITER	AMOUNT	WHICH PERSON DELIVERD	RECEIPT	BOOKING DATE	DELIVERY DATE
AKANSHYA DASH	Iyyapanthangal	10	220	300	sankar	yes	Jan. 2, 2018	Jan. 5, 2018
SRIRAM	Porur	15	330	450	kishore	no	Jan. 3, 2018	Jan. 6, 2018
GOURAV VERMA	Guindy	17	374	510	ram	yes	Jan. 4, 2018	Jan. 7, 2018
SANJAY KV	AshokNagar	25	425	750	samy	no	Jan. 5, 2018	Jan. 8, 2018
INIYAN SENTHIL KUMAR	Triplicane	18	396	540	anand	yes	Jan. 6, 2018	Jan. 9, 2018
S KRISHNA KUMAR	Iyyapanthangal	13	286	390	siva	no	Jan. 7, 2018	Jan. 10, 2018
ANHAD SARAN	Porur	12	264	360	kumar	yes	Jan. 8, 2018	Jan. 11, 2018
ASHUN MITESH KOTHARI	Guindy	48	1056	1440	illyas	no	Jan. 9, 2018	Jan. 12, 2018
SHUBHAM SEMWAL	AshokNagar	6	132	180	velu	yes	Jan. 10, 2018	Jan. 13, 2018
AJAY KUMAR	Triplicane	9	198	270	saravanan	no	Jan. 11, 2018	Jan. 14, 2018
PRABHAV DOBHAL	Iyyapanthangal	17	374	510	mohan	yes	Jan. 12, 2018	Jan. 15, 2018
KATARI MEGHA VARUN S	Porur	22	484	660	balu	no	Jan. 13, 2018	Jan. 16, 2018
CHAKKA.DHEERAJ KUMAR	Guindy	21	462	630	sathis	yes	Jan. 14, 2018	Jan. 17, 2018

Figure 2: User information page

Figure 2 shows the user information page where user's name, branch, no of water canes he used etc., is displayed.

NAME	BRANCHES	RATING	MOBILENUMBER	FEEDBACK
AKANSHYA DASH	Iyyapanthangal	5	8754732040	#Iyyapathangal_branch branch is good really good water supplies
SRIRAM R	Porur	2	8939479762	#Porur_branch branch good but litlle time late delivered
GOURAV VERMA	Guindy	1	9791574650	#Guindy_branch branch brach is worst
SANJAY KV	AshokNagar	4	9003118430	#AshokNagar_branch branch is nice branch on time delivered
INIYAN SENTHIL KUMAR T	Triplicane	5	8248145339	#Triplicane_branch is some late , service is bad
S KRISHNA KUMAR	Iyyapanthangal	5	8939751320	#Iyyapathangal_branch branch is good really good water supplies

Figure 3: User feedback page

Figure 3 is the user feedback page where user has given feedbacks for the water and service he had taken

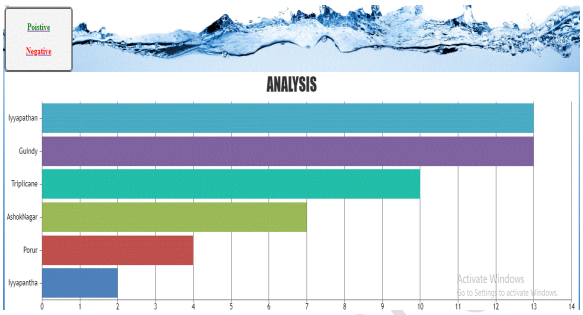

Figure 4: Positive Analysis Graph

Figure 4 shows the positive analysis graph where it shows the analysis of positive feedbacks.

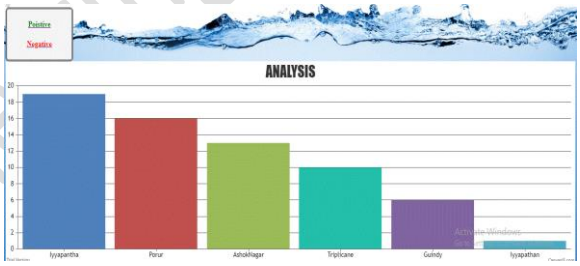

Figure 5: Negative Analysis Graph

Figure 5 shows the negative analysis graph where it shows the analysis of negative feedbacks.

V. CONCLUSION

In this research, we applied the data mining classification techniques for the purpose of detecting customers' with fraud behaviour in water consumption. We used SVM and KNN classifiers to build classification models for detecting suspicious fraud customers. The models were built using the customers' historical metered consumption data; the Cross Industry Standard Process for Data Mining

(CRISP-DM). This phase took a considerable effort and time to pre-process and format the data to fit the SVM and KNN data mining classifiers.

VI. FUTURE WORK

The conducted experiments showed that a good performance of Support Vector Machines (SVM) and K-Nearest Neighbours (KNN) had been achieved with overall accuracy around 70% for both. In Future accuracy of the same can be improved with the help of improved techniques. With the use of the proposed model, the water utilities can increase cost recovery by reducing administrative Non-Technical Losses (NTL's) and increasing the productivity of inspection staff by onsite inspections of suspicious fraud customers.

REFERENCES

[1] AihuaShen, Rencheng Tong "Application of classification Models on credit card Fraud Detection", 2007.

[2] AnastassiosTagaris "Implementation of Prescription Fraud Detection Software Using REDBMS Tools and ATC Coding", 2009.

[3] N/A, "Jordan Water Sector Facts & Figures, Ministry of Water and irrigation of Jordan". Technical Report. 2015.

[4] N/A, "Water Reallocation Policy, Ministry of Water and irrigation of Jordan". Technical Report. 2016.

[5] C. Ramos, A. Souza , J. Papa and A. Falcao, "Fast non-technical losses identification through optimum-path forest". In Proc. of the 15th Int. Conf. Intelligent System Applications to Power Systems, 2009, pp.1-5.

[6] E. Kirkos, C. Spathis and Y. Manolopoulos, "Data mining techniques for the detection of fraudulent financial statements", Expert Systems with Applications, 32(2007): 995–1003.

[7] Juan Ignacio, Carlos Leon "Real Application on Nontechnical losses detection", The 2011 World Cogress in Computer Science, Computer Engineering, and Applied Computing (WORLDCOMP 11), Volume: The 2011 International Conference on Data Mining.

[8] Ishmael S. Msiza, Fulufhelo V. Nelwamondo and TshilidziMarwala "Artificial Neural Networks and Support Vector Machines for Water Demand Time Series Forecasting", 2007.

[9] Jeyaranjani J and, Devaraj D "Machine Learning Algorithm for efficient power theft detection using smart meter data" International Journal of Engineering & Technology, 7 (3.34) (2018) 900-904.

[10] "Monedero, Félix Biscarri, Juan I. Guerrero, MoisésRoldán, Carlos León "An Approach to Detection of Tamperingin Water Meters" Procedia Computer Science 60 (2015) 413 – 421.

AUTHORS

Ghantasala Venu Gopal has received his B.Tech in Computer Science and Engineering from Institution of Engineers India, Kolkata in 2002 and M.Tech degree in Computer science and Engineering from Allahabad Agricultural Institute, Allahabad in 2006. He is pursuing Ph.D from Rayalaseema University, Kurnool. He is dedicated to teaching field from the last 18 years. He has guided 35 P.G and 15 U.G students. His research areas included Data Mining, Image Processing. At present he is working as Associate Professor in PBR VITS, Kavali, Andhra Pradesh, India.

V. Balaji has Received his B.Sc Degree in Computer Science from MSR Degree COllege, Kavali affiliated to Vikrama Simhapuri University, Nellore in 2017 and pursuing PG Degree in Master of Computer Applications (M.C.A) from PBR VITS, Kavali affiliated to JNTU Anantapur, Andhra Pradesh, India.