Detection Of Fraudulent Behaviour In Water Consumption using Data Mining Based Model

Ramesh Babu M[1], Dr. P Pedda Sadhu Naik[2]
[1]M. Tech(CSE), [2]Professor & HOD, Dept. of CSE
Dr. Samuel George Institute of Engineering & Technology, Markapuram, A.P., India.

Abstract: *Data mining is defined as a process used to extract usable data from a larger set of any raw data. Fraudulent behavior in drinking water consumption is a significant problem facing water supplying companies and agencies. This behavior results in a massive loss of income and forms the highest percentage of non-technical loss. Finding efficient measurements for detecting fraudulent activities has been an active research area in recent years. Intelligent data mining techniques can help water supplying companies to detect these fraudulent activities to reduce such losses. This paper explores the use of two classification techniques (SVM and KNN) to detect suspicious fraud water customers. The main motivation of this research is to assist Yarmouk Water Company (YWC) in Irbid city of Jordan to overcome its profit loss. The SVM based approach uses customer load profile attributes to expose abnormal behavior 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. The accuracy of the generated model hit a rate of over 74% which is better than the current manual prediction procedures taken by the YWC. 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, SVM, KNN, CRISP-DM.

1. INTRODUCTION:

Data mining is the process of uncovering patterns and finding anomalies and relationships in large datasets that can be used to make predictions about future trends. The main purpose of data mining is to extract valuable information from available data. 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.

According to Jordan ministry of water and irrigation, this issue always has been one of the biggest barriers to the economic growth and development for Jordan. This crisis situation has been aggravated by a population increase that has doubled in the last two decades. Efforts of the ministry of Water and irrigation to improve water and sanitation services are faced by managerial, technical and financial determinants and the limited amount of renewable freshwater resources.

To address these challenges, Jordan ministry of water and irrigation as in many other countries is striving, through the adoption of a long-term plan, to improve services provided to citizens through restructuring and rehabilitation of networks, reducing the non-revenue water rates, providing new sources and maximizing the efficient use of available sources. At the same time, the Ministry continues its efforts to regulate the water usage and to detect the loss of supplied water. Water supplying companies incur significant losses due to fraud operations in water consumption. The customers who tamper their water meter readings to avoid or reduce billing amount is called a fraud customer. In practice, there are two types of water loss: the first is called technical loss (TL) which is related to problems in the production system, the transmission of water through the network (i.e., leakage), and the network washout.

The second type is called the non-technical loss (NTL)which is the amount of delivered water to customers but not billed, resulting in loss of revenue. The management of the Yarmouk Water Company (Jordan) has a significant concern to reduce its profit losses, especially those derived from NTLs, which are estimated over 35% in the whole service area in the year 2012[1]. One major part of NLT is customer's fraudulent activities; the commercial department manages the detection processes with the absence of an intelligent computerized system where the current process is costly, not effective nor efficient. NTL is a serious problem facing Yarmouk Water Company (YWC). In 2012 the NTL reached over 35%, ranging from 31% to 61 according to districts, which results in a loss of 13 million dollars per year[2]. Currently, YWC follows random inspections for customers, the proposed model in this paper provides a valuable tool to help YWC teams to detect theft customers, which will reduce the NTL and raise profit. Literature has abundant research for Non-Technical Loss (NTL) in electricity fraud detection, but rare researches have been conducted for the water consumption sector. This paper focuses on customer's historical data which are selected from the YWC billing system. The main objective of this work is to use some well-known data mining techniques named Support Vector Machines (SVM) and K-Nearest Neighbor (KNN) to build a suitable model to detect suspicious fraudulent customers, depending on their historical water metered consumptions.

2. LITERATURE REVIEW:

This section reviews some of the applications of data mining classification techniques in fraud detection in different areas such as Detection of Fraudulent Financial Statement, Fraud Detection in Mobile Communication Networks, Detecting Credit Card Fraud, and Fraud Detection in Medical Claims. For example, **Kirkos et al. [4]** proposed a model for detecting fraud in financial statements, where three data mining classifiers were used, and namely Decision Tree, Neural network and Bayesian Belief Network. **Shahine et al.[5]** Introduced a model for credit card fraud detection; they used decision tree and support vector machines SVM. In addition, **Panigrahi et al. [6]** proposed a model for credit card fraud detection using a rule-based filter, Bayesian classifier, and Dempsters-Shafer adder. **Carneiro et al.[7]** developed and deployed a fraud detection system in a large e-tail merchant. They explored the combination of manual and automatic classification and compared different machine learning methods. **Ortega et al.[8]** proposed a fraud detection system for Medical claims using data mining methods. The proposed system uses multilayer perceptron neural networks (MLP). The researchers showed that the model was able to detect 75 fraud cases per month. **Kusaksizoglu et al.** [9] introduced a model for detecting fraud in mobile communication networks. The results showed that the Neural Networks methods MLP and SMO found to give best results. In addition, **CHEN et al.[10]** proposed and developed an integrated platform for fraud analysis and detection based on real time messaging communications in social media. **Nagi et al. [11]**[12][13] introduced a technique for classifying fraudulent behavior in electricity consumption. The proposed method is a combination of two classification algorithms, Genetic Algorithm (GA) and Support Vector Machine (SVM), which yield a hybrid model (named GASVM). The technique processed the past customers 'consumption profile to reveal abnormal consumptions of the customers of

Tenaga Nasional Berhad (TNB) electricity utility in Malaysia. After an investigation, four categories were found (change of tenant, replaced the meter, faulty meter, and abundant house). An expert system was designed to remove such customers by considering characteristics that distinguish between these four customer's categories and theft customers. This intelligent system hit rate reached 60% where they indicated that this model raised the detection of fraud activities from 3% using current procedures in the company to, a hit rate of 60% after onsite inspection. C.**Ramos et al.[3]** presented optimum-path forest classifier to detect fraud customers in electricity consumption. The classifier was compared with other robust classifiers ANN, SVM-Linear, SVM-RBF. The results showed that OPF accuracy is similar to SVM-RBF but superior in training time, which enables real-time classification. The other two classifiers accuracy were not comparable.

León et al.[14] suggested a model that can reveal electricity fraud customers. The data was obtained from the Spanish Endesa Company. The classification model is based on Generalized Rule Induction (GRI) and Quest Decision Tree methods. Furthermore, they introduced two statistical estimators, which are used to weigh customers' trend and the non-constant consumption. The model assists in the identification of abnormal consumption which may arise from abnormal with no fraud so they can easily be re-billed, and fraud customers where adequate procedures can take place. The accuracy of the model reached 22%.

Filho et al. [15] implemented decision tree classification technique in the detection of suspected fraud customers and corrupted measurement meters. They used five months customers' consumption data, where a classification of customers to fraud and non-fraud were applied. The technique raised the hit rate a hit rate of 5% using current techniques to 40%.

Jiang et al.[16] suggested an approach using Wavelettechniques and a group of classifiers, to automatically detect fraud customers in electricity consumption. The wavelet technique was used to express the properties of the meter readings. These readings were used to build models using several classifiers, based on the assumption that abnormalities in consumption appear when fraud occurs. **Cabral et al.[17]** introduced a fraud detection system using data mining techniques for high-voltage electricity customers in Brazil. The used techniques used customers' historical data to be compared with the current consumption and present the possible fraud status. The customers are marked as below regular consumptions and used to be investigated by company inspection team.

De Faria et al.[18] presented a use case of forensics investigation procedures applied to detect electricity theft based on tampered electronic devices.**Viegas et al.** [19] provided an extended literature review with an analysis on a selection of scientific studies for detection of non-technical losses in the electric grid reported since 2000 in three well know databases: Science Direct, ACM Digital Library, and IEEE Xplore. **Coma-Puig et al.** [20]developed a system that detects anomalous meter readings on the basis of models that are built using some machine learning techniques using past data. The system detects meter anomalies and fraudulent customer behaviour (meter tampering), and it is developed for a companythat provides electricity and gas. **Richardson et al.[21]** introduced a novel privacy preserving approach to detecting energy theft detection in smart grids. Malicious behaviour is detected by calculating the Euclidean distance between energy output measurements from installation over a day. These distances are then clustered to identify outliers and potentially malicious behaviour. 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. For example, **Monedero et al.[22]** developed a methodology consists of a set of three algorithms for the detection of meter tampering in the Emasesa Company (a water distribution company in Seville)

3. EXISTING SYSTEM

Literature has abundant research for Non-Technical Loss (NTL) in electricity fraud detection, but rare researches have been conducted for the water consumption sector. Water supplying companies incur significant losses due to fraud operations in water consumption. The customers who tamper their water meter readings to avoid or reduce billing amount is called a fraud customer. In practice, there are two types of water loss: the first is called technical loss (TL) which is related to problems in the production system, the transmission of water through the network (i.e., leakage), and the network washout. The second type is called the non-technical loss (NTL) which is the amount of delivered water to customers but not billed, resulting in loss of revenue. To address these challenges, Jordan ministry of water and irrigation as in many other countries is striving, through the adoption of a long-term plan, to improve services provided to citizens through restructuring and rehabilitation of networks, reducing the non-revenue water rates, providing new sources and maximizing the efficient use of available sources. At the same time, the Ministry continues its efforts to regulate the water usage and to detect the loss of supplied water.

Drawbacks of Existing Problem:

The customers who are recorded in the municipality of Gazaas water theft have been manually marked with the label 'YES'in the field Fraud Status, and the rest were labeled as 'NO.' The data was filtered to remove inconsistency and noise cases. The data is normalized using z-score to fit the SVM model. Similarto all fraudulent cases, the data classes are unbalanced. SVM has a parameter set that can be used to the weight and balance the two classes. The ratio for each class was calculated to compute the parameter values. The values were multiplied by 100 to achieve a suitable ratio weight for SVM. The random subsampling was applied to the samples with class label "NO" to weight the classes for KNN and ANN. The results showed that SVM classifier has the best accuracy over the other two classifiers for the balanced samples either with consumption feature alone or with all selected features, The season consumption dataset was the most suitable for monthly and yearly datasets because it takes into consideration the seasonal consumption changes where the intelligent model raised the hit rate from 10% random inspection to 80%. This research showed that unbalanced samples gave the best accuracy for all classifiers.

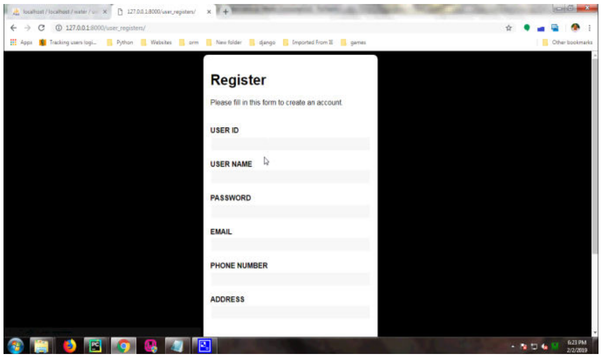
4. PROPOSED SYSTEM

This paper focuses on customer's historical data which are selected from the YWC billing system. The main objective of this work is to use some well-known data mining techniques named Support Vector Machines (SVM) and K-Nearest Neighbor (KNN) to build a suitable model to detect suspicious fraudulent customers, depending on their historical water metered consumptions. The CRISP-DM (Cross Industry Standard Process for Data Mining) was adopted to conduct this research. The CRISPDM is an industry standard data mining methodology developed by four Companies; NCR systems engineering, DaimlerChrysler AG, SPSS Inc. and OHRA. The CRISP-DM model consists of business understanding, data understanding, data preparation, model building, model evaluation and model deployment. To extract the fraud customers' profile, a new table is created containing the client's number, the water consumption, and a new attribute for fraud class.

Benefits of Proposed System:

This attribute is filled with a value of 'YES'. Another table for the normal clients is created, and the fraud class attribute is filled with the value "NO". The two tables are then consolidated into one table containing the customer ID, consumption profile, and fraud class attributes. To filter the data, some preprocessing operations were performed such as Eliminate redundancy, Eliminate customers having zero consumption through the entire period, Eliminate new clients who are not present during the whole targeted period, and Eliminate customers having null consumption values. Filtering the data resulted in a reduced original dataset of the non-fraud customer to 16114 record and the fraud customers to 647 records.

5. EXPERIMENTAL RESULTS:



Register

Please fill in this form to create an account

USER ID

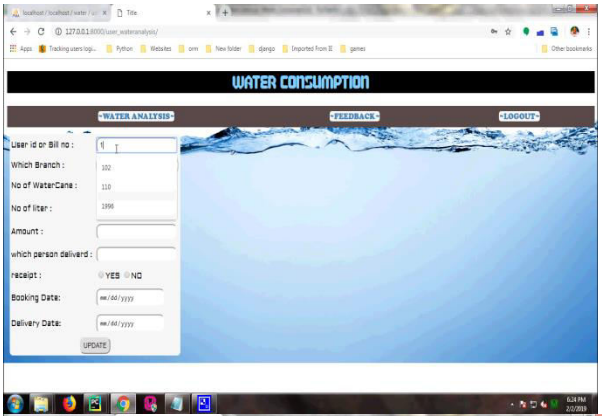
USER NAME

PASSWORD

EMAIL

PHONE NUMBER

ADDRESS



User id or Bill no : 1

Which Branch : 102

No of WaterCane : 110

No of liter : 1998

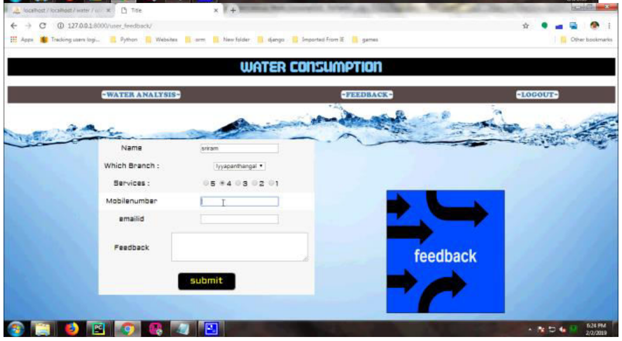
Amount :

which person deliverd :

receipt : ○ YES ○ NO

Booking Date: mm/dd/yyyy

Delivery Date: mm/dd/yyyy



Name sriram

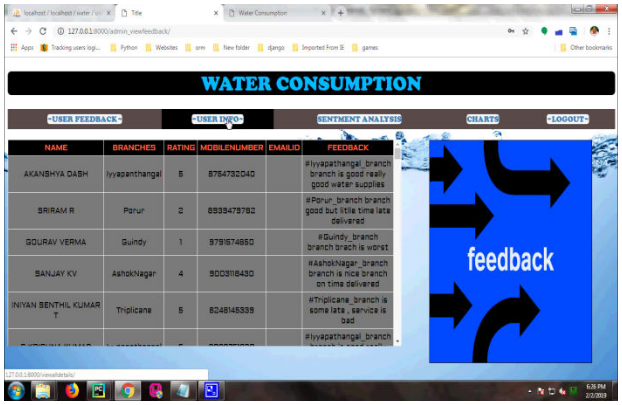
Which Branch : Iyyapanthangal ▾

Services : ○ 5 ● 4 ○ 3 ○ 2 ○ 1

Mobilenumber

emailid

Feedback



NAME	BRANCHES	RATING	MOBILENUMBER	EMAILID	FEEDBACK
AKANSHYA DASH	Iyyapanthangal	5	8764732040		#Iyyapathangal_branch branch is good really good water supplies
SRIRAM R	Porur	2	8939479782		#Porur_branch branch good but litlle time late delivered
GOURAV VERMA	Guindy	1	9791674850		#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

6. CONCLUSION

In this Paper, 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). The data used in this research study the data was collected from Yarmouk Water Company (YWC) for Qasabat Irbid ROU customers, the data covers five years customers' water consumptions with 1.5 million customer historical records for 90 thousand customers. This phase took a considerable effort and time to pre-process and format the data to fit the SVM and KNN data mining classifiers. The conducted experiments showed that a good performance of Support Vector Machines (SVM) and 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. The model hit rate is 60%-70% which is apparently better than random manual inspections held by YWC teams with hit rate around 1% in identifying fraud customers. This model introduces an intelligent tool that can be used by YWC to detect fraud customers and reduce their profit losses. The suggested model helps saving time and effort of employees of Yarmouk water 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.

References:

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

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

[3] 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.

[4] 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.

[5] Y. Sahin and E. Duman, "Detecting credit card fraud by decision trees and support vector machines", IMECS, 2011, Vol I, pp. 16 – 18.

[6] S. Panigrahi, A. Kundu, S. Sural and A. Majumdar, "Credit card fraud detection: a fusion approach using dempster–shafer theory and Bayesian earning, information fusion", 2009, 10(4): 354–363.

[7] N. Carneiro, G. Figueira and Costa M., "A data mining based system for credit-card fraud detection in e-tail decision support systems", Decision Support Systems, 2017, 95(C): 91-101.

[8] Ortega P., Figueroa C., and Ruz G. "A Medical Claim Fraud/Abuse Detection System based on Data Mining: A Case Study in Chile", Inproc of DMIN, 2006.

[9] B. Kusaksizoglu, "Fraud detection in mobile communication networks using data mining", Bahcesehir University, The Department of computer engineering, Master Thesis. 2006.

[10] C. Liang-Chun, H. Chien-Lung, L.Nai-Wei, Y. Kuo-Hui and L. Ping- Hsien, "Fraud analysis and detection for real-time messaging communications on social networks", IEICE Trans. Inf. & Syst., 2017,Vol. E100–D, No.10, pp: 2267-2274.

[11] J. Nagi, K. Yap, S. Tiong, S. Ahmed and A. Mohammad. "Detection of abnormalities and electricity theft using genetic support vector machines", In Proc. IEEE TENCON Region 10 Conf., 2008, pp.1-6.

[12] J. Nagi, Mohammad A., Yap K., Tiong S., Ahmed S. "Non-Technical Loss Analysis For Detection Of Electricity Theft Using Support Vector Machines", In Proc IEEE 2nd International Power and Energy Conference, 2008, pp. 907-912.

[13] J. Nagi, K. Yap, S. Tiong., S. Ahmed, M. Mohamad. "Nontechnical loss detection for metered customers", IEEE Transactions on Power Delivery, 2010, 25(2): 1162-1171.

[14] C. León, F. Biscarri, I. Monedero, J. Guerrero, J. Biscarri and R. Millán, "Variability and trend-based generalized rule induction model to ntl detection", IEEE Transactions on Power Systems, 2011, 26(4):1798 -1807.

[15] J. Filho, E. Gontijio, A. Delaiba, E. Mazina., J. Cabral, J and Pinto. "Fraud identification in electricity company customers using decision tree", Systems, Man and Cybernetics, IEEE International Conference, 2004, 4: 3730 – 3734.

[16] R. Jiang, H. Tagiris, A. Lachsz. and M. Jeffrey "Wavelet-based features extraction and multiple classifiers for electricity fraud detection", In Proc. IEEE/PES Transmission and Distribution Conf. Exhibit. 2002.

[17] J. Cabral, J. Pinto. E. Martins and A. Pinto, "Fraud detection in high voltage electricity consumers". 2008.

[18] R. De Faria, K. Ono Fonseca, B. Schneider and S. Nguang, "Collusion and fraud detection on electronic energy meters - a use case of forensics investigation procedures", in 2014 IEEE Security and Privacy Workshops, pp. 65-68.

[19] J. Viegas, P. Esteves, R. Melicio, V. Mendes and S. Vieira, "Solutions for detection of non-technical losses in the electricity grid: a review", Renewable and Sustainable Energy Reviews, 2017, 80: 1256-1268.

[20] B. Coma-Puig, J. Carmona, R. Gavald, S. Alcoverro, and V. Martin, "Fraud detection in energy consumption: a supervised approach". In Proc IEEE Intl. Conf. on DSAA, 2016, pp. 120-129.

[21] C. Richardson, N. Race, and P. Smith, "A privacy preserving approach to energy theft detection in smart grids", 2016 IEEE International Smart Cities Conference (ISC2), Trento, pp. 1-4.

[22] Monedero I., Biscarri F., Guerrero J., Roldán M., and León C. "An Approach to Detection of Tampering

in Water Meters", In Procedia Computer Science, 2015, 60: pp 413-421.