Energy efficient heating systems using IOT and machine learning

MSc Research Project
Data Analytics

Naga Srikanth Janjanam
x15028836

School of Computing
National College of Ireland

Supervisor: Mr Michael Bradford

National College of Ireland
Project Submission Sheet – 2015/2016
School of Computing

Student Name:	Naga Srikanth Janjanam
Student ID:	x15028836
Programme:	Data Analytics
Year:	2016
Module:	MSc Research Project
Lecturer:	Mr Michael Bradford
Submission Due Date:	12/12/2016
Project Title:	Energy efficient heating systems using IOT and machine learning
Word Count:	7171

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are encouraged to use the Harvard Referencing Standard supplied by the Library. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action. Students may be required to undergo a viva (oral examination) if there is suspicion about the validity of their submitted work.

Signature:	
Date:	21st December 2016

PLEASE READ THE FOLLOWING INSTRUCTIONS:
1. Please attach a completed copy of this sheet to each project (including multiple copies).
2. **You must ensure that you retain a HARD COPY of ALL projects**, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer. Please do not bind projects or place in covers unless specifically requested.
3. Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Contents

Energy efficient heating systems using IOT and machine learning

Naga Srikanth Janjanam
x15028836
MSc Research Project in Data Analytics

21st December 2016

Abstract

District heating is widely used in cold countries and it is a way of maintaining centralised heating systems instead of traditional way of using multiple smaller heating devices. It would reduce the cost of setup, maintenance and also would deliver high efficiency. This research focuses on reducing energy wastage and building energy efficient heating systems in district heating industry. With the help of modern technical advancements, like using IoT, MBus protocols, Radio communications, Java based services, Databases and Machine Learning, it is possible to gather various data points like current customer usage, weather at the customer location, time at the customer location, etc. Such factors can be used with machine learning algorithms to predicted the next 1 hours customer heat requirement and it can be used to build efficient district heating system. An efficient heating system would consume less energy thus reducing energy wastage. So far this method has achieved 88 percent accuracy with this method and it is expected to grow with some future work as mentioned in future works section. Gubbi et al. (2013)

1 Introduction

District heating is a way of implementing centralised heating systems usually referred as district heating boiler, instead of a traditional way of using multiple smaller boilers. Their are cheaper to setup and known to delivery higher efficiency and easier maintenance. It is majorly used in apartment blocks, shared heating environments and for commercial purposes in cold countries. The major bottle neck in district heating systems is that the energy requirements of using them are very high and the customer usage needs varies on various factors like time, weather, season, etc. Currently there is no way of figuring out the variations in customer usage and optimising the heating systems. So there is huge wastage of energy in maintaining them.

Questions

1. How can we reduce the wastage in energy spent on district heating systems by predicting the customers demand based on their current usage and weather?

2. How can data mining methods be applied on district heating industry to predict the customers heat usage based on their current usage and weather conditions?

Such energy wastage can be minimised if we can predict the customer usage. This research proposes and critically evaluates a way of predicting customers heat usage every hour using advancements in modern technology.

With the rapid growth in Internet of Things (IoT), information and communication equipment and better radio communications, it has became possible to use IoT devices to remotely read customers heat meter readings from the meters installed in their apartment blocks. Such implementation is currently available with an Irish district heating company Prepago Platform Limited and they are primary source of meter data in this project. And with the help of hundreds of weather reporting systems installed across Ireland by the Irish National Meteorological Service, it is possible to exactly get the weather at the district heating systems location. Such information can be regularly fetched using Java based services, stored in MySQL database, further ETL process performed using database aggregate queries, cleaned using PHP scripts, stored in intermediate storage and further with the help of Machine Learning, it is possible to predict the next 1 hours hot water demand of the customers. Such prediction can be used to reduce the energy usage by district heating systems by optimising them accordingly. Karimi and Atkinson (2013) Oldewurtel et al. (2010)

This research paper would cover in detail the related work, methodology, implementation procedure and would evaluate the results at the end.

2 Related Work

It is assumed that the audience of this paper know about Hadoop, HDFS, Map reduce, machine learning, predictive analytics and big data processing in cloud. The definitions and detailed description of the technologies used are covered in the next section.

This section starts with the area and covers the existing theories proposed by various authors, factors influencing consumer behaviour and practices, current methods used to saving energy in heating systems, available tools and methods for smart meter data analytics and their current commercial implementations.

2.1 Area

This paper proposes a way of predicting the customers hot water usage and using such information to control boiler rooms to prevent energy waste. To achieve this, few basic data points like customers current usage, time and date information, weather information are needed. The best way to acquire customer current usage is by reading meter when needed, that could be possible with modern advancements in electronic and communications technologies like IoT devices, information on weather can be achieved using weather prediction systems. Once data is available the customer hot water usage prediction can be done by smart meter analytics which is a way of reading water usage meters using latest advancements in technologies like IoT and performing analytics on such data.

IoT is known for elevating the possibilities in developing smart cities, buildings and grids to make them the future for smart environments. And an estimation of an event before occurring, enabling actions such mitigation of the occurrence or be prepared for its eventuality is Predictive Analytics. It conceptualises the application in developing energy management systems which could provide a forecast of the usage of energy that can be enabled by efficient and effective scheduling of building operations. Derguech et al. (2014)

2.2 Existing Theories

This section discusses the existing theories proposed by various authors on the effective use of smart maters, using such data to predict customer hot water usage, using weather data to build a prediction model and a methodology to build a complete smart metering system.

The theory proposed by the author Zavala et al preempts weather conditions in adjusting the heating or cooling based on a numerical weather prediction model that uses thermal resistance and inertia of a building. The proposed framework is a stochastic dynamic based on the solution of real-time optimization using single weather data source. Zavala et al. (2009)

Another author Liu and Nielsen (2015) proposed a methodology which states that the smart meter analytics system must be efficient, available, must allow real time processing and must be a complete solution. His technical methodology to achieve a complete solution is as given below.

Efficiency: The system provides great browsing and analysis experience as it supports low latency queries, and better service quality by enabling in-memory based databases.

High performance NoSQL databases (with trading off consistency for availability, fault tolerance and performance)

OR

Traditional relational database management systems with their support low latency queries would be exploited. Support of common operations over time series data such as temporal range queries by the selected data storage system should be available.

Availability: Only a highly available system captures near real time data and provide data service and analytics, A cluster environment, e.g., a private or public cloud, which offer fault tolerance as a service to the system which requires reliability and availability to enable this system.

Real time and batch data ingestion: Two data ingestion mechanisms will be implemented by the system that is real time and batch data ingestion. Real-time data ingestion which will provide near real time analytics results to customers by integrating the data stream into the system directly from smart meters, while batch ingestion will be used to handle data in bulk mode, e.g., the data uploaded by users, or data from legacy systems.

Solution: A complete solution is enabled by the system to smart meter data analytics including data ingestion, transformation, loading, data analytics and visualization. The system will aim in supporting big data and streamline of smart meter data analytics.

2.3 Customer behaviour change

This section would provide a short summary of the important factors influencing consumer behaviours and practices. Technological interventions are involved in most energy efficiency measures implemented (or yet to be implemented) in Europe, but people adjusting their energy consumption behaviour are equally in observation as behavioural models are necessary to comprehend consumers activities such as what and why they do. Such models tend to vary widely by theory, concepts and applications (Axsen and Kurani (2012)). The relationships are dynamic, and not static between various factors that influence behaviour and consumption practices and the human element, as they are assumed to be in literature available largely on this topic. They may change periodically, rendering the consumer behaviour and further the process of consumption practices development somewhat irrational and unpredictable to some extent.

Recent social science research reveals some of the limitations of the work carried out in this area on consumer behaviour over the past decades. Social science can improve our understanding of individual and societal responses to their environments; this has been used to investigate public's relationship with energy, energy use and energy efficiency behaviour change initiatives/measures.

There is evidence that routine utilisation is controlled to a large extent by social standards, and is profoundly shaped by cultural and economic factors as argued by Elizabeth Shove (Chappells and Shove (2005)). According to her not only are the habits changing, but they are doing so in a way that often turns into a standardisation of consumption patterns (via commercial interests, more frequently than not) favourable to an escalation of resource consumption and environmental degradation. Shove challenges the conventional wisdom whereby adoption of more sustainable ways of life depends on the dispersal of 'green' beliefs and actions through society in her work entitled Converging Conventions of Comfort, Cleanliness and Convenience. The current consumption patterns are demonstrated pertaining to energy and water, reflect that we are generally the routines and habits are not easily available.

2.4 Energy Savings in domestic heating systems

Below are the few methods proposed by in the refereed papers on a way of reducing energy waste in domestic water heating systems like Geysers and Solar water heaters.

In Geysers the heating elements activity is controlled by geyser thermostats. Energy savings could be achieved If the thermostat setting was reduced as it uses less electricity to heat the water. The element will be turned on if the temperature of water in the tank falls below a certain threshold, until it reaches a desired level. This includes the additional fact that increases in the rate of heat loss as hotter water creates a higher temperature gradient with the ambient. Dent et al. (2013)

In Solar water heating system, the energy could be saved if it is used in the after noon. It might have little to no effect if most hot water is used in the morning in comparison to hot water usage that peaks in the evenings, especially during the summer months. This is due to the availability of useful solar energy only in the noon. Observation on the impact of changing the time of hot water consumption, simulations were carried out for a hypothetical situation, where the peak period over 5 hours a day on water usage in the given day and the concentration on a specific hour (x) with the distribution. In general, it was observed that peak hot water consumption took place early in the afternoon as the maximum energy savings was received. It can be concluded that users of solar water heating systems can be benefited by hot water usage to afternoon. Due to commitments such as work and school it may not be realistic. Hence consumption of water should be around the closest time to afternoons in order to reduce the spending.Ijumba and Sebitosi (2010)

2.5 Smart Meter Data Analytics

As mentioned before smart meter analytics is a way of reading water usage meters using latest advancements in technologies like IoT and performing analytics on such data. As meters are used in every household with hot water usage, the volume of data of all the houses of a city would be high. This section would cover the existing tools, technologies used for smart meter analytics and benchmarking, existing commercial implementations.

2.5.1 Tools and Platforms

Traditional technologies such as R (S-PLUS), Matlab, SAS, SPSS and Stata can be applied in smart meter data analytics for numeric computing and comprehensive statistical analysis. The analytics technologies are to support in-memory, in-database, and in parallel on a cluster is trending. They use synthetic electricity data, and evaluates six queries for smart meter data analytics algorithms on HP Vertica cluster platform. Arlitt et al. (2015)

2.5.2 Commercial Implementations

In the current commercial implementations, smart meter analytics is highly in demand due to the huge increase in data acquisition through various government implementations like smart meter roll-out programme in Sweden which introduced 5.2 million customers and Smart Metering Implementation Programme in UK which involves upgrading or replacement of the existing meters which are over 53 million in UK during the years 2015 and 2020 in domestic and other buildings.

3 Methodology

This section covers the strategic list of various modules needed to execute this project. It starts with the process flow diagram of the modules needed and discusses a gist of methods used for data acquisition, technologies used, flow of data between various modules, data stores used, prediction models used to achieve the target of predicting customer hot water usage every hour.

3.1 Process flow diagram

I have designed the below given process flow based on the points from the papers Kamdi and Bhoyar (2014) Zavala et al. (2010) .

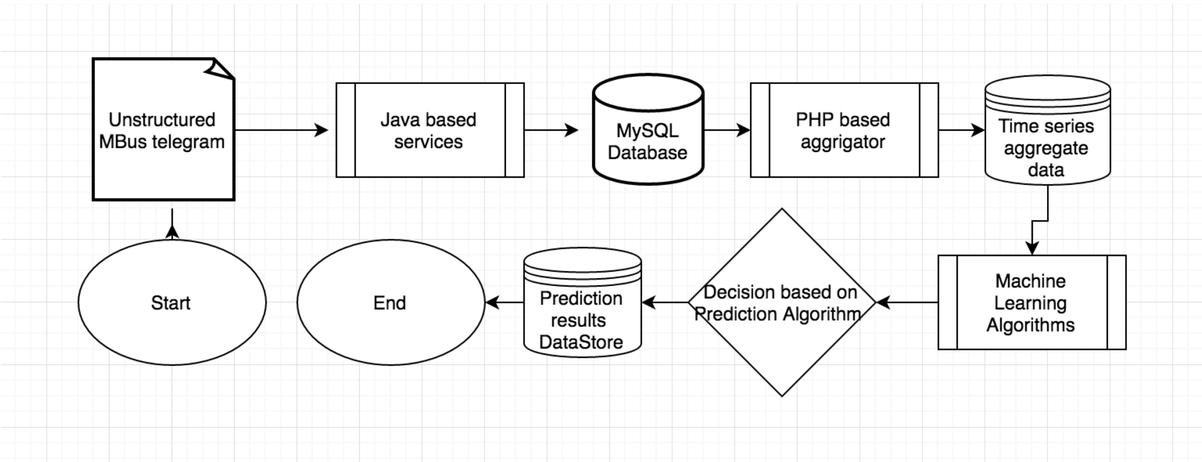

Figure 1: Process flow of various modules.

Below is the step wise discussion on the various modules as per the process flow.

3.1.1 Data volume

As data is the key for deriving value and the strategy chosen needs to be able to process it fast and instantaneously even if data volume is huge. Below points covers the details of volume of the data this strategy is designed for.

The volume and the scale of the data which is generated from IOT Heat meters is very vast and a brief stats for the same are mentioned below.

Lets assume a data set of 10000 heat meters.

Heat meters = 10,000 units

Number of readings per day = 96 (Every 15 mins)

This gives us with a data set of 96*10000 = 960,000 MBus telegrams per day

As described in the above mentioned research, in the years of 2015 and 2020, UK is expected to upgrade their Heat, gas and electricity meters for more than 53 million which was announced under the The Smart Metering Programme and this smart reading implementation already done in Ireland and official programme for which conversion of all the old and existing meters is also under consideration. This smart reading system is also been deployed in most of the European cities and which would help build energy efficient systems. The information would show us how scalable the data could be.

This live implementation across Ireland, UK, USA, Canada and South Africa can give access to a minimum of billion smart meters.

Thus in production the data set can be of

Heat meters = 1 Billion units

Number of readings per day = 96 (Every 15 mins)

This leaves us with a data set of 96*1,000,000,000 = 96 Billion MBus telegrams per day

3.1.2 Data acquisition process

As the research mentioned above states the need for huge volume of data, the system needs to be designed in a way where it could handle it and scale up over the time.

As stated before the main source of data acquired with the help of an Irish district heating company Prepago Platform Limited. The data of heat meters installed in every apartment block are requested through IoT devices which are communicated via M Bus protocol. IoT devices return semi structured M Bus telegram of meter reading information. The length and format of the the M Bus telegram would vary for different heat meters. This is how the semi structured which has been cleaned would be load into the data warehouse.

Another useful input would be the weather data which would be playing a major role in this research project would be downloaded and loaded into the data warehouse with necessary useful attributes of the data.

Kimball and Ross (2011)

The process of gathering data will go through the standard process of Extract, Transform and Load.

3.1.3 MySQL

There are many factors that are to be considered when performing a database operation and running queries on them, major ones are the features, support and also price and licensing. While this is a major concern, there are several open source products which are free and aren't expensive. MySQl is a open source relational database management system RDBMS. It is very widely used for web and embedded applications as it is fast and reliable. For the similar reasons it was made a popular alternative to other database systems. Bell (2016)

3.1.4 PHP

The data aggregation and processing is needed to generate the required subset of data from the above step of MySQL and we have used PHP scripting language for writing the program to do the same. PHP is a general purpose scripting language. It is widely used for web development. Generally, all the web based power systems applications for analysis are delivered through web servers that would run scripting languages. One of those important scripting languages is PHP. Gubbi et al. (2013)

3.1.5 Data Warehousing

The data from the above section needs to be stored in a data warehouse in a time series manner. As the systems needs to operate every hour the data needs to be a live stream (Polyzotis et al. (2007)) and as the volume grows, the theories of big data needs to be implemented. In the present-day scenario, the word Big Data is used and is very widely distributed over many years and one major efficient tool to handle this huge amount of data is a fast-growing data warehousing technology. Data warehouse understands and implements the data integration solutions. This concept of building a data warehouse is originally based on two of main concepts integrating separated data that is detailed by facts, properties and events within a single storage and separation of data sets and the applications that are used for processing the operations and for solving the problems faced during the analysis. Inmon in his monograph gave a very first definition of data warehouse where he has defined it as a subject-oriented, integrated and contains a historical data which is a indestructible collection of data in the support of managements decisions. Inmon and Hackathorn (1994)

3.1.6 Machine Learning

Machine learning is a branch in information technology field that studies the design of algorithms. Typical Machine learning queries are conceptual based tasks, function learning or " Predictive modeling", Clustering and finding predictive patterns. These tasks are Learned through available data that were observed through experiences. For instances, It hopes that including the experience into its tasks will eventually improve learning. Machine learning is a key concept to enhance the skills and improve learning methods in such a way that it becomes more automatic, so that humans do not need to involve further. Machine learning has close ties with knowledge discovery data, Data Mining, Artificial intelligence, Neural Networks and statistics. Typical applications of machine learning can be classified into scientific knowledge discovery and more commercial applications. R can be used for various algorithms like KNN, Logistic Regression, Vector scale etc.

3.1.7 Machine Learning Algorithms

Designing a very efficient, more accurate and much low complex system is a challenge with the technologies used currently. Several machine learning algorithms have been developed so that they can overcome these challenges and build a model suitable for the environments in the training phase.

3.1.8 Predictive Analytics

It is the technological usage of the business techniques, data modelling, algorithms and machine learning approach for finding future perspectives, prediction and trends. It utilizes big data available on the web and social networking sites for critical insights and building predictive models on top of it. The theories related to predictive analytics are taken from Ismail et al. (2011).

4 Implementation

The implementation of this project aims to reduce the energy waste in boilers rooms. Boiler rooms in district heating industry are described using a word "Scheme" and are used to describe a set of apartment blocks using a boiler. For example, 80 apartment blocks might be using 1 boiler. So it would be having 1 scheme for those 80 apartment blocks.

This section covers the project implementation in depth.

Below given are the list of stage involved in this project implementation.

1. Technical Environment
2. Data source
3. MySQL: Initial processing and data storage
4. PHP: Data aggregation and processing
5. Data warehouse: Time series data store in an intermediate storage
6. Create validation data set
7. Data analysis and visualisation using R programming
8. Building Machine Learning models using Caret library
9. Build a custom streaming engine
10. Make predictions
11. Incrementally store predictions to target data store
12. Use predictions to district heating systems

4.1 Environment

Below are the various components needed in the development environment.

1. Linux server
2. MySQL
3. PHP
4. Excel
5. R library installed
6. R Programming to create a custom script
7. Caret library

4.2 Data source

As stated before the main source of data acquired is with the help of an Irish district heating company Prepago Platform Limited.

Technically, it is read using IoT Systems, in which the live readings from the heating systems are acquired through MBus protocol using IOT devices. This data will be a long unstructured Telegram and it is further processed to calculate and extract meter reading values from it using Java services.

4.3 MySQL: Initial processing and data storage

The data acquired from IOT devices are cleaned and a simplified version of it is stored in MySQL Database. The data from all the IOT devices like heat usage meters are associated with their appropriate building locations and are stored as a record in MySQL table.

4.4 PHP: Data aggregation and processing

The individual meter readings data is processed and is further aggregated into application specific data like a data set per building. It calculates various data fields like usage, previous usage average for last 10 times, usage for a particular hour, Percent of Difference last time over average usage and it also includes weather information.

The system is designed for district heating industry and heat is highly dependent on weather conditions. So various live weather information along with temperature, wind speed are included in the data set.

4.4.1 Weather Data

Weather information included are

1. Temperature: in degrees
2. Humidity
3. Wind String. A string describing the details of the wind conditions. Eg: Calm, From the East at 3.8 MPH Gusting to 7.6 MPH
4. Wind Direction: SW, East, NNE
5. Wind Speed
6. Wind Pressure
7. Temperature Feels Like
8. Visibility
9. Season. Can be spring, summer, fall or winter.
10. Date in dd/mm/yyyy format.
11. isWeekend: 0 (for no) or 1 (for yes)

4.5 Data warehouse: Time series data store in an intermediate storage

The PHP based processor from the above section combined with the weather data of that particular weather data is stored in an intermediate storage of a flat file. In future as the data volumes grow it would be required to build data warehouse replacing the flat file to support faster querying, to better access and data replication.

4.6 Create validation data set

The final processed data need to be further cleaned sub divided to form a training data set for the machine learning model training purpose.

4.6.1 Summary of validation dataset

The data from the above step is an aggregated data of a scheme for a time. As described at the be beginning of this section, a scheme is a set of apartment blocks using a boiler room in district heating industry. The Scheme's variation in heat usage is made into a variable called percentOfDiff and a classified version of it is called POD as in the below given table. The project's prediction algorithms needs to predict the classifier variable POD_Predicted. The various other data variables supporting to predict heat usage are as given below.

Data Set	**Data Type**	**Information**
Scheme	integer	A scheme is a set multiple apartment blocks identified using a unique ID number.
Hour	integer	Hour of a day in 0-23 format.
Houses	integer	Count of number of apartment blocks
Previous Reading	integer	Yesterdays aggregate meter reading
Previous Usage	integer	Yesterdays aggregate heat usage
Avg Usage	integer	This Aggregate heat usage of past 10 days before yesterday
percentOfDiff	decimal	Variation in heat usage in percentage
POD	decimal	A classifier describing the recent change in the heat usage
$\text{POD}_{Predicted}$	decimal	A value needs to be predicted and is a classifier describing the next one hours predicted change in the heat usage
Previous Hour	integer	
City	nvarchar	
Local Time	timestamp	
Temperature	integer	Actuall temperature in degrees
Humidity	integer	
Wind String	nvarchar	A string describing the details of the wind conditions. Eg: Calm, From the East at 3.8 MPH Gusting to 7.6 MPH
Wind Direction	nvarchar	A string describing the direction of the wind. Eg: SW, East, NNE
Wind Speed	integer	
Wind Pressure	integer	
Temperature Feels Like	integer	Feels like temperature due to other factors like wind, wind pressure, humidity, etc.
Visibility	integer	
Season	nvarchar	A string representing 4 seasons in an year. Can be spring, summer, fall or winter.
Date	Date	Current date, in which prediction is being performed in dd/mm/yyyy format.
isWeekend	integer	An integer acting as a Boolean representing if the current day is a weekend or not. 0 (for no) or 1 (for yes).

Below given is the screnshot of the data set in Excel.

Scheme	Hour	Houses	Previous Reading	Usage	Avg	percentOfDiff	Previous Hour	City	Local Time	Temperature	Humidity	Wind String	Wind Direction	Wind Speed	Wind Pressure	Temperature Feels Like	Visibility
3	0	76	1032970	243	56	333.9285714	20	Dublin	Wed, 02 Nov 2016 23:00:19 +0000	9.8	92%	Calm	SW	0	30.24	10	10
5	0	70	156622	745	407	83.04668305	20	Dublin	Wed, 02 Nov 2016 23:00:19 +0000	9.8	92%	Calm	SW	0	30.24	10	10
6	0	19	146271	1497	690.5	116.7994207	20	Dublin	Wed, 02 Nov 2016 23:00:19 +0000	9.8	92%	Calm	SW	0	30.24	10	10
7	0	55	930145	496	256.5	93.37231969	20	Dublin	Wed, 02 Nov 2016 23:00:19 +0000	9.8	92%	Calm	SW	0	30.24	10	10
14	0	54	236882	76	42	80.95238095	20	Dublin	Wed, 02 Nov 2016 23:00:19 +0000	9.8	92%	Calm	SW	0	30.24	10	10

Figure 2: Validation dataset screenshot.

4.7 Data analysis and visualisation using R programming

The above data is processed in R studio using R programming language. The dependency relationship between data variables is analysed and to find the dependent and independent variables.

So the dependent variables are

1. PoD
2. Temperature
3. Hour

So the independent variable is

1. isWeekend

4.8 Building Machine Learning models using Caret library

Based on the relationship between the variables the next step would be to train a machine learning model build using with the help of caret library in R.

4.9 Build a custom streaming engine

Once the model is build, on the real time implementation of this system needs to process a live stream of data to achieve complete automation. With in R programming, we have to build a custom streaming engine as existing models doesn't serve the purpose. The steaming engines responsibility is to read the new records that needs machine learning and sends them for prediction.

4.10 Predictions, storage and end result

The above streaming input fetched a row of all the data variables mentioned in the section 4.6.1 . It explains a classifier variable POD_Predicted that needs to be predicted. We also know the dependent and independent variable as mentioned in the section 4.7 . I have used the below mentioned formulae to train the model as the impact of the variables POD, isWeekend, Temperature, Hour is very high on the target variable of POD_Predicted.

$$POD_Predicted\ PoD + isWeekend + Temperature + Hour \quad (1)$$

The result is further appended in a target data store by the R script. The target data store is directly used with the Prepago system to further control heating systems.

5 Evaluation

Below given are the few kinds of evaluation needed to measure the accuracy and usefulness of this research.

1. Data evaluation
2. Model evaluation

5.1 Data evaluation

The data is acquired directly from the IOT devices which are read using MBus protocol from Heat meters.

They fetch a long MBus telegram which is processed using Java based services and the calculated usage is further stored in a MySQL data base along with a latest time stamp. As the data has a valid time stamp and it is coming directly from the servers of the data provider, it can be considered as a trusted data.

5.2 Model Evaluation

This process involves the process of selecting the best model for the process and to achieve that we need to go thorough few prerequisites as given below.

1. Prepare validation data set
2. Use test harness with 5-fold cross validation
3. Test and try different models
4. Select the most accurate model and prediction

5.2.1 Prepare validation dataset

To find the accuracy of the models we create so that we can know on which algorithm is working fine.

Accuracy of the models will be calculated using statistical methods, but the results generated by the statistical methods needs to be compared against actual valid data to be 100% sure.

So to do that we need to introduce 2 new variables called "Actual value", "Actual increment" . This variable gets its information from future records of similar scheme and hour information.

We can do this because while training we have a time series data set of a time period.

The variable the "Actual increment" would be the actual percent of increment or decrement in the heat usage. It can be classified into few ranges like (lowest to -0.03,-0.03 to -0.005, -0.005 to 0.005, 0.005 to 0.03 and 0.03 to highest). By doing so it gets simplified for the data model and also for the heat systems as heat control levels can be classified.

5.2.2 Use test harness with 5-fold cross validation

From the above dataset we need to hold back some portion of the data that will not be used with the training of the algorithm. It can be further used by the same algorithm to calculate the accuracy of itself.

And to achieve that we will be splitting the date into two 5 folds, that is if the data is split into 5 portions, then we will be giving 4 portions of it to training and 1 portion of it to cross validate and calculate the accuracy of the model.

5.2.3 Test and try different models

So far we have a clear understanding of the data, identified dependent and independent variables as mentioned in section 4.7 and also developed a formulae to try with as mentioned in section 4.10. We are still unclear of which model to try. So lets choose few related machine machine algorithms that could work in our scenario of classification type of data and try them out.

Below given are the algorithms that are tried.

1. Boosted logistic regression
2. Random forest
3. Linear Discriminant Analysis
4. K-Nearest Neighbors (KNN)
5. Classification and Regression Trees (CART).
6. Gaussian Naive Bayes (NB).
7. Support Vector Machines (SVM).

It is recommended that we try both simple linear and non linear algorithms. We also should reset the random number seed every time we try a new algorithm so that the results would be directly comparable.

The above algorithms when used with formulae mentioned in section 4.10 to predict the data variable POD_Predicted mentioned in section 4.6.1 have generated the below given accuracies as given below.

Algorithm	**Accuracy**
Boosted logistic regression	0.8814706
Random forest	0.8269697
Linear Discriminant Analysis	0.7588745
K-Nearest Neighbors	0.7598268
Classification and Regression Trees	0.603773585
Gaussian Naive Bayes	0.778355
Support Vector Machines	0.7779221

5.2.4 Select the most accurate model

We see that the Boosted logistic regression has generated the most accurate results in our model so far.

Why Boosted logistic regression? The reason Boosted logistic regression has generated a better accuracy is that it is a generalized linear model, unlike other linear models. It creates an ensemble classifier, it combines the output of weak learners to generate a weighted sum that represents a final output.

5.2.5 Prediction

Now we need to setup the same algorithm on the real time processing as well. The results generated from the prediction will be further send to the heat systems to control them to optimise them based on the predicted needs. It helps reduce energy waste and also better hot water supply based on predicted demand. Further the actual usage of the heat by the end customers will be read from the IoT devices are will be returned to the prediction system as another data record.

In the previously mentioned step of "Prepare validation dataset" we have mentioned on how to use the future data from the time series data to calculate the actual usage and which can be further used to accuracy of the model.

Thus we have observed an accuracy of 88% and thus it has a significant result on the energy saving due by adjusting the heat on the heating systems.

5.3 Case Study 1: Predict the heat usage per schemes

As mentioned in the section 4, a scheme is a set of apartment blocks using district heating from a boiler. This project's purpose is to predict the demand on the heat systems based on the data from the IOT devices. With the help of modern technologies we were able to get live meter readings. Such data can be further processed to find trends and with the help of live streaming and the capability to process big data in real time we are able to develop machine learning systems that can predict the heat usage. Such prediction was possible with the methodology stated in the section 3 and the implementation procedure mentioned in section 4 and the above section of 5.2.3 concludes the same with results.

5.4 Case Study 2: Energy savings on district heating systems

This is primary purpose of this project as it helps optimise the energy consumption of the heat systems. In the current district heating industry there is no way to predict the changes in the demand for heat, so the boiler systems are always kept on. As a result when they are not needed they are considered to be under used and when there is more demand then production, they fail to produce heat as per demand. Also keeping those systems live all the time, is energy consuming, so when there is low demand we are just wasted the energy spent on them. So with the help of this project as mentioned in the section 5.2.5 we can eliminate the wastage and there by save energy spent. This also helps us optimise district heating systems to be ready when the demand is more.

5.5 Case Study 3: Demand and supply management

As we are able to predict the heat usage we can exactly know the demand of the heat, so we can optimise the heat systems to produce as per the requirement. Then with the help of this project district heating systems are able to better serve the demand and supply of customers. The prediction possibility as mentioned in the section 5.2.3 combined with the process of using this prediction control the heating systems for more demand as mentioned in section 5.2.5 we can achieve better heat supply.

5.6 Discussion

The above discussed case studies cover the three primary uses of this project. They take the completed advantage of all the technological advances implemented in this project like Internet of Things, Big data processing, Using the advantages of cloud based processing, Machine learning and Prediction.

6 Conclusion and Future Work

The need for automated systems is increasing day by day and the modern technical advancements allow us to build systems that are capable to handle increasing technical complexities.

In this project we have leveraged the the combination of this modern technical advancements to achieve automation in energy savings in the district heating industry.

To summarise, this research covers the use of the IOT technology to get the ability to read live meter readings remotely, instantaneously and fast . It discusses a method of using services in Java that can acquire MBus telegrams and process them to derive meaningful information from them. The data bases are used to store such data. The big data processing technical advancement is used to process huge amounts of data to generate the required output from them. The data warehouses are used to store processed data in an incremental and time series manner. The Live streaming services are used to gather and process live flow of data across multiple modules of this system. The machine learning is used to build models based on the data from the data warehouse and prediction is used to predict the demand for heat usage based on the demand form the live stream of data.

Thus the combination of technologies enable us to build the machine learning systems and its prediction further help us to optimise the energy consumption in the district heating industry and help build energy efficient systems thus saving vast amounts of energy across the globe.

6.1 Future work

The system can always be improved, its accuracy in prediction can be further increased. The current system only deals with the general use, but the exceptional non common use like changes in demand during festivals and holidays needs to be added to the machine learning.

We can do the below given improvements to the system in future

1. Add more data variables like holidays and festivals
2. Add the trends of month and season and customise the weather prediction accordingly and observe the trends
3. Use the next hour weather temperature, that is weather forecast
4. Add the ability to predict to for next hour weather prediction as in the current system next hours usage will be predicted based on the live weather report of the current hour.
5. More real time features, trends of more frequent readings and also with trends of both past records of same time
6. Implement complex trends: A mix of multiple trends together to form a single trend like trends from
   (a) more frequent readings
   (b) trends from the past records of same time holiday and festival impact on peoples usage
7. Integration into other prediction systems like Local holiday and festival impact prediction.

Acknowledgements

I would like to thank my supervisor **Mr Michael Bradford** for the great skills and knowledge he has imparted on me. His guidance towards this successful project is highly acknowledged.

I would also like to thank **Prepago Platform Limited** for granting me the access with their servers, databases and for the continuous support to make this project successful.

References

Arlitt, M., Marwah, M., Bellala, G., Shah, A., Healey, J. and Vandiver, B. (2015). Iotabench: an internet of things analytics benchmark, *Proceedings of the 6th ACM/SPEC International Conference on Performance Engineering*, ACM, pp. 133–144.

Axsen, J. and Kurani, K. S. (2012). Social influence, consumer behavior, and low-carbon energy transitions, *Annual Review of Environment and Resources* **37**: 311–340.

Bell, C. (2016). High availability iot solutions, *MySQL for the Internet of Things*, Springer, pp. 251–271.

Chappells, H. and Shove, E. (2005). Debating the future of comfort: environmental sustainability, energy consumption and the indoor environment, *Building Research & Information* **33**(1): 32–40.

Dent, I., Aickelin, U. and Rodden, T. (2013). The application of a data mining framework to energy usage profiling in domestic residences using uk data, *arXiv preprint arXiv:1307.1380* .

Derguech, W., Bruke, E. and Curry, E. (2014). An autonomic approach to real-time predictive analytics using open data and internet of things, *Ubiquitous Intelligence and Computing, 2014 IEEE 11th Intl Conf on and IEEE 11th Intl Conf on and Autonomic and Trusted Computing, and IEEE 14th Intl Conf on Scalable Computing and Communications and Its Associated Workshops (UTC-ATC-ScalCom)*, IEEE, pp. 204–211.

Gubbi, J., Buyya, R., Marusic, S. and Palaniswami, M. (2013). Internet of things (iot): A vision, architectural elements, and future directions, *Future Generation Computer Systems* **29**(7): 1645–1660.

Ijumba, P. and Sebitosi, A. B. (2010). Evaluating the impact of consumer behaviour on the performance of domestic solar water heating systems in south africa, *Journal of Energy in Southern Africa* **21**(1): 25–34.

Inmon, W. H. and Hackathorn, R. D. (1994). *Using the data warehouse*, Wiley-QED Publishing.

Ismail, M. J., Ibrahim, R. and Ismail, I. (2011). Adaptive neural network prediction model for energy consumption, *Computer Research and Development (ICCRD), 2011 3rd International Conference on*, Vol. 4, IEEE, pp. 109–113.

Kamdi, M. A. A. and Bhoyar, M. P. V. (2014). Autonomic computing.

Karimi, K. and Atkinson, G. (2013). What the internet of things (iot) needs to become a reality, *White Paper, FreeScale and ARM* pp. 1–16.

Kimball, R. and Ross, M. (2011). *The data warehouse toolkit: the complete guide to dimensional modeling*, John Wiley & Sons.

Liu, X. and Nielsen, P. S. (2015). Streamlining smart meter data analytics, *10th Conference on Sustainable Development of Energy, Water and Environment Systems.*

Oldewurtel, F., Gyalistras, D., Gwerder, M., Jones, C., Parisio, A., Stauch, V., Lehmann, B. and Morari, M. (2010). Increasing energy efficiency in building climate control using weather forecasts and model predictive control, *Clima-RHEVA World Congress*, number EPFL-CONF-169735.

Polyzotis, N., Skiadopoulos, S., Vassiliadis, P., Simitsis, A. and Frantzell, N.-E. (2007). Supporting streaming updates in an active data warehouse, *2007 IEEE 23rd International Conference on Data Engineering*, IEEE, pp. 476–485.

Zavala, V. M., Constantinescu, E. M. and Anitescu, M. (2010). Economic impacts of advanced weather forecasting on energy system operations, *Innovative Smart Grid Technologies (ISGT), 2010*, IEEE, pp. 1–7.

Zavala, V. M., Constantinescu, E. M., Krause, T. and Anitescu, M. (2009). On-line economic optimization of energy systems using weather forecast information, *Journal of Process Control* **19**(10): 1725–1736.