National College of Ireland

Predictive Modelling of Home Appliances Energy Consumption in Belgium

MSc Research Project
Data Analytics

Anant Chaudhary
x17131511

School of Computing
National College of Ireland

Supervisor: Dr. Catherine Mulwa

National College of Ireland
Project Submission Sheet – 2017/2018
School of Computing

Student Name:	Anant Chaudhary
Student ID:	x17131511
Programme:	Data Analytics
Year:	2018
Module:	MSc Research Project
Lecturer:	Dr. Catherine Mulwa
Submission Due Date:	13th September 2018
Project Title:	Predictive Modelling of Home Appliances Energy Consumption in Belgium
Word Count:	8054

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:	13th September 2018

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):	

Predictive Modelling of Home Appliances Energy Consumption in Belgium

Anant Chaudhary
x17131511
MSc Research Project in Data Analytics

13th September 2018

Abstract

The increasing trend in energy consumption is becoming cause of concern for the entire world, as the energy consumption is increasing year after year so is the carbon and greenhouse gas emission, the majority portion of the electricity generated is consumed by industrial sector but a considerable amount is also consumed by residential sector. It is important to study the energy consuming behaviour in the residential sector and predict the energy consumption by home appliances as it consume maximum amount of energy in the residence. The European energy goal of 2020 is to reduce the energy consumption and carbon emission by 20%. This project focuses on predicting the energy consumption of home appliances based on humidity and temperature. It has resulted in implementation of five prediction regression models, i.e. multiple regression, lasso regression, ridge regression, SVM regression and Random Forest are developed and results are presented based on RMSE, MAE and MAPE, the dataset for the analysis was taken from a house located in Stambruges (Belgium), keeping European goal of 2020 in mind. In addition the results of reviewed literature of home energy consumption in Europe is also presented.

Keywords: *Energy Consumption, Home Appliances, Multiple Regression, Ridge Regression, Lasso Regression, SVM Regression, Random Forest, Europe*

1 Introduction

In today's world where the technology is increasing at an alarming rate, making human life easier and comfortable and so is the global warming, the major reason behind the increasing rate of global warming is the C02 and greenhouse gas emission which is generally emitted by burning fossil fuels in order to produce electricity, other energy sources such as solar, wind and hydropower is also used to generate electricity but till now most of the countries are still dependent on thermal power, the major portion of this electricity generated is used by industrial sector which have strict laws for carbon emission in the atmosphere but a considerable amount of electricity is also used by residential sector, so it important to study the energy consumption behaviour in this sector which can help to

conserve energy and minimise wastage of it, which in terms results in reducing emission of CO2 and greenhouse gases. One such initiative taken by government of most of countries in Europe to meet the energy goals of 2020 (reduction of total energy consumption, CO2 and greenhouse gas emission by 20% and increase the use of renewable energy sources by 20%) is installation of Smart Metering System (SMS) in residential sector which is capable of monitoring individual appliance energy consumption that are used in the house (Rodriguez-Diaz et al.; 2016). The appliances include HVAC's (Heating, Ventilation and Air Conditioning System), television, washing machine, refrigerator etc. All these appliances are connected to the smart meter via communication network which allows user to monitor and manage energy consumption based on their expected comfort, energy price variation and equivalent CO2 emission (Basu et al.; 2013). The smart meter data can be used for variety of application such as predicting future energy consumption of the household, to create DR (Demand- Response) system in which consumers can schedule the task based on electricity price at a given time, to create load desegregation system etc. All this can be attained by simply extracting the data from smart meters and train them using machine learning algorithm according to the requirement.

The evolution in the study of energy consumption is based on mechanism which are, Smart Metering System (SMS): It is a metering system in which two-way communication between consumer and electricity provider is established, it is installed in consumer's household and is connected to electrical appliances via communication network, the energy consumption of these appliances is then recorded by this smart meter and the total energy consumption data is sent to service provider for regulating energy requirements of the consumer (Niyato et al.; 2011). Home Energy Management System (HEMS): HEMS consist of smart meter, sensors and application connected to control center which is responsible for monitoring and optimizing the power consumption by controlling the home appliances and lighting which in terms saves electricity and hence minimizes the electricity bill of the consumer. It includes variety of feature in it such as scheduling the HVACs (home heating, ventilation and air- conditioning), turning on/off the appliance with the help of smart-phone etc. There are number of HEMS services in the market such as Microsoft Holm, Apple Smart-Home Energy Management and Google Power meter (Niyato et al.; 2011).

There are two major factors on which the total energy consumption of the household depends upon, the number of appliances and the occupants using the appliance, the electrical appliances used by the occupants in a household leave accountable signals such as temperature, humidity, vibrations etc. in the nearby environment in which the appliance is being used. This project focuses on predicting the total energy consumption of home appliances based on indoor and outdoor environmental conditions using machine learning regression algorithms namely Multiple Regression, Lasso Regression, Ridge Regression, Support Vector Machine and Random Forest, RMSE (Root Mean Squared Error), MAE (Mean Absolute Error) and MAPE (Mean Absolute Percentage Error) evaluation parameters are used to select the best model.

1.1 Motivation and Background

The European policies and target of 2020 regarding reduction of energy consumption and increase the use of renewable energy resources has motivated many researchers to carry out their study in this field, some have predicted the total energy consumption based on

smart meter data while others have use smart meter data for various other application, as energy consumption is a vast domain still there is lot of scope of new findings in this area. The government of European countries such as Spain, Italy, France etc. have already started taking the initiative to meet the goal of 2020 by installing smart meters in the households and creating awareness among the people to reduce wastage of it, but still they are lagging behind as the consumers are resistive towards the change, so there is a need to study energy consuming behaviour of the consumers and predict the energy consumption based on their usage pattern, energy consumption in household is basically because of two components, lighting and electrical appliances which include HVAC's, television, refrigerator, microwave etc. and all appliance leave traceable signals such as temperature, humidity vibration etc., so the aim of this project is predicting total energy consumption of home appliances based on humidity and temperature of indoor, that is where the appliances are used and the outdoor temperature. This will help to give prior information to consumer about the energy consumption by these appliance so that they can reduce the wastage of energy. The data is taken from a house located in Stambruges (Belgium), keeping the European goal of 2020 in mind, the model can be generalized for other European countries as well.

1.2 Research Question

"Can prediction(RMSE, MAE and MAPE) of energy consumption for home appliances based on humidity and temperature using regression techniques and models(Lasso, Ridge, Multiple Regression, SVM, Random Forest) support consumers in Belgium?"

To address the research questions, the following objectives have been specified and implemented

1.3 Research Objectives

Objective1- A review of literature on house energy consumption from 2007-2018.

Objective2- Implementation, evaluation and results of prediction models for energy consumption in Belgium using regression modelling techniques.

Sub-Objective2(a): Implementation, evaluation and results of multiple regression model.

Sub-Objective2(b): Implementation, evaluation and results of lasso regression model.

Sub-Objective2(c): Implementation, evaluation and results of ridge regression model.

Sub-Objective2(d): Implementation, evaluation and results of SVM regression model.

Sub-Objective2(e): Implementation, evaluation and results of random forest regression model.

2 Literature Review on Home Energy Consumption (2008-2018)

2.1 Introduction

This section presents a review of energy consumption for residential sector in Europe, which will give an idea of energy consuming behaviour of users in Europe and helps to

study their usage pattern as it can vary from places to places and Belgium being in Europe will have similar energy consuming behaviour compared to other European countries . A review of data mining prediction techniques used during predictions of energy consumption is also presented which covers the algorithms used by other researchers for energy consumption prediction.

2.2 A Review of Home Energy Consumption in Europe

Europe has set new goals to reduce carbon emission, greenhouse gases and increase the use of renewable energy resources and cut down energy production by fossil fuels by 2020 whereas powerful countries such as United States and China heavily rely on fossil fuels, the European Union has set a policy which forces countries in the European Union to reduce 20% emission of greenhouse gases, increase 20% renewable energy consumption and reduction of 20% in the total energy consumption by 2020. Most of the EU countries have submitted their action plan which basically includes installation of photovoltaics and wind power plants but to optimize energy consumption in residential sector it is important to provide accurate information of the energy consumed by the devices in household, for that smart meters are installed which is connected to devices with internet and monitors the energy consumption, by this consumers will be aware of the energy consumption and the electricity board can have the information regarding energy demand of the consumer (Rodriguez-Diaz et al.; 2016). Another author Miezis et al. (2016) studied the behaviour of house owners in a multi-family apartment building in Eastern Europe country Latvia and observed that only 1% of country's population is concerned about renovating the houses due to excessive energy consumption by old space heating equipment, this is mainly because house owners in a multi-family apartment building don't interact with each other and the tenants living in the building are afraid that house owners will increase the rent when asked to renovate the house, out of 38,000 houses in Latvia, only 711 buildings were renovated which results in 1.9% according to the data of Ministry of Economics in Latvia, so there stands a lot of scope to minimize energy consumption by renovating these old buildings. Bardazzi and Pazienza (2017) discusses the increase in energy consumption of household due to population aging in Italy, the author believes that the energy dependency in Italy is almost complete and with the increase in population aging and life expectancy, the number of old age people staying at home is also increasing which might be a problem as they are totally dependent on thermal comfort (heating and air conditioning) than to environmental temperature and the younger population boost the energy consumption by using new gadgets that are introduced. Bardazzi and Pazienza also mentioned that the energy efficiency in Italy improved by 10% from 2000-2012 whereas European average was 19% , so it is important to study the need of energy consumption for different age groups in Italy in order to increase energy efficiency in Italy. A similar theory was proposed by Balta-Ozkan et al. (2014) in which the need for smart grid was discussed in Europe, unlike Bardazzi and Pazienza whose focus was only on population aging factor in Italy, the author here discusses the socio- economic and demographic factors preventing the countries to build more smart grids, the author also discusses the change in the policies done by the government of Germany and Italy to reduce energy consumption, in Germany the reliability of smart homes is the main concern of the people whereas in Italy the lack of technology to build smart homes is the key issue and the common concern in both the countries among people are the aesthetic values and sentiments towards the buildings, so according

to the author, first step should be to make people aware about the benefits to convert their heritage homes to smart homes. The energy consumption problem in Italy was also showcased by Felicetti et al. (2015) in which the author believes that energy consumption in residential sector will increase by 40% in Italy and there is a need to optimize this by introducing CSE (Collaborative Smart Environment) and HEMS (Home Energy Management System) which will reduce the wastage of electricity and therefore reduce greenhouse gas emission. Gonzalez-Lezcano and Hormigos-Jimenez (2016) discusses a solution for Southern Europe specifically Madrid where the temperature is too hot in summers and too cold in winters, so the inhabitants are very much dependent on heating devices and air-conditioning systems which results in the increase of energy consumption to a greater extent, the author believes to design natural ventilation in such a way that the dependency on HVAC's (Heating, Ventilation and Air- Conditioning Systems) can be minimized which consumes the maximum energy in the household and the indoor temperature can be regulated using the natural ventilation, it can be achieved by studying the direction and wind speed in the area, the holes in the facade and the structural model of the entire house. A similar study was taken by Sanchez-Guevara et al. (2017) in which the author focuses to establish a minimal energy requirements to achieve minimal habitability conditions in Spain as the summer heat wave of 2003 killed 35,000 people across Europe, Gonzalez-Lezcano and Hormigos-Jimenez paper mainly focused to reduce energy consumption by building proper natural ventilation whereas Sanchez-Guevara et al. paper aims to establish minimal energy requirements to low-income household in order to survive extreme weather condition so this paper seems to be more reliable for the survival of the people as some countries in Europe have extreme weather conditions and depending on Natural Ventilation might not be the solution, the author has also termed the people who can't meet the energy requirement to sustain as "Fuel Poverty" and European countries such as Ireland and France are already taking actions to provide electricity to the Fuel poverty class. Another Author Casals et al. (2017) provides a gamification technique EnerGAware (Energy game for Awareness of Energy Efficiency) as a solution to make people aware about the wastage of electricity in residential sector, the author believes that 40% of the totally energy consumption in Europe is in building sector and it has a lot of potential to reduce this consumption by 1509 million tonnes of oil by 2050. It achieves it's motive of reducing energy consumption by developing an energy game called EnergyCat for android users in which a cat observes the behaviour of people living the house and the energy utility devices can be choose according to the person playing the games, this shows the behaviour of the person towards saving energy, this technique of the author resulted a good way to create awareness among the people. Caffarel et al. (2013) discusses the benefits of the European project of smart city "MALAGA" , each house were installed with smart metering system, monitoring devices and android application for reporting the energy consumption, it resulted in the reduction by 10% of energy consumption in the houses for 42% of the people living in the smart city, 33% of the people kept their previous consumption level and 25% of the people increased their energy consumption, it resulted a good initiative by government of Spain to increase use of renewable energy resources and reduces the emission of CO2. A common problem in all the European countries is addressed by Rovsing et al. (2011) which is the kind of technology at affordable price and the ease of installation to reduce the energy consumption of a household and the consumer's tendency towards the change, only a fraction of people which includes technological freaks and wealthy people who can afford to pay for the installation go forward with the home automation, the government needs to come

with a solution which should be reliable as well as cost efficient. Aune (2007) discusses energy consumption problem in Norway during the year 2002/2003 where Norway faced heavy energy crises due to the shortage of rainfall during that year and the country is mostly dependent on Hydroelectricity, the energy prices were raised by 43% and people suffered by using minimum heating devices at home or using less number of rooms, still the average reduction in 2003 of total energy consumption was just 2.3%, this is due to the structure of Norwegian home which is not designed to meet the minimum energy requirements, the author also states two factors which can result in reducing the energy consumption of the household are to reduce wasteful behaviour and to make people aware about buying energy saving equipment.

2.3 A review of Machine Learning Techniques used for Home Energy Consumption

A model was proposed by Basu et al. (2013) to predict the energy consumption of home appliance for the next hour, the IRISE dataset used for the analysis was taken from REMODECE (Residential Monitoring to Decrease Energy Use and Carbon Emission) database which is used to store energy consumption and carbon emission of European countries. The working mechanism is built on ORACLE which is a knowledge system which takes past history of appliance energy consumption, hour of the day, day of the week and season of the year, Oracle's output is then fed to a classification algorithm which classifies the status of the home appliance as ON/OFF for the next hour. Bayesian Network, Decision Tree (C4.5) and Decision table classification algorithms were used to train the models, the model was tested for all the three cases which are by including all the knowledge information, by only including the past consumption and not including any knowledge on all the three classifiers. Decision Tree (C4.5) resulted in the highest accuracy when all the knowledge information was inserted. The use of EDHMM-di (Explicit-Duration Hidden Markov Model with differential observations) was proposed by Guo et al. (2015) to detect and estimate individual home appliance loads from aggregated power data which is collected using smart meters. Candanedo et al. (2017) proposed a model to predict total energy consumption of home appliances for the next 24hrs based on past energy consumption data, weather data from nearby airport and energy use of lighting, unlike Basu et al. whose focus was predicting the on/off state of the appliance, this paper focused on predicting overall consumption which is more effective in terms of saving electricity, data filtering and feature selection was used to remove non-predictive parameters, four regression models namely Multiple Regression, Random Forest, Support Vector Machine with Radial Kernel and Gradient Boosting Machine (GBM) were used to train the models, all models were trained by 10 fold cross validation to select the best. The GBM model overfitted the training set by giving 97% accuracy but giving only 57% accuracy in the test dataset. Another researcher Zeng et al. (2016) used optimum regression method to study the weather influence on energy consumption, unlike Candanedo et al. who used weather data from the nearby airport for the prediction, this author used actual weather data of the residential location further increasing the accuracy, machine learning algorithms such as SVM (Support Vector Machine) and ANN (Artificial Neural Network) were used to train the models.

A wide area of research was also conducted to optimize DR (Demand Response) mechanism and create HEMS (Home Energy Management System) using machine learning algorithms, one such researcher was Zhang et al. (2016) who proposed a model to

create a learning based DR system which takes user preferences and price of energy at a particular time into account and schedule task of the user accordingly, the motive is to minimize the energy bill of the user keeping consumer comfort in mind, Neural Network and Regression Based Learning were used to train the model, the model resulted to be successful by correctly scheduling user's task. Another model in which user preferences were used to minimize energy bill was proposed by Jin et al. (2017), the base of the model was MPC (Model Predictive Controller) that is used for generating control decision for individual devices based on user preferences, utility interfaces, weather services, system identification, statistical learning, individual devices and submeter sensors, Gaussian mixture method were used to learn the usage pattern of the customer, the model proposed by Zhang et al. used to cut-off HVAC's energy when the cost and demand of energy is high whereas Jin et al. model proposed the used of PV (Photo-voltaic) generation and home battery system, this model resulted in more efficiency than the previous model. Veras et al. (2018) proposed a model which used load shifting method, that is the demand during the peak period is shifted to another time of lower consumption maintaining the balance of daily energy consumption of the household, the data was taken from 10 families located in Brazil containing 29 appliances in each household, genetic algorithm was used to train the model and the proposed method resulted in saving the energy bill to a great extent, the use ANN (Artificial Neural Network) to separate controllable and non- controllable loads was proposed by Ponocko and Milanovic (2018), the aggregated data in smart meters were decomposed to controllable and non-controllable loads and then the energy consumption of both the type of loads were forecasted and compared to the total energy consumption, the model resulted to be useful for electric providers to include only those loads in DR which have high energy consumption, to optimize DR and design HEMS in order to save energy and reduce electricity bill was carried out by many researchers and turns out be a great progress in the study of energy consumption.

Lot of other researchers also contributed towards energy consumption reduction, one such researcher was Huebner et al. (2016) who studied the factors responsible for the increasing energy consumption due to socio-demographics, appliance ownership and use, attitudes and self-reported behaviour, it then shows that which variables have high explanatory power, the data from 845 homes were taken for the analysis, 4 regression models were trained taking different variables, appliance ownership and use resulted in the highest variability with 34% , 21% resulted for the size of household and a combined model taking all the predictors resulted in explaining 39% of the variability, Lasso regression algorithm was used to train all the models.Du et al. (2017) studied the energy rebound effect in construction industry in China by using Ridge regression machine learning algorithm, China being the highest emitter of CO2 and greenhouse gases has made a great effort to save energy, the paper is focused on logical relationship among capital input, technical change, economic growth and energy consumption, the results obtained from the ridge regression shows that the development of industrial sector is China is mostly depend upon labour input, the only way to reduce CO2 emission and energy consumption is by increasing investment on technical inputs which will bring new machinery and technology that will consume less energy resulting in low energy consumption.

2.4 Conclusion

The above section presents an overview of literature that were referenced during this research, the section is divided into two parts, the first part goes through the behaviour

of people and countries in Europe towards energy consumption, it mentions the initiative taken by European Union to reduce carbon emission and Energy Consumption by 2020, it also reflects different countries in Europe such as Spain, France, Latvia etc. and there problems and measures towards achieving the goal. The second section focus on the use of machine learning techniques in Energy consumption domain, some researchers have used past energy consumption data to forecast the future energy consumption, some have proved the influence of outdoor weather on energy consumption but including indoor as well as outdoor conditions has not been focused by many researchers, this project intends to do so and find out the results.

3 Scientific Methodology Approach Used, Design and Data Preparation

3.1 Introduction

This section explains the methodology followed to execute the project including the process flow diagram, the cleaning, preparing and transformation techniques applied in order to model machine learning algorithms followed by conclusion at the end of the section.

3.2 Scientific Methodology Approach Used

Any research in the field of data analytics is carried out by using one of three methodologies, KDD (Knowledge Discovery Data) (Fayyad; 1996), CRISP-DM and SEMMA, the difference between the three models was studied by Azevedo and Santos (2008), KDD and SEMMA is similar to one another, both the methodologies contain 5 stages but have different name of the stages and CRISP- DM methodology involves understanding the business perspective of the analysis before proceeding towards the other stages and the deployment of application at the end of analysis, this project has been developed using KDD methodology.Figure 1 represents the scientific methodology approach used during the implementation of prediction models.

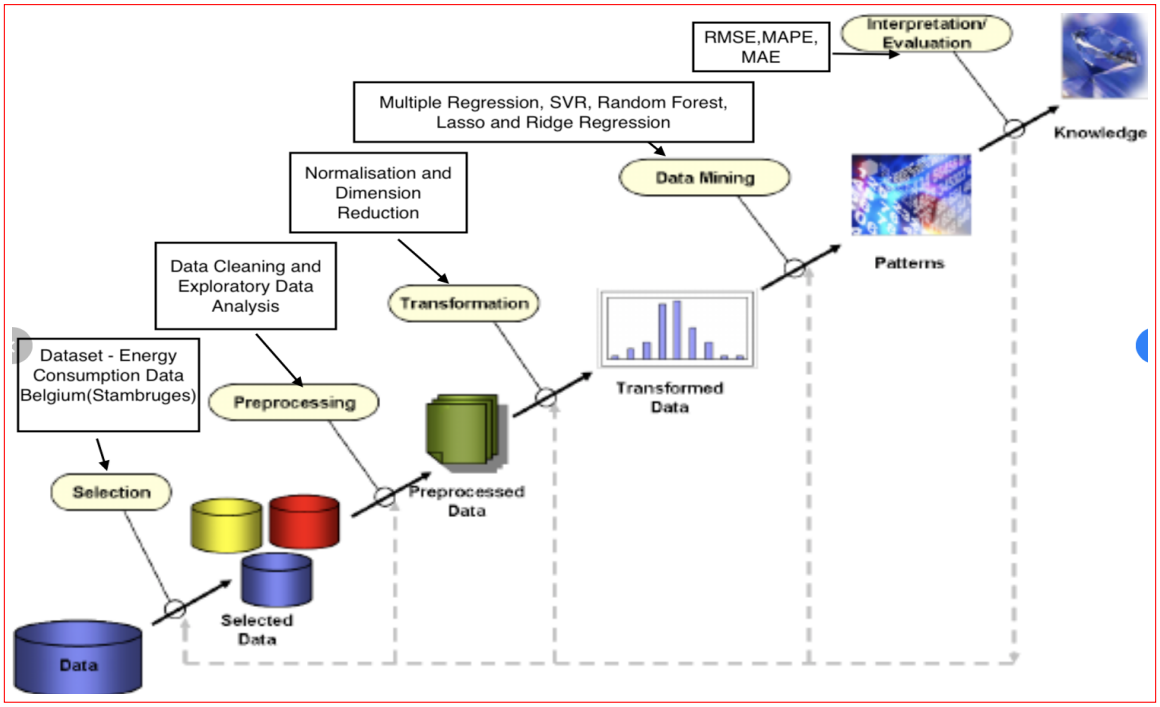

Figure 1: Scientific methodology approach used

The phases involved during the implementation of prediction models for energy consumption of home appliance :

Stage 1 Selection : During this stage, the dataset used for this research for extracted from UCI machine learning website.The background of the data used for this research is explained in the later part of this section.

Stage 2 Pre-Processing : In this stage, during this project, data cleaning and exploratory data analysis were performed, this stage is one of the most important part of the process as it involves cleaning data such as removing null values, unwanted data and also to explore data to gain insights from it in order to proceed towards the next stage of the analysis. The steps taken in this stage to make data consistent is explained further in the later part of this section.

Stage 3 Transformation : During the research, in this stage, data is transformed by various methods such as dimension reduction, normalization, feature extraction etc.

Stage 4 Data Mining : In this stage various machine learning models are applied to find patterns in the data for various application such as predictions , forecasting, classification etc. This project uses regression machine learning algorithms namely Multiple Regression, SVR, Random Forest, Ridge Regression and Lasso Regression which are explained in the next section of this report.

Stage 5 Interpretation/Evaluation : This stage mainly deals with finding the results of the model and defining parameters based on which different models can be compared, this project uses RMSE (Root Mean Squared Error), MAE (Mean Absolute Error) and MAPE (Mean Absolute Percentage Error) for the evaluation.

3.3 Project Process Flow Diagram

The below Figure 2 of process flow diagram explains the overall steps of the research, it consist of three layers, data layer is a layer in which all the processes will be done in

order to make data ready for building models, the first stage of this layer is selection in which, data from the database will be extracted and directly fed in R, the second stage of this layer consist of pre-processing techniques which will be applied to data in order to making data consistent, this stage comprises of data cleaning and exploratory data analysis, the third stage of this layer consist of data transformation techniques which will applied to data according to the model being built, the next layer is business logic layer which comprises of two stages, data mining stage in which actual models will be built using machine learning algorithms and the evaluation stage in which results of the model built will be discussed. The final layer of the process flow is client side in which the results will be presented in the form of graphs and visuals which is understandable by the business user.

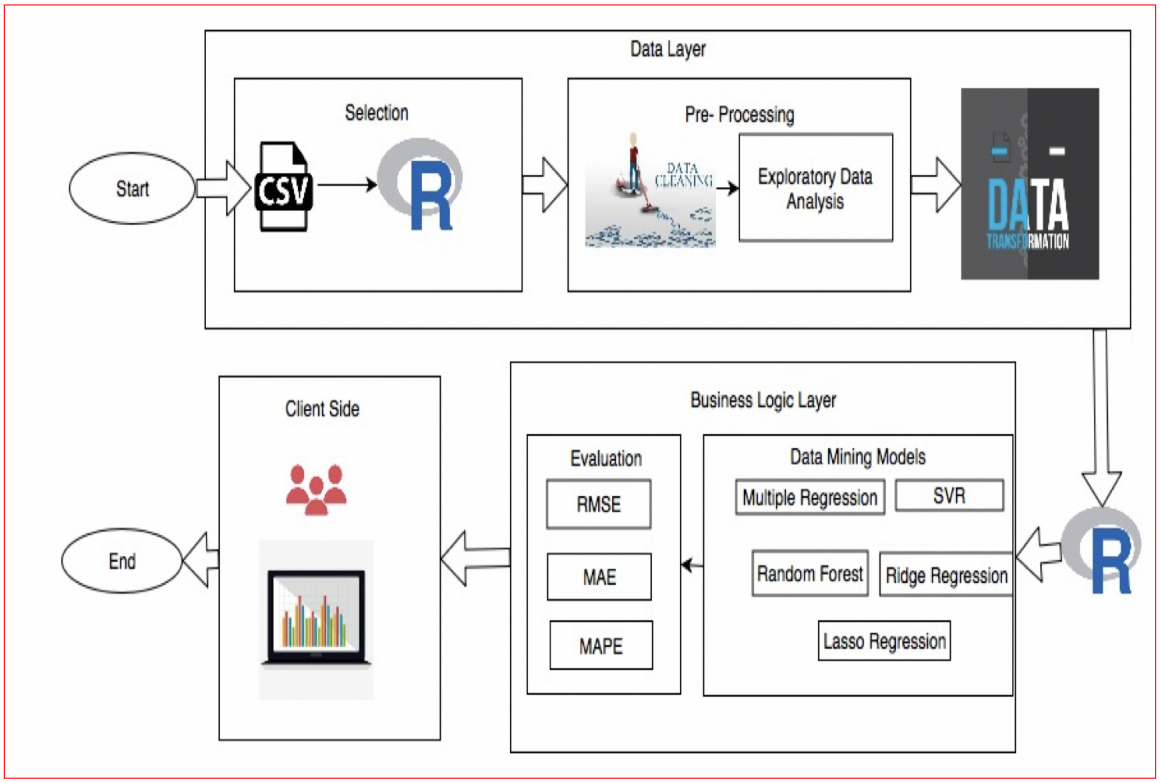

Figure 2: Project process flow diagram

3.4 Data Preparation

3.4.1 Background of Dataset

The dataset used for the research is taken from UCI machine learning database[1], the dataset is extracted from a house located in Stambruges (Belgium) by installing smart meter and ZigBee smart meter sensors which is capable of detecting energy consumption, humidity and temperature of each room where the appliances are used, the house contains three rooms, two bathrooms, living area, dining room, office, garage, kitchen, laundry, ironing room and game room, each of this room is equipped with electrical appliances and sensors to detect the energy consumption, humidity and temperature at a given time, the

[1]Data Source- EnergyConsumption Dataset website: `https://archive.ics.uci.edu/ml/datasets.html`

data from this sensor is then sent via internet to energy monitoring system where it is stored and reported, so the dataset contains 19735 rows and 29 attributes, the humidity and temperature of each area along with the total energy consumption of the house is present in the dataset, except for that, the outdoor climatic condition such as humidity, temperature, windspeed etc. is also present. This dataset has been chosen for the research keeping the European goal of 2020 in mind as the data belongs to a house situated in a European country so similarly the Energy behaviour of other European countries can also be studied.

3.4.2 Data Pre-processing

This phase of data preparation consist of two processes :

Data Cleaning :

The data taken from UCI machine learning website, fortunately it does not contain any missing values but still just to be assured, missing values were checked by using the code as shown in figure 3

```
# checking the missing values in the dataset
sapply(energydata, function(x)sum(length(which(is.na(energydata)))))
```

Figure 3: Code to check missing values

After that row names of the dataset were changed for better understanding of the data, the unwanted attributes which was not required for the analysis were removed from the dataset and the datatype of the remaining attribute were then checked and corrected according to the data, all of the attributes in the dataset is numerical so the datatype of all the variables were changed numerical and a final dataset was prepared for the next step.

Exploratory Data Analysis :

After cleaning the data, one of the most important step in any data analytics research is to gain insights to data by performing various exploratory data techniques such as checking the trend of the data as shown in figure 4, here the first 100 values of energy consumption have been plot with the linear regression line to check the trend, it can be observed from the plot that there is no particular trend and the dots are scattered in all the direction moreover the energy consumption values of some of dots are quite higher than regression line proving that the data is not fit for predicting energy consumption based on past energy consumption using time series analysis and regression algorithms should be used in order to deal with such data.

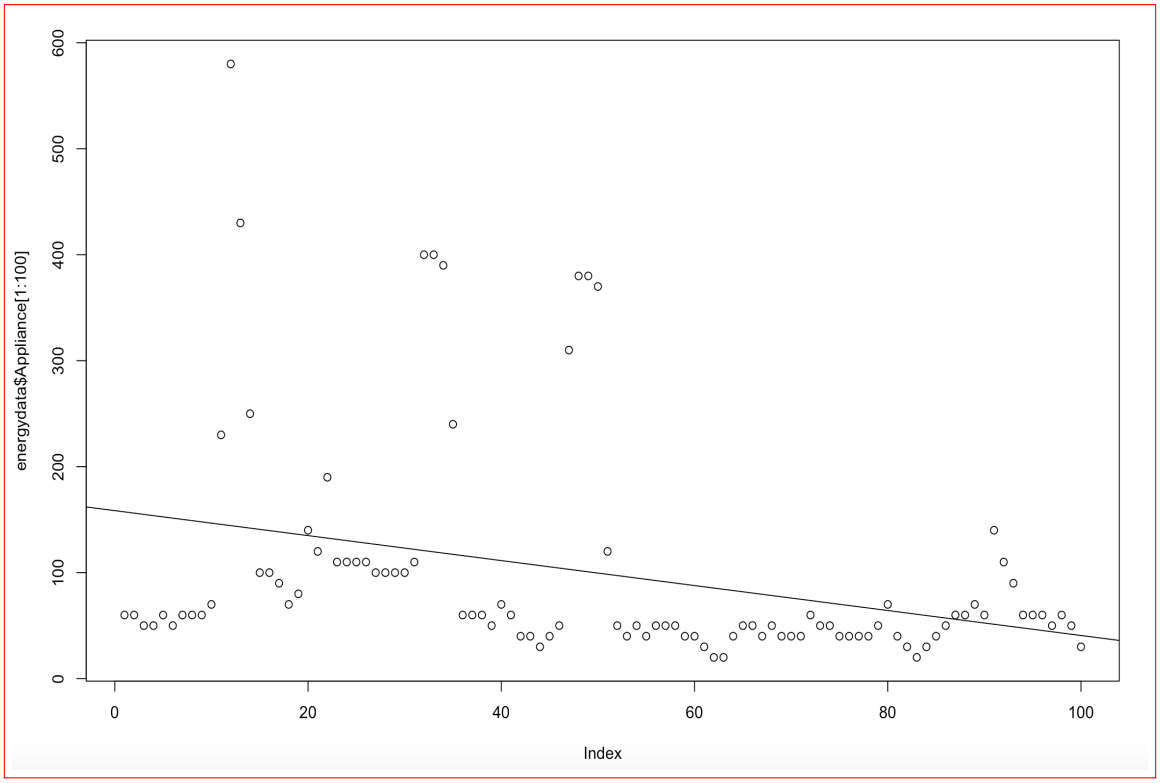

Figure 4: Plot of first 100 rows to check trend in dataset

Correlation plots to check correlation between predictor variables and the dependent variable was also used as shown in below figure 5, here blue dots represent the positive correlation and red dots represent the negative correlation and the shade of colour represents the strength of correlation between the variables. The scale from 1 to -1 is also represented at the right hand size of the plot.

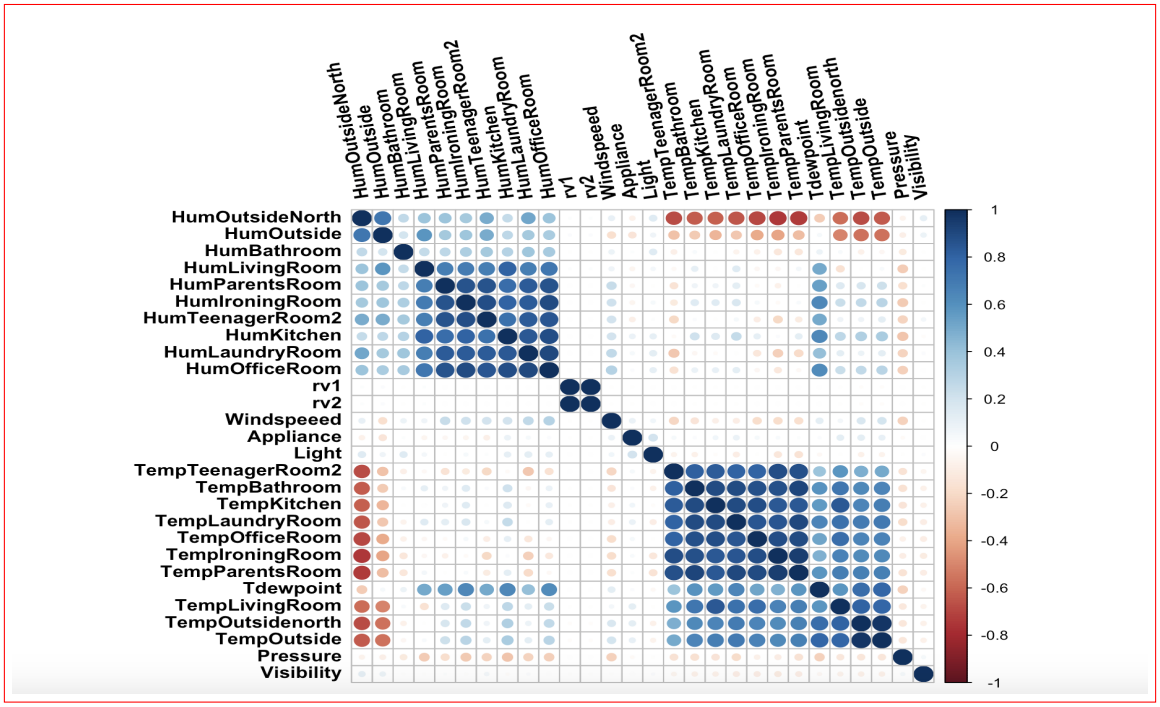

Figure 5: Correlation plot

The variable importance plot generated by applying random forest algorithm is also an important part while exploring the data, the algorithm is implemented at the initial stage of the analysis to study the predicting capacity of the predictor variables as shown in the figure 6, it can be observed from the plot that energy consumption of light is explaining the highest variance while prediction followed by the remaining variables.

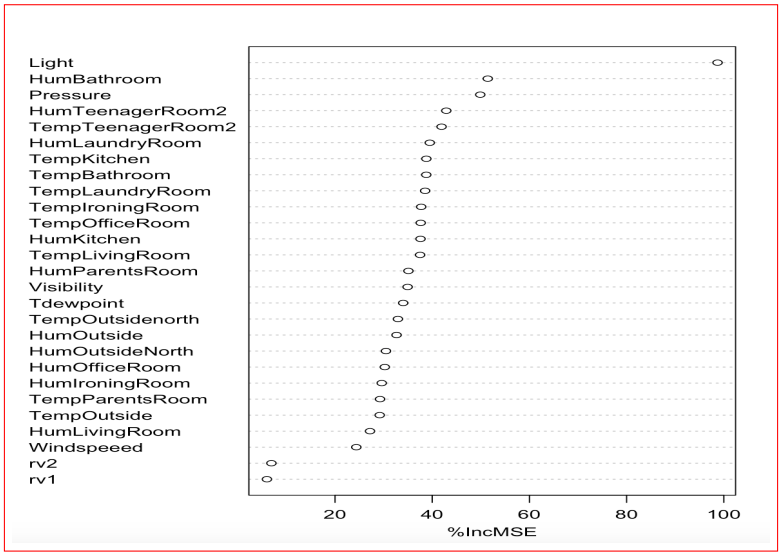

Figure 6: Variable importance plot

According to the information retrieved from exploratory data analysis, various transformation techniques are applied and models and build accordingly.

3.4.3 Data Transformation

This is a stage in which data is transformed from the original dataset to a more suitable dataset which is required in order to build machine learning model, this research is based on regression algorithms so the most important data transformation required is to normalize the data in order to avoid biasing, other than that other transformation methods includes parameter reduction which was done based on variable important plot generated by random forest algorithms as discussed in the previous sub-section, the least significant predictors which was not affecting much variance was removed from the analysis and the model was built again. After data transformation process the final data is ready to build machine learning models which will be individually discussed in the next section of the report.

3.5 Conclusion

The scientific methodology approach discussed in this section and the project process flow diagram design are used during the implementation of prediction models in chapter 4 and the methods used for preparing the data such as removing missing values, exploring the data by plotting correlation and variable importance plot, data transformation techniques namely normalisation and parameter reduction were used in order to build the models.

4 Implementation, Evaluation and Results of Prediction Models for Energy Consumption of Home Appliances in Belgium

4.1 Introduction

This section covers the implementation and evaluation of regression models which were built for analysis based on the data which was prepared in the previous section, before moving forward towards the actual implementation, the prepared data was divided into training and testing dataset in the ratio of 80% and 20% respectively, volunteer sampling method, that is first 15788 rows was used as training set and the remaining were used as test set, this is done because target variable (Energy Consumption) is a time dependent data. After creating the training and testing set for the analysis, regression models namely Multiple Regression, Lasso Regression, Ridge Regression, Random Forest and Support Vector Machine with Radial kernel are implemented which are explained separately in the following subsection and based on below parameters, regression models are evaluated and their result are discussed.

i) RMSE (Root Mean Squared Error) – $\frac{\sqrt{\sum_{i=1}^{n}(p_i - a_i)^2}}{n}$

$a = actual\ target \quad p = predicted\ target$

The RMSE value gives the error rate of regression models and it can be used to compare different regression models with the same unit, as given by the formula it is computed by subtracting the predicted value of energy consumption by the actual value of energy consumption and the difference was squared to avoid negative values, further the mean of all the values was taken and the obtained mean was further square rooted to give RMSE. All these steps are done in R programming language by just installing metrics package.

ii) MAE (Mean Absolute Error) - $\frac{\sum_{i=1}^{n}(|p_i - a_i|)}{n}$

$a = actual\ target \quad p = predicted\ target$

The MAE value is measured in the same unit as original data and gives the absolute error rate, it is similar to RMSE just little smaller, as given by the formula it is computed by subtracting the predicted value of energy consumption by the actual value and taking absolute value of the difference to avoid negative values further obtaining the mean of all values. This evaluation parameter is also included in metrics package of R.

iii) MAPE (Mean Absolute Percentage Error) - $\frac{1}{n}\sum\frac{|a_i - p_i|}{|a_i|} * 100$

$a = actual\ target \quad p = predicted\ target$

The MAPE gives the percentage error of the absolute difference of actual energy consumption and predicted energy consumption divided by the mean. It helps is evaluating the results in terms of percentage. This evaluation parameter is also present in metrics package of R.

4.2 Implementation, Evaluation and Results of Multiple Regression model

4.2.1 Implementation

The multiple linear regression model is used for predicting dependent variable (energy consumption) based on predictor variables (indoor and outdoor environmental conditions) and is given by$formula Y = b_0 + b_1X_1 + b_2X_2.....b_pX_p$

where Y is the dependent or target variable, 1 is p distinct predictor variables, 0 is value of Y when all the independent variable is 0 and 1 are estimated regression coefficient, for this research all the implementation steps has been carried out in R programming language, multiple linear regression was implemented initially by taking all the parameters as show in figure 7

```
#Implementation of Multiple regression  taking all parameters
multipleReg <- train(Appliance~Light
                     +TempKitchen
                     +HumKitchen
                     +TempLivingRoom
                     +HumLivingRoom
                     +TempBathroom
                     +TempOfficeRoom
                     +HumOfficeRoom
                     +HumBathroom
                     +TempOutsidenorth
                     +HumOutsideNorth
                     +TempIroningRoom
                     +HumIroningRoom
                     +TempTeenagerRoom2
                     +HumTeenagerRoom2
                     +TempParentsRoom
                     +HumParentsRoom
                     +TempOutside
                     +Pressure
                     +HumOutside
                     +Windspeeed
                     +Visibility
                     +Tdewpoint,
                     TrainData,
                     method ='lm')
```

Figure 7: Implementation of multiple regression with all parameters

and then the parameters were reduced according to the variable important plot discussed in exploratory data analysis section 3.4.2 generated by random forest algorithm as shown in figure 8

```
# Multiple regression with reduced parameters
multipleRedReg <- train(Appliance~Light
                        +TempKitchen
                        +HumKitchen
                        +TempLivingRoom
                        +TempBathroom
                        +TempOfficeRoom
                        +HumBathroom
                        +TempIroningRoom
                        +TempTeenagerRoom2
                        +HumTeenagerRoom2
                        +Pressure,
                        TrainData,
                        method ='lm')
```

Figure 8: Implementation of multiple regression with reduced parameters

Both the models were then evaluated based on the evaluation parameters as discussed in the next sub section.

4.2.2 Evaluation and Results

Table1: Results of multiple regression

Model	RMSE	MAE	MAPE
Including all parameters	85.72	49.20	53.6%
Reduced parameters	88.60	50.92	56.1%

The RMSE, MAE and MAPE value of both the models including all parameters and model built by reducing the parameters are depicted in the above table 1, from the table we can see that all the three values of the model when included all parameters is lower than that of the model with reduced parameters which shows that the former model is better prediction model than the latter model having low error rate in all the three evaluation parameters, this is true in the real life scenario as energy consumption of appliances is dependent on all the indoor and outdoor environment condition, no parameter can be reduced in order to improve the prediction model. The absolute error of model including all the parameters and model with reduced parameters is 49.20 and 50.92 respectively which shows that model will predict with +/- 49.20 when including all parameters and +/- 50.92 for reduced parameters, it is not a very good model but as the energy consumption has high range of values it can be considered as moderate performing model.

4.3 Implementation, Evaluation and Results of Lasso Regression model

4.3.1 Implementation

Lasso regression is a shrinkage model, it causes regression coefficient of non or less predictive parameters to move towards 0 which help in increasing the overall predictability of

the model and it is also used as a variable selection method for linear regression models, the goal of Lasso regression is to minimize

$$\sum_{i=1}^{n}(y_i - \sum_{j} x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{p} |B_j|$$

λ is a tuning parameter which is used to control penalty, it is basically amount of shrinkage, in order to implement lasso regression in this research, custom control parameter technique has also been applied to further improve the predictive capacity of model and avoid over-fitting. As shown in figure 9, the number of folds is equal to 10 with 5 repeat in custom control parameter.

```
#Custom Control parameter
install.packages("glmnet")
library(glmnet)
library(caret)
library(psych)
custom <- trainControl(method = "repeatedcv",
                       number = 10,
                       repeats = 5,
                       verboseIter = T)
```

Figure 9: Custom control parameter

Lasso regression uses number of tuning parameters to train the model, in the below figure 10 alpha = 1 shows that the implemented algorithm is Lasso regression, if the value of alpha = 0, then the implemented algorithm is Ridge regression.

```
#-----implementation of Lasso regression
set.seed(17654)
LassoReg <- train(Appliance~Light
                  +TempKitchen
                  +HumKitchen
                  +TempLivingRoom
                  +HumLivingRoom
                  +TempBathroom
                  +TempOfficeRoom
                  +HumOfficeRoom
                  +HumBathroom
                  +TempOutsidenorth
                  +HumOutsideNorth
                  +TempIroningRoom
                  +HumIroningRoom
                  +TempTeenagerRoom2
                  +HumTeenagerRoom2
                  +TempParentsRoom
                  +HumParentsRoom
                  +TempOutside
                  +Pressure
                  +HumOutside
                  +Windspeeed
                  +Visibility
                  +Tdewpoint,
                  TrainData,
                  trControl=custom,
                  method ='glmnet',
                  tuneGrid= expand.grid(alpha = 1,
                                        lambda= seq(0.001,0.1, length=5)))
```

Figure 10: Implementation of lasso regression

The value of lambda was taken in the sequence of 0.001 to 0.1 to get the best shrinkage value that will give lowest RMSE and best predictive model as shown in figure 11, here the best RMSE of the training set is when value of lambda is between 0.01 to 0.025.

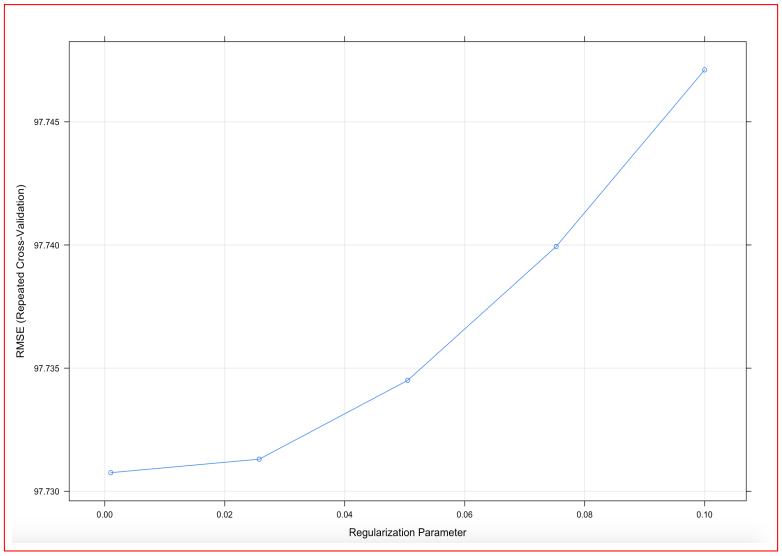

Figure 11: best value of lambda

4.3.2 Evaluation and Results

Table2: Results of lasso regression

Model	RMSE	MAE	MAPE
Lasso Regression	85.93	49.55	54.27%

The RMSE, MAE and MAPE value of lasso regression is depicted in the above table 2 , from the table it can be clearly seen that the MAE and MAPE value of the model is 49.55 and 54.27% which shows that the model will predict with an error rate of predicted value 49.55 and the percentage of error will be 54.27%, which is not a great model but when considering our predictor variable (indoor and outdoor environment conditions) namely humidity and temperature, its bit difficult to predict energy consumption based on that. Lasso Regression also gives variable important plot as an output as shown in the below figure 12, it can observed that predictor variable which have high variability towards the model have been given high importance according to the lasso regression model and variables which have low variance have been move towards zero.

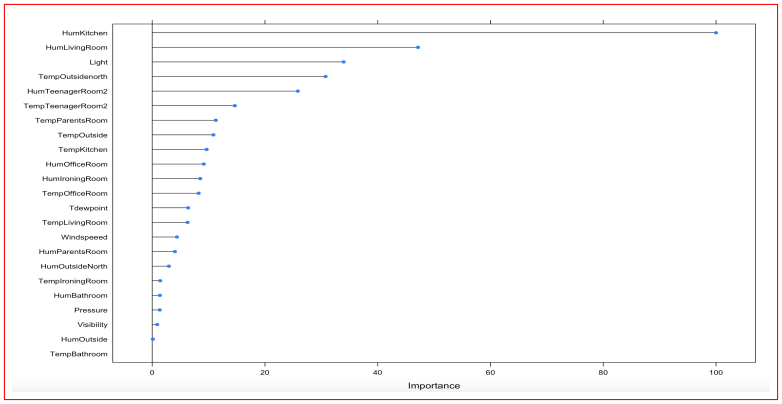

Figure 12: Variable importance plot

4.4 Implementation, Evaluation and Results of Ridge Regression model

4.4.1 Implementation

Ridge Regression is a lot similar to lasso regression, unlike lasso regression where shrinkage method was used, ridge regression uses an estimator to penalize the predictor variables that are less significant to build the model, the goal of ridge regression is same as lasso regression which is to minimize :

$$\sum_{i=1}^{n}(y_i - \sum_{j} x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{p} |B_j|$$

but instead of βj Ridge regression uses $\beta ridge = (X'X + \lambda lp) - 1X'Y$ putting further constraint on the $\beta j's$. All this computation is done by R, keeping things simple and easy to understand, for this model also custom controlled parameter has been used as discussed in the previous subsection. Ridge regression works on the same principle as lasso regression and uses some tuning parameters which is shown in the below figure 13, here trControl = custom shows that the model is using custom parameters to train, alpha = 0 means that the trained model is ridge regression, if the value of alpha = 1 , then the model is Lasso regression.

```
# Implementation of Ridge Regressio#
set.seed(18765)
ridgeReg <- train(Appliance~Light
                  +TempKitchen
                  +HumKitchen
                  +TempLivingRoom
                  +HumLivingRoom
                  +TempBathroom
                  +TempOfficeRoom
                  +HumOfficeRoom
                  +HumBathroom
                  +TempOutsidenorth
                  +HumOutsideNorth
                  +TempIroningRoom
                  +HumIroningRoom
                  +TempTeenagerRoom2
                  +HumTeenagerRoom2
                  +TempParentsRoom
                  +HumParentsRoom
                  +TempOutside
                  +Pressure
                  +HumOutside
                  +Windspeeed
                  +Visibility
                  +Tdewpoint,
                  TrainData, method ='glmnet',
                  trControl=custom,
                  tuneGrid= expand.grid(alpha = 0,
                                        lambda= seq(0.001,0.1, length=10)))
```

Figure 13: Implementation of ridge regression

The value of lambda was taken between 0.001 to 0.1 to get the best estimated value that will give low RMSE value and have best predictive capacity as shown in figure 14, here the value of RMSE is constant when value of lambda is between 0.01 to 0.1.

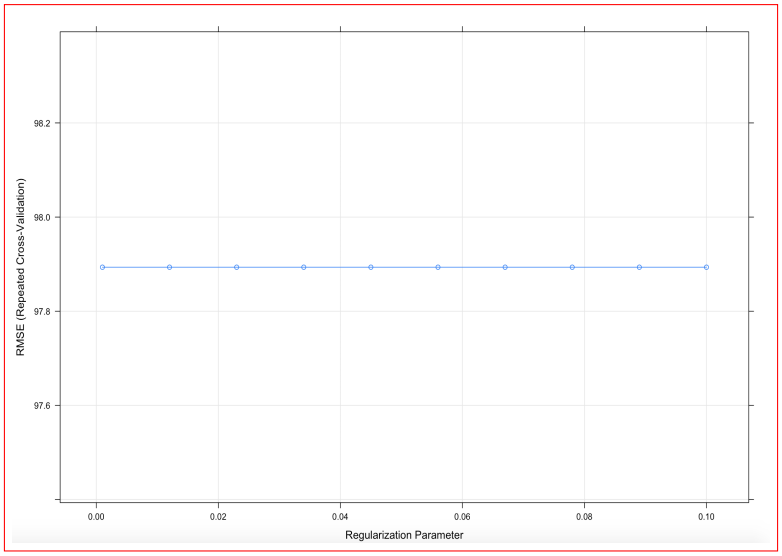

Figure 14: Best value of lambda

4.4.2 Evaluation and Results

Table3: Results of ridge regression

Model	RMSE	MAE	MAPE
Ridge Regression	86.46	49.26	53.6%

The RMSE, MAE and MAPE value of Ridge regression is depicted in the above table 3, it can be seen that the MAE and MAPE value of the model is 49.26 and 53.6% which shows that the model will predict with an error rate of predicted value 49.26 and the percentage of error will be 53.6%, which is also not a great model. Ridge regression also gives variable important plot like lasso regression as an output as shown in the below figure 15, it can observed that plot is similar to lasso regression but the sequence of some parameters have been changed according to the importance of parameters as used by the Ridge regression.

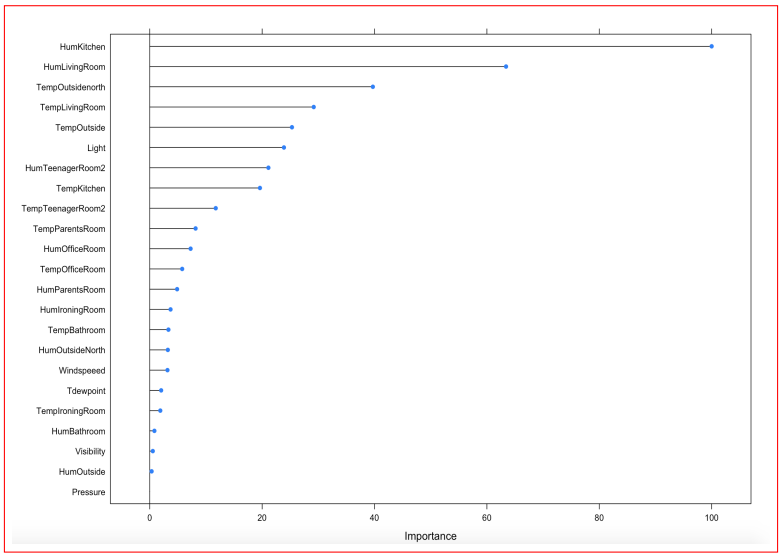

Figure 15: Variable importance plot

4.5 Implementation, Evaluation and Results of SVM Regression model

4.5.1 Implementation

The Support Vector machine is a machine learning algorithm that can be used for classification as well as regression, the basic principle being the same for both the methods, it creates a hyperplane separating different classes, but for regression its bit difficult as the prediction is based on real numbers, the objective of the algorithm is to set margin of tolerance (epsilon), the support vector machine regression for non-linear data is given by $y = \sum_{i=1}^{N}(\alpha_ı - \alpha_i^*). < \phi(x_i), \phi(x) > +b$, which can be reduced to $y = \sum_{i=1}^{N}(\alpha_ı - \alpha_i^*).k(x_i, x) + b$, here k is a kernel which transform the data into higher dimensional space to make the data possible for linear separation, there are many types of kernel that are used to transform the data which entirely depends upon the data used for the analysis, for this research, radial kernel has been used to build the prediction model as in the below figure 16.

```
#Implementation of SVM model
SVMmodel <- train(Appliance ~ .,
                  data=TrainData[-1],
                  method = "svmRadial",
                  verbose=TRUE)
```

Figure 16: Implementation of SVM

SVM regression model is also implemented for two scenarios, initially all the parameters were included to build the model and then the parameters were reduced based on the variable important plot generated by random forest algorithm as discussed in 3.4.2 .

4.5.2 Evaluation and Results

Table4: Results of SVM regression

Model	RMSE	MAE	MAPE
SVM with all parameters	88.93	55.38	64.87%
SVM with reduced parameters	91.44	57.08	66.49%

The RMSE, MAE and MAPE value of both the models with all parameters and model built by reducing the parameters are depicted in the above table 4, it can be observed that all the three values of the model when included all parameters is lower than that of the model with reduced parameters which shows that the SVM model considering all the parameter is better prediction model than the model with reduced parameter, having low error. The absolute error of model including all the parameters and model with reduced parameters is 55.38 and 57.08 respectively showing that model will predict with 55.38 when including all parameters and 57.08 for reduced parameters, the mean percentage error of both the models is very high with 64.87% error when including all the parameters and 66.49% when reducing the parameters thus making the model not fit for predicting energy consumption.

4.6 Implementation, Evaluation and Results of Random Forest model

4.6.1 Implementation

Random forest is an algorithm which is very flexible and easy to use, that can used for classification and well as regression, it gives promising result even without any hyper tuning parameter, it works on the principle of decision trees by creating the number of specified branches, the number of branches is generally specified while implementing the model otherwise the branch creation will never stop resulting in overfitting of data, this algorithm is also used in the initial stage of the analysis to explore the important variables having higher predicting capacity, the random forest modelled for this research is displayed in below figure 17 , to implement random forest algorithm in R programming language, a package called "randomForest" needs to be installed, the ntree = 500 restricts the random forest algorithm to grow more that 500 branches thus avoiding overfitting, important = TRUE gives the variable important plot which was discussed in section 3.4.2 .

```
#implementation of random forest
#install.packages("randomForest")
library(randomForest)

RFModel <- randomForest(Appliance ~ .,
                        data=TrainData[-1],
                        importance=TRUE,
                        ntree=500)
```

Figure 17: Implementation of random forest

4.6.2 Evaluation and Results

Table5: Results of random forest

Model	RMSE	MAE	MAPE
Random Forest	197.48	177.33	72%

The above table 5 depicts the RMSE, MAE and MAPE value of Random forest algorithm, from the table it can be observed that absolute error and the mean percentage error of the model is 177.33 and 72% respectively making it a worst model for predicting energy consumption.

4.7 Comparison of Developed Models and Conclusion

Table6: Comparison of all the models implemented for prediction of energy consumption of home appliances.

Model	RMSE	MAE	MAPE
Multiple Regression with all parameters	85.72	49.20	53.6%
Multiple Regression with reduced parameters	88.60	50.92	56.1%
Lasso Regression	85.93	49.55	54.27%
Ridge Regression	86.46	49.26	53.6%
SVM with all parameters	88.93	55.38	64.87%
SVM with reduced parameters	91.44	57.08	66.49%
Random Forest	197.48	177.33	72%

The above table 6 shows the comparison of all the models implemented for the analysis, from the table it can observed that RMSE, MAE and MAPE value of Multiple regression model when all the parameters are included is the lowest, making it best model for the analysis and the values for Random Forest model is the lowest, making it the worst model for predicting energy consumption based on humidity and temperature, the values for the remaining models Lasso Regression, Ridge Regression and SVM lies in between, the Lasso and Ridge regression models works on the mechanism of penalising and reducing the predictors which are less significant in predicting the energy consumption of home appliances, making the models less significant in predicting the energy consumption as all the indoor and outdoor environment conditions are equally important for prediction, according to the analysis, on the other hand SVM works on the principle of separating hyperplanes and transforming the data into higher dimensional space with the help of kernel trick did not give the desired outcome as the energy consumption data is scattered making it difficult to separate by hyperplane. This section completes objective 2 as specified in the research objectives.

5 Conclusion and Future Work

The goal of this research was to find out whether we can predict energy consumption of home appliances based on humidity and temperature which can support consumers in Belgium, the energy consuming behaviour of people in Europe and the initiative taken European countries to meet the energy goals of 2020 was discussed in section 2.2 completing the objective 1 of the research, KDD methodology was used to implement this research, the data was taken from a house located in Stambruges (Belgium), all the steps of the research was implemented in R programming language, correlation plots were used to find the correlation between predictor and dependent variable, variable importance plot was used to reduce the non-significant parameters and data was normalized to avoid biasing, regression models namely Multiple Regression, Lasso Regression, Ridge Regression, SVM Regression and Random Forest were implemented and their results were evaluated based on RMSE, MAE and MAPE completing the objective 2 of the research, The Multiple Regression model when all the parameters were included resulted in the best model by giving lowest RMSE = 85.72, MAE = 49.20 and MAPE = 53. 6% values, although Multiple Regression resulted in the best model but still it is not very good for prediction of energy consumption of home appliances based on humidity and temperature as the values of RMSE, MAE and MAPE is on the higher side, we need more parameters other than humidity and temperature to accurately predict energy consumption of home appliances, so answering to the research question, based only on humidity and temperature, we cannot get promising predictions of energy consumption of home appliances to

support consumers in Belgium, more predictive parameters will be required for a good prediction model.

Future work includes adding more predictive parameters such as occupancy information, the area of the house, day to day activities performed by the occupant, more indoor and outdoor environment conditions, by using different datasets can also improve the prediction. Training other machine models such as ANN (Artificial Neural Network) can further boost the predictive capacity, energy consumption is a vast domain and have lot of scope in future .

Acknowledgement : I would like to thank my parents to support me financially and mentally throughout my course, my supervisor Dr.Catherine Mulwa for guiding me throughout the project and the UCI machine learning team to make the data available for analysis.

References

Aune, M. (2007). Energy comes home, *Energy Policy* **35**(11): 5457–5465.

Azevedo, A. and Santos, M. (2008). *KDD, semma and CRISP-DM: A parallel overview*, p. 182–185.

Balta-Ozkan, N., Boteler, B. and Amerighi, O. (2014). European smart home market development: Public views on technical and economic aspects across the united kingdom, germany and italy, *Energy Research and Social Science* **3**(C): 65–77.

Bardazzi, R. and Pazienza, M. (2017). Switch off the light, please! energy use, aging population and consumption habits, *Energy Economics* **65**: 161–171.

Basu, K., Hawarah, L., Arghira, N., Joumaa, H. and Ploix, S. (2013). A prediction system for home appliance usage, *Energy and Buildings* **67**: 668–679.

Caffarel, J., Gomez, I., Del, C., Martinez, R. and Lastres, C. (2013). Lessons learned on home energy monitoring and management: Smartcity málaga, p. 263–264.

Candanedo, L., Feldheim, V. and Deramaix, D. (2017). Data driven prediction models of energy use of appliances in a low-energy house, *Energy and Buildings* **140**: 81–97.

Casals, M., Gangolells, M., Macarulla, M., Fuertes, A., Vimont, V. and Pinho, L. (2017). A serious game enhancing social tenants' behavioral change towards energy efficiency.

Du, Q., Li, Y. and Bai, L. (2017). The energy rebound effect for the construction industry: Empirical evidence from china, *Sustainability (Switzerland)* **9**(5).

Fayyad, U. (1996). Data mining and knowledge discovery: Making sense out of data, *IEEE Expert-Intelligent Systems and their Applications* **11**(5): 20–25.

Felicetti, C., De, R., Raso, C., Felicetti, A. and Ammirato, S. (2015). Collaborative smart environments for energy-efficiency and quality of life, *International Journal of Engineering and Technology* **7**(2): 543–552.

Gonzalez-Lezcano, R. and Hormigos-Jimenez, S. (2016). Energy saving due to natural ventilation in housing blocks in madrid, *IOP Conference Series: Materials Science and Engineering* **138**.

Guo, Z., Wang, Z. and Kashani, A. (2015). Home appliance load modeling from aggregated smart meter data, *IEEE Transactions on Power Systems* **30**(1): 254–262.

Huebner, G., Shipworth, D., Hamilton, I., Chalabi, Z. and Oreszczyn, T. (2016). Understanding electricity consumption: A comparative contribution of building factors, socio-demographics, appliances, behaviours and attitudes, *Applied Energy* **177**: 692–702.

Jin, X., Baker, K., Christensen, D. and Isley, S. (2017). Foresee: A user-centric home energy management system for energy efficiency and demand response, *Applied Energy* **205**: 1583–1595.

Miezis, M., Zvaigznitis, K., Stancioff, N. and Soeftestad, L. (2016). Climate change and buildings energy efficiency - the key role of residents, *Environmental and Climate Technologies* **17**(1): 30–43.

Niyato, D., Xiao, L. and Wang, P. (2011). Machine-to-machine communications for home energy management system in smart grid, *IEEE Communications Magazine* **49**(4): 53–59.

Ponocko, J. and Milanovic, J. (2018). Forecasting demand flexibility of aggregated residential load using smart meter data.

Rodriguez-Diaz, E., Vasquez, J. and Guerrero, J. (2016). Intelligent dc homes in future sustainable energy systems: When efficiency and intelligence work together, *IEEE Consumer Electronics Magazine* **5**(1): 74–80.

Rovsing, P., Larsen, P., Toftegaard, T. and Lux, D. (2011). A reality check on home automation technologies, **1**: 303–327.

Sanchez-Guevara, S., Mavrogianni, A. and Neila, G. (2017). On the minimal thermal habitability conditions in low income dwellings in spain for a new definition of fuel poverty, **114**: 344–356.

Veras, J., Silva, I., Pinheiro, P. and Rabelo, R. (2018). Towards the handling demand response optimization model for home appliances, *Sustainability (Switzerland)* **10**(3).

Zeng, Q., Zhang, N., Wang, Y., Liu, Y., Kang, C., Zeng, Z., Yang, W. and Luo, M. (2016). *An optimum regression approach for analyzing weather influence on the energy consumption*.

Zhang, D., Li, S., Sun, M. and ONeill, Z. (2016). An optimal and learning-based demand response and home energy management system, *IEEE Transactions on Smart Grid* **7**(4): 1790–1801.