José Pedro Figueiredo Machado

Anomalia e deteção de fraude em dados de telemetria

U.PORTO

FACULDADE DE CIÊNCIAS
UNIVERSIDADE DO PORTO

Departamento de Ciência de Computadores
Faculdade de Ciências da Universidade do Porto
julho 2022

José Pedro Figueiredo Machado

Anomalia e deteção de fraude em dados de telemetria

Relatório de Estágio Curricular

Orientador: Pedro Faria (nome completo)
Co-orientador: Tânia Carvalho (nome completo)

Departamento de Ciência de Computadores
Faculdade de Ciências da Universidade do Porto
julho 2022

Acknowledgments

I would like to express my special thank of gratitude to my teacher as well as our university that gave me the opportunity to do this project on the topic Machine learning.

In addition I would also like to thank my supervisors Pedro Faria and Tânia Carvalho for all their help and advice with this work. Last but not least a big thanks to everyone that allowed me to succeed, including, my family, freinds and girlfriend.

I am really thankful to all of them.

Prefácio

The desire for this research work has come from the university subject Project/Internship, this aims to assess the students' ability to face research challenges in a university environment and to promote the students' curricular enrichment with a view to their professional integration.

The project in question was carried out for and with the supervision of TekPrivacy, a company with a mission to develop innovative support technology always linked to data.

This work would also not have been completed without the participation of all the parties who were directly or indirectly involved in the research for this paper.

Abstract

Energy fraud detection is a critical aspect of smart grid security and privacy preservation. Machine learning and data mining have been widely used by researchers for extensive intelligent analysis of data to recognize normal patterns of behavior such that deviations can be detected as anomalies. This paper discusses a application of machine learning technique for examining the energy consumption data to report energy fraud using Isolation Forest model and smart meters data. The approach detection identifies divers form of fraudulent activities resulting from unauthorized energy usage.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Machine learning plays a key role in predicting fraud and ensuring correct process performance upon anomaly detection. Current studies and applications are mostly conducted in an unsupervised manner. However, the best approach to solving these risks that organizations are exposed to, particularly those dealing with large temporal data, has yet to be found.

Data from utility meters such gas, electricity, water is a rich source of information for distribution companies, beyond billing. In this paper we present a unsupervised technique, which primarily but not only feeds on meter information, to detect meter anomalies and customer fraudulent behavior. Our system detects anomalous meter readings on the basis of models built using machine learning techniques on past data.

The full system has been developed with a refactorised version of the data related to energy consumption in London.

1.1 Ambit

It was decided within the scope of the Intership program of the faculty of cience from the university of Porto, and with the assistance of the Tekprivacy company, to create a system for anomaly and fraud detection in telemetry data.

1.2 Objectives

The project began with the following objectives proposed:

1. Study state-of-art techniques for fraud detection in temporal data sets.
2. Experiment methods and train unsupervised learning models.
3. Apply the techniques and models to detect anomalies and/or potential cases of fraud.
4. Evaluate and assess each model.

1.3 Methodology

The adopted methodololy was CRISP DM [3] which stands for "Cross Industry Standard Process for Data Mining". It is a process model that serves as the base for a data science process and its divided in six sequential phases:

1. Business understanding: What does the business needs
2. Data understanding: What data do we have or need and if it is clean
3. Data preparation: Organization of the data for modeling
4. Modeling: Modeling techniques that should be apply
5. Evaluation: Which model best meets the business objectives or model analisys
6. Deployment: How do stakeholders access the results

Like anything else CRISP has both benefits and weaknesses, but making the balance of both this methodology seemed very promissing and solid. It has benefits such as being generalizeble, adoptable and flexible and on the other hand problems may emerge because it could be considered a model to rigid, antique and a little heavy on documentation.

Chapter 2

State of the art and learning summary

Energy and water fraud are a very common problems and currently lack the detection accuracy that is needed to minimize it. Researchers have proposed and applied various machine learning techniques to address this issue. In order to choose the model that best fits our data, several possibilities have been studied.

2.1 Related studies

The processes of committing both energy and water fraud, as well as the way of measuring consumption, are very identical, so studies of one type of data can be applied to the other.

A study by Christa Cody and Vitaly Ford [5] applied Decision Trees to detect fraudulent activity. The proposed system includes:

Pre-processing of data: Create an uncorrupted dataset with no missing values.

Feature selection: Limit selected features

Data aggregation: Manipulate measurement intervals

Indexing and compressing: Provide faster data retrieval

After using the model to predict results the anomaly detection was done using the Root Mean Squared Error formula:

$$\text{RMSE} = \sqrt{\frac{\sum_{t=1}^{n}(y'_t - y_t)^2}{n}}$$

y'_t predicted energy
y_t energy consumption
n light number of instances

Care had to be taken to ensure that the model did not over-fit or under-fit but this one successfully applied the Decision Tree model to accurately predict possible anomalies.

Also Vitaly Ford together with Ambareen Siraj and William Eberle [6] investigated the field of energy fraud detection, in this case using Neural Networks. For data processing, strategies such as Data cleaning to eliminate inaccuracies, incompleteness and inconsistencies, Feature selection and both Indexing and compressing were also used. The model construction was divided into five steps:

Parameter selection of input/output for the neural network: Choose the value to use in the input layer, the results of the output layer and the number of hidden layers needed.

Generation of Training and Validation Datasets: Generate various unique training instances using the training set by selection data points one by one and marking them as output nodes for the training instance.

Training the Neural Network: Using the already generated training dataset the neural network is trained resulting in the learning of the consumer consumption behavior.

Prediction: Run the model applying the validation dataset. The output is the predicted value of the expected data.

Detection of Deviation: Working with the RMSE we verify if its result is above a certain threshold, and if it is then the data in the validation set could possibly correspond to an energy fraud.

This research demonstrated that a neural network can be used as a computation intelligent tool for detecting different types of anomalies. for more specific results:

Table 2.1: Confusion matrix.

TN	75.00%
TP	93.75%
FN	6.25%
FP	25.00%

There were also a research paper made by Assia Maamar and Khelifa Benahmed [4] where comparisons between models were made:

Table 2.2: Performance measures between best models.

Ref	Model	Accuracy %	Recall %	Data set
CIT	Neural networks	83.5	29.8	Light S.A Company, Brazil
CIT	Neural networks	93.75	78.94	Collected by CER^b, Ireland
CIT	Support Vector Machines	72.70	53.0	TNBD[a] data set
CIT	SVM (Gauss)	77.41	64.0	TNBD[a] data set

2.2 Practical study

An analysis that attempted to detect anomalies in a dataset referring to energy consumption using different algorithms was also reviewed and put into practice [1]. The algorithms implemented were:

1. Re-partition of data into categories: The dataset was separated into four different categories, weekdays and weekends, and then each of these into days or nights.

2. Cluster based anomaly detection (K-mean): Number of clusters calculated using the elbow method, after that the points that were to far from the centroid of its group were considered anomaly's.

3. Gaussian/Elliptic on each categories separately: Gaussian method applied for each category and envelope to get the anomaly's.

4. Markov Chain: Define 5 levels of value (very low, low, average, high, very high) and use Markov chain to calculate the probability of sequence. If the probability is very weak we consider the sequence as an anomaly.

5. Isolation Forest: Simple to apply, works well with different data re-partitions and it is efficient with high dimension data.

6. One class SVM: This algorithm performs well for multi-modal data and had a similar result to isolation forest but fond some anomalies in average values.

7. RNN: Learn to recognize sequence in the data and then make prediction based on the previous sequence. We consider an anomaly when the next data points are distant from RNN prediction.

All the algorithms used were able to predict anomalies with more or less accuracy. Having isolation forest, SVM and RR proved to be much more preferable than the others.

Chapter 3

Business Understanding

The outdated electrical grid is undergoing a slow transition to the new Smart Grid technology. This allows the electrical grid to be utilized in numerous ways to become more efficient, reliable, and beneficial to the consumer as well as the service providers. The stream of data generated by the smart meter can provide a log of fine-grained energy consumption that allows the consumers to monitor their energy usage for numerous reasons including financial and environmental. Although the accessibility of the data creates many opportunities for improvements, the same accessibility creates an avenue for potential privacy and security violations. However, security violations related to fraudulent activities can be reduced by intelligent analysis of the fine grained data. By learning the characteristic patterns within the fine-grained data, normal energy usage can be predicted and as such, any anomalies can be reported to be potential energy fraud.

In this paper, we discuss the use of energy fraud detection to address some of these security violations. There are two types of energy fraud to consider:

1. The consumer's smart meter reports less energy consumption than actually consumed.
2. The consumer's smart meter reports more energy consumption that actually used due to rogue connections.

These types of fraud can be created by numerous mechanisms such as:

- Unauthorized tapping to electricity line.

- Bypassing the smart meter to report customized energy consumption.
- Tampering with the smart meter by implanting chips to slow down its readings.

Energy theft was the major issue in traditional power systems in the global. In the U.S. individually, the lowest estimate show electricity theft still costs consumers and utilities well over $1 billion each year. Theft of electricity and gas cost UK energy consumers $299 million every year. Therefore, It is substantial to develop efficient and credible detection systems, which can reveal the energy theft threats.

Chapter 4

Data Understanding

Before the approach is described, it is imperative to introduce the data to be analyzed for deriving computational intelligence.

4.1 Data Collection

The dateset [2] under investigation consists of smart meter consumption data from approximately 5.567 residential London households.

These representative samples were collected between November 2011 and February 2014 by acron group and darksky and the data values were captured every 3O minutes. Raw data was stored in different files int a total of around 26 million entries having the following parameters:

1. Information on the households panels.
2. The block files with the half-hourly smart meter measurement.
3. Detailed Daily information.
4. Details on the acorn groups and their profile of the people in the group.
5. Daily data from the weather conditions.
6. Hourly data from the weather conditions.
7. Holidays.

4.2 Data description

The files used for our study were the ones containing the data of the half-hourly/daily measurements, weather and holidays.

Table 4.1: Daily data file structure after treatment

House	Data	Energy consumption Kw/h
MAC000116	2011-12-14	21.846
MAC000116	2011-12-15	32.334
...	...	...
MAC003133	2014-02-27	21.072
MAC003133	2014-02-28	0.160

1.979.209 entries

Table 4.1 represents a small sample of the daily measurements file where the data is represented in three columns. The first column represents the household id. The second column shows the date and time associated with the meter reading and the third column is the energy consumption measurement in kW.

Table 4.2: Weather data file structure

Date	Maximum	Minimum	Mean
2011-11-11	11.96	8.85	10.405
2011-12-11	8.59	2.48	5.535
...	...	...	...
2014-02-12	8.83	3.03	5.930
2014-02-15	9.90	5.38	7.640

882 entries

Table 4.2 is relative to the weather data. The first column shows the day in question. The second and third columns represents maximum and minimum temperature, respectively and the last column mean temperature of the day. The measurements are made in C$^{\underline{o}}$.

Table 4.3: Holidays data file structure

holidays	Type
2012-12-26	Boxing Day
2012-12-25	Christmas Day
2012-09-04	Easter Monday
2012-06-04	Good Friday
...	...
25 entries	

Table 4.3 relative to the holidays, is constituted by only 2 rows, the day of the holiday and what holiday it represents.

4.3 Data exploration

Explore data: Dig deeper into the data. Query it, visualize it, and identify relationships among the data.

Data exploration is the first step in data analysis involving the use of data visualization tools and statistical techniques to uncover data set characteristics and initial patterns.

For organization and processing reasons it was decided to use the energy data grouped by days. In this set there are 3,510,433 entries, if we use the set with records every half hour, considering that a day has 24h, we would have a total of MATH(2*24*n^{o}of daily entries).

4.3.1 Energy

Table 4.4: Energy description

	Energy Kw/h
min	0.0
max	332.55
mean	10.12
std	9.12

We decided to divide the energy into categories to better understand and visualize this parameter:

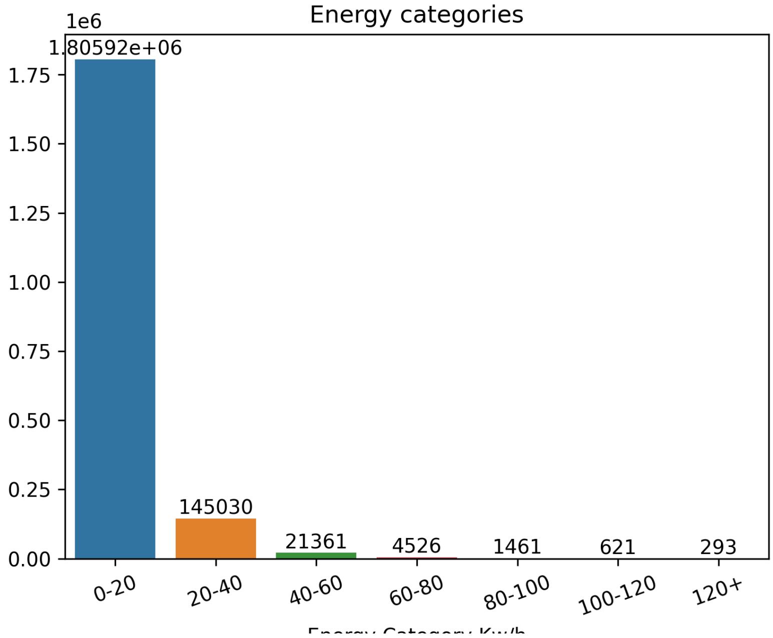


Figure 4.1: Energy Catgory distribution

As we can see in the figure above, the most common daily consumption is between 0 and 20 Kw/h. From this information we can expect that the categories with the highest number of anomalies would be exactly others than that one.

4.3.2 Year

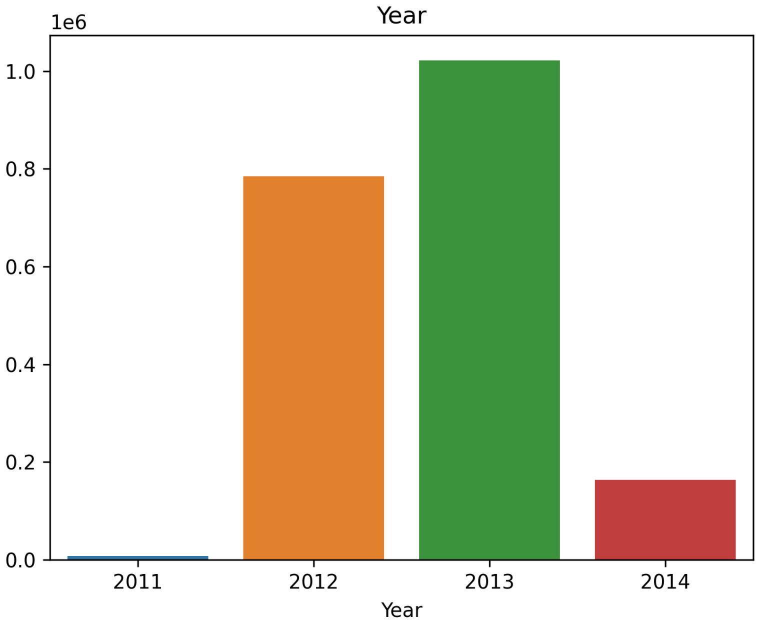


Figure 4.2: Year distribution

In the figure above we can see the distribution of records over the four years. The years 2012 and 2013 contain a much larger amount of entries than the others. This means that certain months of 2011 and 2014 will not be analyzed by the model, which can negatively alter the influence of possible parameters such as month, season and temperature.

4.4 Data quality

(Verify data quality: How clean/dirty is the data? Document any quality issues.)
Data quality is the measure of how well suited a data set is to serve its specific purpose. Measures of data quality are based on data quality characteristics such as accuracy, completeness, consistency, validity, uniqueness, and timeliness.
To deal with this we use a precess called data cleaning, where we fix or remove incorrect, corrupted, incorrectly formatted, duplicate, or incomplete data within a dataset. When combining multiple data sources, there are many opportunities for

data to be duplicated or mislabeled.

4.4.1 Energy

Table 4.5: Data missing values

Parameters (daily)	Missing values
Energy min	30
Energy max	33
Energy mean	30
Energy std	11331

As we can see there are some missing values in the energy data, especially when it comes to the energy std, that we will have to deal with when preparing the data.

Table 4.6: Energy data quality

Measurements	Values Kw/h
Mean	10.124
Std	9.129
Min	0
Max	332.556

Data quality characteristics such as consistency, validity and uniqueness, seem to make sense. One would expect the lowest recorded energy value to be 0 and the highest a clear anomaly. The average of 10 Kw/h appears to make sense given the data as well as the standard deviation.

4.4.2 Weather

Table 4.7: Data missing values

Parameters (daily)	Missing values
Temperature min	0
Temperature max	0
Temperature day	0

There were no missing values in the weather data, but some days were missing, for example, we only have 28 days recorded in the month of January 2012, in this case we would have day 1, 2, 4, 5, with day 3 missing.

Table 4.8: Energy data quality

Measurements	Max	Min
Mean	13.660	7.414
Std	6.184	4.889
Min	-0.06	-5.66
Max	32.4	20.54

As we can see the values seem to make sense overall. The mean values are within the minima and maxima, the standard deviation is lower than the mean, and the maxima and minima are also in line with the spectables values.

4.4.3 Holiday

The holiday data is quite short, there are no missing values, and none of it shows irregularities.

Chapter 5

Data Preparation

5.1 Energy

Since the present values seemed to make sense we can only deal with the non-existent ones. It was decided to remove the null energy records, as it would not make sense to replace them with an average as this would alter our models.

It would make no sense to give the models houses with only one entry when they could not learn with such a low number.

Table 5.1: Number of records per house

House	Count
MAC005563	1
MAC005560	1
...	...
MAC000147	829
MAC000149	829

In view of these numbers it was decided that it would be best to leave out houses with a number of registrations less than 650. This will help to reduce the extensive amount of entries as well as improve the accuracy of our models.

From the given data it was possible to create parameters in order to help our model better recognize an anomaly.

First, from the "day" attribute it was possible to retreat the day of the week it represents:

Monday=0 → Sunday=6

```
data.frame['week_day'] = data.frame['time'].dt.dayofweek
```

Also using the "day" attribute, a binary parameter referring to the weekend was created:

Weekday = 1 and Weekend = 0

```
data.frame['weekend'] = (
        data.frame['week_day'] < 5).astype(int)
```

To facilitate the construction of graphs and the reading of data from the model the date of the days in question were passed to numeric.

```
data.frame['day_int'] = (
        data.frame['time'].astype(np.int64)/100000000000).astype(np.int64
```

For the same purposes the name of the house in question was changed from text to numeric.

```
#House name example MAC005563
data.frame['house_int'] = data.frame['house'].apply(
        lambda x: int(x.lstrip("MAC")))
```

Finally we extracted the months of the years and the years themselves:

January=1 → December=12

```
data.frame['month'] = data.frame['time'].dt.month
```

2011, 2012, 2013, 2014

```
data.frame['year'] = data.frame['time'].dt.year
```

5.2 Weather

As verified in the data quality section the weather dataset had a problem with missing values. To solve this problem we had to either create the missing days ourselves or ignore the weather parameter altogether. We decided then that to the missing values we would associate the average of the general temperature.

```
data.frame['weather_mean'].fillna(
        value=data.frame['weather_mean'].mean(),inplace=True)
```

With that done we just have to create the parameter with the average temperature of the day:

```
weather.frame['weather_mean'] = (
        weather.frame["max"] + weather.frame["min"]) / 2
```

and attach it to the main dataframe:

```
data.frame = data.frame.merge(
        weather.frame,how='left' , left_on='time', right_on='time')
```

5.3 Holidays

There was no need to create new parameters or clear data referencing the holidays, so we added the data to the main dataframe directly.

holiday=1 and Normalday=0

```
data.frame['holiday'] = data.frame['time'].apply(
        lambda x: (x in holidays.frame.time.values))
```

Chapter 6

Modeling

6.1 Choosing the model

With the anomaly detection process in mind two approaches were chosen, Isaolation Forest and SVM (Support Vector Machinhes).Isolation Forest because it is an unsupervised learning algorithm for anomaly detection, meaning that it identifies the anomaly by isolating outliers in the data. This seemed to be a good option given the data we had. As for the Support Vector Machine the supervised machine learning algorithm is often cited and used in classification problems. However, SVM is also increasingly being used in a one-class problem which makes it more suitable for the job.

6.2 Isolation Forest

Isolation Forest is a machine learning algorithm for anomaly detection. It is an unsupervised learning algorithm that identifies the anomaly by isolating outliers in the data.

This model is based on the Decision Tree algorithm. It isolates outliers by randomly selecting a feature from the given set and then randomly selecting a value divided between the maximum and minimum values of that feature. This random splitting of features will produce shorter paths in the trees to the outliers data points, that will distinguishing them from the rest of the data.

In general, the first step in anomaly detection is to build a profile of what is "normal" and then report anything that cannot be considered normal as anomalous. However, the forest isolation algorithm does not work on this principle, it does not first define "normal" behavior, and it does not calculate point-based distances. Isolation Forest works by explicitly isolating anomalies by isolating anomalous points in the data set. It is based on the principle that anomalies are few and different observations, which should make them easier to identify. The model uses a set of Isolation Trees for the data points given to isolate the anomalies.

Isolation Forest recursively generates partitions in the data set by randomly selecting a feature and then randomly selecting a split value for the feature. Presumably anomalies require fewer random partitions to isolate compared to the "normal" points in the dataset, so anomalies will be the points that have a shorter path length in the tree, path length being the number of edges traversed from the root node.

Using Isolation Forest, not only can we detect anomalies faster, but we also require less memory compared to other algorithms.

Model build

```
model=IsolationForest(max_samples='auto',
contamination=float(0.0005), random_state=random_state)
model.fit(data.frame[features].values)
```

6.3 Support Vector Machines

One of the most important qualities of SVM is that it creates nonlinear boundaries by projecting the data with higher dimensions in the space. This is done using its nonlinear function, it uses the function to lift the feature space F of the observations from the I space which cannot be separated by a linear function even straight line. This hyperplane is used to separate the data of one class from the other classes being also possible to be in the form of a nonlinear curve.

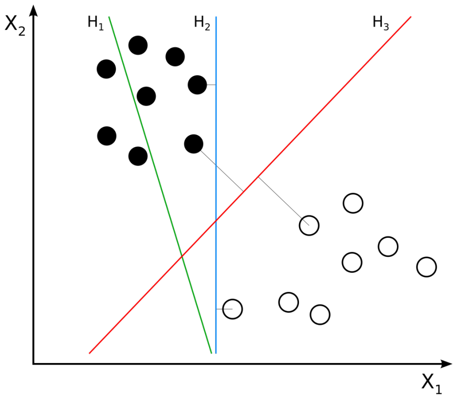


Figure 6.1: SVM example

The above image represents the hyperplanes H1, H2, H3 used to separate the data points of two classes. We can see that H1 is not diving them but H2 and H3 are, fitting a hyperplane between the data points calculates the distance that can be considered from hyperplane to the closest point. It should be both equaland maximized for each category data. In the image we can see that H3 is a better fit option than the H2 where H1 doesnt fit at all. To avoid overfitting of the model slack variables may be introduced which allow some data points to lie within the margin.

In the One-Class version all data belong to a single class. In this case, the algorithm is trained to learn what is “normal”, so that when a new data is shown the algorithm can identify whether it should belong to the group or not. If not, the new data is labeled as out of ordinary or anomaly.

Model build

```
model=OneClassSVM(contamination=float(0.0005),
nu=0.5, max_iter= -1)
model.fit(data.frame[features].values)
```

6.4 Model decision

After the aplication of the models we could see that both are capable of properly modeling multi-modal data sets. One class SVM is sensitive to outliers, making it more appropriate for novelty detection. But only when the training data is not contaminated with outliers which is the case.

An example fo this would be:

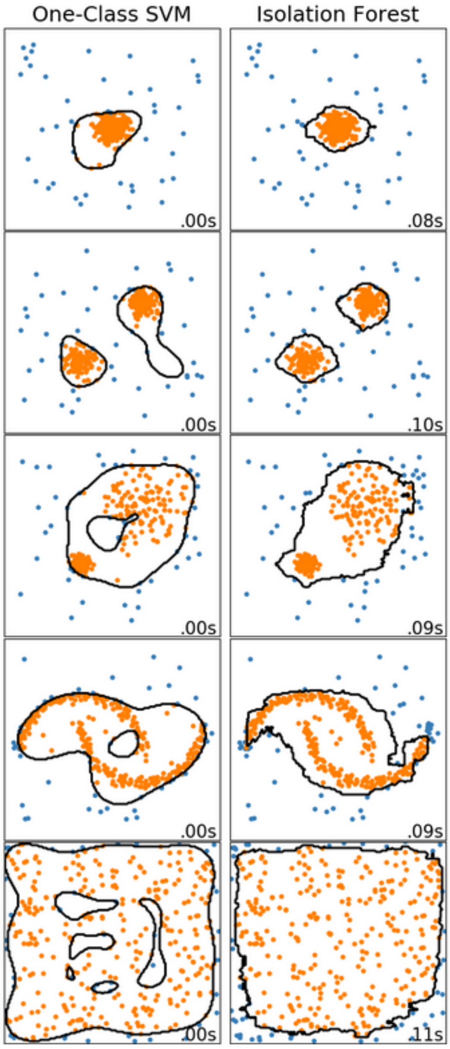


Figure 6.2: One-Class SVM and Isolation forest comparation

Also since the splits of the decision tree are chosen at random, isolation forest is faster to train, this is an important factor givent he amount of data we are working with. Having all this in mind it was then decided to further explore the Isolation Forest algorithm instead of SVM one.

Chapter 7

Evaluation

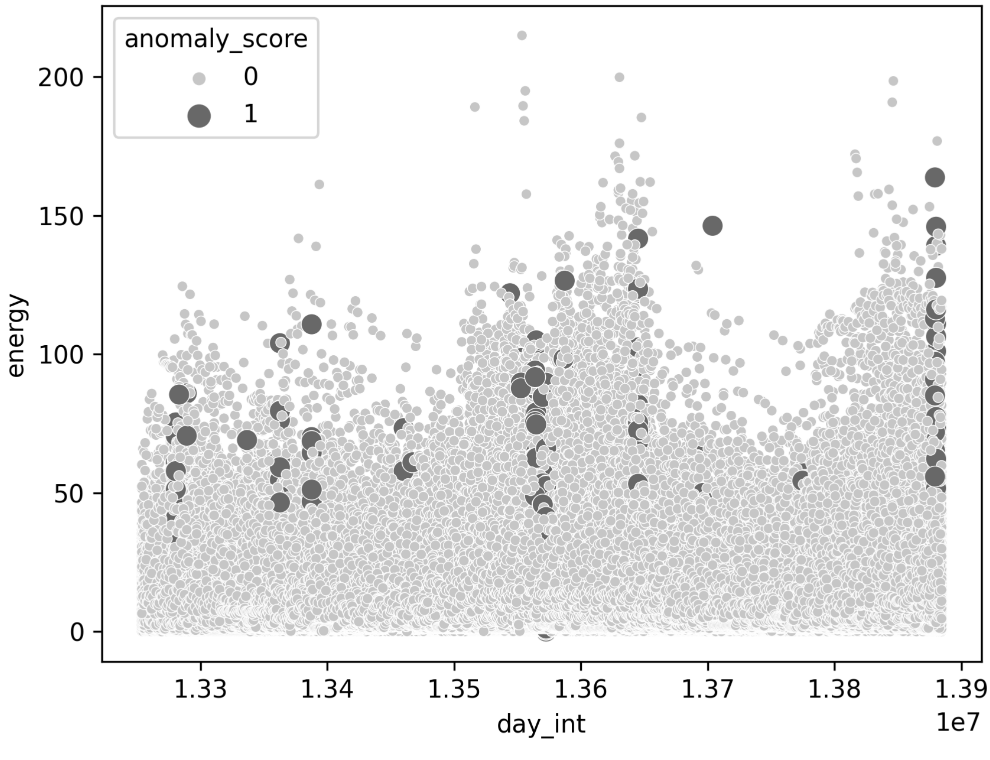


Figure 7.1: Isolation forest classification

Looking at Figure 7.1 the anomalies detected are focused on temporal groups. These also appear in a larger dimension when the energy expenditure on that day is higher than 50 Kw/h.

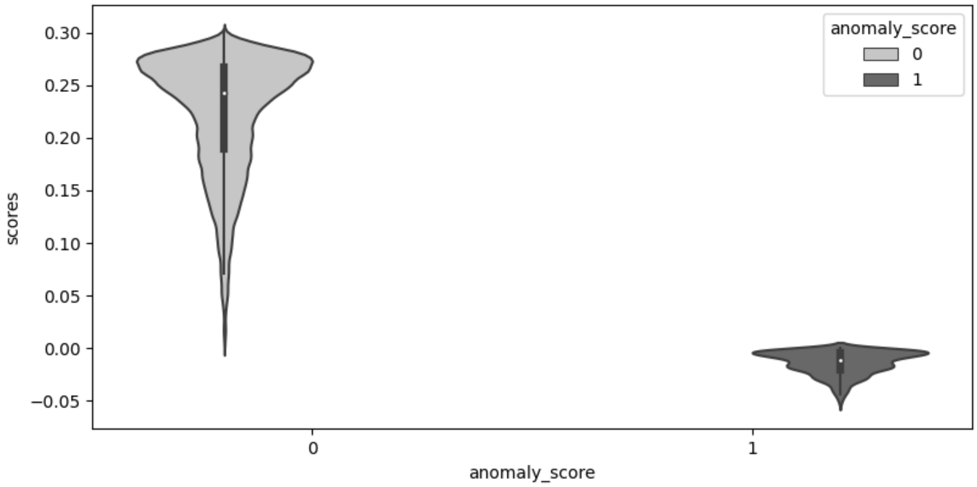
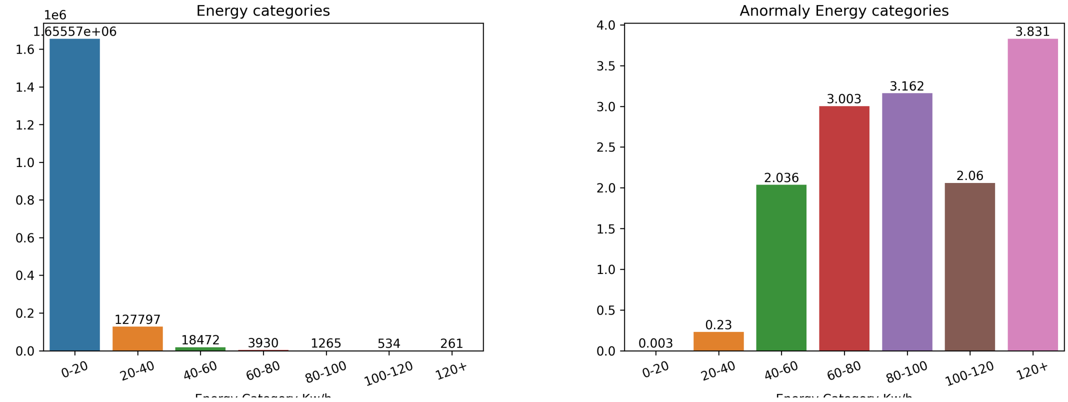


Figure 7.2: Isolation forest scores

For the model, Isolation Forest, to classify a record as an anomaly or not a score is acquired from each of these. In Figure 7.2 we can see the variation of these. We can see that negative scores are automatically called anomalies, but the existence of negative scores classified as normal is verified, this happens due to the contamination given to the model that limits the number of points that the model can call anomalies.

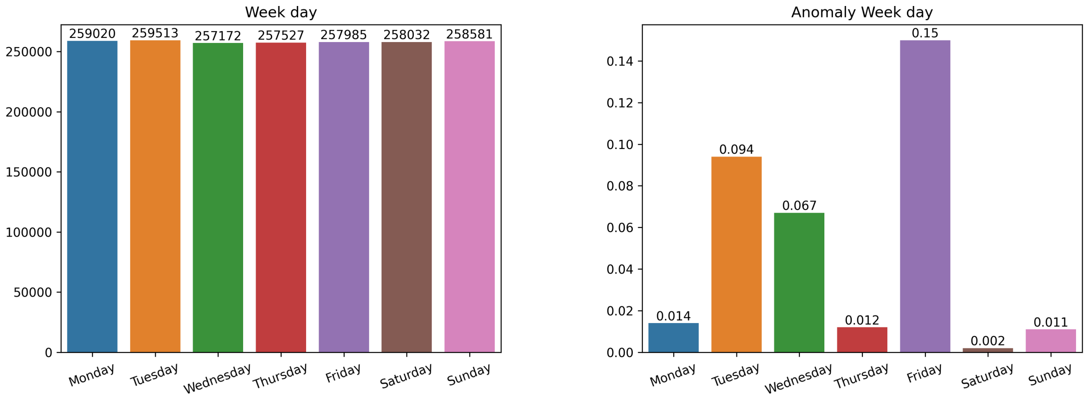


(a) General

(b) Anomalys

Figure 7.3: Energy Categories

Checking the distribution of anomalies and taking into account the energy spent, we can see in Figure 7.3 (b) that, as shown in 7.1 the categories with the highest volume are from the 40-60 category upwards, comparing to the amount of existing records

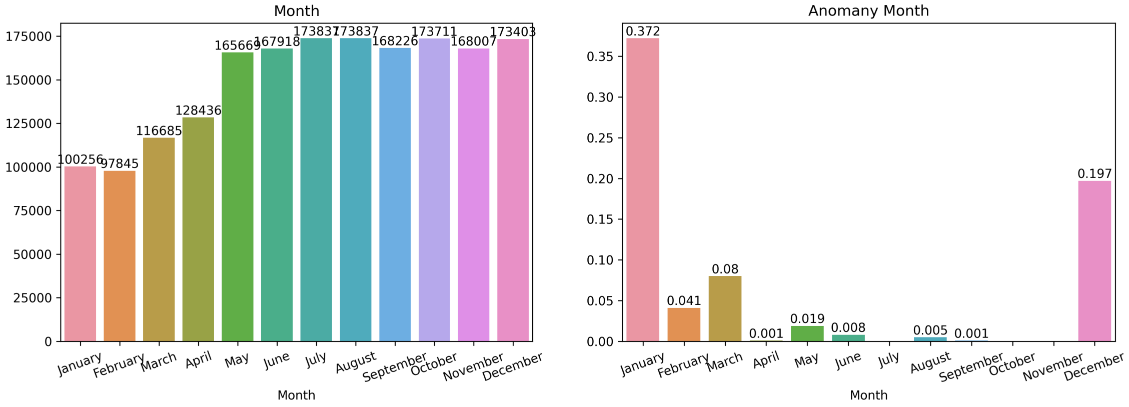


(a) General

(b) Anomalys

Figure 7.4: Days of the week

In Figure 7.4, evaluation of the days of the week parameter we can see that, strangely, there is an accumulation of anomalies on Fridays, with Wednesdays, Thursdays also having relevant values. MAY OR MAY NOT MEAN SOMETh!ING.



(a) General

(b) Anomalys

Figure 7.5: Months

After listing the anomalies taking into account the month in which they are located we can see a clear preference of the model for January and December. Two possible reasons for this have been put forward:

1. As the temperature in these months is lower, houses with central heating will use it to maintain a comfortable temperature, increasing energy expenditure, while other houses without it will maintain an identical energy consumption as they invest in clothing or wood-based heating. This contrast can alert the model of possible anomalies. To resolve this would require more information about the house in question, specifically whether it has central heating.

2. Another alternative would be that being these festive months, most people are on vacation. This is important because if the inhabitants of a house decide to go away for a few days, the energy of the house will be almost null, and if there are groups of family or friends in the house, certain houses will have a drastic increase in energy. To verify if this is the case we decided to look at the energy distribution for these two months, if this is what is happening the amount of houses with low and high energy will increase in relation to the others, that is, an increase in the extremes would have to be verified. As we can see in Figure 7.6 the expected did not happen.

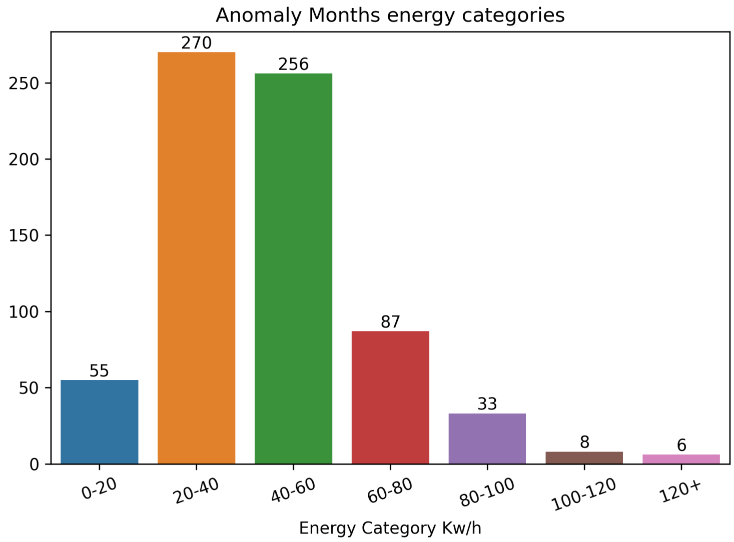
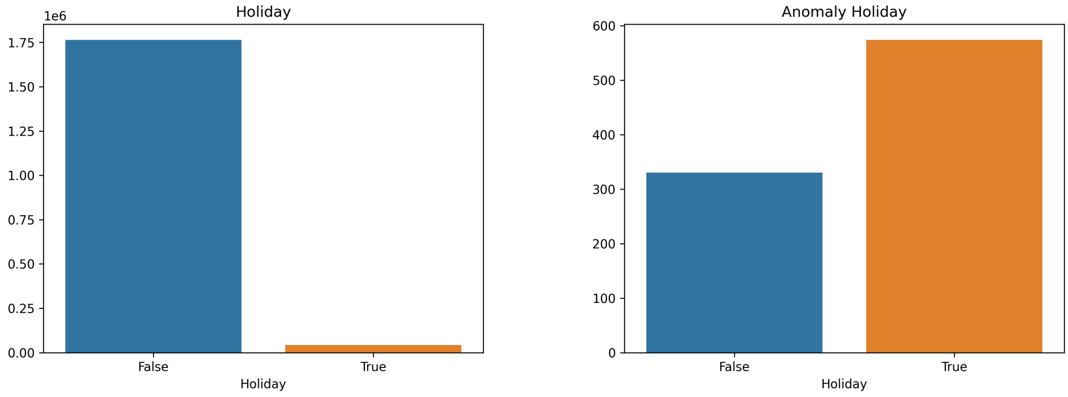


Figure 7.6: Energy distribution for January and December

Looking at the number of anomalies detected on the holidays we can see that they predominate in the model. This is clearly a strange event, once again the theory that in festive days most people are on vacation meaning the energy of the houses will be almost null or it will have a drastic increase. To verify if this is the case we again looked at the energy distribution for those day. In Figure 7.8 we can see that the



(a) General

(b) Anomalys

Figure 7.7: Holidays

expected did not happen again.

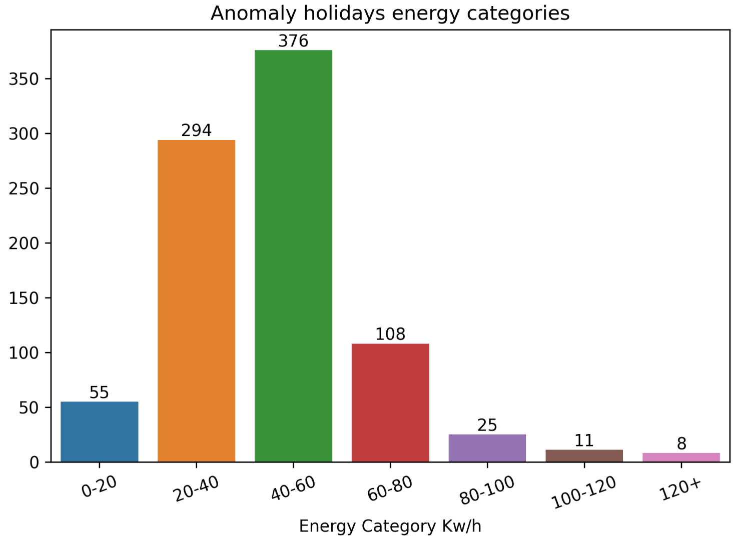


Figure 7.8: Energy distribution for holidays

Chapter 8

Conclusion

We can conclude that, after studying the current state of the field, various methods and models being tried and applied, the use of the choosen model, Isolation Forest, is promising and capable of classifying records as anomalies or not. The result of the experiments conducted reveals that it can draw conclusions with the data provided by smart meters, which shows that it is also possible to apply it to groups of data to predict future anomalies.

8.1 Limitations

There were some limitations that prevented the work from reaching further stages. Time would be one of these, and given the size of the opportunities to learn, fields to explore, and the time needed to learn/research, it seemed few and far between.

Another important factor was the fact that initially the work would be done for the company "Águas de Gaia". Both data and objectives would be suggested by them as well, but due to bureaucratic problems this was not possible in time, also increasing the waiting to start the project.

8.2 Trabalho futuro

Future work may focus on a better understanding of why a point is classified as an anomaly and the weight that each of the given parameters has, it would also be interesting to have the model classify points as, in case of anomaly, fraud type 1, 2 or unknown anomalies. The goal is that in the future, through this study, it will be possible to do forecasting and predict eventual crimes.

Appendix A

Acrónimos

RMSE	Root Mean Squared Error
SVM	Support Vector Machines
RNN	Recurrent neural network
KW	kilowatts per hours
Cº	Celsius temperature
std	Standard deviation
CRISP DM	Cross Industry Standard Process for Data Mining

Appendix B

Código

For reproduction purposes all code used to be found at:

https://github.com/Josmachd/smart_meters_in_london

References

[1] Victor Ambonati. Unsupervised anomaly detection, 'https://www.kaggle.com/code/victorambonati/unsupervised-anomaly-detection/notebook'.

[2] Jean Michel D. Smart meters in london, 'https://www.kaggle.com/datasets/jeanmidev/smart-meters-in-london'.

[3] Nick Hotz. Cross industry standard process for data mining, 'https://www.datascience-pm.com/crisp-dm-2/'.

[4] Maamar A. Benahmed K. Machine learning techniques for energy theft detection in ami. *Proceedings of the 2018 International Conference on Software Engineering and Information*, 2018.

[5] Cody C. Ford V. and Siraj A. Decision tree learning for fraud detection in consumer energy consumption. *2015 IEEE 14th International Conference on Machine Learning and Applications (ICMLA)*, 2016.

[6] Ford V. and Siraj A. Eberle W. Smart grid energy fraud detection using artificial neural networks. *2014 IEEE Symposium on Computational Intelligence Applications in Smart Grid (CIASG)*, 2014.