Orlando Utilities Commision

Aravind Kumar Sankara Narayanan
University of Central Florida
aravindkumar@knights.ucf.edu

Vishnu Prasanth Reddy Patur
University of Central Florida
vipra9896@knights.ucf.edu

Vipul Bhogendra Kota
University of Central Florida
vipulbkota@knights.ucf.edu

Disha Bhatnagar
University of Central Florida
dishabhatnagar@knights.ucf.edu

ABSTRACT

Electricity theft is something that has become very common nowadays. It is kind of an invisible crime nobody talks about. With the advancement of technology, we now have smart meters, which are fast and reliable that gives more information than was ever thought of before. Smart technology can easily detect if someone is bypassing the meter. Smart meters have an ability to trigger alarms if it is tampered, which are very helpful. These alarms tell us when and where the theft is happening. These real time alerts can help authorities be alert if there is a sign of theft. But, even with the implementation of smart meters, sometimes theft events cannot be detected effectively. It is difficult at times, to know the nature of a theft, how complex it is, and the perpetrators now have even more advanced ways of stealing power, which cannot be detected by the tools available to us. Detecting such thefts is a costly and time-consuming process. Such sophisticated ways of stealing power have made the utility authorities feel that we need even more sophisticated and advanced analytics to deal with such activities. It becomes an exhausting task for the utility companies to detect and fight the people responsible for theft. With most utility companies thinking of ways to detect such thefts, OUC, Orlando utilities commission has suggested to develop an analytical system that monitor thefts in different locations and predict any potential thefts that are bound to happen in future. This report discusses the how such a system will be built and what features it will provide. According to another survey, theft is more rampant in residential and small commercial sectors. It is interesting to know that some locations are more prevalent to power thefts than the others. There should be a system where it shows how many power thefts are happening based on a particular location. In this project, we will be looking into some of these aspects also. We will help to build a system that correlates data, analyses and visualizes it. Visualization is a huge tool if we want to see the location based or meter-based power thefts. We will be showing effective and advanced visualizations in the project so that the OUC can easily detect thefts and take quick actions for it.

KEYWORDS

Machine learning, Xgboost, Quora, Features, Classifiers

1 PROPOSED SYSTEM

The product built provides an automated solution for the customers that completely reduces manual intervention like field visits. The product provides remote monitoring and detection irrespective of locations across regions/premises. The minimum viable product classifies whether the particular record is a theft and predicts the probability of potential theft getting occurred in and around Orlando region. The end product is achieved by building a complete state of the art of machine learning model that has the ability to identify complex electricity thefts that reports the thefts to the customers accurately and efficiently followed by an analytics dashboard that tracks the current electricity usage based on the historical data and thereby providing insights.

2 DATA ARCHITECTURE

The following diagram shows the design view of the data architecture that is followed throughout the project.

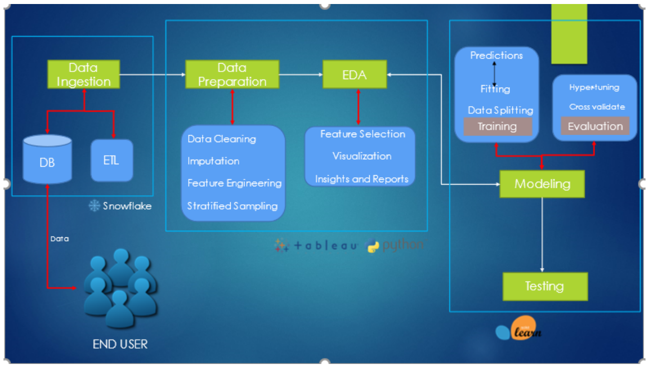


Figure 1: Block Diagram of workflow to detect data architecture

2.1 Data Ingestion :

The primary process involved in Data ingestion process is to move the data from one or more sources to a destination where it can be stored and further analyzed. Here the data are stored in the data warehouse AWS snowflake which needs to be extracted, transformed and loaded to perform analytical operations upon it.

2.2 Establishing database connectivity from snowflake:

The database connectivity has been established between the snowflake and local work station using python connector package. This helps performing the ETL analysis locally instead of doing on cloud platform.

2.3 Entity Relationship Diagram :

The data hosted in snowflakes has been distributed across number of tables, Hence it is important to identify the relationship between each tables that contribute to the main analysis. Below is the Entity Relationship Diagram

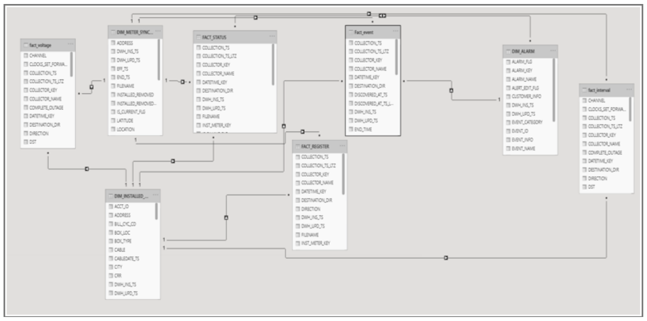


Figure 2: Entity Diagram

1. **Dim meter installed and Fact interval:**
The meter records the energy consumption in interval basis(15 min) the facts of interval energy consumption is present in fact interval . The tables are connected through installed meter key.
2. **Dim meter installed and Fact voltage:**
The meter records the voltage consumption in interval basis the facts of interval voltage consumption is present in fact voltage . The tables are connected through Installed meter key.
3. **Dim meter installed and Fact events :**
Each meters will have multiple events like Time Changed, Reset Error facts regarding this events are present in fact events . The tables are connected through Installed meter key.
4. **Dim meter installed and Fact status :**
Each meters will have multiple status for events like Time Changed, Reset Error facts regarding this status are present in fact status . The tables are connected through Installed meter key.
5. **Dim meter installed and Fact register :**
Each meters will register records based on time of use bucket. Facts of the register records are present in fact register . The tables are connected through Installed meter key.
6. **Dim Alarm and Fact event and Fact status :**
The event or status triggers alarm they are connected through event Id or status ID.

2.4 ETL Analysis :

Data from multiple tables hosted in AWS snowflake are extracted by using SQL queries. Used multiple joins and filter statement based on the primary keys to extract data and transform the data in a structured format. For actual SQL queries performed refer in the python notebook named ‘datapreparation.pynb’ in the deliverables.

2.5 Data Preparation:

The following methods has been incorporated on to the data inorder to transform the raw data into much more structured dataset to make good predictions.

2.6 Data Labeling :

In order to classify whether there is an energy theft happening, it is important to have a dependent variable which is the answer you want your machine learning model to predict. Hence to score the data, the target variable has been labeled as status category with values 1 and 0 where 1 represents the particular record is a theft, whereas 0 represents the particular record is not a theft. Since the data is based on interval records of every 15 minutes, if any one of the status category appears as tampered alert amongst all the 96 intervals in a day, all the energy records recorded by the meter on that particular day has higher probability of being a theft. Based on that analysis, the target variable is labeled.

2.7 Exploratory Data Analysis :

Exploring the data with different attributes and joining different tables gave us insights in the form of graphs and tables that helped us to know the relationship and statistics of the data. We created interactive Tableau dashboards to view the analysis. Below are some of the interesting insights about the data that are visualized in the form of plots.

1. **Energy Usage by month :**
The meter records the energy consumption in interval basis(15 min) the facts of interval energy consumption is present in fact interval . The tables are connected through installed meter key.

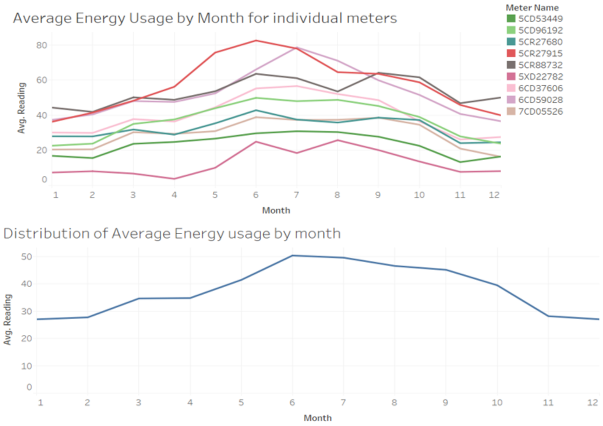


Figure 3: Monthly Usage

From the above line chart, the insights gained is that the average energy consumption has been considerably low in winter season (November – March) in comparison with other months. Hence during winter consumers tend to use less of utilities especially the air conditioning due to the temperature. The first plot visualizes the average energy usage for sample individual meters selected randomly, and the second plot showcases the overall cumulative average consumption by month. These plots serves as a baseline for our model.

(2) **Energy Usage by Hour :**
The above plot visualizes the energy consumption by hourly usage. From the line chart we get an insight about the data is that there has been high amount of energy consumption starting from morning 10 AM in the morning until midnight 12 AM where different places across the region starting from restaurants, hospitals, private and public sectors start functioning their daily work which leads to high energy consumption, whereas after 12, the entire city goes under sleep with less use of energy overall. This visualization gives an interesting insight and serves as an baseline about hourly usage of consumers.

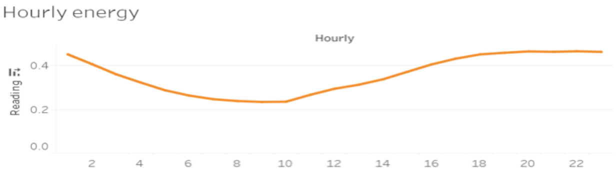


Figure 4: Hourly Usage

(3) **Tampered Meter Vs Non- Tampered Meters :**
After labelling the data as tampered and non tampered, we wanted to visualize how the energy usage varies between tampered and non-tampered meters. Below line chart showcases the difference in energy consumption between tampered and non tampered meters visually. The insights gained is that in non-tampered/original meter there has been a gradual decrease and increase in each hour of the day. Whereas in Tampered meters we can see frequent fluctuations with energy consumption decreasing and increasing abruptly.

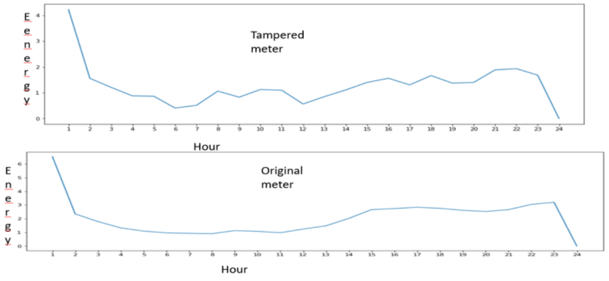


Figure 5: Hourly Usage

3 TABLEAU DASHBOARDS:

All the visualizations and insights have been grouped together to provide a collated view which can be further extended as a real time interactive dashboard that can be served as an monitoring tool to keep track of energy usage across the region. The minimum viable product produced in this project is three interactive tableau dashboards that provides different insights within the region with respect to county wise, premise wise, zip code ,etc . Note these dashboards are built with respect to sample data that serves as a baseline of the entire dataset.

Dashboard 1 : The following Interactive dashboard showcases the Energy usage by location, Date and the Status category by means of geographical location, time series with trend lines and bar chart

Here is the tableau link to view the dashboard : https://public.tableau.com/profile/vishnu.prasanth.reddy#!/vizhome/dash_16049735669410/Dashboard2

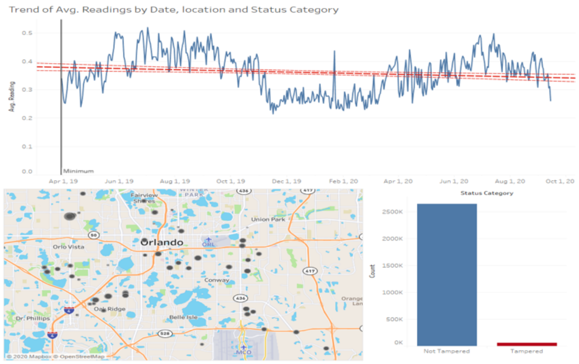
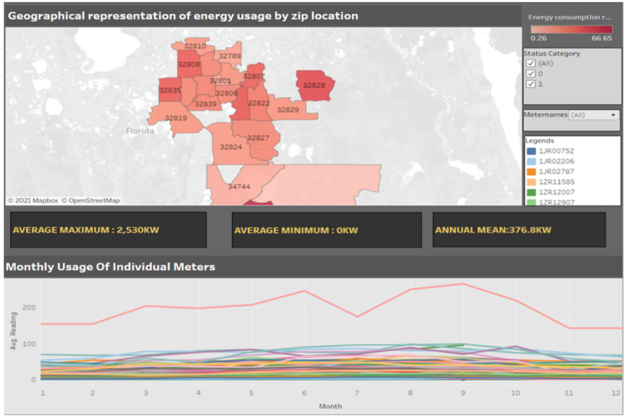


Figure 6: Dashboard 1

Dashboard 2 : The following dashboard showcases the Energy usage by zip location and the individual meters average monthly energy usage by means of hue geographical maps and line charts. Followed by the average maximum, minimum and the annual mean energy consumed per region are projected.

Here is the tableau link to view the dashboard : https://public.tableau.com/profile/aravind.kumar.sankara.narayanan7651#!/vizhome/dashboard2_16050278543240/Dashboard12?publish=yes beginfigure[h]


Geographical representation of energy usage by zip location

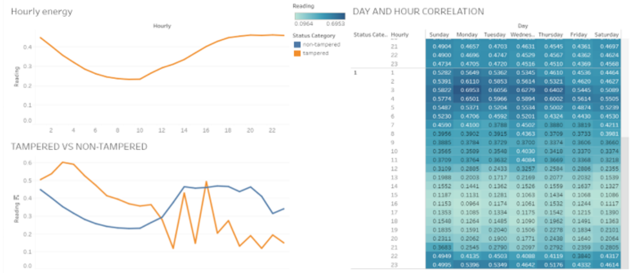
AVERAGE MAXIMUM : 2,530KW

AVERAGE MINIMUM : 0KW

ANNUAL MEAN:376.8KW

Monthly Usage Of Individual Meters

Dashboard 3 : The following dashboard shows average energy usage by hour , shows comparison of tampered and non tampered meters and It shows correlation heat map to understand energy usage with respect day of week and hour of day. beginfigure[h]



DAY AND HOUR CORRELATION

Status Cate..	Hourly	Sunday	Monday	Tuesday	Wednes..	Thursday	Friday	Saturday
	21	0.4904	0.4657	0.4703	0.4631	0.4545	0.4361	0.4697
	22	0.4900	0.4696	0.4747	0.4529	0.4567	0.4362	0.4624
	23	0.4734	0.4705	0.4720	0.4516	0.4510	0.4369	0.4568
1	1	0.5282	0.5649	0.5362	0.5345	0.4610	0.4536	0.4464
	2	0.5391	0.6110	0.5853	0.5614	0.5321	0.4620	0.4627
	3	0.5822	0.6953	0.6056	0.6279	0.6402	0.5445	0.5089
	4	0.5774	0.6501	0.5966	0.5894	0.6002	0.5614	0.5505
	5	0.5487	0.5371	0.5204	0.5534	0.5002	0.4874	0.5239
	6	0.5230	0.4706	0.4592	0.5201	0.4324	0.4430	0.4530
	7	0.4590	0.4100	0.3788	0.4502	0.3880	0.3819	0.4211
	8	0.3956	0.3902	0.3915	0.4363	0.3709	0.3733	0.3981
	9	0.3885	0.3784	0.3729	0.3700	0.3374	0.3606	0.3660
	10	0.3565	0.3589	0.3548	0.4030	0.3418	0.3370	0.3374
	11	0.3709	0.3764	0.3632	0.4084	0.3669	0.3368	0.3218
	12	0.3109	0.2885	0.2433	0.3257	0.2584	0.2806	0.2355
	13	0.1988	0.2003	0.1717	0.2169	0.2077	0.2032	0.1539
	14	0.1552	0.1441	0.1362	0.1526	0.1559	0.1637	0.1327
	15	0.1187	0.1131	0.1281	0.1063	0.1434	0.1068	0.1086
	16	0.1153	0.0964	0.1174	0.1061	0.1532	0.1244	0.1117
	17	0.1353	0.1085	0.1334	0.1175	0.1542	0.1215	0.1390
	18	0.1548	0.1264	0.1485	0.1090	0.1962	0.1491	0.1363
	19	0.1835	0.1591	0.2040	0.1506	0.2278	0.1834	0.2101
	20	0.2311	0.2062	0.1900	0.1771	0.2438	0.1640	0.2064
	21	0.3683	0.2545	0.2790	0.2097	0.2792	0.2359	0.2805
	22	0.4949	0.4135	0.4503	0.4088	0.4119	0.3840	0.4317
	23	0.4995	0.5396	0.5349	0.4642	0.5176	0.4332	0.4614

Here is the tableau link to view the dashboard :
https://public.tableau.com/profile/vipul.bhogendra.kota#!/
vizhome/ouc_16074432180080/Dashboard1?publish=yes

4 MODELING:

After cleaning the data and labeling the target variable explicitly, now the aim is to build a machine learning model that identifies the meter fraud records. Since the data has unnecessary columns we have eliminated few and sampled 750k instances randomly to produce unbiased estimate and to reduce time complexity. Now the data set is ready to split into training and testing set.

4.1 Data Splitting:

The data set has been split in to 3 parts say training, validation and testing set. First the dataset has been split into 80% train set and 20% testset. 80%training set is further divided into trainset which is 80%of training set and 20% of training set which is validation set. The objective of splitting the data like this is to train the model, validate it using validation set and evaluate using test set. This was we can consider overfitting.

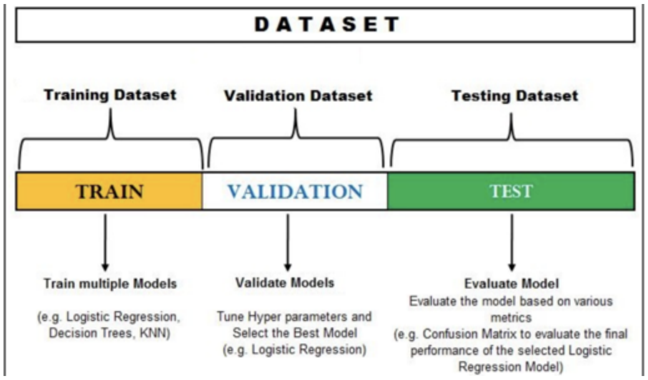

DATASET
Training Dataset
Validation Dataset
Testing Dataset
TRAIN
VALIDATION
TEST
Train multiple Models
(e.g. Logistic Regression, Decision Trees, KNN)
Validate Models
Tune Hyper parameters and Select the Best Model
(e.g. Logistic Regression)
Evaluate Model
Evaluate the model based on various metrics
(e.g. Confusion Matrix to evaluate the final performance of the selected Logistic Regression Model)

Figure 7: data split

4.2 Training:

We have trained 4 different machine learning models. To reduce the time complexity, we have manually set the parameters by using trial and error method instead of hyper parameter tuning. Since it's an imbalanced dataset, where the target variable's minority class is only 2%. we can have used SMOTEEN (Synthetic minority over sampling technique) to improve the performance of the model. SMOTENN :A widely adopted technique for dealing with highly unbalanced datasets. It consists of removing samples from the majority class (under-sampling) and / or adding more examples from the minority class (over-sampling).

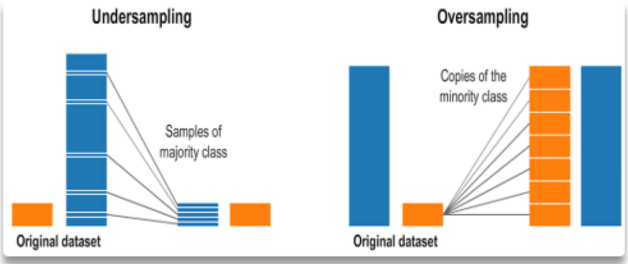


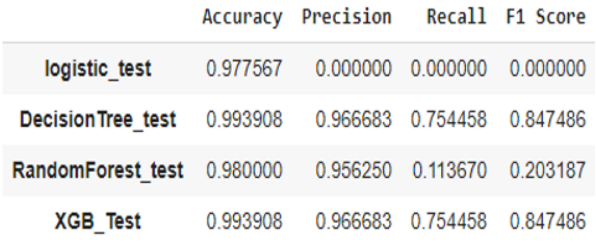
Figure 8: SMOTEEN

5 EVALUATION

Considering the business value in the problem statement handled, we consider precision and recall as one of the important metrics while validating our performance of the model built. Predicting a non-tampered meter as tampered is acceptable as it can be rectified after further investigation but missing out on a tampered case is unacceptable. We are considering F1-score metric is to find an equal balance between precision and recall.

5.1 Without SMOTEEN

The performance metrics for all the 4 models were mentioned below. We can observe that the accuracy of all the models is above 97.7% because the data set is imbalanced, and the minority class variables were also classified as majority class which can be explained by logistictest. All the instances here were classifies as non-tampered. So, we can't consider accuracy as our evaluation metric. To overcome this, we are considering F1-score as our evaluation metric. We can observe that both the decision tree and XGboost have similar f1 scores this can be explained by ROC and AUC curves. From the figure we can observe that XGboosts AUC is grater than any other model.



	Accuracy	Precision	Recall	F1 Score
logistic_test	0.977567	0.000000	0.000000	0.000000
DecisionTree_test	0.993908	0.966683	0.754458	0.847486
RandomForest_test	0.980000	0.956250	0.113670	0.203187
XGB_Test	0.993908	0.966683	0.754458	0.847486

Figure 9: Without SMOTEEN

We can observe that both the decision tree and XGboost have similar f1 scores this can be explained by ROC and AUC curves. From the figure we can observe that XGboosts AUC is grater than any other model.

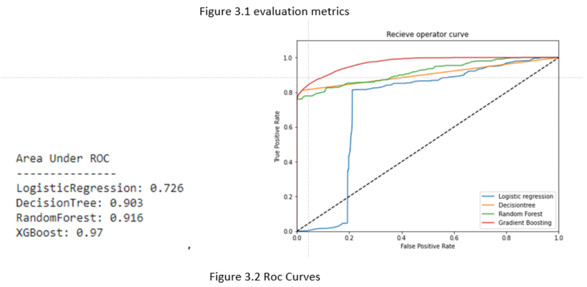


Figure 3.1 evaluation metrics

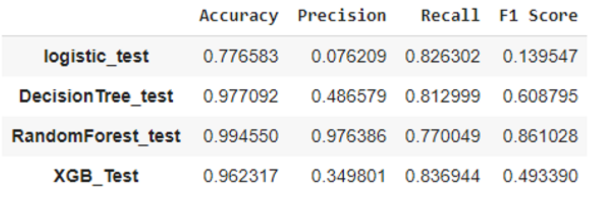
```
Area Under ROC
---------------
LogisticRegression: 0.726
DecisionTree: 0.903
RandomForest: 0.916
XGBoost: 0.97
```

Figure 3.2 Roc Curves

Figure 10: With SMOTEEN

5.2 With SMOTEEN

with smoteen we can observe that there a increased performance of the models, Random Forest has delivered highest accuracy, precision and F1-score. So we can consider this as a final model for prediction.



	Accuracy	Precision	Recall	F1 Score
logistic_test	0.776583	0.076209	0.826302	0.139547
DecisionTree_test	0.977092	0.486579	0.812999	0.608795
RandomForest_test	0.994550	0.976386	0.770049	0.861028
XGB_Test	0.962317	0.349801	0.836944	0.493390

Figure 11: With SMOTEEN

6 ISSUES ENCOUNTERED WHILE DEVELOPING MVP

Dealing with huge amount of data was a challenge while developing our MVP. With over 6 billion records spread across different tables in the Datawarehouse snowflake. Extracting data from the data warehouse into the local stations was a huge challenge were we encountered higher run time during extraction. Since local stations such as laptop having minimal CPU configurations, there were times were run time memory issues was encountered.

7 ACCOMPLISHMENTS

As per the problem statement and the expected solution, we as a team accomplished the expectations and come out with a better solution that can be leverage across globally. From the requirements standpoint the solution has been achieved. But from the quality standards it require lot of time with more data to provide the best MVP.

8 IMPROVING THE ACCURACY OF THE SOLUTION

The product was tested with sample data which is a representation of the entire data, however it is important to train on massive amount of data to make sure the model accurately classifies. This can be achieved only through neural networks which identifies complex patterns and decreases the error rate close to zero. But these kind of neural network training require higher amount of GPUs to train on. Thus our next progress would be to try neural networks on higher configuration super power stations.

9 SOFTWARES USED

a. Snowflake b. SQL c. Python d. Tableau e. MSOFFICE f. Google colab g. Scikitlearn

REFERENCES

[1] https://www.distributedenergy.com/home/article/13024533/energy-theft-a-growing-problem: :text=Industry%20experts%20estimate%20that%20%246,credit%20card%20data%20

[2] https://www.sciencedirect.com/science/article/abs/pii/S030142151000861X.

[3] https://hiokiusa.com/electricity-theft-the-crime-that-nobody-talks-about/

[4] https://machinelearningmastery.com/what-is-imbalanced-classification/ 05.09090, 2016.