ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

School of Computer and Communication Sciences IC

Communication Systems Section

Master Thesis Project Report

Data analysis and Anomaly Detection in Buildings using Sensor Data

Author: Deepak Karkala

EPFL Supervisor
Prof. Martin Rajman
Nano-Tera.ch Center

Company Supervisor
Fabrizio Lo Conte
eSMART Technologies Ltd

March 2018

Data Analysis and Anomaly Detection in Buildings using Sensor Data

Abstract

The eSMART-building solution provides home automation functionalities such as heating regulation, energy consumption monitoring, lighting accessible through a centrally mounted touchscreen. The micro-modules installed behind the electric switches sense data such as heating energy consumption, temperature, water consumption and the sensed data is communicated to a central database. The goal of this thesis project is to detect the presence of abnormalities in the heating energy and water consumption in the apartments. The project involved pre-processing the recorded data in order to validate it before further processing. This was followed by the selection of the most appropriate methods for detecting abnormalities in each of the three defined detection scenarios. Two complementary methods (Linear Regression and Local Outlier Factor) were employed in order to obtain more diverse information and for reliable detection of abnormalities. The selection of the algorithms were justified by a thorough qualitative evaluation of the results produced by the detection methods. The application of the methods posed two specific issues which were addressed by the use of Bayesian model and a weighted version of the Local Outlier Factor method. The anomaly scores produced by the detection methods were used to raise alerts and the interactive dashboard developed as part of this project was used to visualize and interpret the abnormalities in energy or water consumption. The alerts raised could then be used by the maintenance team to trigger required corrective actions in advance, leading to predictive maintenance of the system. The project illustrated the effectiveness of data analysis techniques in detecting anomalies even with the lack of labeled dataset and the techniques developed were shown to be readily extensible to other similar application domains.

Acknowledgments

I would like to thank my thesis advisor Prof. Martin Rajman of EPFL for guiding me throughout this project. His inputs were invaluable in carrying out the project.

My sincere thanks to Fabrizio Lo Conte for giving me an opportunity to pursue my Master thesis project at eSMART Technologies and for all the useful suggestions during the project. I would like to acknowledge Laurent Fabre of eSMART Technologies for taking his valuable time to evaluate the results. I would also like to thank Florent Eherler for all the useful discussions with the advisor and helping me with the dataset used in the project.

Finally I must express my profound gratitude to my parents and to my sister for providing me support and encouragement throughout my years of study. This accomplishment would not have been possible without them. Thank you.

Contents

List of Figures

Chapter 1

Introduction

1.1 Context

The eSMART-building solution brings together all of an apartment's basic functions such as heating regulation, energy consumption monitoring, lighting, blind management accessible through a wall mounted touchscreen and also remotely via a smartphone application. The modules installed in apartments report the measured data to a central database. The data collected over the years includes heating energy consumption, room temperature, target temperature set by residents, volume of hot water consumption, orientation of blinds, states of lighting.

As part of this solution, the company also provides maintenance support to the residents of the apartments. Consequently when the system malfunctions (most often the heating system), the company would receive a call from the residents and based on the severity of the issue, the company would then send the support team to the site of the apartment for repairing the system.

The above process is time consuming and expensive from the company's perspective and it would be beneficial to the company if such faults in the system be detected in advance using data analysis techniques such that appropriate actions can be taken before the system breaks down further. This is commonly referred to as predictive maintenance in literature. Accordingly the goal of this thesis project is to analyze the collected data, detect the presence of such abnormalities in the data and raise alerts which could then be used by members of support team for further inspection.

1.2 Problem statement

The procedure mentioned above falls under the general category of anomaly detection systems. As described in Chandola et al. [10], anomaly detection refers to the problem of finding patterns in data that do not conform to expected behavior. Such deviant points are referred to as anomalies, outliers, abnormalities, aberrations in different application domains.

In order to detect anomalies, the first step involves defining what is normal behavior. Only then a data point can be classified as abnormal. Accordingly the first step of the project involved discussions with the domain experts to define the normal behavior and thereby define the nature of abnormalities in the data. In this context, abnormalities are defined with respect to normal behavior of the system in the past. A list of scenarios were then defined and detecting outliers in these scenarios formed the research objective of this project. The following section describes these scenarios in detail.

1.2.1 Scenario 1: Abnormal heating energy consumption in buildings with respect to external temperature

The heating energy consumption in a building is directly related to the external temperature with higher energy being consumed on colder days. Accordingly it will be of concern to people managing the building heating system to know if the total heating energy being consumed in buildings is proportional to the external temperature. Any deviations from the expected energy consumption can be used to raise alert to prompt further maintenance actions. Scenario 1 deals with detecting such abnormalities in heating energy consumption in buildings with respect to the external temperature.

1.2.2 Scenario 2: Abnormalities in heating regulation in apartments with respect to ambient and target temperatures

In addition to the external temperature, the heating energy consumption in each individual apartment is also dependent on the target temperature set by the residents in the apartment. The higher the target temperature set, higher will be the heating energy consumption. Under normal functioning of the system, it is expected that the room temperature follows the target temperature. However faults in heating regulation system could lead to unexpected behaviors. Such abnormal behaviors include continuation of heating even when the room temperature is higher than the target temperature leading to unnecessary excessive use of heating energy or there could be cases where despite setting a higher target temperature, the room temperature stays low and does not reach the target temperature over several days leading to uncomfortable conditions for the residents in the apartment. Detecting such abnormal conditions at the earliest using data analysis can lead to predictive maintenance and is of crucial importance both to the company as well as the residents of the apartments.

1.2.3 Scenario 3: Abnormalities in hot water consumption

The third detection scenario deals with consumption of hot water in apartments. While not as critical as the functioning of the heating regulation system, it would be of interest to detect abnormalities in the volume of hot water being consumed in apartments. Alerts can be raised to notify the residents of the apartment in case of abnormally high consumption over several consecutive days.

1.3 Related Work

Having defined the scenarios, the next step of the project involved reviewing the existing anomaly detection methods. Such methods are used in wide variety of applications such as detecting fraudulent credit card transactions, intrusions in network security, detecting tumors in medical images, fraudulent cases in insurance or health care system. Although the underlying principles of such methods are similar, in general there will be a number of domain specific assumptions which restrict the generalizability of the algorithms to different domains. The goal of this step of the project is to review the existing methods and determine the relevance of the existing techniques to this research problem. Due to the lack of labeled data, only unsupervised techniques are considered.

The survey by Chandola et al. [10] provides a comprehensive overview of the existing methods for anomaly detection in a number of different domains. Two broad categories of methods are used for anomaly detection in sensor data. The first class of methods consider the temporal nature of the data and aims at detecting unusual patterns in the time-series data whereas the second class of methods treat data as points in multi-dimensional feature space. These two class of methods are reviewed in this section.

Liu et al. [13] demonstrated the use of regression and ARIMA models for detecting abnormal energy consumption in buildings. Although similar to the problem under consideration, tuning the parameters of ARIMA model turned out to be impractical for data from large number of apartments in different buildings. Aytekin et al. [9] and Shipmon et al. [17] explored the use of deep neural networks such as Recurrent Neural Networks (RNN) and Long Short Term Memory Networks (LSTM) for anomaly detection in time-series data. The idea involves training the networks on normal sequences of past data, predict the future values and then consider the difference between the predicted and the actual value. More recently Ahmad et al. [7] applied the concept of Hierarchical Temporal Memory to detect anomalies in streaming data. Each of the above mentioned works aimed to detect abnormal patterns in a single stream of time-series data. However the problem being addressed in the current work involves a large number of different time-series data corresponding to different apartments. Although the overall nature of data is similar for all the apartments, there are a number of apartment specific nuances in the data. This leads to different apartments possibly requiring different time domain models to capture the abnormalities resulting in limited applicability of time-series based methods for this research work. This particular problem was addressed in the work by Liu et al. [13] by including apartment specific characteristics such as floor area, occupant density, types of appliances etc which are not part of the dataset used in this project.

Considering industrial applications of recent past, Twitter's open source system Kejariwal [11] employs time series decomposition with a statistical test known as Seasonal Hybrid ESD for detecting anomalies in time-series social network data. Similar statistical tests were also used by Intel's Amitai et al. [8] to detect anomalies in data streams produced by Internet of Things (IoT) devices. However the tests used in these works assume normal distribution of the data which is not the case for the data used in this project. Netflix illustrated the use of Robust Principal Component Analysis (RPCA) to perform robust anomaly detection Wong et al. [19]. Although the data was seasonal and non-normally distributed, unlike the current problem the data used was high dimensional and was therefore more suited for methods such as RPCA.

Neighborhood based methods such as Local Outlier Factor are one of the popular class of methods and was explored by Paulauskas and Bagdonas [15], Ma et al. [14] among others for detecting outliers in network traffic and traffic signals respectively. These methods work with the assumption that normal data instances occur in dense neighborhoods whereas anomalies occur far from their closest neighbors. Although promising, this method encountered difficulties when applied to the current problem when for example a group of outliers themselves formed a cluster. This issue also restricted the use of clustering based methods used in Kiss et al. [12] and Syarif et al. [18]. The results with these models revealed that clustering or neighborhood based methods alone could not be used for reliable detection.

The literature review thus revealed that despite the presence of a wide variety of methods for anomaly detection, no single method can be generalized for all the application domains without explicitly considering the domain specific subtleties. In general the problem of anomaly detection is challenging due to the rare occurrences of anomalies which can further be compounded by the lack of labels as in the current work. Further static rules based system are not applicable because the abnormalities can occur in widely different circumstances with vastly different properties which could change over time and it is not possible to define rules for all such cases. The problem also gets exaggerated during the initial days of operation of a building when only limited data is available. All these factors combined makes the problem of anomaly detection in energy and water consumption in buildings a challenging one with no trivial solutions.

1.4 Structure of the report

Chapter two describes the data collection procedure, different types of data collected and the nature of the data. The chapter also provides a detailed explanation of data pre-processing which involves performing various tests in order to validate the data.

Chapter three describes the methodology used to detect outliers in the data. It describes the different choices available, motivations for the final choice of the algorithms and the improvements to the existing methods.

Chapter four provides the results and a detailed qualitative analysis of the results produced by the selected anomaly detection algorithms. The most representative cases of abnormalities in the data are presented for each scenario. It also includes a discussion on the interpretation of results, applicability of the techniques developed for other application domains.

Chapter five describes the interactive dashboard developed for the end user to interact with the anomaly detection system. Features of the dashboard are described detailing how the alerts raised by the detection methods could be exploited by the support team for predictive maintenance.

Chapter 2

Data description and pre-processing

With the project overview described in the previous chapter, this chapter describes the first step of any data analysis project, description of the data and the pre-processing steps performed before further processing of the data. The chapter begins with describing the data collection procedure. The different types of data collected and their statistics is described. It is then followed by a detailed explanation of data pre-processing which involves performing various tests in order to validate the data. The chapter concludes by describing the contribution of the pre-processing stage in locating faults in data collection and recording procedure.

2.1 Data collection

This section describes the data collection procedure. The data is recorded using sensors installed either in individual houses or in buildings consisting of many apartments. Figure 2-1 shows the overview of the data recording procedure. Each apartment is equipped with a central tablet and a group of modules installed behind the standard electric switchboard. The modules sense data such as amount of energy being consumed, volume of hot water consumption, internal room temperature and the target temperature set by the user. These modules then communicate the sensed data to the tablet using power line communication. The tablets then periodically send the collected data to a central database.

2.2 Data description

This section describes the dataset used in the project. There are a total of 44 buildings with heating solution deployed. The corresponding number of apartments is 506. The

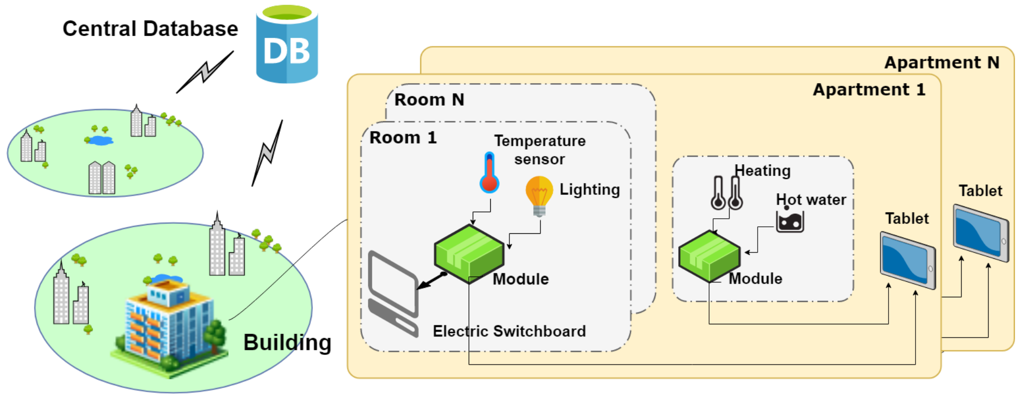


Figure 2-1: Overview of data collection procedure: The system is installed in buildings consisting of many apartments each of which is installed with modules and a centrally located tablet. The modules placed between electric switchboard and the sensors, heating valves sense data and transmit the sensed data to the tablet. The tablets then transmit the collected data to the remote database.

earliest buildings has data starting from early 2013. Over the years, data is available for more buildings. Thus for the earliest buildings, data is available for close to five years. The collected data includes heating energy consumption, internal temperature, target temperature (referred to as set-point), volume of hot water consumption, orientation of blinds, position of windows and doors, position of blinds, electricity consumption. The data used in this project is limited to heating energy consumption, internal temperature, target temperature, volume of hot water consumption.

2.3 Data pre-processing

The data collection procedure as described in the previous section is error prone due to several reasons.

1. **Faulty sensors**: The sensors could malfunction resulting in unreasonable values or no value being sensed.

2. **Data lost during transmission**: This could occur due to the loss of communication between modules and the tablet or the tablet and remote database.

3. **Data recording errors**: Finally there could also be errors when the data is being recorded in the database.

Instances of each of the above cases are described in the following section. If not handled, such erroneous data will corrupt the learning process and lead to meaningless results. As a result, it is vital to remove such data before further processing. This step is commonly referred to as pre-processing. The following section describes the different tests and conditions that were employed in order to validate the sanctity of the data.

2.3.1 Heating Energy consumption

The heating energy refers to the energy meter readings recorded in all the apartments. The heating energy data is recorded once every 10 minutes. In case the data is not received, a NULL value is inserted. Further the value is recorded in the database only when there is a change from the previous value. The daily energy consumption (in kWh) in apartments is then calculated as the difference between the last and the first meter readings for the day. During this process, there are a number of ways in which corrupt data can end up getting recorded. The following steps describe in detail the tests performed to validate the heating energy data.

1. **Missing time-stamps**: For the missing time-stamps, the last written value (NULL or last valid value) is propagated. With this, the total number of days for which energy data was available for all the apartments was 443,175.

2. **Days starting or ending with NULL value**: The energy consumption per day is computed as the difference between the last and the first meter reading for the day. As a result, days where starting or ending value was NULL were removed before further processing. There were 51,437 such days which constitute for 11.6% of the total number of days.

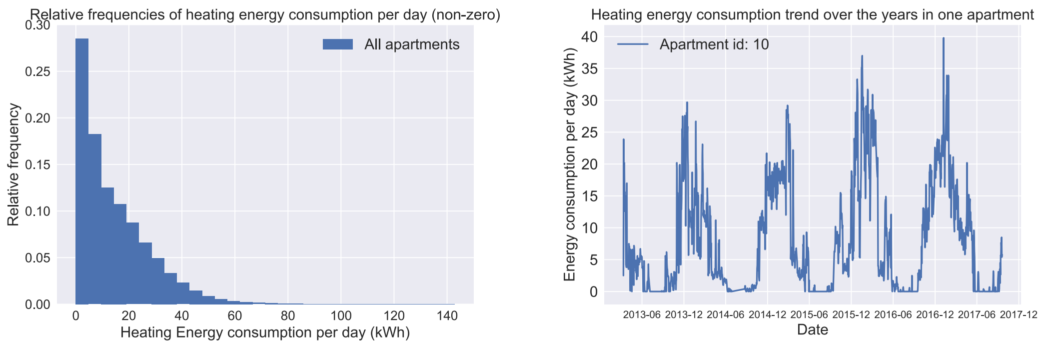
3. **Missing NULL**: These are cases where a NULL should have been recorded but was not. Such cases are characterized by consecutive days with no values and then followed by a huge increase in energy meter readings in very short time. If not handled, this would result in very high energy consumption values which is not the case since the NULL was not recorded.

4. **Duplicate time-stamps**: These are cases when there is more than one value recorded at the same minute within a span of few seconds. This violates the condition that energy data is expected to be recorded once every 10 minutes. There were 156 such instances on 37 days constituting for 0.01% of total number of days. Such days were not considered for further processing.

5. **Decrease in energy meter readings**: The energy meter readings are expected to always increase or stay at the same value. But it was found that there are cases where there was a decrease in the readings. Including such data will result in negative energy consumption values. There were 1813 such days which constitute for 0.54% of the total number of days.

6. **Valid range**: Finally a maximum value was set for the heating energy to be considered as valid. This value was set after discussion with domain experts. Except few apartments which had additional large factors on the meter readings, all other apartments are expected to have energy consumption within this range. The maximum valid energy consumption per day in an apartment was set at 200 kWh. There were 14 days (0.007% of total days) with measurements outside valid range and these cases were verified to be due to data error.

After pre-processing and validation, there were a total of 389,647 number of days with valid heating energy data and this was used for further processing in the learning algorithms. The distribution of daily energy consumption values (non-zero) of all the apartments after validating the data is as shown in Figure 2-2a. The heating energy



(a) (b)

Figure 2-2: (a) Distribution of daily heating energy consumption values in all the apartments over the years. (b) The heating energy consumption trend over the years in one of the apartments.

consumption trend over the years in one of the apartments is shown in Figure 2-2b A strong presence of seasonality can be seen in the heating energy data. During summer the heating system is not in use resulting in zero energy consumption and during winter the usage reaches peak values resulting in seasonal trends in the data.

2.3.2 Room temperature and setpoint temperature

Room temperature: These are the internal temperature recorded in the rooms of the apartment. These are recorded once every 10 minutes only if there is a change by more than 0.5 degrees from the previous recorded value.

Set-point temperature: This is the target internal temperature set by the residents in the apartment. These are recorded whenever the value changes since the user can set different values at anytime. Under normal functioning of heating regulation in the apartment, the room temperature is expected to follow the set-point temperature during winter months. The following tests were performed to validate the temperature data.

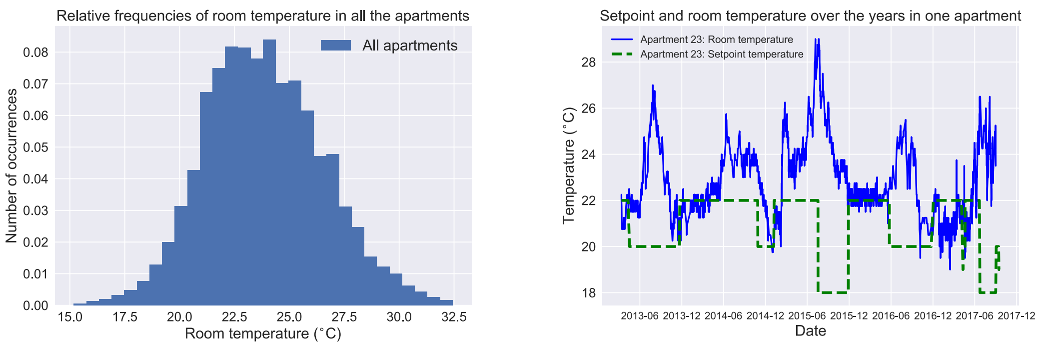
1. **Duplicate time-stamps**: These are cases when there is more than one value recorded at the same minute within a span of few seconds. There were 698 such days constituting for 0.001% of total number of days. Such days were not considered for further processing.

2. **Spatial continuity**: The temperatures in the adjacent room can be expected to be approximately close to each other. If there are cases, where temperature inside the same apartment at the same time are wildly different from each other, it is safe to conclude that it may be due to data error. Such cases were

identified by using the Local Outlier Factor method. There were 2363 such days constituting for 0.012% of total number of days. Such days were not considered for further processing.

3. **Valid range**: +10 to +35 degree Celsius was set as the valid range for the recorded temperature values. 5.6% of total days had values outside the valid range.

The distribution of room temperature values of all the apartments after data validation is as shown in Figure 2-3a The set-point and room temperature values over the



(a) (b)

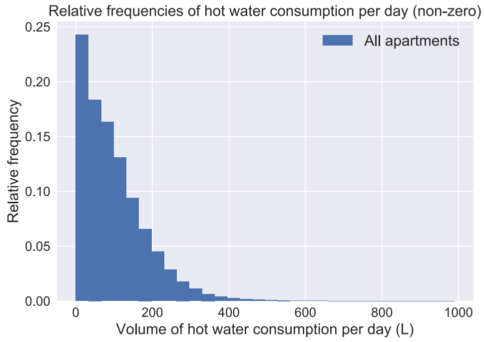
Figure 2-3: (a) Distribution of room temperature values in all the apartments over the years. (b) The set-point and room temperature over the years in one of the apartments.

years in one of the apartments is shown in Figure 2-3b. It is evident from the figure that during the winter months, the room temperature follows the target set-point temperature set by the residents in the apartment.

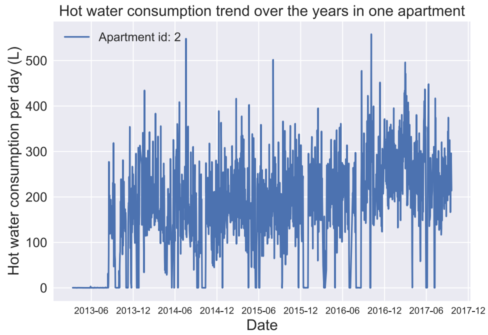
2.3.3 Volume of hot water consumption

The volume of hot water consumption is recorded once every 10 minutes. In case the data is not received, a NULL value is inserted. Further the value gets recorded in the database only when there is a change from its previous value. The pre-processing steps involved in validating the volume of hot water data is same as that of pre-processing heating energy data and hence not repeated again. The maximum valid volume of hot water consumption per day in an apartment was set at 1000 litres.

The distribution of daily consumption of volume of hot water of all the apartments after validating the data is as shown in Figure 2-4a. The hot water consumption trend over the years in one of the apartments is shown in Figure 2-4b. It can be observed that unlike heating energy there is no seasonality in the hot water usage data.



(a)



(b)

Figure 2-4: (a) Distribution of hot water consumption values in all the apartments over the years. (b) The hot water consumption trend over the years in one of the apartments.

2.4 Heating Degree Days

As mentioned in the problem statement section of the report, Scenario 1 aims to detect abnormalities in heating energy consumption in buildings with respect to external temperature. However external temperature is a continuously varying quantity and it is difficult to quantify the extent of how cold or hot it was on any given day with measures such as mean external temperature being poor estimates. In order to avoid this, professionals and companies providing heating (or cooling) solution for buildings use the measure "Degree Days". As will be discussed in the future chapters, Degree Days play a crucial role in detecting heating energy abnormalities in buildings and as a result, the concept of Degree Days is explained in detail in this section.

As described in [2], Degree Days are a measure of external temperature which can be accumulated over a period of days, months etc at a given location. Heating degree day (HDD) is a measure used to quantify the heating energy demands of a building. HDD is derived from outside air temperature measurements. The heating requirements for a given building at a specific location are considered to be directly proportional to the number of HDD at that location. Heating degree days are defined relative to a base temperature-the outside temperature above which a building requires no heating. HDD provides a measure of how much (in degrees), and for how long (in days), the outside temperature was below that base temperature.

To illustrate the concept of HDD, an example computation is presented here. If the outside temperature was constantly 2 degrees below the base temperature for 1 entire day, then over that period, there would be a total of 2 heating degree days (2 degrees * 1 day = 2 degree days). However in reality, computing HDD is more involved since the external temperature varies throughout the day.

Figure 2-5 shows the variation of external temperature and the corresponding Heat-

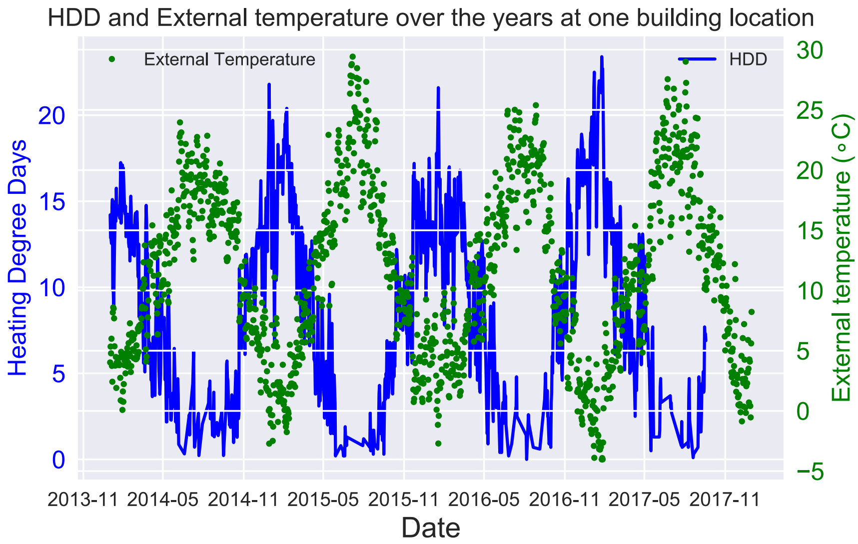


Figure 2-5: External temperature and the corresponding Heating Degree Days values over the years at one of the building locations.

ing Degree Days values over the years. It can be noticed that lower the external temperature, higher is the Heating Degree Days value. Unlike other data described in the previous sections, Heating Degree Days values were not sensed and recorded by the system. Instead these values were obtained using an external API [1] and no pre-processing steps were necessary to validate the data.

This concludes the description of data and the pre-processing steps. The pre-processing stage of the project was vital in multiple aspects. Firstly it helped in validating the data without which the results of the learning algorithms described in following chapters would be meaningless. Further the pre-processing steps of detecting duplicate time-stamps, decrease in meter readings proved to be useful in locating errors in data collection and recording procedure. This turned out to be the first major contribution of the project from the industrial point of view. With all the data having been validated, the next step involved the use of different algorithms to identify outliers. The following chapter describes the various detection algorithms used to identify abnormalities in data.

Chapter 3

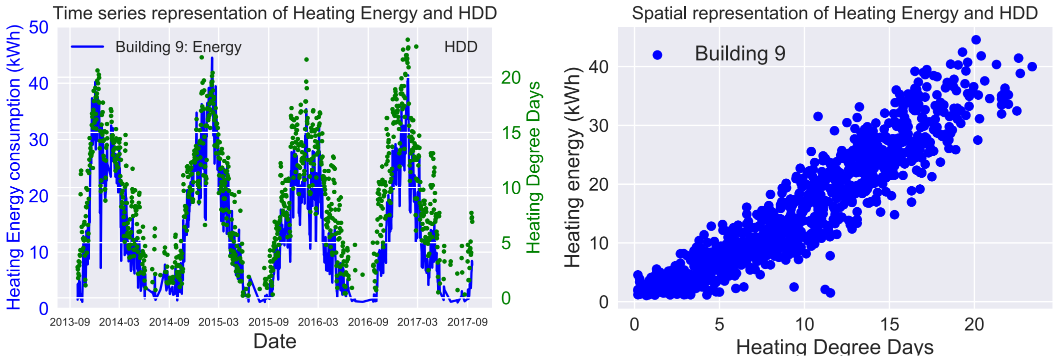
Anomaly detection methods

The previous chapter dealt with data pre-processing. With the data validated, the following step is to determine the most appropriate algorithm to detect abnormal patterns in the data. This chapter describes the methodology used in determining the most relevant anomaly detection algorithm for each of the scenarios. The different choices for the outlier detection algorithms are described. The motivations for the final choice of the algorithm is then explained in detail. The chapter ends with the description of the improvements to the selected algorithms in order to solve certain domain specific issues.

3.1 Methodology

The choice of the algorithm for anomaly detection is influenced by few key factors such as nature of the data involved, definition of normal and abnormal behavior and the availability of labeled data. The data involved in this project is continuous time-series. Due to the seasonal nature of the data, the abnormalities are of contextual nature. For example, while a low value of heating energy consumption is not considered abnormal in summer, a similar value occurring in winter would be classified as abnormal. Thus the value of heating energy alone does not determine the abnormality of the data instead the context in which the value occurs determines the abnormality of the data. Such anomalies are commonly referred to as contextual anomalies. Finally there are no labels defined for normal and abnormal categories of the data as a result of which a large class of supervised algorithms are not applicable.

With these constraints and assumptions, the following section describes in detail the different contrasting alternatives for anomaly detection in each scenario and the rationale behind the selection of the final choices of algorithms.



(a) Temporal representation (b) Spatial representation

Figure 3-1: (a) Heating energy consumption and HDD as a function of time. (b) Heating energy and HDD represented in two-dimensional feature space.

3.2 Sceneario 1

As discussed in section 1.2.1, Scenarios 1 involves analysis at building level with heating energy considered as a function of Heating Degree Days and the goal is to locate the abnormalities in heating energy consumption with respect to Heating Degree Days. There is a trade-off in terms of how much past data should be used in order to evaluate the anomaly score on any given day. While using more data helps in building better models, any changes in trend may not be reflected if very old data is used. As a result, a sliding window technique was used such that on any given day only points belonging to this window will be considered as input to the detection algorithms. The sliding window length was set to duration of 365 days.

Figures 3-1 shows the two contrasting representations of the data used in Scenario 1. Figure 3-1a shows the time-series representation of data for scenario 1. It is evident from the figure that each of these are continuous time-series data and are highly seasonal in nature. Figure 3-1b shows the spatial representation of the same data. Each of these points refer to one day's data namely heating energy consumption in a building per day which is computed as the average heating energy consumed in all the apartments in the building. The two dimensional plot shows the heating energy of a building as a function of heating degree days.

3.3 Scenario 2

As discussed in section 1.2.2, Scenario 2 involves analysis at individual rooms in the apartment with heating energy considered as a function of both Heating Degree Days and the difference between set-point and room temperature. The goal is to identify abnormalities in heating regulation system. In order to accomplish this, the following tasks are performed.

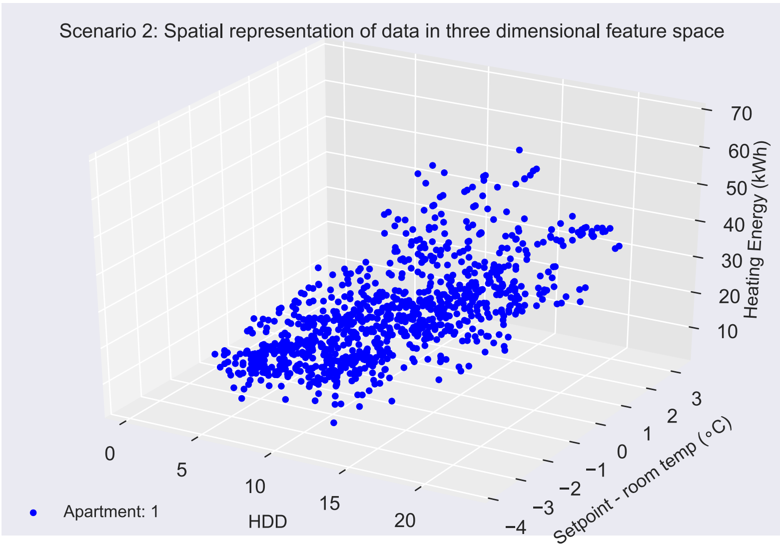


Figure 3-2: Heating energy of an apartment shown as a function of Heating Degree Days and the difference between set-point and room temperature in three dimensional feature space.

1. The per day energy consumption in each apartment is computed.
2. The daily Heating Degree Day value is computed.
3. The average difference between the set-point and the room temperature is calculated for each day.

These will form the three dimensional feature space for Scenario 2. Similar to Scenario 1, a sliding window of length 365 days was used for processing. Since the spatial representation is similar to that of Scenario 1, it is not repeated again. Figure 3-2 shows the spatial representation of data for scenario 2. Each of these points refer to one day's data namely heating energy consumption in an apartment per day, the average difference between set-point and room temperature for the day and the HDD for the day. The three dimensional plot shows the heating energy of an apartment as a function of heating degree days and the difference between set-point and room temperature.

3.4 Anomaly detection methods for Scenario 1 and Scenario 2

As illustrated in the previous section, there are two broad categories of algorithms applicable to the data in Scenario 1 and Scenario 2. The first option is to treat them as temporal data and detect the presence of unusual patterns in the time-series data. The alternative option is to ignore the temporal nature of the data and treat them as points in multi-dimensional space. The following section describes the different methods under these categories.

3.4.1 Time series based methods

Time series based methods work by forecasting the future values of the sequential data and then comparing the difference between the actual and the predicted value to determine the extent of abnormality of the data point.

One such class of methods are Markov chains and Hidden Markov Models which measure the probability of a sequence of events happening. These approaches build a Markov chain for the underlying process, and this can then be used to measure the probability of new sequences occurring, and thereby detect the presence of any unusual patterns in the data.

The ARIMA (Autoregressive Integrated Moving Average) models are another class of models for time-series forecasting in which a future value is expressed as a function of past values of time series and/or past values of errors in forecasting. These models have three parameters p,d,q where p is the order of the autoregressive model, d is the degree of differencing and q is the order of the moving-average model. In order to get accurate estimates of future values, these parameters will need to be meticulously tuned. However since the dataset involved time-series data belonging to large number of apartments, tuning these parameters to capture the underlying model was found out to be unfeasible and as a result this method was not explored further.

The other class of time series based models is the Recurrent Auto-Encoder model. Given time series data, these models will encode the data in low dimensional representation and then try to reconstruct the original sequence. The idea behind such models is that the anomalous data points will not be reconstructed resulting in significant difference between the original and reconstructed sequence. Although applicable to detect outliers in heating energy data, due to the presence of large number of apartments with each one possibly requiring a different model, this class of models was not a suitable option for the dataset used in this project.

3.4.2 Clustering based methods

Clustering based models are one of the popular methods to detect the presence of outliers in the data. These models are based on the assumption that normal data will belong to clusters while anomalies will not belong to any clusters or belong to small clusters. However determining the right number of clusters is very challenging particularly with different number of data clusters present in different buildings and apartments. Consequently this class of models were not considered for further analysis.

3.4.3 Distribution Model based methods

These models consider the probability distribution of the data, with the assumption that outliers will have a low probability of occurrence. Statistical inference tests are then used to determine if an unseen data point belongs to the model fit using the normal data. Box plot and Grubb's test are the popular choices detect outliers in univariate data. However these methods assume that the data samples come from a normally distributed population. As seen in Figure 2-2a this is not the case and as a result such methods cannot be used to detect outliers in Scenarios 1 and 2.

3.4.4 Classification based methods

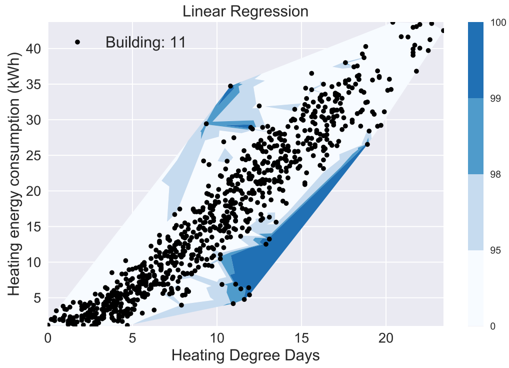
Due to the lack of labeled data, a large class of classification methods were not applicable. One of the relevant method is the One-class SVM (Support Vector Machine). Given a set of points, this method will determine a boundary of that set and classify new points as belonging to that set or not [6]. The percentile anomaly scores obtained using this method was used to derive contour plots for Scenario 1 and is shown in Figure 3-3b. Another class of classification based model is the Isolation Forest. This method is based on random forests and works by isolating observations by randomly selecting a feature and then randomly selecting a split value between the maximum and minimum values of the selected feature [4]. The corresponding contour plot for Scenario 1 is shown in Figure 3-3c.

3.4.5 Nearest Neighbor based methods

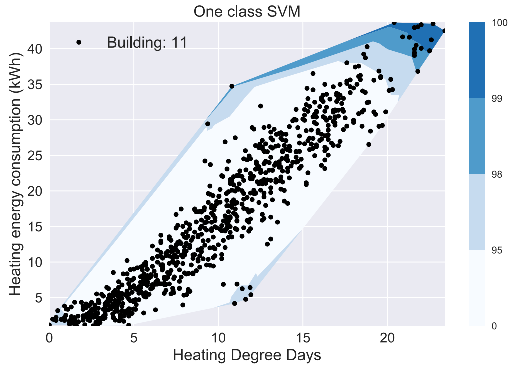
Similar to clustering Local Outlier Factor (LOF) is a density based approach. It measures the local density of a given sample with respect to its neighbors. The anomaly score is based on how isolated the object is with respect to the surrounding neighborhood. The points which have significantly lower density than their neighbors will be considered anomalous [5]. The anomaly score for Local Outlier Factor is computed as follows,

1. **kDistance**$(\mathbf{A})$ is defined as the distance of object A to its k-th nearest neighbor.

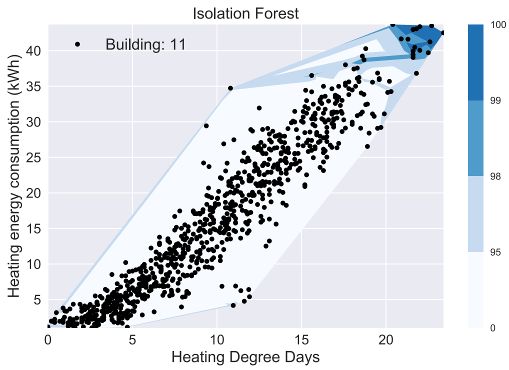
2. The set of k nearest neighbors is denoted by $\mathbf{N_k(A)}$



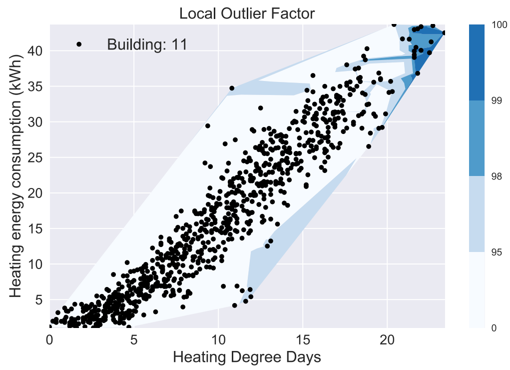
(a) Linear Regression



(b) One-class SVM



(c) Isolation Forest



(d) Local Outlier Factor

Figure 3-3: Contour plots of percentile anomaly scores for different methods. These figures illustrates which points are considered as anomalous by different detection methods. Generally such contour plots are drawn with uniform color scale. However since the task here is to determine the most anomalous points, the percentile cut-offs for the four different colors are set at 95, 98 and 99. Thus the black points lying in the white region fall under less than 95 percentile scores. Points lying in progressively darker shades of blue are more abnormal in nature. It can be deduced that the three methods one-class SVM, Isolation Forest and Local Outlier Factor detect outliers of similar nature (points with very high values of HDD where the density is low). Linear Regression on the other hand treats points which deviate from the linear trend as more anomalous.

3. $\mathbf{reachabilityDistance_k(A, B)}$ is computed as $max\{kDistance(B), d(A, B)\}$ where $d(A, B)$ is the distance of B from A

4. Local reachability density (lrd) of A is then given by

$$lrd(A) = \frac{1}{\sum_{B \in N_k(A)} \frac{reachabilityDistance_k(A,B)}{|N_k(A)|}} \tag{3.1}$$

5. Finally Local Outlier Factor anomaly score is computed as

$$lof_k(A) = \frac{\sum_{B \in N_k(A)} \frac{lrd(B)}{lrd(A)}}{|N_k(A)|} \tag{3.2}$$

which is the ratio of local density of neighbor points with respect to a point's own local density.

The number of neighbors was set to 20 for the Local Outlier Factor method. The corresponding contour plot for Scenario 1 is shown in Figures 3-3d.

3.4.6 Linear Regression

It is evident from Figure 3-1b that the heating energy consumption closely follows a linear trend with respect to Heating degree days the fact which is also illustrated in [3]. This fact strongly encourages the use of linear regression. Further as explained in the next section, linear regression provides complementary information compared to the methods discussed earlier thereby leading to a promising choice for detecting outliers in Scenarios 1 and 2. The anomaly score for linear regression is computed as follows,

1. Maximum likelihood estimate $\hat{y}$ is obtained by minimizing the least square difference between observed points x and the fitted value.

2. Residuals are defined by $res = y - \hat{y}$

3. Standardized residuals are then computed as

$$sres = \frac{(y - \hat{y})}{\sigma_{res}} \tag{3.3}$$

where σ_{res} is the standard deviation of the residual terms.

These standardized residuals are the anomaly scores for Linear Regression. The corresponding contour plot for Scenario 1 is shown in Figure 3-3a.

3.5 Ensemble of complementing methods

Figures 3-3 shows the contour plots derived using percentile anomaly scores from different methods. It can be observed that the three methods isolation forest, one-class SVM and local outlier factor provide very similar information in classifying which points are anomalous (points with high values of HDD where the density is low). Among these methods, isolation forest is more suited to high dimensional data and therefore is not preferred for the dataset used in this project. Further the decision boundary of one class SVM is very sensitive to the number of outliers in the data. As a result, of the three methods, local outlier factor was the preferred method of choice. Contrastingly linear regression provides complementary information when

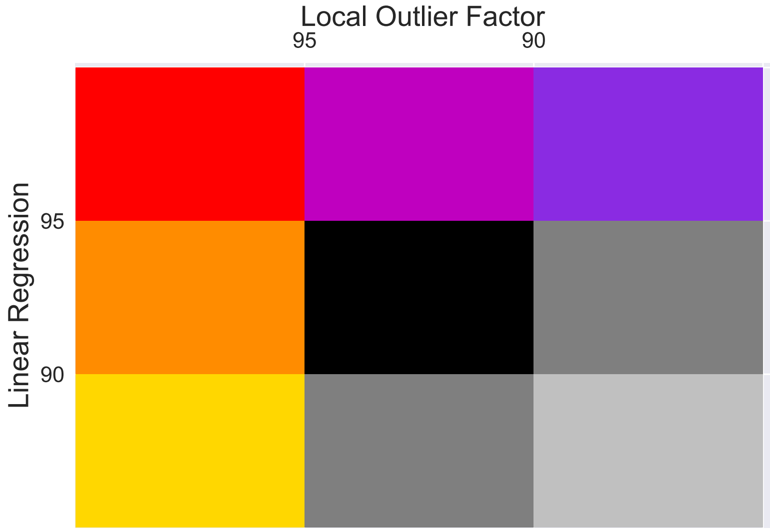


Figure 3-4: Color code based on percentile anomaly scores of the two methods. The percentile cut-offs used are [90, 95][1]s specified in the figure. This scheme implies that points in red and black are the ones where the percentile scores for both methods were higher than 95 and 90 respectively. As a result, points in shades of yellow are the ones where LOF considers more anomalous than LR while the opposite is true for points in shades of blue.

compared to these other methods. This is illustrated by inspecting the anomaly scores for linear regression and local outlier factor in further detail. As described in Sections 3.4.5 and 3.4.6, the anomaly scores for local outlier factor and linear regression are given respectively by

$$lof_k(A) = \frac{\sum_{B \in N_k(A)} \frac{lrd(B)}{lrd(A)}}{|N_k(A)|} \tag{3.4}$$

$$sres = \frac{(y - \hat{y})}{\sigma_{res}} \tag{3.5}$$

From the above equations, it can be noted that the definition of anomaly is different for the two methods. While local outlier factor only considers the local neighbors of a point to compute anomaly score, linear regression considers all the data points in determining the anomaly score. This naturally leads to the idea of combining these two diverse methods since the overall anomalous score from the two methods will be more informative. This phenomenon is illustrated in Figure 3-5 where the anomaly scores from the two methods are shown on the same plot using color codes. Based on the anomaly scores of each point, the corresponding percentile scores are computed.

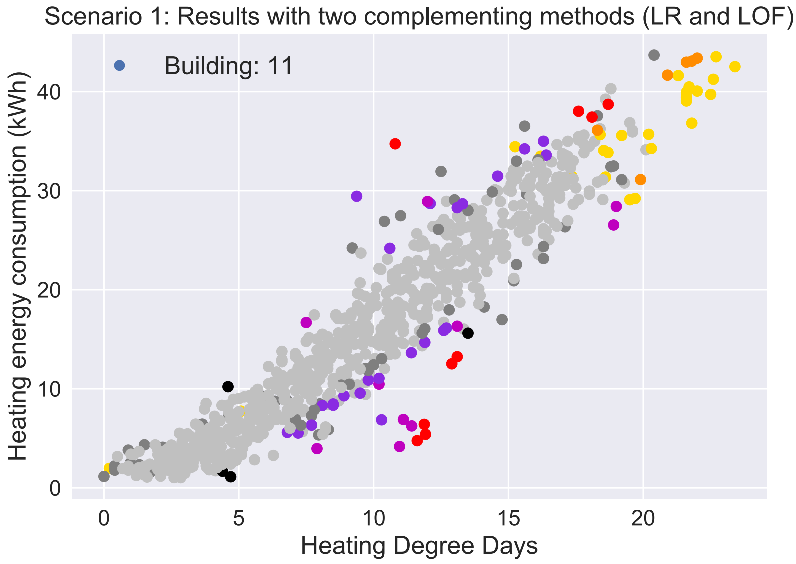


Figure 3-5: Combined results of Linear Regression and LOF for Scenario 1. Points in shades of yellow are considered more anomalous by LOF as compared to Linear Regression whereas points in shades of blue are considered more anomalous by Linear Regression and not so much by LOF. This illustrates the diversity of the selected methods

Different percentile cut-offs are then set to determine the color of each point. The color code is shown in Figure 3-4. The numbers on the axis indicate the percentile cut-offs for the two methods. [1].

Figure 3-5 reveals how LOF considers points with lower density (at high values of HDD) as anomalous (denoted by shades of yellow). In contrast, Linear Regression considers points which deviate from the linear trend as more anomalous (points in shades of blue). This also leads to the conclusion that LOF alone cannot be used for Scenario 1 due to false positives at high values of HDD. This phenomenon shows the significance of using multiple methods for detecting anomalies.

Similar conclusions can be drawn from Scenario 2 result shown in Figure 3-6. The color code is same as that in Scenario 1. In this case the yellow points on the right side of the figure are identified as anomalous by LOF and not so much by Linear Regression. Discussions with experts revealed that these points are indeed anomalous and should be identified as abnormalities. Thus in this case LOF was able to offset the limitations of Linear Regression, proving again the importance of using multiple methods for reliable detection.

[1] It is to be noted that while in practice in order to detect abnormalities, the cut-offs used are much higher (99.9, 99.8), however for the sake of illustrating the diversity of the two methods, lower cut-off values (95, 90) were used

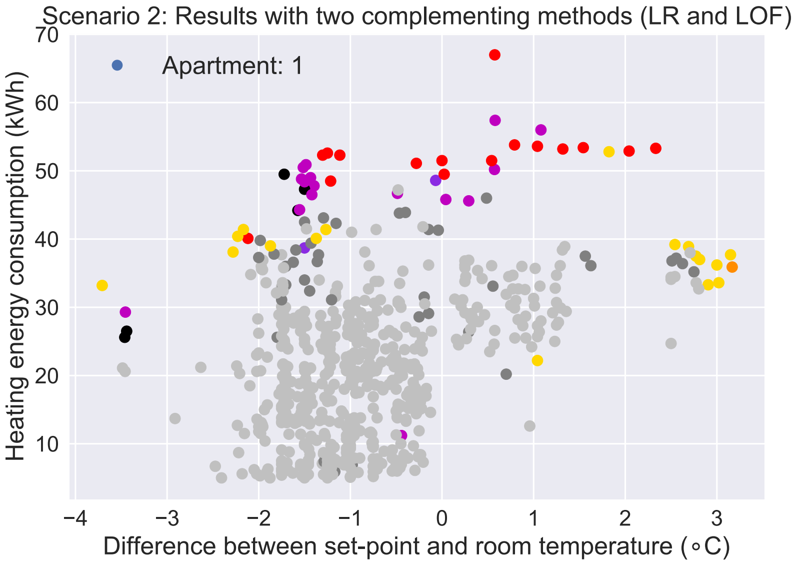


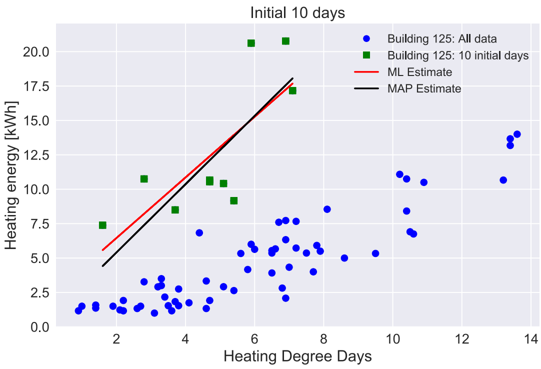
Figure 3-6: Combined results of Linear Regression and LOF for Scenario 2. Points in shades of yellow are considered more anomalous by LOF as compared to Linear Regression whereas points in shades of blue are considered more anomalous by Linear Regression and not so much by LOF. Similar to Scenario 1, this illustrates the diversity of the selected methods.

3.6 Improvements to methods

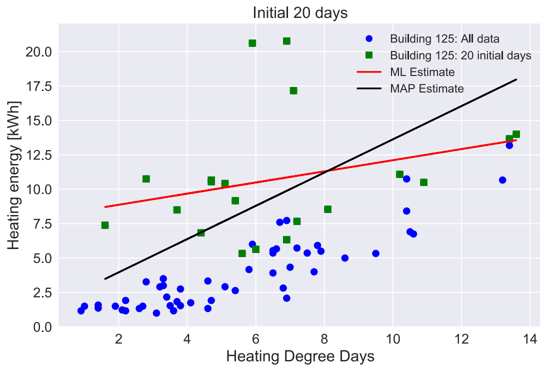
Having selected linear regression and local outlier factor for anomaly detection, this section will describe and address the two issues encountered during the application of the methods to Scenarios 1 and 2. Addressing these issues lead to improved methods and was one of the contributions of this research project.

3.6.1 Bayesian model to handle new buildings

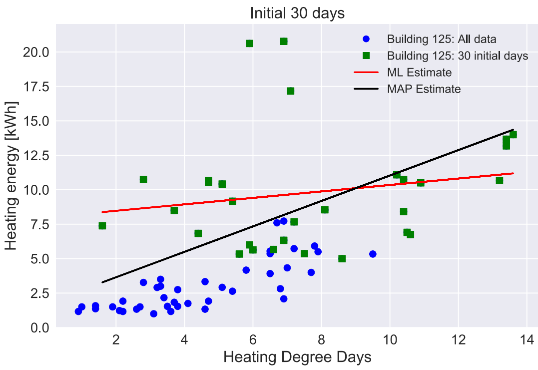
One of the problems encountered when applying linear regression was the lack of data points for newly installed buildings. This implied that there were no reliable estimate of the building's heating energy consumption and therefore no reliable way of determining the presence of abnormalities for the initial six months. However this was unacceptable from the company's point of view since it was vital to determine the presence of any abnormalities particularly in the new buildings. Ideally estimates from other buildings of similar size could have been used but the surface area of the buildings was not available as part of the dataset. Consequently in order to address this, Bayesian approach was used where during the initial days of operation of buildings instead of considering the few unreliable data points, the average energy consumption of all the buildings in the database was used as an estimate and this estimate formed the prior information for the Bayesian method.



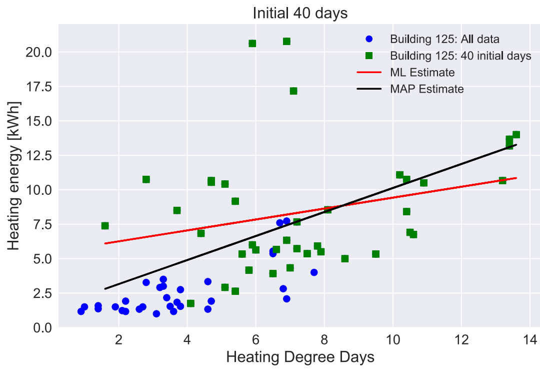
(a) Initial 10 days



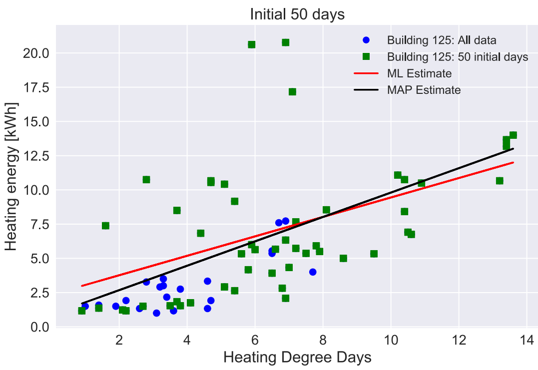
(b) Initial 20 days



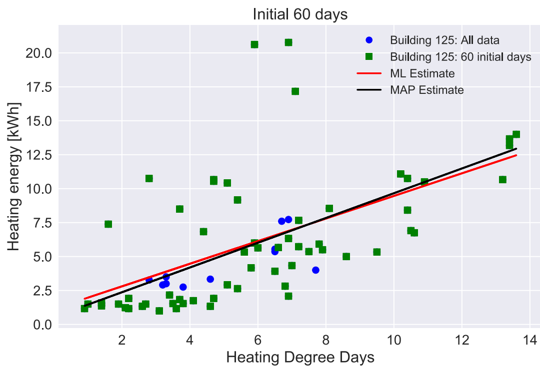
(c) Initial 30 days



(d) Initial 40 days



(e) Initial 50 days



(f) Initial 60 days

Figure 3-7: The figure shows the progression of linear regression model without and with Bayesian model (red and black lines respectively) over the first 60 days. The blue circles refer to all the available days for the building. The green squares denote the building data for the first 'N' days where 'N' is varied from 10 in (a) to 60 in (f) in steps of 10 days.

The heating energy consumption can be modeled as

$$\sim \mathcal{N}(\mu, \sigma^2) \tag{3.6}$$

$$\mu = \alpha + \beta * HDD \tag{3.7}$$

The priors α and β are then set as the mean intercept ($\mu_{intercept}$) and slope (μ_{slope}) of energy consumption of all the buildings respectively with respect to HDD (Heating Degree Days).

$$\alpha \sim \mathcal{N}(\mu_{intercept}, \sigma_1^2) \tag{3.8}$$

$$\beta \sim \mathcal{N}(\mu_{slope}, \sigma_2^2) \tag{3.9}$$

Figure 3-7 illustrates the advantage of using such prior information during the initial days of operation of a building. The figure shows the progression of linear regression model without and with Bayesian model (red and black lines respectively) over the first 60 days. The blue circles refer to all the available days for the building. The green squares denote the building data for the first 'N' days and 'N' is varied from 10 in Figure 3-7a to 60 in Figure 3-7f in steps of 10 days. It can be noted from the figure that without the Bayesian prior information, the linear estimate can be widely inaccurate during the first few days. The Bayesian model acts as a trade off between the prior information and the maximum likelihood estimate and as more data points are available, the two models converge to the same value as seen in Figure 3-7f.

3.6.2 Weighted local outlier factor

As described in Section 3.4.5, the anomaly score for Local Outlier Factor is computed using the following equations,

Local reachability density (lrd) of A is defined as

$$lrd(A) = \frac{1}{\sum_{B \in N_k(A)} \frac{reachabilityDistance_k(A,B)}{|N_k(A)|}} \tag{3.10}$$

Local Outlier Factor anomaly score is then computed as

$$lof_k(A) = \frac{\sum_{B \in N_k(A)} \frac{lrd(B)}{lrd(A)}}{|N_k(A)|} \tag{3.11}$$

It can be observed that the anomaly score is computed based on the density of the point relative to the density of its neighbor points. However there will be cases when most neighbors of an anomalous point are themselves outliers. This results in lower anomaly score for the point and the standard Local Outlier Factor fails to identify such outliers. In order to counter this, the Weighted Local Outlier Factor considers the anomaly scores of all the neighbor points as weights while computing the local reachability density of the point. The intuition behind this idea is that when the neighbor points are themselves outliers, they should be weighed by their anomaly score such that the test point gets relatively higher anomaly score. Accordingly the anomaly score computation for WLOF was updated as follows,

The weighted local reachability density (lrd) of A is defined as

$$wlrd(A) = \frac{1}{\sum_{B \in N_k(A)} \frac{anomalyScore(B) * reachabilityDistance_k(A,B)}{\sum_{B \in N_k(A)} |anomalyScore(B)|}} \tag{3.12}$$

And the Weighted Local Outlier Factor anomaly score is then given by

$$wlof_k(A) = \frac{\sum_{B \in N_k(A)} \frac{wlrd(B)}{wlrd(A)}}{|N_k(A)|} \tag{3.13}$$

Figure 4-4a in Section 4.2.2 of next chapter illustrates one of the cases where the Weighted LOF outperforms LOF.

3.7 Scenario 3

As discussed in section 1.2.3, Scenario 3 aims at identifying abnormalities in the hot water consumption in apartments. In order to accomplish this, volume of daily hot water consumption is computed for all the apartments. Similar to Scenarios 1 and 2, detections are performed on daily data and the sliding window of length 365 days is used for detecting abnormalities.

3.7.1 Anomaly detection methods for Scenario 3

Unlike Scenarios 1 and 2, Scenario 3 involves a single univariate dataset of volume of hot water consumption. As a result, a number of classification based and nearest neighbor based methods are not applicable. One of the popular methods for outlier detection in univariate data is the Robust z-score method [16]. Further due to the temporal nature of hot water consumption values, the other commonly used method is exponential smoothing. Consequently these two methods were used to detect abnormalities in hot water consumption in the apartments.

Exponential smoothing with control limits

This is a time-series based method, where the temporal data is filtered using the following equation,

$$y_t = \alpha * x_t + (1 - \alpha) * y_{t-1} \tag{3.14}$$

where α is the smoothing factor and $0 < \alpha < 1$.

The filtered output will mask the short term variations and captures the long term

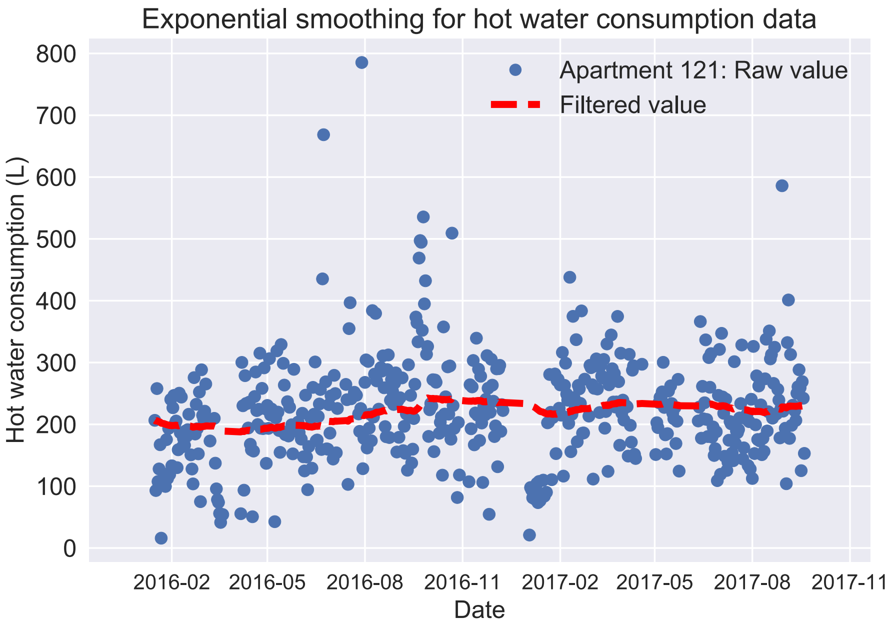


Figure 3-8: Exponential smoothing for hot water consumption data

trend of the time series data. Following this, an upper control limit is defined and the difference between the daily usage value and the upper control limit determines abnormality of the data point. This is illustrated in Figure 3-8. The blue points corresponds to daily usage of hot water, the dashed red line denotes the filtered value. The most common way of using this to detect outliers is by setting a higher control limit such that all points above this upper control limit will be classified as anomalous. Instead of using such fixed value, in this project the difference between the observed and the filtered value was used to derive an anomaly score based on which a suitable threshold could then be defined by the end user.

Robust z-score

This method falls under the statistical based methods. Considering the deviation of a data point from the mean is a commonly used method to determine the abnormality of a data point [2]. However such measures which uses mean to measure abnormality are highly sensitive to the presence of outliers in the data and are thereby not very reliable. The standard z-score is given by

$$Standardzscore_i = \frac{(x_i - \mu)}{\sigma} \tag{3.15}$$

where μ and σ are the mean and standard deviation of the population. In [16], the authors illustrate the concept of robust statistics which aims at searching for models fitted by the majority of the data instead of all the data points thereby reducing the effect of outliers. In contrast to the standard z-score the robust z-score is given by

$$Robustzscore_i = \frac{(x_i - median_{j=1,2..,n}(x_j))}{MAD} \tag{3.16}$$

where MAD refers to median absolute deviation and is given by

$$MAD = 1.483 * median_{i=1,2...n}|x_i - median_{j=1,2...n}(x_j)| \tag{3.17}$$

It can be noted from Equations 3.15 and 3.16 that the mean and the standard deviation in the standard z-score are replaced by median and MAD respectively. This will make the method more robust in the presence of outliers. Results using this technique are presented in the next chapter.

This concludes the methods section of the project. The chapter began with the description of the methodology used to select the most appropriate detection methods for scenarios. This was followed by a detailed inspection of the suitability of different methods. A thorough justification was provided for the reasoning behind the selection of algorithms. The advantages of using multiple detection methods was then illustrated. The chapter also discussed the improvements to the existing methods done as part of the project. The next chapter will provide the results for the methods presented in this chapter.

[2] Although not applicable in the strict sense due to non-Gaussian nature of the data, nevertheless it was found out that the method provides a good estimate of the abnormalities in the data

Chapter 4

Results and Discussion

Having selected the most appropriate methods for anomaly detection, this chapter describes the results obtained using the selected methods for all the scenarios. The chapter begins with the description of the evaluation methodology followed by qualitative analysis of the results for all the scenarios. The chapter concludes with the discussion of the contributions of this research project and the justification of extensibility of the techniques developed in this project to other applications.

4.1 Evaluation methodology

This section describes the methodology used in evaluating the results of the anomaly detection methods. As described earlier in the report, the data used in this project is unsupervised, there are no labels available for normal and abnormal cases in the dataset. As a result of which commonly used classification evaluation metrics such as precision, recall, ROC curves cannot be used to evaluate the quality of the results. Instead qualitative analysis is used to evaluate the results obtained. For each of the scenarios, results for few specific cases which are representative of all the cases in the dataset will be presented. The cases presented will illustrate the detected true positives and also include the false positive cases as determined by the domain experts. The procedure involved experts going through top most detected abnormalities in the past data and categorizing them as true and false positives.

4.2 Results

As mentioned in sections 3.2, 3.3 and 3.7, the data for all the scenarios were processed on a daily basis. As a result, Scenario 1 produces an anomaly score for one building for each day. Similarly Scenario 2 would produce an anomaly score for each of the rooms in an apartment for each day and Scenario 3 produces an anomaly score for each apartment every day. The anomaly detection methods were applied for each day in the dataset with the sliding window length of 365 days. Results presented here includes the use of Bayesian Linear Regression and Weighted Local Outlier Factor described in Sections 3.6.1 and 3.6.2 respectively. For each of the scenarios, results were sorted in the order of abnormalities. Further based on the discussion with domain experts, the most concerning cases of abnormalities is when there is a sequence of consecutive days with unusual behavior. Accordingly the cases presented here consists of several consecutive days of abnormal behavior and are amongst the top most abnormalities evaluated by the experts. These cases represent the most commonly occurring abnormalities in the dataset.

4.2.1 Scenario 1

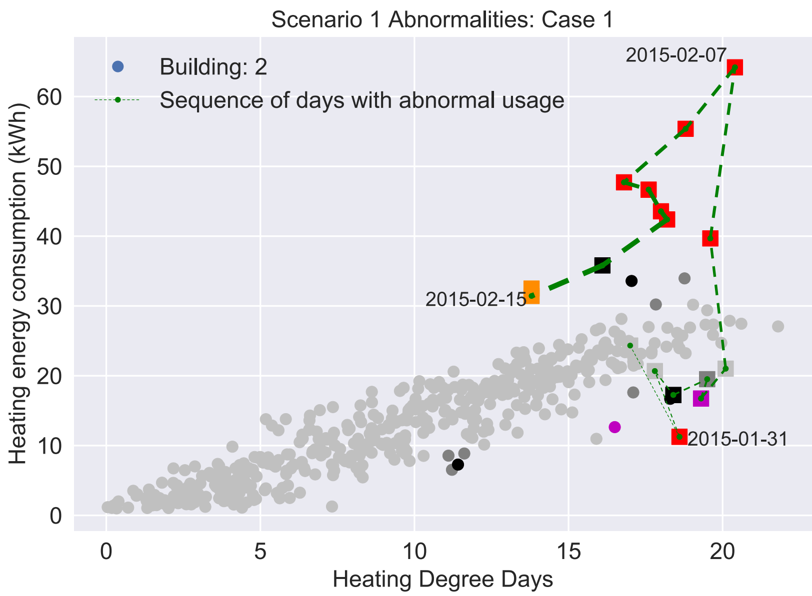


Figure 4-1: Abnormal case 1 for Scenario 1. Each point refers to one day's data. The abnormal sequence of days (square markers) are connected by dashed green line. Colors are based on percentile anomaly scores produced by Linear Regression and LOF.

To reiterate, Scenario 1 aims to detect abnormal heating energy usage in buildings with respect to the Heating Degree Days. Three different abnormal cases are presented for Scenario 1. In each case, the Figures show the heating energy consumption in buildings as a function of Heating Degree Days. As illustrated in previous chapters, a linear trend is expected. Each data point in the following Figures represent one day's data. In order to show results from two methods (Linear Regression and Local Outlier Factor) the same coloring scheme introduced in Figure 3-4 is used here. Since only abnormal points are of interest, the percentile cut-offs were set to 96 and 98. With this, the points in black and red are the ones for which percentile scores from both methods are higher than 96 and 98 respectively. The two methods strongly agree that such points are indeed anomalous. The green dashed line represent the sequence of days (the thickness of dash increases with the progression of days) with abnormal usage (denoted by square markers).

Case 1: Long sequence of days with abnormal heating energy usage

Case 1 illustrates the abnormal heating energy usage in one of the buildings. The green dashed line in Figure 4-1 shows a sequence of 15 consecutive days with abnormalities. It can be noted that the abnormal behavior begins with lower than expected energy consumption. The points with square markers connected by green lines represent these days. This is then followed by a huge increase in heating energy consumption for several days for which the percentile scores for both methods are higher than 98 resulting in red colored points. This case illustrates a typical behavior where some malfunctioning in the system leads to lack of heating energy for few days which is usually followed by maintenance team visiting the site of the building for repair and as a result the next few days is characterized by very high energy consumption. With the use of anomaly detection it can be seen that an alert could have been raised as soon as days with low energy consumption were identified. This alert could then have been used to prompt necessary actions, thus preventing the system going into further degradation.

Case 2: Long sequence of days with high heating energy usage

Case 2 depicts a similar sequence of days in another building. Unlike Case 1 there are no days of lower heating energy consumption instead there is a high energy consumption consistently for over 15 days. Like in Case 1, both the detection methods strongly indicate the presence of abnormal usage (colored in red) which leads to raising an alert during early stage of the abnormal sequence and thereby resulting in early detection of abnormal situation.

Case 3: False positive due to latency in the system

Case 3 illustrates a case where the raised alert was classified as false positive. As mentioned before, the processing for all the scenarios is done on a daily basis (each point represents a day). However there may be sequence of days when the external temperature may vary widely from day to day. Case 3 shows such behavior where as seen by the dashed green line, there is a significant decrease in HDD for two days before it increases again. The heating energy consumption nevertheless stayed high for those two days resulting in an alert. However in reality this could be due to the latency in the system. Such behavior is very rare and this was the only such case observed in the dataset. Further this can be avoided by raising an alert only when abnormalities are detected over several consecutive days.

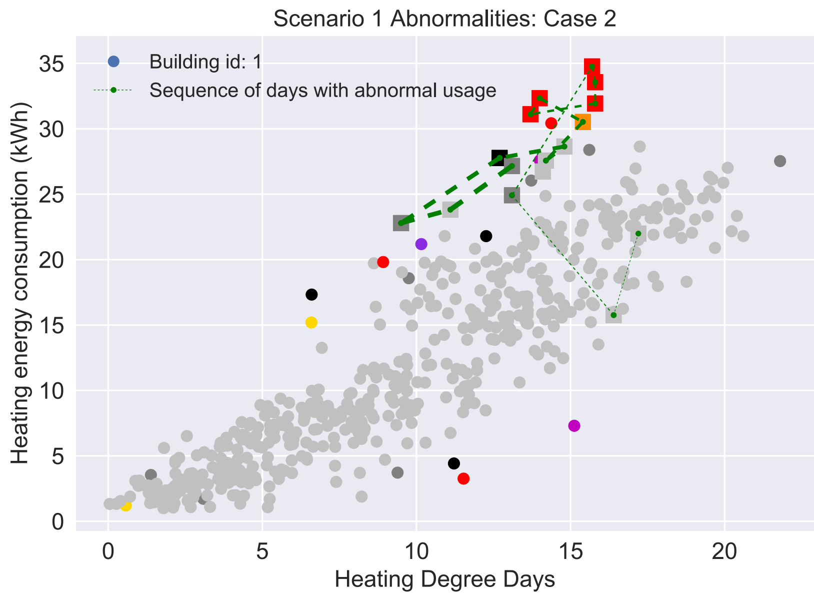


Figure 4-2: Abnormal case 2 for Scenario 1. The points in square markers show the consistently high energy consumption for several days and the red colored points indicate that both methods agreed that these points were indeed anomalous.

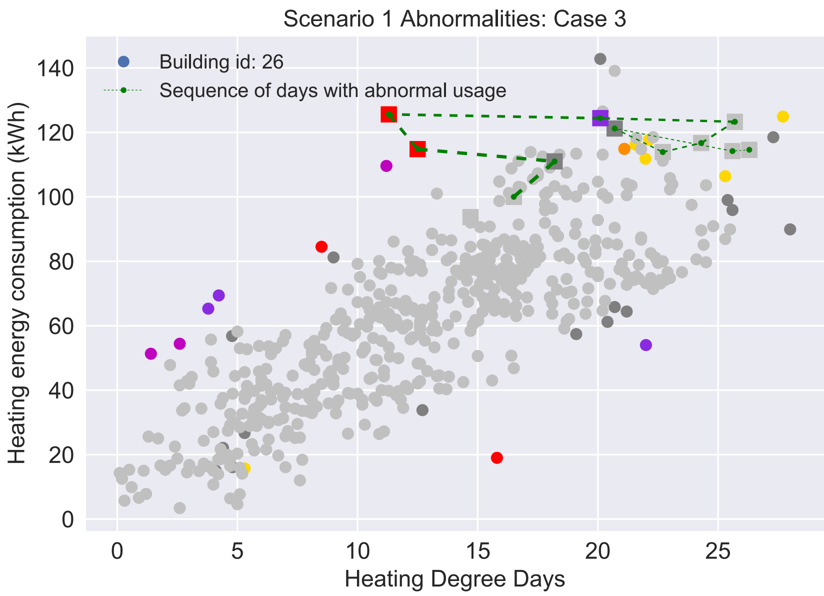


Figure 4-3: Abnormal case 3 for Scenario 1. This case illustrates the false positive raised by the system.

4.2.2 Scenario 2

Scenario 2 aims at detecting abnormalities in heating regulation system in individual rooms of the apartments. The Figures shown for the two cases of Scenario 2 plots heating energy consumption as a function of difference between set-point and room temperature. It is to be noted that while detection methods for Scenario 2 uses three dimensional feature space with HDD as the other feature, for the sake of easier visualization of abnormalities only two features are shown. Like Scenario 1, each point represents a day and the coloring scheme is also identical to that of Scenario 1. Finally the green dashed line represents the sequence of days with abnormal usage.

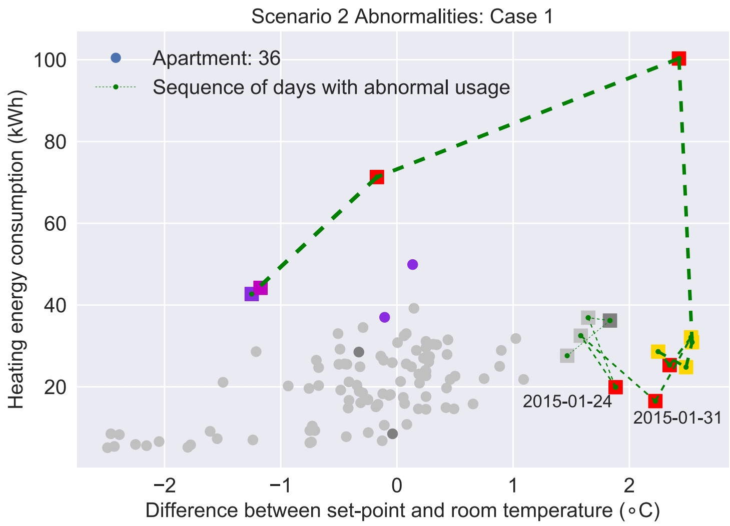
Case 1: Abnormal heating energy consumption in apartment

Case 1 in Scenario 2 shows the abnormalities in heating energy consumption in one of the apartments in the building shown in Case 1 of Scenario 1. As noted in Section 2.3.2 of data description, under normal functioning of the system, the room temperature is expected to follow the set-point temperature. Figure 4-4a illustrates a case where this is violated. The points connected by dashed green line denotes the days when the room temperature drifts apart from the set-point instead of following it. This can be seen in Figure 4-4b which shows the set-point and the room temperature for the sequence of days with abnormal usage. It is to be observed as soon as the system enter this phase, Local Outlier Factor signals the presence of (points in yellow). This is then followed by days of very high energy consumption where both Linear Regression and Local Outlier Factor indicate the presence of abnormality (points in red). Accordingly the temperature now reaches the set-point value as seen during the last few days in Figure 4-4b. Similar to Case 1 of Scenario 1, it can be inferred that the anomalies detected during the early stage of abnormal sequence can be used to raise an alert for further inspection.

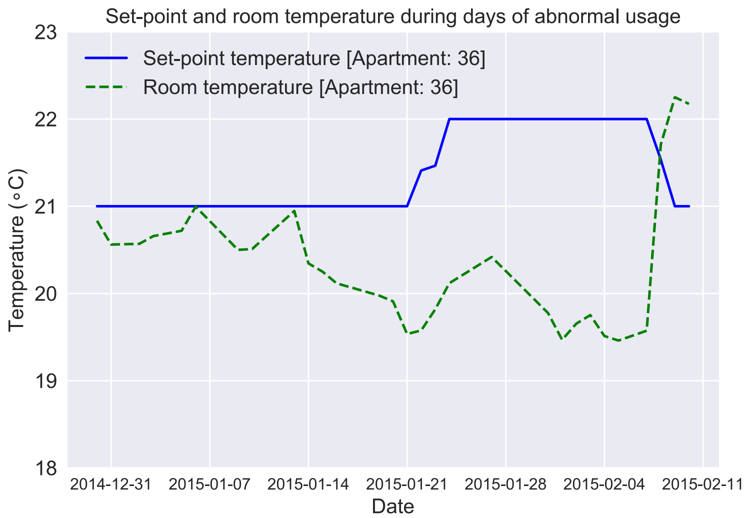
This was also one of the cases where Weighted LOF outperformed the standard LOF. As evident in Figure 4-4a, the anomalies to be detected are themselves surrounded by other anomalous points making the anomaly scores produced by LOF less effective. Weighted LOF on the other hand produced relatively higher anomaly scores for such points.

Case 2: Abnormal heating regulation in apartment

Case 2 illustrates a case with irregularities in heating regulation in the apartment. Similar to Case 1, the room temperature fails to follow the set-point temperature over lot of days (nearly a month) as shown in Figure 4-5b. The corresponding days are detected as anomalous (points in yellow and red) as seen in Figure 4-5a. Since the abnormalities are detected during early stage, the raised alert can prompt the necessary corrective measures.

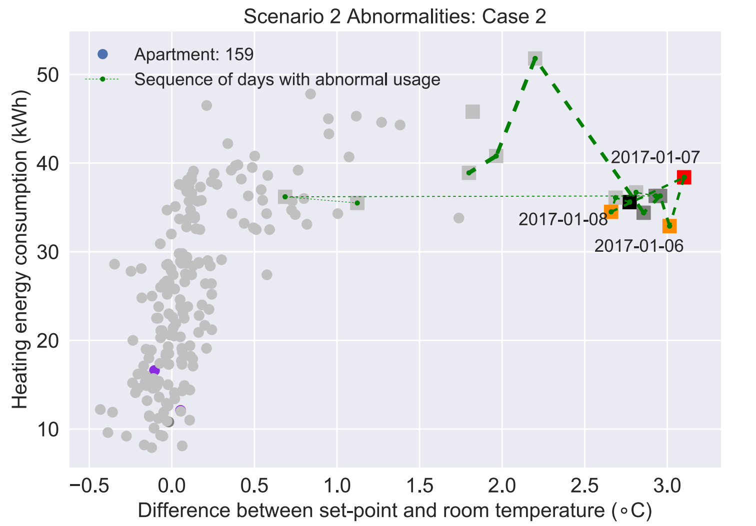


(a) Heating energy as a function of difference between set-point and room temperature.

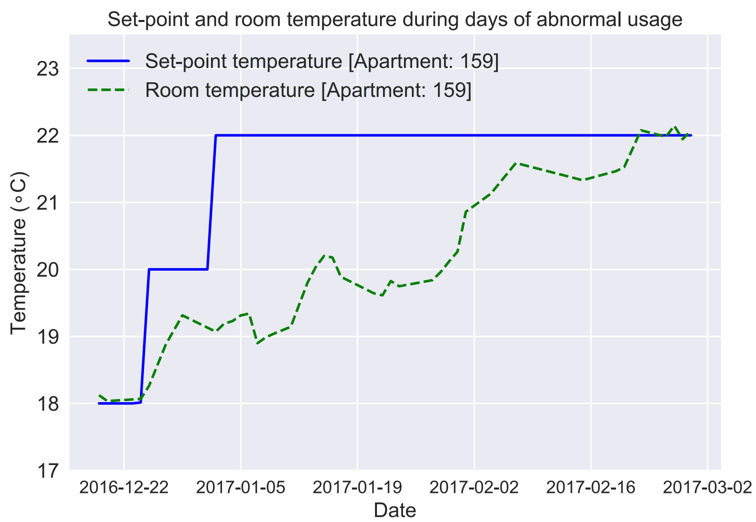


(b) Set-point and room temperature trend for the days of abnormalities

Figure 4-4: (a) Heating energy as a function of difference between set-point and room temperature. The points in square markers denote days with abnormal usage. The colors are based on the percentile anomaly scores produced by Linear Regression and LOF. (b) Set-point and room temperature for the days of abnormalities. It can be seen that as soon as the room temperature drifts away from set-point temperature, the corresponding days in (a) are identified as anomalies.

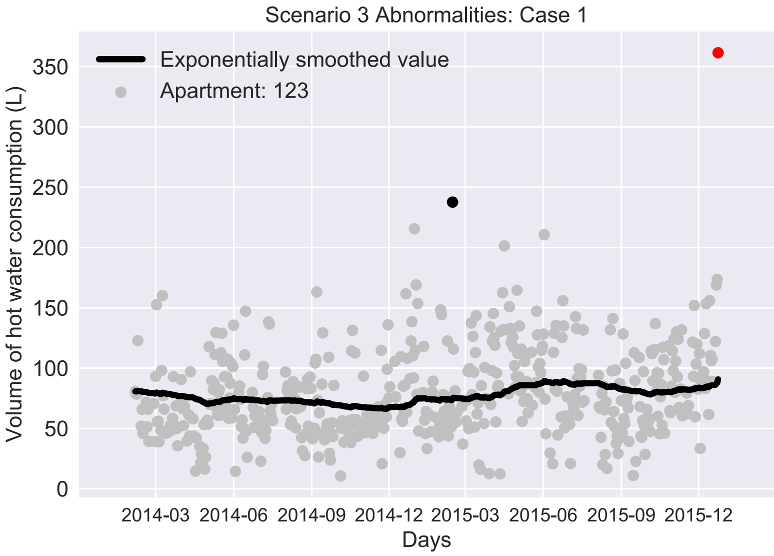


(a) Heating energy as a function of difference between set-point and room temperature.

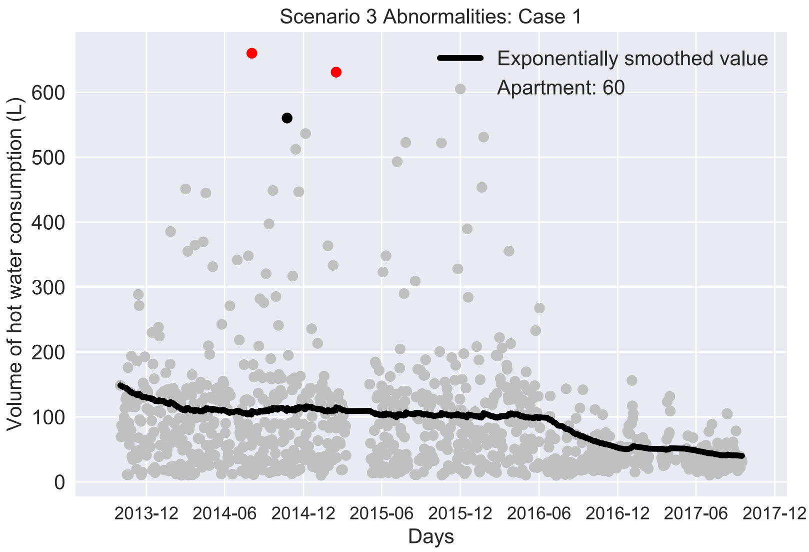


(b) Set-point and room temperature trend for the days of abnormalities

Figure 4-5: (a) Heating energy as a function of difference between set-point and room temperature. (b) Set-point and room temperature trend for the days of abnormalities. It can be seen that it takes nearly a month for the room temperature to reach the set-point value and these days are detected as anomalies as indicated by the dates specified in (a)



(a) Abnormal Case 1 for scenario 3



(b) Abnormal Case 2 for scenario 3

Figure 4-6: (a) and (b) show two examples of abnormal hot water consumption in apartments. The colors are based on the percentile anomaly scores of the two detection methods. The coloring scheme is same as that in 3-4, the percentile cut-off values were set at 99.9 and 99.8. The red colored points denote the detected abnormalities.

4.2.3 Scenario 3

Figures 4-6a and 4-6b illustrate two cases of abnormal hot water consumption in apartments. Both Exponential Smoothing and Robust z-score were used to detect outliers. The colors are based on the percentile anomaly scores of the two methods. The coloring scheme is same as that in Figure 3-4, however since the goal here is to detect abnormalities, the percentile cut-off values were set at 99.9 and 99.8. As a result only a few points are not in grey. The red colored points are the ones where both the detection methods strongly agree on the abnormality of the data point.

4.3 Discussion

The previous section illustrated the most representative cases of abnormalities for each of the scenarios. The qualitative analysis of the results proved the capabilities of the selected detection methods in effectively detecting the presence of abnormalities in the data. As illustrated by the cases, the early detection of abnormalities can be invaluable in triggering the required maintenance actions and preventing the system from any further degradation. In general this ultimately leads to the less equipment downtime and less expensive measures from the company's perspective. This in essence was the goal of this project and the results in the previous sections conclusively proved the capabilities of the selected anomaly detection methods for predictive maintenance.

The literature review revealed that despite the presence of plenty of different anomaly detection techniques, there were a number of restrictions in applying such methods for this research problem. The results from previous sections illustrated how such restrictions were overcome and thereby filling the gap shown in the literature. In particular, the use of multiple complementing methods which produce diverse results improved the reliability of the detections. It was found out that the use of a single method resulted in higher number of false positives. Further the Bayesian model helped solve the issue of limited data during initial days of operation and the Weighted Local Outlier Factor improved the performance in the presence of a cluster of outliers. All these factors contributed to the development of a reliable anomaly detection system.

Finally although the last two chapters primarily focused on detecting anomalies in energy and water consumption in buildings using sensor data, the methods and the results discussed in this project are not restricted to this particular problem and instead can be applied to a number of other domains. For example the idea of using two complementary methods can be easily extended to other application domains. Further time-series based models are one of the most commonly used methods for anomaly detection in sensor data streams. However as discussed due to the presence of large number of apartments possibly with different time-domain characteristics, it may not always be possible to use such time-series based methods. This project showed how that limitation can be overcome by using methods that only considers spatial representation of the data. Lastly the idea of Weighted Local Outlier Factor can be used in other systems where the outliers can potentially form a closely grouped cluster and where the standard Local Outlier Factor is not very effective in identifying such outliers.

This chapter provided the qualitative evaluation of the results of the anomaly detection methods. The most representative cases were presented for each of the scenarios and effectiveness of the methods in detecting abnormalities were illustrated. The chapter also included a discussion of the contributions of this research project and justified the use of techniques developed as part of this project in other domains. The final part of the research project is to present these results in an exploitable way such that the end user can take necessary actions on detecting the presence of such abnormalities. This is described in the following chapter.

Chapter 5

Deployment

This chapter describes how the results from the detection methods are presented to the end user (the maintenance team) such that the results are interpretable and exploitable in order to trigger necessary actions. An interactive dashboard was built for this purpose and the key features of the system are presented in this chapter.

5.1 Anomaly detection system

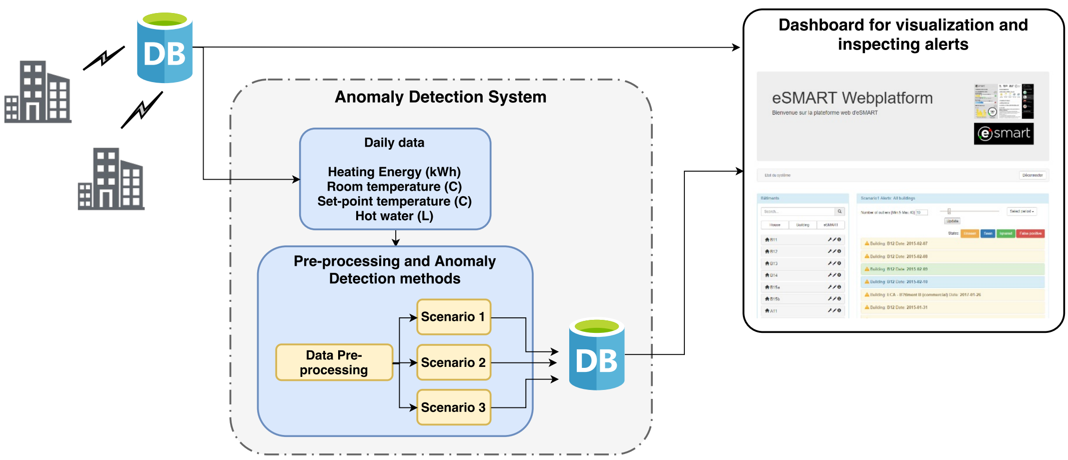


Figure 5-1: Overview of Anomaly Detection System. The data recorded is input to the system which performs pre-processing checks and apply anomaly detection methods for all the scenarios, the results of which can then be visualized in the dashboard.

Figure 5-1 shows the overview of the deployed anomaly detection system. The data collected in all the apartments is posted on a daily basis as an input to the anomaly detection system. This data includes heating energy consumption, set-point and room temperature, hot water consumption. Pre-processing checks are performed to validate the data which is then followed by the application of the detection methods for all the three scenarios. The anomaly scores produced by the methods are then stored in the database which can then be accessed through the dashboard.

5.2 Dashboard for predictive maintenance

An interactive dashboard was developed in order for the maintenance team to exploit the results of anomaly detection methods. The goal here is to make it easier for the end user to understand and interpret the alerts raised by the system. Accordingly the developed dashboard offered the following functionalities. Figure 5-2 illustrates these features.

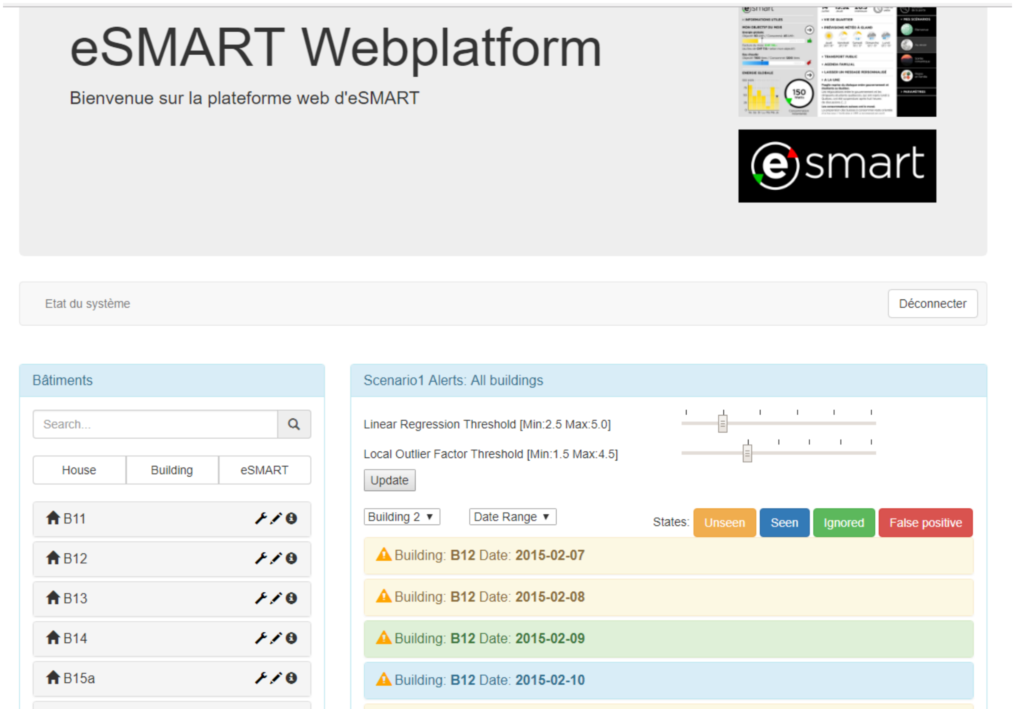


Figure 5-2: Dashboard for interacting with Anomaly Detection System. It offers features such as visualizing the alerts raised, setting different states for alerts, tune the system sensitivity.

1. **View past outliers sorted by order of abnormalities**: Alerts would be raised when the anomaly score produced by the methods exceed certain threshold. These alerts were arranged in the order of abnormality such that the most abnormal data point would be featured at the top.

2. **Visualize outliers detected by methods**: In order to provide more details about the alert, the dashboard also included spatial representation of data with

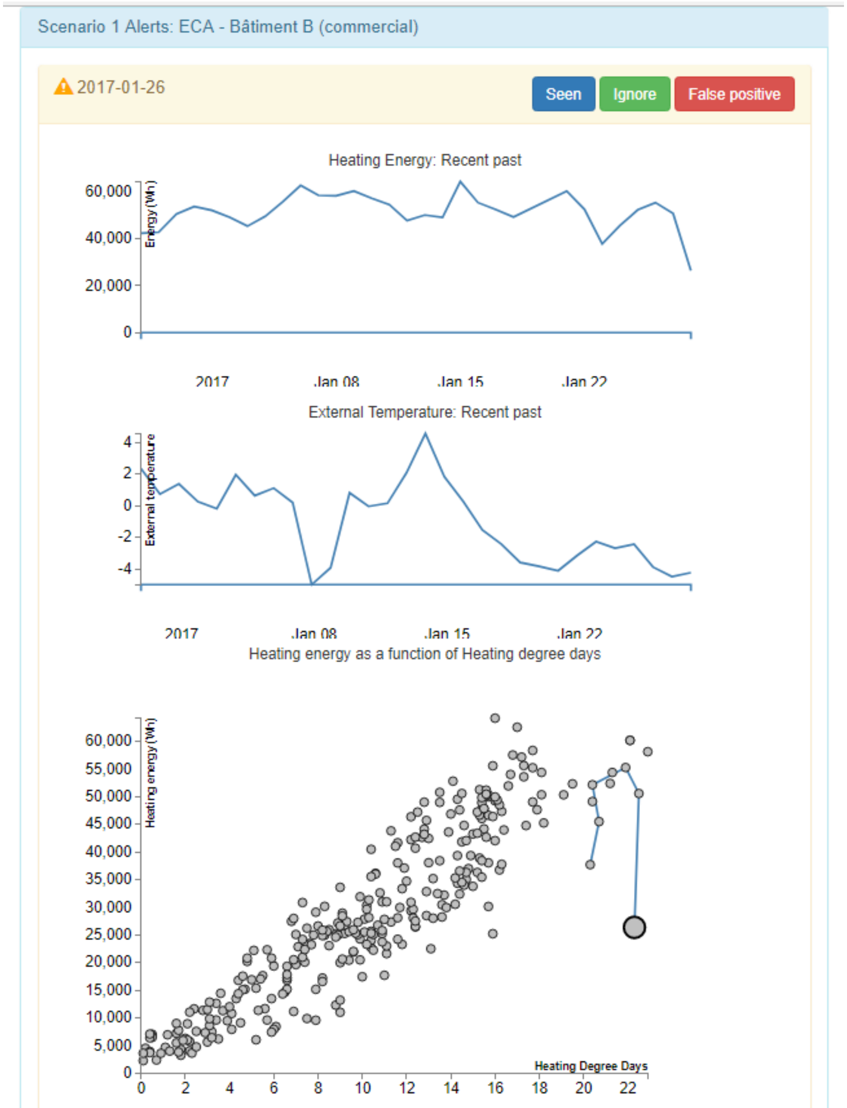


Figure 5-3: In order to interpret the alerts, the dashboard offers plots of spatial representation of data with the abnormalities highlighted. Further data such as Heating energy, HDD can also be visualized.

the abnormal data point highlighted. Further other data such as heating energy consumption, temperature, set-point were also shown for data from recent past. This can be seen in Figure 5-3.

3. **Set detection thresholds of methods to control sensitivity of the system**: It is very critical that the system does not raise false positives. The framework includes the option for setting detection thresholds of the methods. This can be used to tune the sensitivity of the system.

4. **Set states for each listed anomaly**: As part of inspecting alerts, it is essential that the user can set different states such as seen/ignore/false-positive for each

of the raised alarm. Further the list of alerts can then be filtered by the set states.

5. **Filter by building/apartment and date range**: Finally the framework also offered the feature of filtering abnormalities by a given building or apartment and by the range of dates during which the abnormalities occurred.

These features will not only aid in identifying abnormalities on a daily basis, but should ideally also help the end user to interpret and understand the reason for the raised alert. Necessary actions can then be taken leading to predictive maintenance of the system.

This chapter described the deployment details of the anomaly detection system. A web framework was developed for the maintenance team to interact with the system and its features were presented in this chapter. These features illustrated how the anomaly detection system can be used for predictive maintenance which was the primary objective of this research project.

Chapter 6

Conclusion

The eSMART-building solution brings together all of an apartment's basic functions such as heating regulation, energy consumption monitoring, lighting accessible through a wall mounted touchscreen. The modules installed in the apartments record the measured data such as heating energy consumption, room temperature, target temperature set by residents, volume of hot water consumption. As part of the maintenance support of such a system, the company receives complaints from the residents of the apartments in case of the malfunctioning of the system which would then lead to the company dispatching its maintenance team to the site of the apartment for repair. This is an expensive procedure and it would be beneficial for the company if such faults in the system can be detected in advance using data analysis techniques. Accordingly three scenarios were defined and to detect abnormalities in these scenarios formed the research objective of this thesis project. Scenario 1 involved finding abnormalities in heating energy consumption in buildings relative to the external temperature. Scenario 2 involved detecting abnormalities in the heating regulation system in the apartments and Scenario 3 dealt with finding abnormalities in hot water consumption in apartments.

The project produced several valuable contributions. The project began with pre-processing of the collected data in order to validate it before further processing. The pre-processing of the heating energy data provided useful insights in locating certain errors in the data collection procedure. This constituted the first major contribution of the project. The pre-processing step was followed by the selection of the most appropriate anomaly detection algorithms for each of the scenarios. Two complementing methods namely linear regression and local outlier factor were used to detect abnormalities in Scenarios 1 and 2. The global and local nature of the two methods respectively provided diverse information about outliers and the effectiveness of the methods in detecting abnormalities in heating energy consumption was justified by a thorough qualitative analysis of the results. The anomaly scores produced by the methods were used to raise alerts which could then be accessed by the support team through an interactive dashboard that was developed as part of the project. This information could then be used to take up further necessary actions.

The data analysis techniques developed in this project can be extended to other applications in similar domains. Most often in practice the data available in domains such as sensor networks, fraudulent credit card transactions, network security do not have labels for normal and abnormal classes and as a result, the well established supervised classification methods discussed in the literature would no longer be applicable. However this project illustrated the capability of the unsupervised algorithms in detecting abnormal patterns in the data. Further by using two complementing methods, the project also provided the additional flexibility for the domain experts to tune the detection thresholds such that abnormalities of certain type could be given preference over other types. The domain experts can then analyze the raised alerts and decide accordingly the preferred values of detection thresholds for the two methods. This idea can be easily extended to other similar applications. The project also successfully illustrated the use of Bayesian Linear Regression for a practical application. Similarly the weighted Local Outlier Factor method was able to overcome certain limitations of the standard Local Outlier Factor method.

Despite favorable results, this work opened up interesting new opportunities for future work. The project used unsupervised methods to detect abnormalities in the dataset since there were no labels available. However the abnormalities detected by the methods were evaluated by the domain experts and classified as true and false positives. These labels can now be used to apply supervised anomaly detection algorithms. Methods belonging to time series based methods (recurrent auto-encoder), classification based methods (random forests, neural networks) which were discarded due to lack of labels in this project can now be considered for detecting outliers. Further the techniques used in this work treated data as points in multi-dimensional feature space. The future work can explore the use of time series based methods which consider the temporal nature of the data. Also one of the suggestions by the domain experts was to use the dashboard to identify and mark the false positives which could then be an input to the learning algorithms to correct itself such that similar alerts would not be raised in the future. Adding these features will result in an improved anomaly detection system and can be considered as part of the future work.

In conclusion, this thesis project conducted in the industrial framework was successful at detecting abnormalities in heating energy and hot water consumption in buildings using data captured by a network of sensors. The detected abnormalities were used to raise alerts which could then be exploited by the maintenance team to prompt required actions. The project illustrated the effectiveness of data analysis techniques for anomaly detection even with the lack of labeled data and the techniques developed as part of this project can be readily extended to other application domains.

Bibliography

[1] Degree days.net - custom degree day data. URL http://www.degreedays.net/.

[2] Degree days - handle with care! URL http://www.energylens.com/articles/degree-days.

[3] Linear regression analysis of energy consumption data. URL http://www.degreedays.net/regression-analysis.

[4] Isolationforest example, . URL http://scikit-learn.org/stable/modules/generated/sklearn.ensemble.IsolationForest.html.

[5] Anomaly detection with local outlier factor (lof), . URL http://scikit-learn.org/stable/auto_examples/neighbors/plot_lof.html.

[6] Density estimation, novelty detection, . URL http://scikit-learn.org/stable/modules/svm.html#svm-outlier-detection.

[7] Subutai Ahmad, Alexander Lavin, Scott Purdy, and Zuha Agha. Unsupervised real-time anomaly detection for streaming data. *Neurocomputing*, 06 2017. URL http://dx.doi.org/10.1002/widm.2.

[8] A Amitai, W Gilad, and F Lev. Change and anomaly detection framework for internet of things data streams, 2016. URL https://software.intel.com/en-us/articles/change-and-anomaly-detection-framework-for-internet-of-things-data-streams.

[9] C. Aytekin, X. Ni, F. Cricri, and E. Aksu. Clustering and Unsupervised Anomaly Detection with L2 Normalized Deep Auto-Encoder Representations. *ArXiv e-prints*, February 2018.

[10] Varun Chandola, Arindam Banerjee, and Vipin Kumar. Anomaly detection: A survey. *ACM Comput. Surv.*, 41(3):15:1–15:58, July 2009. ISSN 0360-0300. doi: 10.1145/1541880.1541882. URL http://doi.acm.org/10.1145/1541880.1541882.

[11] Arun Kejariwal. Introducing practical and robust anomaly detection in a time series, 2015. URL https://blog.twitter.com/engineering/en_us/a/2015/introducing-practical-and-robust-anomaly-detection-in-a-time-series.html.

[12] I. Kiss, B. Genge, P. Haller, and G. Sebestyn. Data clustering-based anomaly detection in industrial control systems. In *2014 IEEE 10th International Conference on Intelligent Computer Communication and Processing (ICCP)*, pages 275–281, Sept 2014. doi: 10.1109/ICCP.2014.6937009.

[13] Fei Liu, Young Lee, Huijing Jiang, Jane Snowdon, and Michael Bobker. Statistical modeling for anomaly detection, forecasting and root cause analysis of energy consumption for a portfolio of buildings. *IBPSA*, 11 2011.

[14] Mathew Ma, Henry Ngan, and Wei Liu. Density-based outlier detection by local outlier factor on largescale traffic data". *Electronic Imaging*, 2016(14), 2016.

[15] Nerijus Paulauskas and Ažuolas Faustas Bagdonas. Local outlier factor use for the network flow anomaly detection. *Sec. and Commun. Netw.*, 8(18):4203–4212, December 2015. ISSN 1939-0114. doi: 10.1002/sec.1335. URL `http://dx.doi.org/10.1002/sec.1335`.

[16] Peter J. Rousseeuw and Mia Hubert. Robust statistics for outlier detection. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 1(1): 73–79, 2011. ISSN 1942-4795. doi: 10.1002/widm.2. URL `http://dx.doi.org/10.1002/widm.2`.

[17] Dominique Shipmon, Jason Gurevitch, Paolo Piselli, and Steve Edwards. Time series anomaly detection; detection of anomalous drops with limited features and sparse examples in noisy highly periodic data. 08 2017. URL `https://arxiv.org/abs/1708.03665`.

[18] Iwan Syarif, Adam Prugel-Bennett, and Gary Wills. Unsupervised clustering approach for network anomaly detection. In Rachid Benlamri, editor, *Networked Digital Technologies*, pages 135–145, Berlin, Heidelberg, 2012. Springer Berlin Heidelberg. ISBN 978-3-642-30507-8.

[19] Jeffrey Wong, Chris Colburn, Elijah Meeks, and Shankar Vedaraman. Change and anomaly detection framework for internet of things data streams, 2015. URL `RADOutlierDetectiononBigData`.