Energy Theft Detection Using Gradient Boosting Theft Detector With Feature Engineering-Based Preprocessing

Rajiv Punmiya; Sangho Choe, Catholic University of Korea

Published in: IEEE Transactions on Smart Grid (Volume: 10 , Issue: 2 , March 2019). [DOI: 10.1109/TSG.2019.2892595]

Abstract

For the smart grid energy theft identification, this letter introduces a gradient boosting theft detector (GBTD) based on the three latest gradient boosting classifiers (GBCs): 1) extreme gradient boosting; 2) categorical boosting; and 3) light gradient boosting method. While most of existing machine learning (ML) algorithms just focus on fine tuning the hyperparameters of the classifiers, our ML algorithm, GBTD, focuses on the feature engineering-based preprocessing to improve detection performance as well as time-complexity. GBTD improves both detection rate and false positive rate (FPR) of those GBCs by generating stochastic features like standard deviation, mean, minimum, and maximum value of daily electricity usage. **GBTD also reduces the classifier complexity with weighted feature-importance-based extraction techniques**. Emphasis has been laid upon the practical application of the proposed ML for theft detection by minimizing FPR and reducing data storage space and improving time-complexity of the GBTD classifiers. Additionally, this letter **proposes an updated version of the existing six theft cases to mimic real-world theft patterns and applies them to the dataset for numerical evaluation of the proposed algorithm.**

Introduction

- ❑The objective of electricity theft detection is to detect unusual activities in the electricity usage of a smart grid (SG) meter (or simply smart meter). Theft can be detected by checking for abnormalities in the user's electricity consumption patterns. Analysing user behaviour from historical data is the fundamental basis of a data science approach like machine learning (ML).
- ❑In this letter, we aim to provide a thorough comparison of the three latest GBCs including extreme gradient boosting (**XGBoost**), categorical boosting (**CatBoost**), and light gradient boosting method (**LightGBM**), and to propose a gradient boosting theft detector (GBTD) based on these GBCs that has a feature engineering-based preprocessing module to improve detection rate (DR), false positive rate (FPR), and time-complexity.
- ❑In the GBTD classifiers, the preprocessing module has a stochastic feature generating function which improves FPR as well as DR by utilizing combinations of daily electricity usage values as features.
- ❑The preprocessing module is also equipped with feature extraction function using weighted feature importance (WFI) that greatly reduces the training time-complexity by discarding irrelevant features (noise) from the customer's dataset. This also helps in reduced storage space usage for customer data in SG.

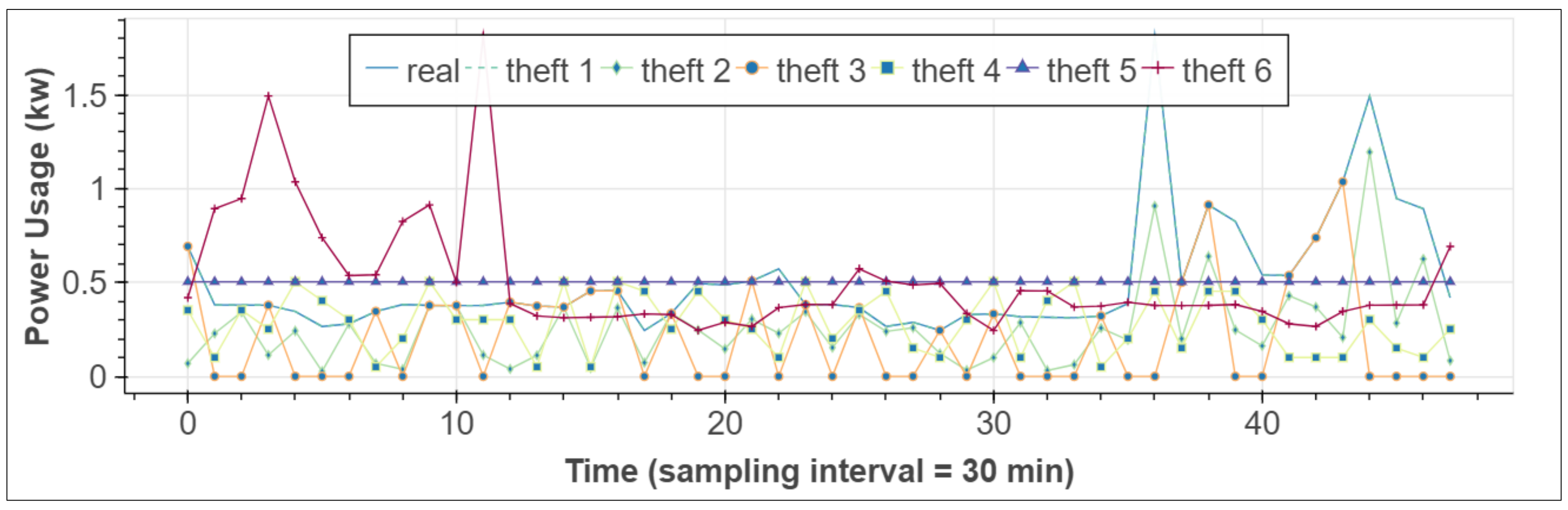


Fig 1. Data generation and visualization of new theft cases.

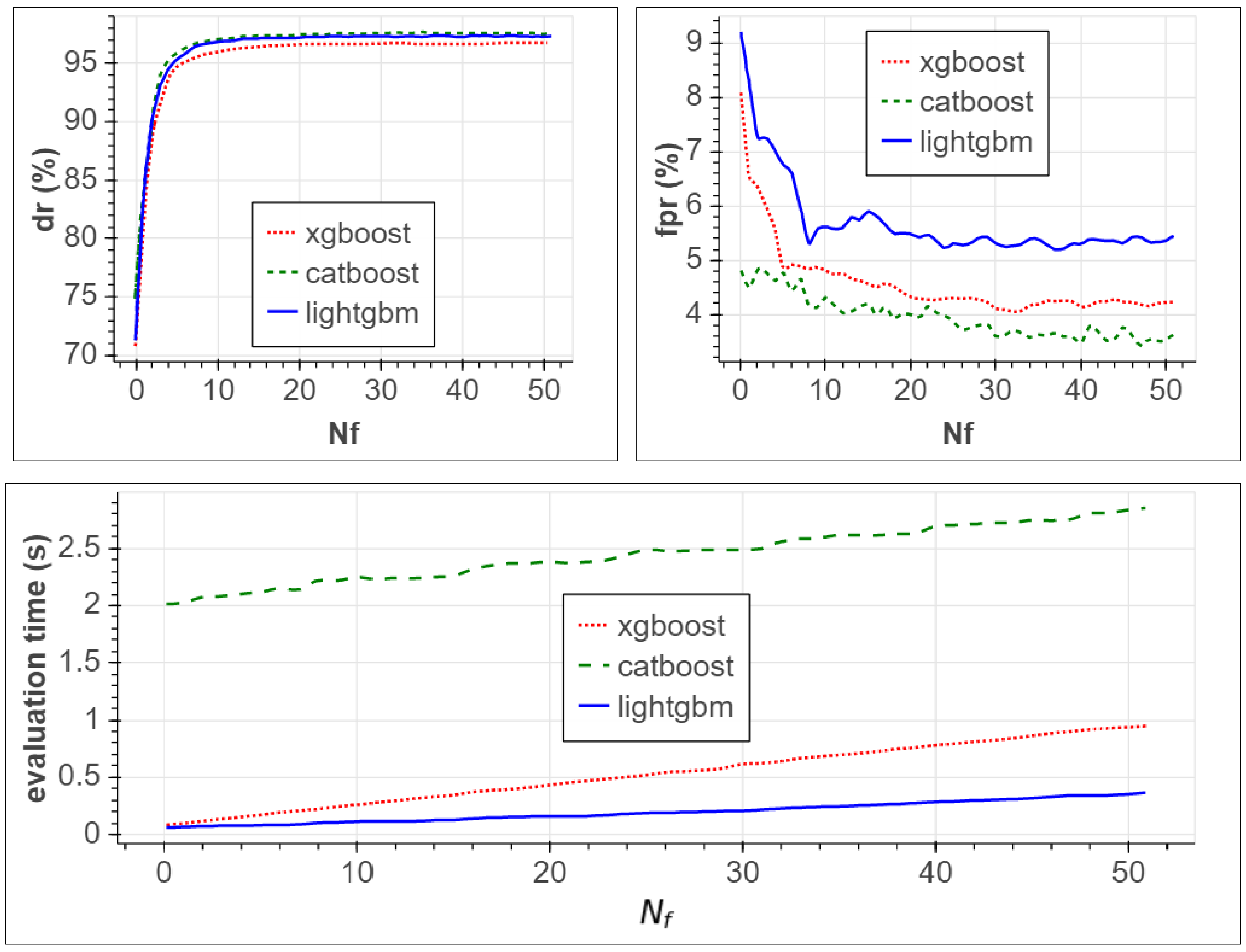
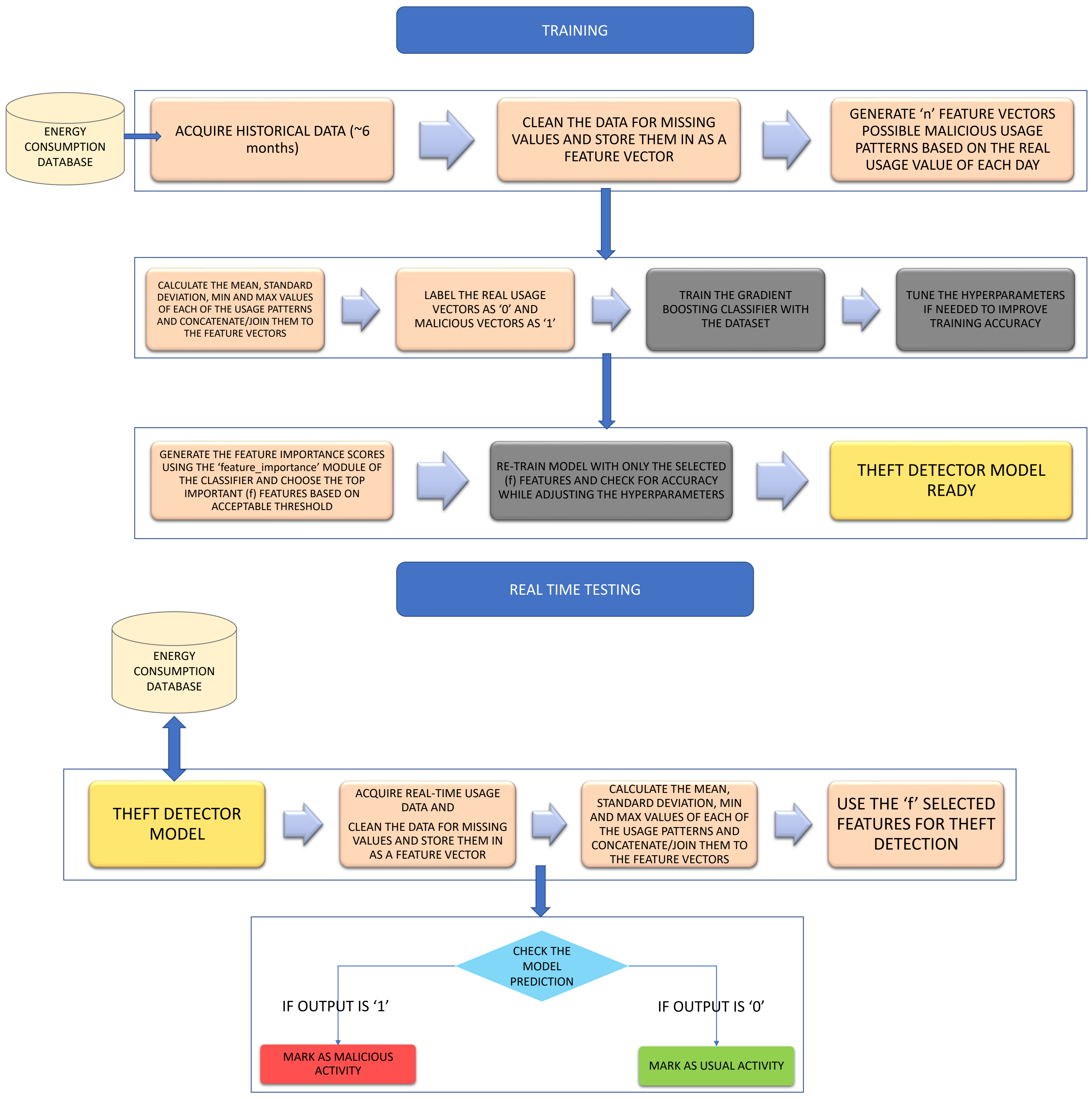


Fig 2. DR, FPR and evaluation time vs the no. of features selected for average of 100 random customers.

Real time model with novel pre-processing techniques



TRAINING

ENERGY CONSUMPTION DATABASE

ACQUIRE HISTORICAL DATA (~6 months)

CLEAN THE DATA FOR MISSING VALUES AND STORE THEM IN AS A FEATURE VECTOR

GENERATE 'n' FEATURE VECTORS POSSIBLE MALICIOUS USAGE PATTERNS BASED ON THE REAL USAGE VALUE OF EACH DAY

CALCULATE THE MEAN, STANDARD DEVIATION, MIN AND MAX VALUES OF EACH OF THE USAGE PATTERNS AND CONCATENATE/JOIN THEM TO THE FEATURE VECTORS

LABEL THE REAL USAGE VECTORS AS '0' AND MALICIOUS VECTORS AS '1'

TRAIN THE GRADIENT BOOSTING CLASSIFIER WITH THE DATASET

TUNE THE HYPERPARAMETERS IF NEEDED TO IMPROVE TRAINING ACCURACY

GENERATE THE FEATURE IMPORTANCE SCORES USING THE 'feature_importance' MODULE OF THE CLASSIFIER AND CHOOSE THE TOP IMPORTANT (f) FEATURES BASED ON ACCEPTABLE THRESHOLD

RE-TRAIN MODEL WITH ONLY THE SELECTED (f) FEATURES AND CHECK FOR ACCURACY WHILE ADJUSTING THE HYPERPARAMETERS

THEFT DETECTOR MODEL READY

REAL TIME TESTING

ENERGY CONSUMPTION DATABASE

THEFT DETECTOR MODEL

ACQUIRE REAL-TIME USAGE DATA AND

CLEAN THE DATA FOR MISSING VALUES AND STORE THEM IN AS A FEATURE VECTOR

CALCULATE THE MEAN, STANDARD DEVIATION, MIN AND MAX VALUES OF EACH OF THE USAGE PATTERNS AND CONCATENATE/JOIN THEM TO THE FEATURE VECTORS

USE THE 'f' SELECTED FEATURES FOR THEFT DETECTION

CHECK THE MODEL PREDICTION

IF OUTPUT IS '1'

IF OUTPUT IS '0'

MARK AS MALICIOUS ACTIVITY

MARK AS USUAL ACTIVITY

Results

- ✓Table 1 shows **that GBTD's superior performance.**
- ✓**Using feature extraction improves GBTD further** .
- ✓Figure 2 shows the general average trend of the 100 customers' theft DR, FPR, and evaluation time (training and testing time) vs N_f while using the proposed feature extraction method which verifies importance of feature extraction.

XGBoost	w/o synth	w/ Mean	w/ Std	w/ Min	w/ Max	w/ All 4
DR (%) FPR (%)	94 6	95 5	95 4	95 4	95 4	**96** **4**
CatBoost	w/o synth	w/ Mean	w/ Std	w/ Min	w/ Max	w/ All 4
DR (%) FPR (%)	97 5	97 6	97 5	97 5	97 5	**97** **3**
Light GBM	w/o synth	w/ Mean	w/ Std	w/ Min	w/ Max	w/ All 4
DR (%) FPR (%)	97 7	97 7	97 6	98 5	97 6	**97** **5**

Table 2. Performance comparison without or with new feature(s) (average of 100 random customers), where revised theft cases are used.

Conclusion

Out of the three GBCs, in terms of DR, both LightGBM and CATBoost outperformed XGBoost. However**, LightGBM appeared to be the fastest classifier with highest FPR while CATBoost performed the slowest with lowest FPR.**

We also numerically proved that GBTD with feature engineering not only minimizes FPR but also reduces customer data storage space as well as processing time.

Choice of the GBC depends on the availability of computation resources and acceptable FPR. As such the proposed algorithm would be beneficial for commercial use.

References

1. P. Jokar, N. Arianpoo, and V. C. M. Leung, "Electricity theft detection in AMI using customers' consumption patterns," *IEEE Trans. Smart Grid*, vol. 7, no. 1, pp. 216-226, Jan. 2016.
2. Irish Social Science Data Archive, [online] Available: http://www.ucd.ie/issda/data/commissionforenergyregulationcer/.
3. J. Heaton, "An empirical analysis of feature engineering for predictive modelling," Proceedings of IEEE SoutheastCon, pp. 1-6, April 2016, Norfolk, VA, USA [DOI 10.1109/SECON.2016.7506650].