Journal of Ambient Intelligence and Humanized Computing (2023) 14:15327–15342
https://doi.org/10.1007/s12652-019-01649-9

ORIGINAL RESEARCH

Performance analysis of machine learning classifiers for non-technical loss detection

Khawaja MoyeezUllah Ghori[1,2] · Muhammad Imran[3] · Asad Nawaz[2] · Rabeeh Ayaz Abbasi[4] · Ata Ullah[2] · Laszlo Szathmary[5]

Received: 5 June 2019 / Accepted: 12 December 2019 / Published online: 18 January 2020

Abstract

Power companies are responsible for producing and transferring the required amount of electricity from grid stations to individual households. Many countries suffer huge losses in billions of dollars due to non-technical loss (NTL) in power supply companies. To deal with NTL, many machine learning classifiers have been employed in recent time. However, few has been studied about the performance evaluation metrics that are used in NTL detection to evaluate how good or bad the classifier is in predicting the non-technical loss. This paper first uses three classifiers: random forest, *K*-nearest neighbors and linear support vector machine to predict the occurrence of NTL in a real dataset of an electric supply company containing approximately 80,000 monthly consumption records. Then, it computes 14 performance evaluation metrics across the three classifiers and identify the key scientific relationships between them. These relationships provide insights into deciding which classifier can be more useful under given scenarios for NTL detection. This work can be proved to be a baseline not only for the NTL detection in power industry but also for the selection of appropriate performance evaluation metrics for NTL detection.

✉ Khawaja MoyeezUllah Ghori
ghori.moiz@inf.unideb.hu

Muhammad Imran
dr.m.imran@ieee.org

Asad Nawaz
asadnawaz316@gmail.com

Rabeeh Ayaz Abbasi
rabbasi@qau.edu.pk

Ata Ullah
aullah@numl.edu.pk

Laszlo Szathmary
szathmary.laszlo@inf.unideb.hu

[1] Doctoral School of Informatics, University of Debrecen, Debrecen, Hungary

[2] Department of Computer Science, National University of Modern Languages, Islamabad, Pakistan

[3] College of Applied Computer Science, King Saud University, Riyadh, Saudi Arabia

[4] Department of Computer Science, Quaid-i-Azam University, Islamabad, Pakistan

[5] Department of IT, Faculty of Informatics, University of Debrecen, Debrecen, Hungary

1 Introduction

Power supply companies are considered the backbone for any country. These companies use kilometers of lines that can transfer the electricity from the production units to individual meters. With the ever growing need of electrical energy around the world, the power companies are bound to produce sufficient electricity that can fulfill the required amount of energy. Along with the challenge to meet the production requirement, these companies face huge setbacks due to non-technical losses (NTL) in distribution networks. NTL is the loss which may be caused by unintentional meter malfunctioning or intentional fraudulent attempts to bypass meters, slowing down or stopping meters, faulty meter readings or even having an illegal connection. NTL in power industry has shaken many economies worldwide. For example, India loses 4.5 billion USD every year on account of NTL. This loss can range upto 50% of the total electricity produced in developing countries (McDaniel and McLaughlin 2009). The developed countries, including USA and UK, also suffer a loss of $1–$6 billion annually (Alam et al. 2004). May it be unintentional or an intentional NTL, any power supply company wants to minimize it by first detecting it and then addressing it properly. If the detected

NTL is unintentional, the company fixes the problem and if the detected NTL is intentional, it issues different levels of penalties to the fraudsters. The techniques used to detect both NTLs are the same. However, it is unrealistic to remove all NTLs from a power utility company. Some companies involve expert technicians to identify potential NTLs. This includes identifying group of households where electricity consumption has decreased or stopped. The on-site inspectors verify these households for a possible NTL detection. This process of on-site inspection incurs a heavy cost to the company and it is practically impossible to inspect a large number of households. This can only be fruitful if the inspection results in a large number of NTL detection. In reality, the ratio of number of NTL detection to the number of inspections is generally very low for companies (Coma-Puig et al. 2016). This can also be explained due to the fact that people may have got their second homes, or they may be on long vacations etc. Shortlisting them for the inspection will only increase the inspection cost.

Over the past decade, the research community has paid attention to detecting NTL with the collaboration of electric suppliers using machine learning classifiers. This includes using support vector machine (SVM), optimum path forest (OPF), random forest, multi-layer perceptron neural network (NN), *K*-nearest neighbors (KNN), Adaboost, naive bayes, decision trees and deep learning. Training sets containing records of NTLs are used to train the models and test sets are used to evaluate them. The list of fraudsters identified by the classifiers is then used for on-site inspection. The hit ratio of NTL detection using machine learning classifiers is very promising as compared to random guessing of potential fraudsters. To compare the performance of these classifiers, different performance evaluation metrics are used. These metrics can help in shortlisting the classifiers for NTL detection under given scenarios.

A detailed comparative study of performance evaluation metrics is still needed to diagnose the relationship between different metrics when used for NTL detection. These relationships have not been discussed sufficiently in the literature of NTL and can be proved to be a baseline for the selection of appropriate performance evaluation metrics for NTL detection. In Pakistan, electric supply companies perform random on-site inspection to identify theft cases. This is the prime reason for a very small success on NTL detection. This paper helps to improve the hit ratio in an electric supply company in Pakistan by identifying and shortlisting the potential theft cases for on-site inspection using machine learning classifiers. This work is performed on a real dataset containing 80,244 monthly consumption records. It will also help to reduce the on-site inspection cost of the company. This paper first uses three classifiers: random forest, KNN and linear SVM to predict occurrences of NTL in the dataset. Then, we compute 14 performance evaluation metrics across the three classifiers to identify the key scientific relationships between these performance metrics with respect to NTL detection. For the appropriate selection of the classifiers, these relationships are crucial. Therefore, in this paper, we have focused on identifying the key scientific relationships between performance evaluation metrics in the domain of NTL detection as shown in Fig. 1. This work can further be extended to predict potential theft in gas sector. Using the consumption pattern of gas consumers, this set of classifiers along with the performance metrics can be used for the identification of gas theft attempts.

The objectives of this work are as follows:

1. Categorize the state of the art NTL detection schemes and present their comprehensive taxonomy.
2. Identify the strengths and weaknesses of the state of the art methods for NTL detection and identify a pool of performance metrics commonly used for NTL detection.
3. Apply machine learning classifiers for NTL detection, and compute and validate their performances on

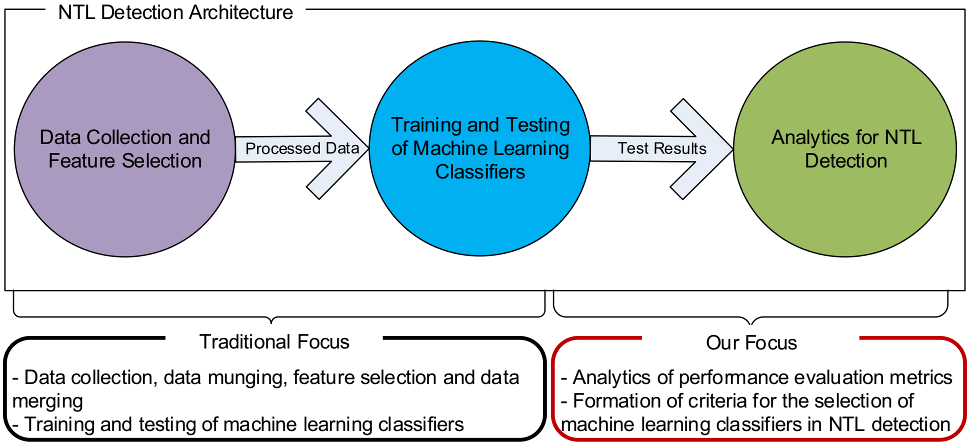
NTL Detection Architecture
Data Collection and Feature Selection
Processed Data
Training and Testing of Machine Learning Classifiers
Test Results
Analytics for NTL Detection
Traditional Focus
- Data collection, data munging, feature selection and data merging
- Training and testing of machine learning classifiers
Our Focus
- Analytics of performance evaluation metrics
- Formation of criteria for the selection of machine learning classifiers in NTL detection

Fig. 1 The NTL Architecture comprising of data collection, feature selection, training and testing of machine learning classifiers and analytics. Note that existing research works mostly focus on NTL detection while analytics of performance evaluation metrics (i.e., our focus in this paper) have often been overlooked

a real dataset containing approximately 80,000 monthly records of electricity consumption.

4. Investigate a pool of the identified performance metrics for NTL detection and highlight the performance metrics that can exhibit the NTL detection in a best and reliable way.

The rest of the paper is as follows: Sect. 2 overviews the existing techniques used for NTL detection. Section 3 introduces the proposed methodology used in this paper for NTL detection. Section 4 describes experiments that are performed on the dataset, the evaluation metrics used and discusses the results. Section 5 describes conclusion and future work. In section 6, acknowledgements are presented.

2 Related work

Big data analytics is frequently used in diverse domains of every-day life. It strives to solve realistic problems by applying machine learning algorithms and data mining approaches. The applications include fraud detection (Jain and Gupta 2019), problem handling of unstructured data (Amalina et al. 2020), disease comorbidity prediction (Lakshmi and Vadivu 2019), Internet of Things (IoT) (ur Rehman et al. 2019), Industrial Internet of Things (IIoT) (ur Rehman et al. 2018), real-time anomaly detection (Ariyaluran Habeeb et al. 2019; Habeeb et al. 2019), preventive medicine using big data (Razzak et al. 2019), and event detection (Saeed et al. 2019).

NTL identification is an application of fraud detection (Han and Xiao 2019). A survey of the existing techniques to detect NTL can be found in Papadimitriou et al. (2017). The study has categorized the techniques handling NTL in to data-oriented, network oriented and hybrid techniques. The data-oriented techniques use consumer's consumption patterns to predict NTL. These techniques can further be divided into supervised, semi-supervised and unsupervised learning paradigms. Supervised learning methods are used when the class label of fraud and no-fraud is provided. Example of supervised learning is SVM. Semi-supervised learning methods are used when only one class label is known and the other label is not definite. Example of semi-supervised learning is anomaly detection. Unsupervised learning methods are used when the class labels are not used at all. Clustering algorithms are the examples of unsupervised learning. Network-oriented techniques include usage of the network data and smart meters which are used to check electric balance with respect to the grid. They have stated that network-oriented techniques are good at detecting NTL in a specific area but fail to identify specific fraudulent consumers. Hybrid techniques use advantages of both techniques where network data is used to locate the fraudulent area and consumption data is used to identify fraudulent consumers. They have listed TP, TN, FP, FN, recall, FPR, recognition rate and Bayesian detection rate as the main performance evaluation metrics. Alongside, they have discussed the roles and responsibilities of the concerned authorities to tackle NTL.

A comprehensive survey for the challenges of NTL detection can be found in Glauner et al. (2017). The authors have compared multiple techniques which are applied in NTL detection. These include expert system, machine learning, SVM, Neural network, fuzzy logic, genetic algorithm, optimum path forest, and rough sets. They have also compared different search techniques for feature selection of customer's master data. The paper identifies some challenges which are still needed to be thoroughly dealt with. For example, the identification of a correct percentage of under sampling of majority class, a need of a thorough comparative study for different techniques dealing imbalance domain, a need of a metric to compare regression with classification problems and creation of a benchmark dataset.

On the basis of the type of data used to detect NTL, the strategies can be categorized into four main types. Consumption-data based techniques involve detecting NTL using only the consumption data collected from the meters installed at the consumer end. Network based techniques involve detecting NTL using the difference between the total electricity supplied and the total electricity billed. These techniques also use the grid data. Hybrid techniques involve using both types of data, that is, consumption data and network data. The fourth category, additional-data based techniques, not only uses consumption data but it also uses other data, for example, climate and temperature data. Efforts have been made in every category to detect NTL. A complete taxonomy for strategies of NTL detection is described in Fig. 2.

2.1 Consumption-data based techniques

Researchers have paid special interest in NTL detection during the past decade. Multiple machine learning techniques have been used to correctly identify NTL. The authors of Nagi et al. (2010) have used SVM to identify NTL in a dataset that is having a highly uneven distribution of class labels. They have claimed a hit rate increase from 3 to 60%. This work is focused on identification of NTL where abrupt changes of users' consumption patterns are found but does not discuss situations where changes are observed gradually. It compares the results of NTL using accuracy and hit rate but does not discuss any relation between them.

Glauner et al. (2016) have used Boolean rules, fuzzy logic and SVM to detect NTL in a dataset of around a million customers while analyzing their monthly consumption patterns. The results show that optimized fuzzy logic and SVM outperformed Boolean rules. They have compared

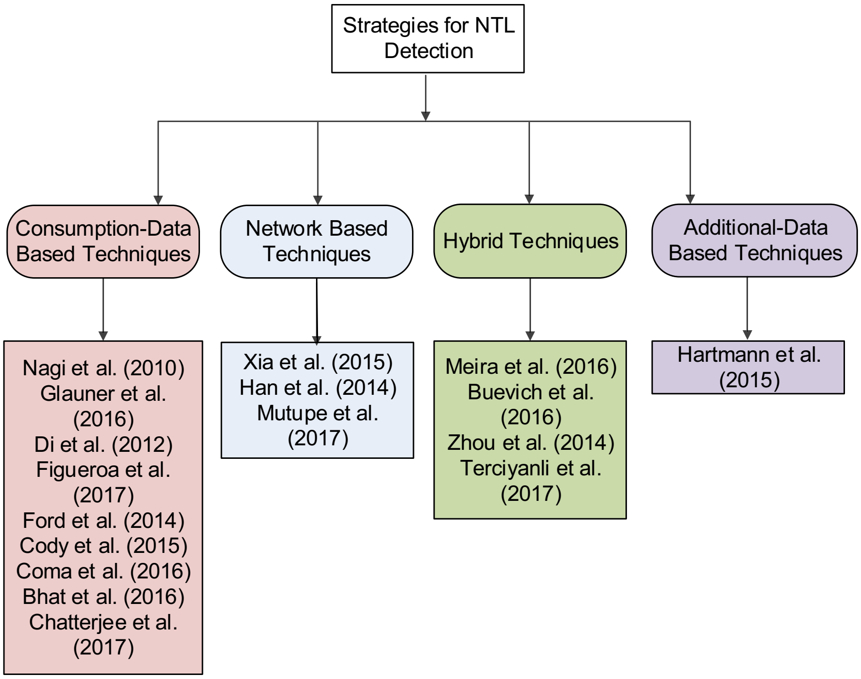

Fig. 2 Taxonomy for strategies of NTL detection

the classifiers with performance evaluation metrics like true positive, true negative, false positive, false negative, recall and specificity. However, the relationships that may exist between these metrics regarding NTL detection are not sufficiently discussed.

A set of classifiers have been used as ensembles to detect frauds in an electricity supply company in Uruguay by Di Martino et al. (2012). They have claimed that a one-class SVM, CS-SVM, Optimum Path Forest (OPF) and C4.5 combined as ensembles have given good measures as compared to applying them individually. The classifiers are compared using accuracy, recall, precision and F value. However, this paper does not discuss the impact of using these metrics for NTL detection.

The authors of Figueroa et al. (2017) have used three classifiers namely linear SVM, non-linear SVM and a multilayer perceptron neural network for NTL detection in a dataset collected from an electric company operating in Honduras. They have used under-sampling and over-sampling strategies to handle the imbalance ratio of fraud and non-fraud instances. Additionally, eight performance evaluation metrics are used to compare performances of the classifiers. The metrics include accuracy, recall, precision, specificity, AUC, F_β, F_1 and Matthews Correlation Coefficient (MCC). However, the paper does not discuss the relationships between these metrics specifically for NTL detection. These relationships may be used to find appropriate combination of metrics for NTL detection.

In Ford et al. (2014), artificial neural networks have been used to predict electric theft detection in a relatively smaller dataset collected from Irish Social Science Data Archive Center. The authors have trained the neural network on three different situations and consequently have stated three observations. One of the observations is that consumer's consumption behavior can be predicted a year ahead. The other observation is that the training of a neural network on the data of three consecutive weeks can predict consumer's consumption behavior for the fourth week. Their final observation is that the consumption patterns can also be predicted in the same weather season. They have used TP, TN, FP and FN to measure the performance of the neural network. The authors extended their work in Cody et al. (2015) to train and test the Irish dataset using M5P decision tree on the same situations. They have used root mean square error (RMSE) to measure the closeness of predicted and actual values. RMSE values are found within the threshold for all three situations.

The authors of Coma-Puig et al. (2016) have used a dataset of a company providing electricity and gas in Spain. They have used a combination of different machine learning techniques to predict NTL in electricity and theft attempts in gas sector. This includes Naive Bayes, KNN, decision trees, neural networks, SVM, random forest and AdaBoost. Their framework has a feature which auto-updates the results of on-field inspection in the database resulting the framework to be adaptive to new theft patterns over a period of time. They have used precision as the only performance evaluation metric.

During the last few years, advancements in deep learning have opened a lot of application areas (Hayat et al. 2019). One of its application areas which still needs attention of research community is NTL detection. The authors of Bhat et al. (2016) have tested convolutional neural network, auto encoders and long short-term memory networks for NTL detection in a relatively smaller dataset containing occurrences of NTL. Experimental results demonstrate that deep learning based strategies have outperformed decision trees, random forest, and neural networks in terms of various performance metrics such as precision, recall, F1 and receiver operating characteristic (ROC) curve.

In Chatterjee et al. (2017), the authors have used deep recurrent neural networks to identify NTL. The data used is related to advanced metering infrastructure (AMI). It is taken from Australian Governments Department of Industry, Innovation and Science. As AMI's data is sequential with respect to time, so recurrent neural network is applied to it. The metric used to evaluate the performance is accuracy which is measured to be 65.3% for a neighborhood. However, it does not use any other performance evaluation metric which may help for a better understanding of NTL detection.

2.2 Network based techniques

In Han and Xiao (2014), the authors have proposed a mathematical expression which calculates the difference between the billed amount of electricity and the total amount of consumed electricity. They have argued that this can help in detecting tempered meters from non-tempered meters. Along with the readings of consumer's meter, an observer meter is also used to calculate the total electricity provided. This approach can be applied in AMI systems as well as in traditional grid systems. A similar approach is used in Mutupe et al. (2017). The authors have used meters at the transformer side and the consumer side. If the total amount of electric power sent by the distributor is not equal to the total amount of electric power received by the consumer, then a possible NTL is marked for on-site inspection. Radio frequency (RF) signals are used to communicate the difference in electric power usage between the consumer and the distribution pole. This work is implemented in Eskom, the electric supply company in South Africa.

Another approach to detect NTL in neighborhood area smart grids is discussed in Xia et al. (2015). The authors have proposed a difference-comparison based inspection algorithm which uses binary inspection tree to calculate the difference in the amount of electricity stolen from a node to its child. The characteristics of binary search tree enables the algorithm to skip large amount of nodes which are useless to check. This helps in quickly identifying malicious meters. The algorithm keeps track of stolen electricity in the associated subtree of a node which helps in probing the next node.

2.3 Hybrid techniques

Feature selection is an important task for the identification of NTL. In Meira et al. (2017), the features are divided into four categories with respect to time, geography, similarity of consumption profile and infrastructure. Random forest, logistic regression and SVM are tested with different proportion of NTL ranging from 10 to 90% across all four categories. Results are compared using area under the curve (AUC). The results obtained from the consumption category are better than the results obtained from infrastructure category. The authors have also claimed that consumption downfall is not the only pattern of NTL rather an increasing consumption pattern can still be a good candidate for NTL. As AUC is the only metric used to evaluate performance of the classifiers, the relationships between different metrics can not be identified for NTL detection.

Buevich et al. (2016) have discussed two different techniques that separate NTL from the overall losses in an electric grid. One of them, the model-driven technique considers the examination of state of meters, packet losses, line losses and consumption of consumers. The other technique, the data-driven one evaluates NTL using a classifier SVM on a synthetic data of different households. The authors have argued that the first technique helps to evaluate the grid, while the second technique gives an estimation of true positive rates (TPR) and true negative rates (TNR) with respect to different levels of NTL. TPR and TNR are the only two performance evaluation metrics compared in Buevich et al. (2016).

Zhou et al. (2014) proposed a load profiling technique which uses advantages of the two approaches for customer classification. One of the approaches is based on geographical location. Customers are grouped together on the basis of similar locality. The other approach is based on similar consumption patterns exhibited by the customers. These customers are then grouped in the same category. The authors have combined these two approaches to categorize customers on the basis of similar customers on similar region using firefly algorithm to detect NTL. They have performed experiments on the data collected from a power supply company in China. Accuracy is the only metric used to evaluate the performance. Thus, no comparison can be made with other metrics used for NTL detection.

A score based approach for NTL detection is applied in Terciyanli et al. (2017). The authors have used three steps for the detection of NTL. The first step comprises of assigning three different scores to each consumer. The first score represents the evaluation of the area in which the consumer is living. The second score represents the change in the usage trend for the consumer. The third score represents the deviation of the monthly consumption from the expected consumption. The second step involves calculating a final score for each consumer using three different weights for the three corresponding scores. As a third step, if the final score is found above a threshold value, an on-site inspection is recommended for a possible NTL detection. This work is performed on a small dataset of an electric supplier in Turkey. However, the paper does not use any of the known performance evaluation metrics.

2.4 Additional-data based techniques

Hartmann et al. (2015) have created different consumption profiles which are based on time, type of customers (i.e., residential or industrial), and weather. Their system uses live machine learning to model the consumption profiles of each customer with respect to time, type of customer and weather information. Based on probability distribution and confidence rate, if a customer's consumption value surpasses the threshold, the system generates an alert for a possible NTL detection. This work is performed on a dataset collected from Creos Luxembourg, the electricity operator in Luxembourg. The results are evaluated using accuracy, precision, recall and $F1$ score.

2.5 Limitations

There have been many attempts to bring down NTL in different companies, regions and countries. A good success in identifying NTL is achieved by applying different machine learning classifiers. Different performance metrics are used to evaluate how good or bad the classifier is in predicting NTL. However, not much has been discussed about the relationships that exist between these performance metrics with respect to NTL. There is still a need to highlight performance evaluation metrics that are specifically suitable in evaluating machine learning classifiers for NTL problem. Table 1 contains the description of all referred papers along with their limitations.

3 Methodology

In this section, we first describe the proposed methodology used for NTL detection in the electric distribution company. Then, we outline the need of separate performance evaluation metrics for NTL detection. Finally, we discuss a number of such existing metrics which proved to be good for NTL detection. The proposed methodology consists of seven steps which are described in the following subsections. The first step is data collection from the company. The data contains monthly consumption records of electric consumers. The data is collected in a comma separated values (CSV) file which needs to be converted into a form suitable for analytic processing. Data munging performs this functionality along with steps like duplicate removal and dealing with NULL values etc. Not all the features are useful for analytic process. Feature selection step shortlists the features which are most useful in predicting NTL. The company separately maintains the data of risky consumers. Data merging step combines the selected features with the the data of risky consumers. Once the features and the records are finalized, the next step is to normalize all features. This is done by scaling step. Next, training and testing of the classifiers is performed. On the basis of the results obtained from testing, different performance evaluation metrics are then calculated which form a strong foundation in identifying different criterion for the selection of suitable classifiers for NTL detection. The complete methodology is shown in Fig. 3.

3.1 Data collection

NTL detection cannot be thoroughly studied without a real dataset. We have collected a dataset from an electric supply company in Pakistan. The collected data contains monthly consumption records of consumers which ranges between January 2016 and March 2017. It comprises of 80,244 monthly consumption records. The data is divided into training set and test set. The training set contains 64,195 monthly consumption records while the test set contains 16,049 monthly consumption records. The training data have 2739 theft instances and 61,456 normal instances. The test data have 683 theft instances and 15,366 normal instances.

3.2 Data munging

The data obtained from the electric company is in a comma separated values (CSV) file. Initially, the raw data collected from the power supply company contained 110 features. Some of the features are redundant and useless. For example, the feature 'Postal Code' contains the same code for all records, the feature 'Meter Number' and 'Registration Number' are both used for unique identification, and the feature 'Write-Off' contains any relief of dues which actually contains all zero entries. After filtering out the useless features, the feature set is reduced to 71 features. It needs a further transformation from a raw format into a form suitable for downstream analytic processing. A number of steps are performed in this process. For example, replacing null values with suitable default values for multiple features. Last disconnection date stores the date on which the electricity is disconnected from a consumer's location. For most of the transactions, this feature contained null value which is replaced with a default future date. The feature 'opening balance' stores the information of the total amount pending to be paid by the consumer. It is converted to 1 and 0. A value 1 indicates that the consumer has not paid the bill of last month and a value

Table 1 Comparison of related work

Category	Articles	Classifiers/strategy	Performance metrics	Compared metrics w.r.t NTL detection
Consumption-data based techniques	Nagi et al. (2010)	SVM	Accuracy, hit rate	No
	Glauner et al. (2016, 2017)	Boolean, fuzzy and SVM	TPR, TNR, FPR, FNR, recall and specificity	Partial
	Di Martino et al. (2012)	One class SVM, CS-SVM, OPF and C4.5	Accuracy, recall, precision and F_{value}	Partial
	Figueroa et al. (2017)	Linear SVM, non-linear SVM, MLP neural network	Accuracy, recall, precision, specificity, AUC, F_{β}, F_1 and MCC	No
	Ford et al. (2014)	NN	TP, TN, FP and FN	No
	Cody et al. (2015)	M5P decision tree	RMSE	No
	Coma-Puig et al. (2016)	Naive Bayes, KNN, decision tree, NN, SVM, random forest and AdaBoost	Precision	No
	Bhat et al. (2016)	Deep learning using convolutional NN, auto encoders and long short-term memory networks	Precision, recall, F1 and ROC curve	Yes
	Chatterjee et al. (2017)	Recurrent NN	Accuracy	No
Network based techniques	Xia et al. (2015)	Comparison based inspection algorithm	None	No
	Han and Xiao (2014)	Difference between billed amount and consumed amount	None	No
	Mutupe et al. (2017)	Difference between billed amount and consumed amount	None	No
Hybrid techniques	Meira et al. (2017)	Random forest, logistic regression and SVM	AUC	No
	Buevich et al. (2016)	SVM	TPR and TNR	No
	Zhou et al. (2014)	Firefly algorithm	Accuracy	No
	Terciyanli et al. (2017)	Allotment of scores based on area, change of usage and deviation of monthly consumption	None	No
Additional-data based techniques	Hartmann et al. (2015)	Live machine learning using weather data	Accuracy, precision, recall and $F1$	Partial

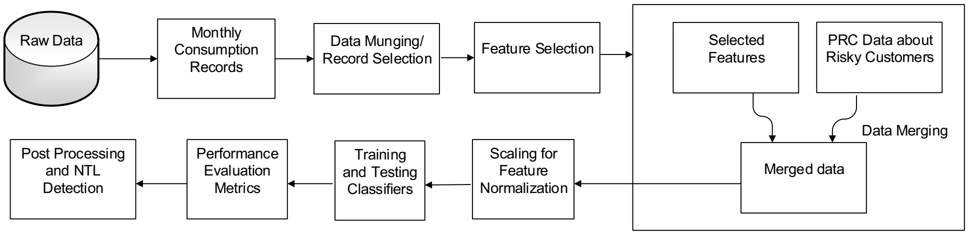
Raw Data
Monthly Consumption Records
Data Munging/ Record Selection
Feature Selection
Selected Features
PRC Data about Risky Customers
Data Merging
Merged data
Scaling for Feature Normalization
Training and Testing Classifiers
Performance Evaluation Metrics
Post Processing and NTL Detection

Fig. 3 Proposed methodology for NTL detection

0 indicates that the consumer has paid the bill. After performing data munging, the total monthly transactions are recorded to be 80,244. Had missing values not been dealt with properly, the number of records would have decreased to an alarming limit.

3.3 Feature selection

From the set of 71 features, a subset of 14 useful features is shortlisted using feature importance. It is a measure of finding the importance of each feature (Breiman 2001). A feature has an importance if the model's error of prediction is increased with the shuffling of the value of the feature. The increase in model's prediction error indicates that the model relies on that feature. Thus, the feature is important. Conversely, a feature has less importance if the model's error of prediction is not changed with the shuffling of the value of the feature. The stability in the model's prediction error indicates that the model does not rely on that feature. Thus, the feature is not important. To obtain the list of useful features, the 71 features are first listed in descending order with respect to feature importance. Then, using the Gini Index, a threshold for the optimum number of features is selected beyond which including any other feature should not affect the F-measure. This way, we have not only found the optimum combination of features for which the F-measure is best but it also has significantly reduced the computational time of the classifiers. The list of shortlisted features, their description and the feature importance is enlisted in Table 2.

3.4 Data merging

Additionally the company provided the data of potential risky consumers (PRC). These consumers are identified during on-site inspection. This data are useful in assigning the values of class labels as true or false. A class label of true indicates an instance of a theft and a class label of false indicates an instance of a normal consumption. This process is shown in Fig. 4. The PRC data is merged with the data shortlisted from the feature selection module.

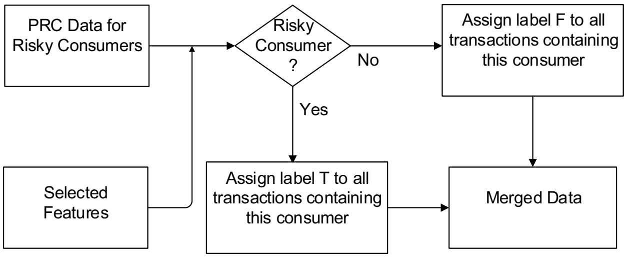
PRC Data for Risky Consumers
Risky Consumer ?
No
Assign label F to all transactions containing this consumer
Yes
Selected Features
Assign label T to all transactions containing this consumer
Merged Data

Fig. 4 Data merging

3.5 Scaling

The data from the selected features is needed to be normalized before applying training and testing. The purpose of applying normalization is to bring all the numerical features in the same scale without disturbing the range differences. The normalized scale for each feature is obtained using Eq. (1) where FV is the current feature value, min(FV) is the minimum feature value in current feature and max(FV) is the maximum feature value in current feature.

$$NS = \frac{FV - min(FV)}{max(FV) - min(FV)} \tag{1}$$

3.6 Training and testing

In this section, we introduce the classifiers that we have used in this paper for training and testing and show how they are applicable to NTL detection. The normalized data is used for training and testing of the three classifiers namely KNN, random forest and SVM.

3.6.1 *K*-nearest neighbors

The main functionality of this classifier works slightly different than other learning techniques. KNN does not use

Table 2 Feature description and feature importance of selected features

Feature	Description	Importances
Units-12Months	Units consumed during last 12 months	0.141807732008
Amount-12Months	Total amount billed during last 12 months	0.116774856878
BilledUnit-YTD	Units billed in current year	0.116751907486
BilledAmount	Amount billed in current month	0.092791430138
1 year LPS	Late payment surcharge in last one year	0.057126632724
Amount-12MonthsAvg	Average monthly amount in last 12 months	0.040861765397
BilledAmount-12MonthsGross	Total payment made in last 12 months	0.038509950395
Units-12MonthsAvg	Average monthly units in last 12 months	0.032045710111
Amount-GrossBilledYTD	Total payment made in current year till date	0.029088274439
Amount-12MonthsAvgGrossBilled	Total average monthly payment made in last 12 months	0.028248805369
Amount-regular	Payable amount for regular units	0.023397470434
1 month LPS	Late payment surchare in last 30 days	0.022879096842
Month-billing	Month of billing	0.016062708402
InstallementNo	Number of installements	0.015831709285

the parameter of weight rather it is a record-based approach which uses k nearest training samples to predict the value of the target variable.

Let $p = (p_1, p_2, \dots, p_n)$ and $q = (q_1, q_2, \dots, q_n)$ be the two samples.

The distance between the two samples is calculated using Eq. (2).

$$d(p,q) = \sqrt{(p_1 - q_1)^2 + (p_2 - q_2)^2 + \cdots + (p_n - q_n)^2} \qquad (2)$$

A test sample is assigned the class which is most frequent in K-nearest samples. The disadvantage associated with KNN is that as it is a record-based learning procedure so with the increase of k the time required to predict a test sample increases. The advantage of using KNN is that as it does not depend on any other element, the runtime for prediction can be decreased by allocating different cores or nodes for parallel execution (Altman 1992).

3.6.2 Support vector machine

Vapinik has proposed a binary classifier (Hearst et al. 1998) that creates a margin between the two classes and tries to maximize this margin. This way it constructs an optimal decision function $f(x)$ that can predict unseen instances with high accuracy as given in Eq. (3) where $sgn(g(x))$ is the boundary between the positive and negative classes (Vapnik 1999).

$$f(x) = sgn(g(x)) \qquad (3)$$

The expected error in classification is calculated using Expression 4 where R is the expected error, t is the training errors, n is the number of training samples, h is the dimension of the set of hyperplanes and η is the confidence metric (Vapnik 1998).

$$R < \frac{t}{n} + \sqrt{\frac{h\left(\ln\left(\frac{2n}{h}\right) + 1\right) - \ln\left(\frac{\eta}{4}\right)}{n}} \qquad (4)$$

SVM needs a comparatively smaller number of training samples. Therefore, unlike neural networks (Cao and Tay 2003), it is less prone to getting struck with the problem of overfitting. Mapping of input to higher dimensions requires setting up of kernel for only a few thousands of training samples (Chang and Lin 2011). This is a major concern in dealing with big data sets. To overcome this problem, we have used linear SVC (Pedregosa et al. 2011a).

3.6.3 Random forest

A random forest comprises of multiple individual decision trees (Liaw et al. 2002). For each tree a separate set of training examples is selected. Using this approach, the problem of over fitting in imbalance datasets is avoided. On the testing phase, the final outcome of a sample is evaluated by using the majority voting scheme from among all the individual decision trees. Another advantage of using this approach is that as different training examples are used in every decision tree, variable number of nodes or cores can be used for training (Ho 1995).

Table 3 Comparison of KNN, SVM and random forest

	KNN	SVM	Random forest
No. of training samples	Very small	Small	Large
Uses distance function	√	×	×
Construct a hyper plane	×	√	×
Uses decision trees	×	×	√
Occurrence of over fitting	√	×	×
Parallel processing	√	×	√

Table 4 Confusion matrix

	Predicted positive	Predicted negative
Actual positive	True positive (TP)	False negative (FN)
Actual negative	False positive (FP)	True negative (TN)

Table 3 shows a comprehensive comparison between KNN, SVM and random forest.

3.7 Post-processing and NTL detection

For the last few years the research community has been focusing on deriving methods which focus on representing the evaluation of classes separately. Table 4 shows the basic confusion matrix which is used to formulate more complex metrics for datasets containing imbalanced class distribution. For NTL, true positive (TP) is the instances of theft cases correctly classified by the classifier and true negative (TN) is the instances of normal cases correctly classified by the classifier. False positive (FP) indicate instances of normal cases identified as theft by the classifier and false negative (FN) indicate the instances of theft cases identified as normal by the classifier.

The metrics are then used to calculate more complex metrics like accuracy, recall, precision, TNR, FPR, FNR, NPV, F_β, arithmetic mean, harmonic mean, G-Mean and dominance. These metrics are discussed in Sect. 3.8. These metrics yield a set of comprehensive observations particularly related with NTL detection. The observations are discussed in Sect. 4.

3.8 Evaluation metrics

Datasets from electric industry have a strong imbalanced distribution of target variable. Doing predictive modeling in these datasets is a challenging task due to the fact that distribution of classes (target variable Y) is non-uniform. It could be a case that training and testing samples contain 99% of total samples belonging to the normal class and the remaining 1% belong to the thieve class. The scenario becomes more complex when the user's choice is biased towards the least represented class i.e. the thieve class. Performance metrics that are used for the balanced datasets can not be efficiently used for datasets with imbalance distribution of target variable as these metrics tend to ignore the thieve class for which the performance measure is actually needed. Thus, giving the performance measures against the unwanted and the most repeated class is not helpful in accessing the performance of the least represented class predictions. Therefore, accuracy and error rate are not the right measures as they are biased towards the normal class (Manning et al. 2008). In fact, we need measures which evaluate the correctness of the normal and the thief class separately. For this, a basic confusion metric is used to calculate true positive (TP), true negative (TN), false positive (FP) and false negative (FN). The basic confusion metric is further used to evaluate more complex metrics which are listed in Eqs. (5)–(9).

One of the most common performance evaluation metric used for classifiers is accuracy. It gives a measure that how accurately the classifier has predicted TP and TN. It tends to get failed for imbalanced datasets where the user preference is towards the FP and FN. For example, talking about NTL 99% of electric consumption is a normal usage and 1% of consumption is a theft case. Now if a classifier correctly predicts all 99% of normal usage and does not predict the remaining 1% of theft usage, accuracy will be measured as 99%. In reality, the classifier was not performing well because it failed to predict the theft class. Equation (5) is used to calculate accuracy.

$$Accuracy = \frac{TP + TN}{TP + FN + FP + TN} \tag{5}$$

Due to lack of handling measure of predicting FP and FN, other measures are derived which either take care of both classes separately or handle the least represented class more accurately. These are discussed through Eqs. (6)–(11).

$$TruePositiveRate(Recall) = \frac{TP}{TP + FN} \tag{6}$$

$$TrueNegativeRate(Specificity) = \frac{TN}{TN + FP} \tag{7}$$

$$FalsePositiveRate = \frac{FP}{TN + FP} \tag{8}$$

$$FalseNegativeRate = \frac{FN}{FN + TP} \tag{9}$$

$$PositivePredictiveValue(PPV)orPrecision = \frac{TP}{TP + FP} \tag{10}$$

$$NegativePredictiveValue(NPV) = \frac{TN}{TN + FN} \tag{11}$$

True positive rate (TPR) or recall is the measure of total number of thieves correctly classified as thieves by the classifier. Recall is also called as sensitivity. As we want to minimize FN, by Eq. (7) minimizing FN will maximize recall. This gives an indication that NTL requires a high recall model. The higher the recall, the better it is for NTL. On the other hand, by Eq. (8) we see that if precision is low, the model can still tolerate because it does not need a high precision. True negative rate (TNR) is also called as specificity. It is a measure that out of total negative instances how many were correctly classified as negative. False positive rate (FPR) is the measure of total number of normal consumers wrongly predicted as thieves. False negative rate (FNR) is a measure that out of total positive instances how many were wrongly classified as negative. Positive predictive value (PPV) or precision is the measure that out of the total predicted positive class instances how many were classified correctly as positive. Negative predictive value (NPV) is the measure that out of total predicted negative class instances how many were correctly classified as negative.

$$F_\beta = \frac{\left(1 + \beta^2\right) Recall \times Precision}{\beta^2 \times Precison + Recall} \tag{12}$$

F_β, as shown in 12, is another metric that is used for evaluation of classifiers in imbalance data sets (Branco et al. 2016). It uses recall (completeness) and precision (exactness) where β is a coefficient used to set the priority between recall and precision. For $\beta = 1$ both recall and precision has the same priority. If β is set to a value greater than 1, recall gets the high weightage and if it is set to a value smaller than 1, precision gets the high weightage. Usually people use value 1 when dealing with imbalance domain. We have tested two different values of β i.e. with $\beta = 1$ and with $\beta = 1.5$. The latter case sets the priority of recall higher than precision. When both recall and precision are high, the value for F_β becomes high.

$$ArithmeticMean = \frac{(Precision + Recall)}{2} \tag{13}$$

$$HarmonicMean = \frac{2 \times Precision \times Recall}{Precision + Recall} \quad (14)$$

$$G - Mean = \sqrt{Sensitivity \times Specificity} \quad (15)$$

Arithmetic mean is the average of precision and recall measure as shown in Eq. (13). It is rarely used as evaluation metric for imbalance datasets as it does not give an insight to the two performance measures. Instead, harmonic mean is preferred which is presented in Eq. (14). Seeing the equation, it is obvious that it is always less than the arithmetic mean of the two. In fact, harmonic mean is closer to the smaller of the two values. So if harmonic mean is high, that is an indication that both precision and recall are high (Sun et al. 2007). $F_\beta = 1$ is the harmonic mean of precision and recall. Geometric mean (G-Mean), presented in Eq. (15), is used when performance measure of both TPR and TNR is of concern. It is a measure that how good the classifier is for both the classes.

$$Dominance = TPR - TNR \quad (16)$$

In García et al. (2008), a new performance measure, Dominance, was proposed. It gives a measure of dominance between the positive and the negative class. Seeing Eq. (16), it can be deduced that it ranges between −1 and +1. A value close to +1 indicates good accuracy of the classifier for the positive class and a value close to −1, depicting good accuracy of the classifier for the negative class.

4 Results and analysis

In this section, we first perform extensive simulation of the random forest, KNN and SVM on training and test data using Python's open source library, scikit-learn (Pedregosa et al. 2011b). Then, we discuss a detailed analysis of the comparison of performance evaluation metrics across the three classifiers along with the comparison of the classifiers. A list of simulation parameters is also presented in Table 5.

4.1 Experimental setup

The experiments were performed on a 64-bit Windows server with Intel Xeon 2.2 GHz processor and 32 GB RAM. All the algorithms were implemented in Python 3. The total number of transaction records is 80,244 out of which 64,195 are selected for training the three classifiers namely random forest, KNN and SVM. The remaining 16,049 records are selected for testing the classifiers. The training time for random forest, KNN and SVM is recorded as 22 seconds, 2 s and 30 s, respectively. The values of TP, TN, FP and FN for the test set across the three classifiers are listed in Table 6. The values of the other complex performance metrics for the three classifiers are listed in Table 7. The performance of the classifiers can vary with the change of dataset as it depends on the selected features. A different dataset with a different set of features can result in increase or even decrease of performance. So, the better the feature set, the higher will be the performance.

KNN out performed random forest and SVM in terms of TP. It has the maximum instances of theft detection which is 678. For random forest, TP is 677 and for SVM, it is 672. Accuracy for the three classifiers are approximated to 99% but seeing precision, it is observed that random forest performed better than KNN and SVM, while KNN outperformed random forest and SVM on the basis of recall as it has the best recall of 99.27%. Random forest has the highest arithmetic mean and harmonic mean.

Table 6 TP, TN, FP and FN values across random forest, KNN and SVM

	Random forest	*K*-nearest neighbors	Support vector machine
TP	677	678	672
TN	15,347	15,343	15,338
FP	19	23	28
FN	6	5	11
Total	16,049	16,049	16,049

Table 5 List of simulation parameters

Classifier	Simulation parameters
Linear SVC	penalty = 'l2', loss = 'squared_hinge', dual = True, tol = 0.0001, C = = 1.0, multi_class = 'ovr', fit_intercept = True, intercept_scaling = 1, class_weight = None, verbose = 0, random_state = None, max_iter = 1000
KNN	n_neighbors = 5, weights = 'uniform', algorithm = 'auto', leaf_size = 30, p = 2, metric = 'minkowski', metric_params = None, n_jobs = None
Random forest	n_estimators = 'warn', criterion = 'gini', max_depth = None, min_samples_split = 2, min_samples_leaf = 1, min_weight_fraction_leaf = 0.0, max_features = 'auto', max_leaf_nodes = None, min_impurity_decrease = 0.0, min_impurity_split = None, bootstrap = True, oob_score = False, n_jobs = None, random_state = None, verbose = 0, warm_start = False, class_weight = None

Table 7 Other complex metrics for the three classifiers

	Random forest	*K*-nearest neighbors	Support vector machine
Accuracy (%)	99.84%	99.83%	99.76%
Precision (%)	97.27%	96.71%	96.0%
Recall (%)	99.12%	99.27%	98.39%
Arithmetic mean (%)	98.20%	98.0%	97.19%
Harmonic mean (%)	98.19%	97.98%	97.18%
NPV	1.0	1.0	0.999
F_β (for $\beta = 1$)	98.2	98.0	97.2
F_β (for $\beta = 1.5$)	98.5	98.5	97.6
G-Mean	99.50	99.56	99.10
Dominance	− 0.008	− 0.006	− 0.014
TPR	0.991	0.993	0.984
TNR	0.999	0.999	0.998
FPR	0.001	0.001	0.002
FNR	0.009	0.007	0.016

4.2 Comparison of precision and recall

An important observation regarding the problem of NTL detection is that the model which has a high recall is most suitable for theft detection. In order to understand this relation, consider the cases of FP and FN. False positives are those normal users that have been predicted by the classifier as thieves whereas false negatives are those thieves that are predicted by the classifier as normal users. Considering the two cases, having a large FP value will only result in increasing the manual effort of on-site inspections whereas a high FN value will result in the failure of the classifier to correctly identify the thieves. Therefore, for NTL it is recommended to promote the classifier which has a low FN value. Now, considering the Eq. (6), it can be observed that recall increases with the decrease of FN. This gives a nice measure of the selection of the classifier for NTL detection that both the precision and the recall should not have equal priority. In fact, for NTL detection, classifiers with high recall are most suitable regardless of what the precision value is. In Table 6, it is observed that KNN has the lowest number of FN, i.e. 5. Consequently, it has the highest recall among the three classifiers as shown in Fig. 5. The lowest recall is observed for SVM which is 98.39%. Thus, the percent increase of recall from using SVM to KNN is 0.89%. This gives a clear indication that for our real dataset, KNN is the better choice for NTL detection. For two classifiers having the same recall but different precision values, the classifier with a high precision should be selected. Precision will increase with the decrease in FP. So, when the two classifiers have equal recalls, the classifier having the lowest FP should be given preference. This observation can be verified by Eq. (10). For all the three classifiers, recall is observed higher than their corresponding precision values. It is further observed that SVM has the lowest precision and recall among the three classifiers as shown in Fig. 5.

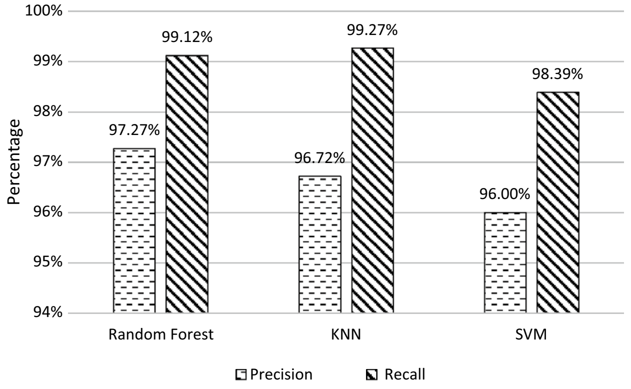

Fig. 5 Comparison of precision and recall

4.3 Comparison of accuracy, FPR and FNR

Total normal users that are predicted as thieves is measured by FPR. A high FPR increases the on-site inspection for theft verification which consequently results in increase of manual efforts. On the other hand, a high FPR also indicates the success of the classifier in identifying thieves that are categorized as normal users in the company. Another measure that gives a close insight of the number of thieves that are wrongly classified as normal users is FNR. A low FNR is desirable in NTL detection. Seeing the accuracy measure, all the three classifiers looked to be performing exceptionally well but observing FPR and FNR, it is found that accuracy is not depicting the facts about FP and FN. Figure 6 shows that FPR is very low for the three classifiers. For KNN, FNR is least among the three classifiers showing that it has the lowest FN value and thus, KNN

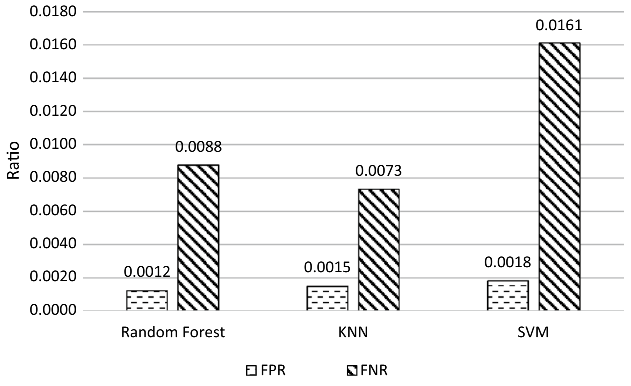

Fig. 6 Comparison of FPR and FNR

is found to be a good choice for NTL in our real dataset. Among the three classifiers, the highest FNR is observed for SVM showing that it has the highest value of FN which can also be verified in Table 6. Thus, for our real dataset, SVM turns out to be the last choice for NTL detection.

4.4 Comparison of $F_{\beta=1}$ and $F_{\beta=1.5}$

$F_{\beta=1}$ and $F_{\beta=1.5}$ $F_{\beta=1}$ and $F_{\beta=1.5}$ are close to each other in all classifier readings. Both are high for the three classifiers depicting that both precision and recall values for the classifiers are considerably high. The lowest reading for the two metrics are observed for SVM showing that precision and recall values for SVM are lower as compared to their counterparts which can also be verified using Table 7. Thus, SVM is the last choice for NTL detection in this real dataset. For $F_{\beta=1}$, random forest has a higher value than KNN and for $F_{\beta=1.5}$, random forest and KNN have equal values. Considering $F_{\beta=1}$, random forest has performed better than KNN and considering $F_{\beta=1.5}$, both random forest and KNN have equal performance. Given that recall has a high weightage in $F_{\beta=1.5}$, for all the classifiers, $F_{\beta=1.5}$ is high as compared to the corresponding $F_{\beta=1}$. This indicates that recall is high for all the classifiers as compared to precision. The percentage increase from precision to recall in random forest, KNN and SVM is 1.9%, 2.65%, and 2.49%, respectively. The highest increase in percentage is observed for KNN and thus, it also has the highest difference of values between $F_{\beta=1}$ and $F_{\beta=1.5}$ i.e, 0.5. This indicates that KNN outperformed random forest and SVM. Also, F_{β} values are observed to be between recall and precision values for all classifiers as shown in Fig. 7. This shows that F_{β} of precision and recall behaves just like the harmonic mean. As discussed in Sect. 4.2, for NTL detection recall should be given high priority as compared to precision. This can be achieved by using F_{β} measure with β value greater than 1.

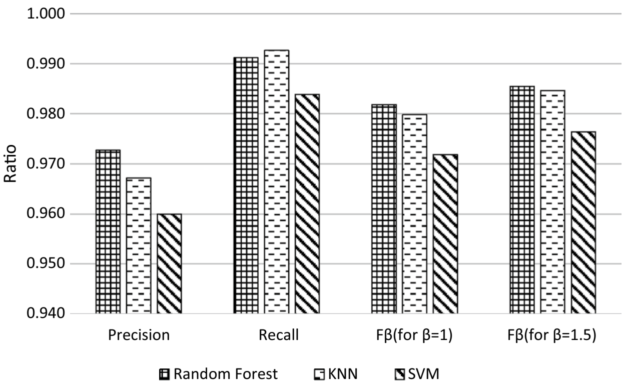

Fig. 7 Comparison of precision, recall, $F_{\beta=1}$ and $F_{\beta=1.5}$

4.5 Harmonic mean

For all the classifiers, harmonic mean is lower than the arithmetic mean. Harmonic mean is also observed to be closer to the smaller of the precision and recall for all classifiers. Random forest has the highest harmonic mean. This indicates that not only the precision and recall values for random forest are high but also they are close to each other. This can be verified by the fact that random forest has the smallest percentage increase from precision to recall i.e, 1.9%. Harmonic mean for SVM is lowest among the three classifiers showing that the corresponding values of precision and recall for SVM are also low, as shown in Fig. 8. Therefore, instead of maintaining both the precision and the recall, harmonic mean can also be used for the evaluation of the classifiers in NTL detection.

4.6 Comparison of TPR, TNR and G-Mean

For all the classifiers, G-Mean is high. This indicates that TPR and TNR for the three classifiers are also high. G-Mean for SVM is lowest among the three classifiers indicating that its TPR is also lowest as shown in Fig. 9. Therefore, it can be deduced that for NTL detection, a classifier with a high G-Mean value is preferable over a classifier with a low G-Mean value. Thus, KNN outperformed random forest and SVM.

4.7 Comparison of TPR, TNR and dominance

A classifier having dominance close to −1 depicts that it has a high TNR but a low TPR. In contrast, a classifier having dominance close to 0 indicates that it is good in predicting both the classes for NTL detection. For NTL detection, TPR and TNR give close insight of the performance of a classifier. Combining TPR and TNR, dominance gives a good choice of a performance evaluation metric for NTL

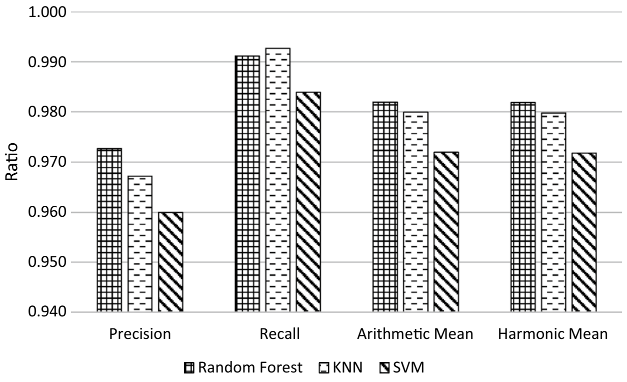

Fig. 8 Comparison of precision, recall, arithmetic mean and harmonic mean

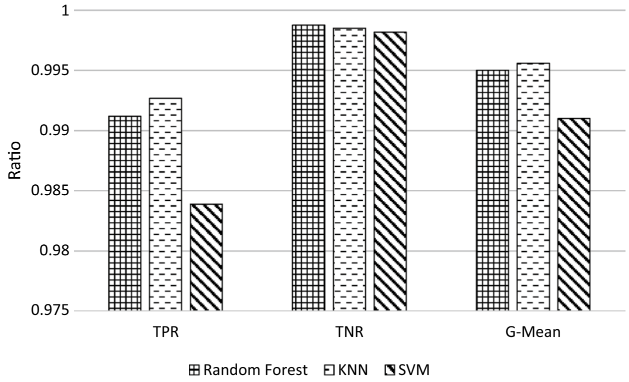

Fig. 9 Comparison TPR, TNR and G-Mean

detection. For our dataset, comparison of TPR, TNR and dominance is shown in Fig. 10. It is observed that among the three classifiers, KNN has the dominance which is closest to 0. This shows that for our dataset, KNN is the best in predicting both classes.

4.8 Comparison of NPV and FNR

For NTL detection, occurrence of theft instances is rare while normal consumers are in huge number. As NPV indicates the number of normal consumers only and ignores the theft cases, therefore for NTL detection, NPV is not a suitable metric. It is observed that for all the classifiers, NPV is close to 100%. A clear reason for this is that NPV is ignoring theft cases and considering normal consumers only. In contrast, FNR is a measure of number of thieves that are predicted as normal consumers. For NTL detection, this ratio is needed to be as low as possible. It is observed that KNN has the lowest FNR. Thus, KNN is a good choice for NTL detection. For the three classifiers, FNR is shown in Fig. 11. The figure shows that SVM has the highest FNR and thus, for our dataset, it is the last option for NTL detection.

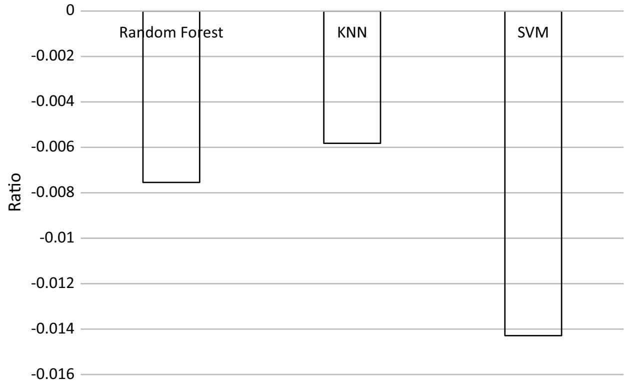

Fig. 10 Dominance of random forest, KNN and SVM

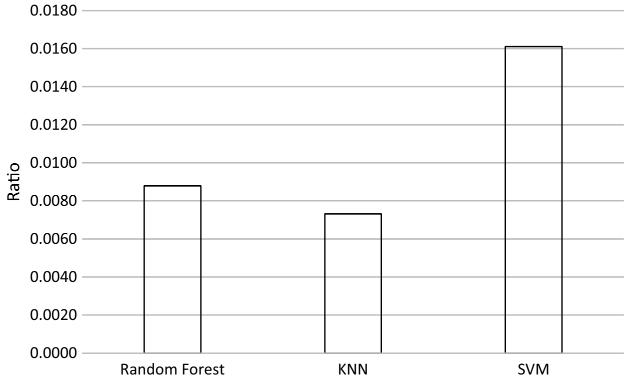

Fig. 11 FNR of random forest, KNN and SVM

5 Conclusion and future work

This paper has used a real-world dataset of an electric supply company in Pakistan to identify the non-technical loss by applying three classifiers namely random forest, *K*-nearest neighbors and linear support vector machine. The aim of the study is to use these classifiers to first identify existing NTL attempts and then predict new theft cases.

It further uses 14 different metrics to perform an in-depth performance analysis of the three classifiers. One of the findings is that for NTL detection, both the precision and recall should not have equal precedence. In fact, the classifier with a higher recall is better. The percent increase of recall from using KNN to random forest is 1.24%. This depicts that random forest is the better choice for NTL detection. This analysis can be used as a baseline for the accurate selection of the classifiers in NTL detection. This work will vastly benefit the electric supplier in detecting NTL. It will not only improve their abilities for NTL detection, but will also save huge amount of monetary losses which they are already bearing.

There is a need to further extend the use of performance evaluation metrics that can estimate and compare error rates on the basis of which a combination of classifiers can be selected for a specified dataset for NTL detection. Currently, there is a small range of graphical metrics used for performance analysis. This includes receiver operating characteristic (ROC) and area under ROC curve (AUC) citeBranco. There is also a need of further exploration in the usage of graphical performance metrics. Furthermore, the performance of classifiers with respect to their categories is another future direction for NTL detection.

Acknowledgements Open access funding provided by University of Debrecen (DE). We are thankful to the electric supply company in Pakistan for giving us a copy of the dataset containing monthly consumption of their customers for the prediction of NTL. The work/publication is supported by the EFOP-3.6.1-16-2016-00022 project.The project is co-financed by the European Union and the European Social Fund. Imran's work is supported by the Deanship of Scientific Research through research group project number RG-1435-051.

References

Alam M, Kabir E, Rahman M, Chowdhury M (2004) Power sector reform in bangladesh: electricity distribution system. Energy 29(11):1773–1783

Altman NS (1992) An introduction to kernel and nearest-neighbor nonparametric regression. Am Stat 46(3):175–185

Amalina F et al (2020) Blending big data analytics: review on challenges and a recent study. IEEE Access 8:3629–3645. https://doi.org/10.1109/ACCESS.2019.2923270

Ariyaluran Habeeb RA, Nasaruddin F, Gani A, Amanullah MA, Abaker Targio Hashem I, Ahmed E, Imran M (2019) Clustering-based real-time anomaly detection—a breakthrough in big data technologies. Trans Emerg Telecommun Technol. https://doi.org/10.1002/ett.3647

Bhat RR, Trevizan RD, Sengupta R, Li X, Bretas A (2016) Identifying nontechnical power loss via spatial and temporal deep learning. In: 2016 15th IEEE International conference on machine learning and applications (ICMLA), Anaheim, CA, 2016, pp 272–279. https://doi.org/10.1109/ICMLA.2016.0052

Branco P, Torgo L, Ribeiro RP (2016) A survey of predictive modeling on imbalanced domains. ACM Comput Surv 49(2):31:1–31:50

Breiman L (2001) Random forests. Mach Learn 45(1):5–32

Buevich M et al (2016) Microgrid losses: when the whole is greater than the sum of its parts. In: 2016 ACM/IEEE 7th international conference on cyber-physical systems (ICCPS), Vienna, 2016, pp 1–10. https://doi.org/10.1109/ICCPS.2016.7479107

Cao L-J, Tay FEH (2003) Support vector machine with adaptive parameters in financial time series forecasting. IEEE Trans Neural Netw 14(6):1506–1518

Chang C-C, Lin C-J (2011) Libsvm: a library for support vector machines. ACM Trans Intell Syst Technol 2(3):27:1–27:27

Chatterjee S, Archana V, Suresh K, Saha R, Gupta R, Doshi F (2017) Detection of non-technical losses using advanced metering infrastructure and deep recurrent neural networks. In: 2017 IEEE international conference on environment and electrical engineering and 2017 IEEE industrial and commercial power systems Europe (EEEIC / I&CPS Europe), Milan, 2017, pp 1–6. https://doi.org/10.1109/EEEIC.2017.7977665

Cody C, Ford V, Siraj A (2015) Decision tree learning for fraud detection in consumer energy consumption. In: 2015 IEEE 14th international conference on machine learning and applications (ICMLA), Miami, FL, 2015, pp. 1175–1179. https://doi.org/10.1109/ICMLA.2015.80

Coma-Puig B, Carmona J, Gavalda R, Alcoverro S, Martin V (2016) Fraud detection in energy consumption: a supervised approach. In: 2016 IEEE international conference on data science and advanced analytics (DSAA). pp 120–129

Di Martino M, Decia F, Molinelli J, Fernández A (2012) Improving electric fraud detection using class imbalance strategies. ICPRAM 2:135–141

Figueroa G, Chen Y, Avila N, Chu C (2017) Improved practices in machine learning algorithms for NTL detection with imbalanced data. In: 2017 IEEE Power & Energy Society General Meeting, Chicago, IL, 2017, pp 1–5. https://doi.org/10.1109/PESGM.2017.8273852

Ford V, Siraj A, Eberle W (2014) Smart grid energy fraud detection using artificial neural networks. In: 2014 IEEE symposium on computational intelligence applications in smart grid (CIASG), Orlando, FL, 2014, pp 1–6. https://doi.org/10.1109/CIASG.2014.7011557

García V, Mollineda RA, Sánchez JS (2008) A new performance evaluation method for two-class imbalanced problems. In: da Vitoria Lobo N et al (eds) Structural, syntactic, and statistical pattern recognition. Springer, Berlin, Heidelberg, pp 917–925. https://doi.org/10.1007/978-3-540-89689-0_95

Glauner P, Boechat A, Dolberg L, State R, Bettinger F, Rangoni Y, Duarte D (2016) Large-scale detection of non-technical losses in imbalanced datasets. In: 2016 IEEE power and energy society innovative smart grid technologies conference (ISGT). pp 1–5

Glauner P, Meira JA, Valtchev P, State R, Bettinger F (2017) The challenge of non-technical loss detection using artificial intelligence: a survey. Int J Comput Intell Syst 10(1):760–775. https://doi.org/10.2991/ijcis.2017.10.1.51

Habeeb RAA, Nasaruddin F, Gani A, Hashem IAT, Ahmed E, Imran M (2019) Real-time big data processing for anomaly detection: a survey. Int J Inf Manag 45:289–307

Han W, Xiao Y (2014) NFD: a practical scheme to detect non-technical loss fraud in smart grid. In: 2014 IEEE international conference on communications (ICC), Sydney, NSW, 2014, pp 605–609. https://doi.org/10.1109/ICC.2014.6883385

Han W, Xiao Y (2019) Edge computing enabled non-technical loss fraud detection for big data security analytic in smart grid. J Ambient Intell Humaniz Comput. https://doi.org/10.1007/s12652-019-01381-4

Hartmann T et al (2015) Suspicious electric consumption detection based on multi-profiling using live machine learning. In: 2015 IEEE international conference on smart grid communications (SmartGridComm), Miami, FL, 2015, pp 891–896. https://doi.org/10.1109/SmartGridComm.2015.7436414

Hayat MK, Daud A, Alshdadi AA, Banjar A, Abbasi RA, Bao Y, Dawood H (2019) Towards deep learning prospects: Insights for social media analytics. IEEE Access 7:36958–36979

Hearst MA, Dumais ST, Osuna E, Platt J, Scholkopf B (1998) Support vector machines. IEEE Intell Syst Appl 13(4):18–28

Ho TK (1995) Random decision forests. In: Proceedings of 3rd international conference on document analysis and recognition, Montreal, Quebec, Canada, 1995, vol 1, pp 278–282. https://doi.org/10.1109/ICDAR.1995.598994

Jain AK, Gupta BB (2019) A machine learning based approach for phishing detection using hyperlinks information. J Ambient Intell Humaniz Comput 10(5):2015–2028

Lakshmi K, Vadivu G (2019) A novel approach for disease comorbidity prediction using weighted association rule mining. J Ambient Intell Humaniz Comput. https://doi.org/10.1007/s12652-019-01217-1

Liaw A, Wiener M et al (2002) Classification and regression by random forest. R News 2(3):18–22

Manning CD, Raghavan P, Schütze H (2008) Introduction to information retrieval. Cambridge University Press, New York

McDaniel P, McLaughlin S (2009) Security and privacy challenges in the smart grid. IEEE Secur Priv 7(3):75–77

Meira JA et al (2017) Distilling provider-independent data for general detection of non-technical losses. In: 2017 IEEE power and energy conference at Illinois (PECI), Champaign, IL, 2017, pp 1–5. https://doi.org/10.1109/PECI.2017.7935765

Mutupe RM, Osuri SO, Lencwe MJ, Daniel Chowdhury SP (2017) Electricity theft detection system with RF communication between distribution and customer usage. In: 2017 IEEE PES power Africa, Accra, 2017, pp 566–572. https://doi.org/10.1109/PowerAfrica.2017.7991288

Nagi J, Yap KS, Tiong SK, Ahmed SK, Mohamad M (2010) Non-technical loss detection for metered customers in power utility using support vector machines. IEEE Trans Power Deliv 25(2):1162–1171

Papadimitriou C, Messinis G, Vranis D, Politopoulou S, Hatziargyriou N (2017) Non-technical losses: detection methods and regulatory aspects overview. CIRED Open Access Proc J 2017(1):2830–2832

Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V et al (2011a) Scikit-learn: machine learning in python. J Mach Learn Res 12(Oct):2825–2830

Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V et al (2011b) Scikit-learn: machine learning in python. J Mach Learn Res 12(Oct):2825–2830

Razzak MI, Imran M, Xu G (2019) Big data analytics for preventive medicine. Neural Comput Appl 33:1123–1131. https://doi.org/10.1007/s00521-019-04095-y

Saeed Z, Abbasi RA, Maqbool O, Sadaf A, Razzak I, Daud A, Aljohani NR, Xu G (2019) What's happening around the world? A survey and framework on event detection techniques on twitter. J Grid Comput 17(2):279–312

Sun Y, Kamel MS, Wong AK, Wang Y (2007) Cost-sensitive boosting for classification of imbalanced data. Pattern Recognit 40(12):3358–3378

Terciyanli E, Eryigit E, Emre T, Caliskan S (2017) Score based non-technical loss detection algorithm for electricity distribution networks. In: 2017 5th international istanbul smart grid and cities congress and fair (ICSG), Istanbul, 2017, pp 180–184. https://doi.org/10.1109/SGCF.2017.7947629

ur Rehman MH, Ahmed E, Yaqoob I, Hashem IAT, Imran M, Ahmad S (2018) Big data analytics in industrial iot using a concentric computing model. IEEE Commun Mag 56(2):37–43

ur Rehman MH, Yaqoob I, Salah K, Imran M, Jayaraman PP, Perera C (2019) The role of big data analytics in industrial internet of things. Future Gener Comput Syst 99:247–259

Vapnik V (1998) Statistical learning theory. Wiley, New York

Vapnik VN (1999) An overview of statistical learning theory. IEEE Trans Neural Netw 10(5):988–999

Xia X, Liang W, Xiao Y, Zheng M, Xiao Z (2015) A difference-comparison-based approach for malicious meter inspection in neighborhood area smart grids. In: 2015 IEEE international conference on communications (ICC), London, 2015, pp 802–807. https://doi.org/10.1109/ICC.2015.7248420

Zhou G, Zhao W, Lv X, Jin F, Yin W (2014) A novel load profiling method for detecting abnormalities of electricity customer. In: 2014 IEEE PES general meeting | conference & exposition, national harbor, MD, 2014, pp 1–5. https://doi.org/10.1109/PESGM.2014.6939307

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.