National College of Ireland

Machine Learning Techniques for Prediction of Electricity Consumption in Buildings

MSc Research Project
Data Analytics

Grzegorz Blaszczyk
Student ID: x21195111

School of Computing
National College of Ireland

Supervisor: Dr. Vladimir Milosavljevic

National College of Ireland
Project Submission Sheet
School of Computing

Student Name:	Grzegorz Blaszczyk
Student ID:	x21195111
Programme:	Data Analytics
Year:	2022
Module:	MSc Research Project
Supervisor:	Dr. Vladimir Milosavljevic
Submission Due Date:	15/08/2022
Project Title:	Machine Learning Techniques for Prediction of Electricity Consumption in Buildings
Word Count:	XXX
Page Count:	18

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 required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature:	
Date:	17th August 2022

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST:

Attach a completed copy of this sheet to each project (including multiple copies).	☐
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	☐
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	☐

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):	

Machine Learning Techniques for Prediction of Electricity Consumption in Buildings

Grzegorz Blaszczyk
x21195111

Abstract

Energy use in buildings is responsible for 40% of total global energy consumption. With this in mind, it is imperative to focus on application of that energy to the most efficient use. This project is focusing on showing which machine learning algorithms are able to perform well to predict energy consumption. This project will utilise CRISP-DM methodology: from business understanding, through data exploration, preparation, and cleaning, to application of various machine learning techniques, and concludes with showing which ones are the best for the task. This work will demonstrate it on the example of one of the Kaggle competitions: ASHRAE - Great Energy Predictor III. Throughout this project a sytematic approach for a full cycle of data science project will be demonstrated, how to handle missing data in a relatively large dataset? Selection of good random samples, and application of machine learning algorithms to the samples, to conclude with comparison of results. The results will show that number of simple techniques achieved very similar results to the advanced ones however, they did that in time only a fraction of the techniques that were suggested in the literature.

1 Introduction

ASHRAE (American Society of Heating, Refrigerating and Air-Conditioning Engineers) is an organisation that operates on a global scale that develops and advances improvements in energy performance and sustainable technologies for buildings' energy performance. It was founded in 1894 to help humanity by developing arts and sciences refrigeration, air-conditioning, ventilation, heating. ASHRAE partners with owners of the buildings under pay-per-performance financing option to improve buildings' energy ratings thus providing sustainable and healthy development for all. In order for pay-for-performance financing model to work they need to predict what will the buildings' performance be in the future after retrofitting. To do that they need to build accurate models that are able to predict meter readings before the retrofitting happens. Kaggle, a platform for data science and analytics enthusiasts, partnered with ASHRAE to help the organisation in the development of those models through launch of competition for the network of data scientists. Since ASHRAE is financed through members' contributions of part what they would spend on building energy consumption, the competition's objective was to forecast energy of buildings that would have been required without retrofitting with more energy efficient materials. Kaggle not only provides help to companies, but also tries to develop skills of data scientists, that's why to make it more difficult, they removed some of the

data. This required a different approach than only applying different machine learning techniques therefore data needed to be cleansed, then the best sample was selected, only then different machine learning models could be applied to the data.

Kaggle attracts data scientist to participate in competitions through financial incentives (e.g. the best model was awarded a prize of $10,000), but mostly through acquired prestige from winning or even participating in one. The competition was highly successful with over 3,500 teams participating from around the world. With more accurate models, more companies and organisations could be convinced to retrofit the buildings and therefore reducing the need for energy. The competition ran in 2019, and while new submissions will not be taken into consideration now, the dataset is available continuously providing opportunity to practice and compare with other data scientists.

1.1 Research Question, Project Objectives and Contributions

The research project will focus on comparison of different machine learning techniques, checking if there is an optimal one, and how different methods affect the forecasted utilisation of energy. Therefore the research and sub-research questions are as follows:

Research Question:"Which machine learning method can be used to predict energy consumption most accurately?"

Sub-Research Question 1:"What are the time costs associated with using more advanced techniques?"

Sub-Research Question 2:"Are there synergy effects between variables?"

The research objectives are listed in Table 1.

Table 1: Research objectives

No.	**Objective**
1	Critically evaluate related work
2	Import the dataset and libraries for data analysis
3	Perform initial exploratory data analysis
4	Identify data issues and perform data cleansing
5	Merge the dataset and check for any other issues
6	Select the best sample from dataset for further data analysis
7	Apply various machine learning techniques
8	Compare the outputs and results
9	Critically evaluate used techniques and suggest further steps

5

2 Related Work

Blaszczyk (2020a) cleansed, prepared, and selected a sample from the data then compared different machine learning methods. He advised in "conclusions and future work" section other methods that were suggested in literature in order to improve accuracy of prediction. Likewise, this project is going to compare application of neural networks however, it will try to improve its below par result (R^2 of 0.07 - 0.81 of ANN vs R^2 of 0.69 - 0.97 of PCR method).

To measure energy consumption of the building Abanda and Byers (2016) focused on its orientation and therefore solar irradiance using Building Information Modelling (BIM) technique. The BIM is an approach that utilises technologies, policies, and processes to oversee the end-to-end building design in a digital manner throughout the lifespan of a building. Unfortunately the building's orientation is not available in the dataset for this analysis. Also, while the BIM approach is not connected directly to my thesis, it served as encouragement to check for potential synergies between impact of various climate components that are available on energy utilisation e.g. is there an impact of combination of wind speed with wind direction or only the components themselves impact energy consumption? Are there additional gains/losses if it is sunny on a warm day or cloudy on a cold day? Or if there is a difference between cloud coverage combined with air temperature? The researchers also pointed out that the closer the building's shape resembles a cube the more energy efficient it is – so could be lead to assumption that there is a correlation of building height (floor count) with its floor area.

While the analysis of Ferrarini et al. (2019) focused on the prediction of energy consumption in one residential building in the north of Italy. The data used covered 6 months only however the weather data came from reliable weather station which was only 10 km away. However, graininess of data of some meters down to 15 min intervals from numerous sensors provided sufficient amount of data points to apply predict energy consumption in a robust way. The researchers also used geometrical data about the building (such as insolation, size, and floor number). Each apartment was divided into 6 zones and sensors were installed in all of them. In their work they mentioned that while the "black-box" approach had the best performance at predicting energy consumption, neural networks could be used to further improve the results.

Wang and Dong (2009) focused on predicting energy consumption in China. They based their research on China's Statistical Yearbook and New China Statistical Data Assembly. They used annualised figures for 50 years and used the following five variables. Dependent variable: Energy Consumption, and independent ones: Gross Domestic Product, Industrial Structure, Total Population, and Technology Progress. They suggested that there was no further improvement in accuracy after introducing more than 3 layers to ANN. They achieved significant accuracy (R^2 of 0.992) using Artificial Neural Networks they further improved it by deploying Genetic Algorithms with optimum number of generations of 9 (to R^2 of 0.996). While this was an outstanding accuracy, the typical datasets that can be used for machine learning are significantly greater than 250 observations which may lead to issue with robustness of the results.

Amin and Khan (2020) focused their work on forecasting energy demand in Bangladesh. Similarly to Wang and Dong (2009) they also used annualised data for their prediction. For independent variables they chose: consumer price index (2010 = base), real income per capita, CO2 emissions, household spending, per capita energy consumption, and population. They took a different approach and performed the analysis using log-linear model. The accuracy, similarly to Wang and Dong (2009) work, was really accurate with R^2 over 0.99. The researchers flagged that a small sample may lead to issues with robustness of the model, and suggested to counterbalance that with deployment of DOLS method. This method permits measuring the impact of various variables using log-linear regression where there is only limited amount of variables available. While the same method was successfully applied by Merlin and Chen (2021). Similarly to previous researchers here also focused on annual economical data of Democratic Republic of Congo. There seem to be number of issues with this approach: first, it focuses on economical

data; second, the energy consumption is explained as a function of population, and energy consumption per capita. These two variables, by design, will be highly correlated with total energy consumption and while they will improve model's accuracy, might not be independent enough from the dependant variable. Researchers (Lü et al.; 2015) focused on prediction of energy consumption in individual buildings by combining building information with weather data. While their accuracy was over 0.9 R^2, they applied this technique only to several buildings. The main limitation of this technique is the necessity to gather vast amounts of data for each building, which typically is not possible to gather on larger-scale projects.

Olu-Ajayi et al. (2022) focused on predicting energy class at a design stage of a building. They suggested that Gradient Boosting technique had the best results, while they predicted energy class rather than energy consumption thus used classification technique rather than a regression one as this project, therefore their results are not comparable to this project, their project inspired comparison of training time required with different techniques.

Pham et al. (2020) applied various machine learning techniques such as Random Trees, M5 Model Trees, and Random Forest to predict the energy consumption. They built their models using monthly data over 3, 6, 9, and 12 months. They used multiple metrics to evaluate the models, one of them was MAPE. Interestingly, using Random Forest for their prediction, the MAPE for prediction 24 steps ahead in number of cases was lower than the prediction for 12 months ahead, this is particularly visible in Table 9 of their article. This could mean that they missed some bi-annual seasonality variable.

Shen et al. (2020) used OLS regression as a benchmark, this project similarly used that method as a base for comparison, due to the easy of use and interpretability. They focused on predicting energy consumption in China's residential market. The data they gathered was not only about the building itself but also contained information about inhabitants (their level of education, income, number of household members) The researchers here found out that OLS had lower performance due to issue with handling non-linear data. This was not the issue with this project as OLS was one of the techniques that achieved highest results.

While there was no consensus when it comes to the best technique used therefore it might be dependent on the dataset that the method is applied to. One thing that researchers (Lü et al.; 2015),(xiang Zhao and Magoulès; 2012),(Robinson et al.; 2017),Lu et al. (2022) agreed about, wherever this information was available, that buildings were responsible for 40% of energy consumption. This fact alone, suggests that proper energy management in buildings should be in the focus.

3 Methodology and Implementation

Can electricity consumption be predicted using machine learning techniques? Are neural regression techniques superior to classical ones for the purpose of estimating retrofitting gains for organisations such as ASHRAE? How accurate these models have to be to justify retrofitting?

To answer these questions understanding of the context of the data is needed. The first step according to CRISP-DM process is Business Understanding. Different accuracy criteria should apply to models involving human participants, such as healthcare inform-

ation. Unfortunately this was not always the case [1]. While models recommending which customers should be targeted for the next renewal campaign, using churn analysis, can have relatively lower accuracy as they will impact finance of a company.

Since the literature shows that buildings are responsible for approximately 40% of all energy consumption. This is a significant share of all energy consumption and will have impact on the demand for it, it will also have impact on life on Earth due to use of fossil fuels for production of it. It is important for organisations such as ASHRAE which are looking at a cost of retrofitting of a building (e.g. if more energy will be spent on retrofitting of a building with more energy efficient solution in its lifespan then it is not profitable to retrofit it. Saved energy could be used for other purposes. Energy production also poses the problem of its storage – it needs constant production in power plants while the storage capacity is severely limited to for example batteries, majority of energy cannot be stored and needs to be utilised when it is produced. This means that the electricity grid has to be designed in a specific way to serve residential, business, and industrial parts of community while continuing its energy flow with limited blackouts. This is the part when machine learning techniques can be deployed to forecast energy demand.

[1] IBM's Watson gave unsafe recommendations for treating cancer: `https://www.theverge.com/2018/7/26/17619382/ibms-watson-cancer-ai-healthcare-science`

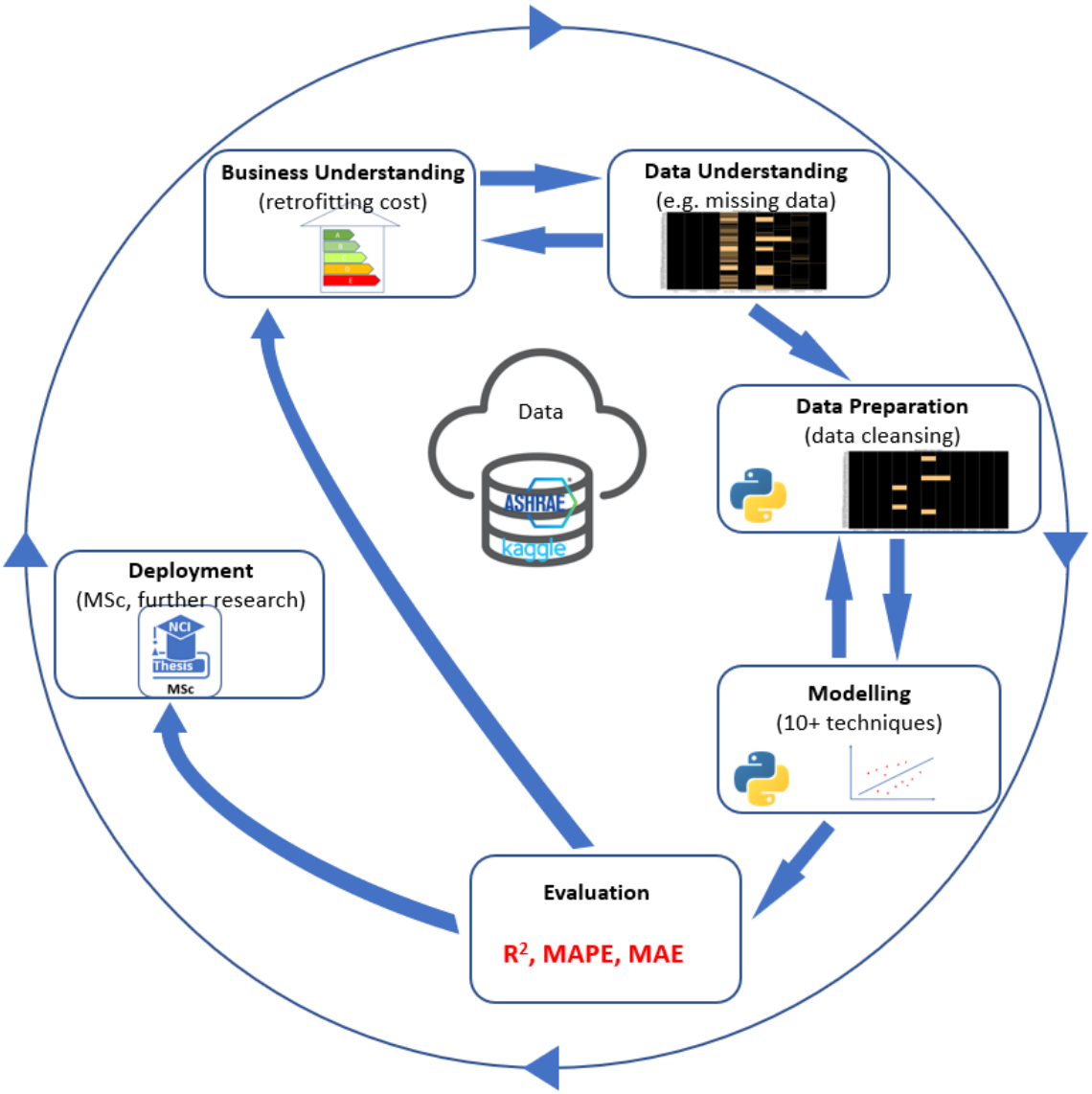

Figure 1: Adapted CRISP-DM Process for Data Mining (Purbasari et al.; 2021)

After initial research into business the next step on CRISP-DM methodology is data understanding. As a result of initial data exploration we notice that there are 6 tables in the dataset, this project focuses on 3 of them:

- train.csv - information about energy consumption, meter type, building id, and timestamp
- building_metadata.csv - contains information about each building, and where the building is located
- weather_test.csv - contains information about weather in a specific location.

The dataset provided by Kaggle consisted of 20 million observations with meter readings, 140 thousands observations for weather dataset, and 1,400 rows of building information. Figure 2 provides information about training dataset.

	building_id	meter	meter_reading
count	2.021610e+07	2.021610e+07	2.021610e+07
mean	7.992780e+02	6.624412e-01	2.117121e+03
std	4.269133e+02	9.309921e-01	1.532356e+05
min	0.000000e+00	0.000000e+00	0.000000e+00
25%	3.930000e+02	0.000000e+00	1.830000e+01
50%	8.950000e+02	0.000000e+00	7.877500e+01
75%	1.179000e+03	1.000000e+00	2.679840e+02
max	1.448000e+03	3.000000e+00	2.190470e+07

Figure 2: Training dataset summary

These tables can be joined on building_id and site_id keys. The first of the tables has all the values, while the last two miss varied amounts of data. First the individual tables will be cleansed, and only then the data will be merged into one table to select the sample.

The data provided by Kaggle includes reading for four meters: electricity, chilled water, steam, and hot water. This project will focus on measuring impact of variables such as building type, weather, and other only on electricity consumption. This is motivated by the fact that the initial data contains the most datapoints for that meter, but also to be able to compare the techniques applied to different samples from the datasets. The data will be first cleansed, then merged, and finally a sample from 3 locations will be selected to verify if the results achieved are robust and comparable in different locations.

Following on from widely accepted George Fuechsel's GIGO rule (Roden et al.; 2022) which states that without good data and its preparation, no matter how good the analysis tools are going to be the results are going to be less than satisfactory. Konstantinou and Paton (2020) claimed that on average data scientist spends 80% of their time preparing data, this project is going to follow similar trajectory in regards to data preparation. To cleanse the data a number of assumptions are required for example to populate missing data in the weather table we infer that there are no significant changes to weather conditions from one hour to the next. This enables the data for the same location to be populated with previous values. This still leaves number of empty fields, therefore the process gets repeated but this time for a day, and then for a month. The output from weather dataframe before and after cleansing is demonstrated in Figure 3 where missing data is highlighted in orange. In the next two Figures, sub-figures A, on the left, represent the original data while Figures B, on the right, represent the data after cleansing.

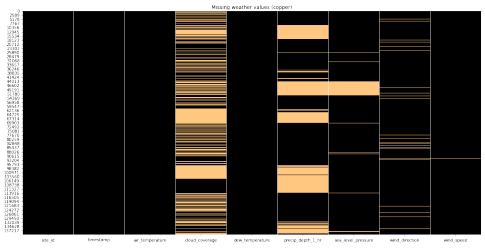

Fig A

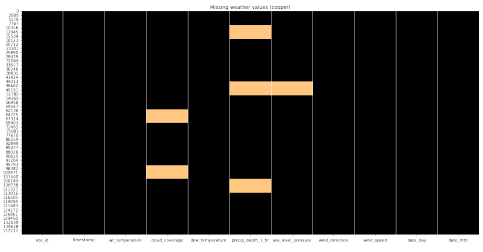

Fig B

Figure 3: Missing values in weather dataframe before (Fig A) and after cleansing (Fig B)

Following on similar logic an assumption can be formed that buildings for the same use have the same floor count and were built in the same year Figure 3

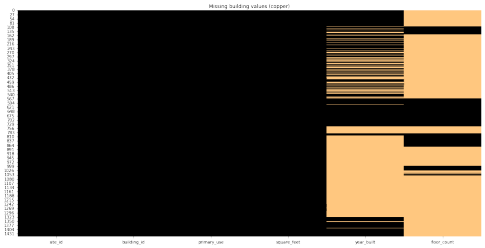
Fig A

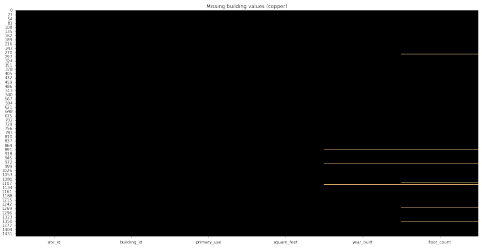
Fig B

Figure 4: Missing values in building dataframe before (Fig A) and after cleansing (Fig B)

Since this project is only going to demonstrate how different machine learning techniques can predict energy consumption and only sample of data is going to be taken, rows where there is missing data can be dropped without impacting on the outcome of this research.

Now that the data has been sufficiently cleansed next thing that needs addressing is data exploration that will also form data understanding. Looking at the building dataset it is noticeable that there is a relationship between the buildings' primary use and the size of the building Figure 5. There is also a link between the year when the building was built and their primary use.

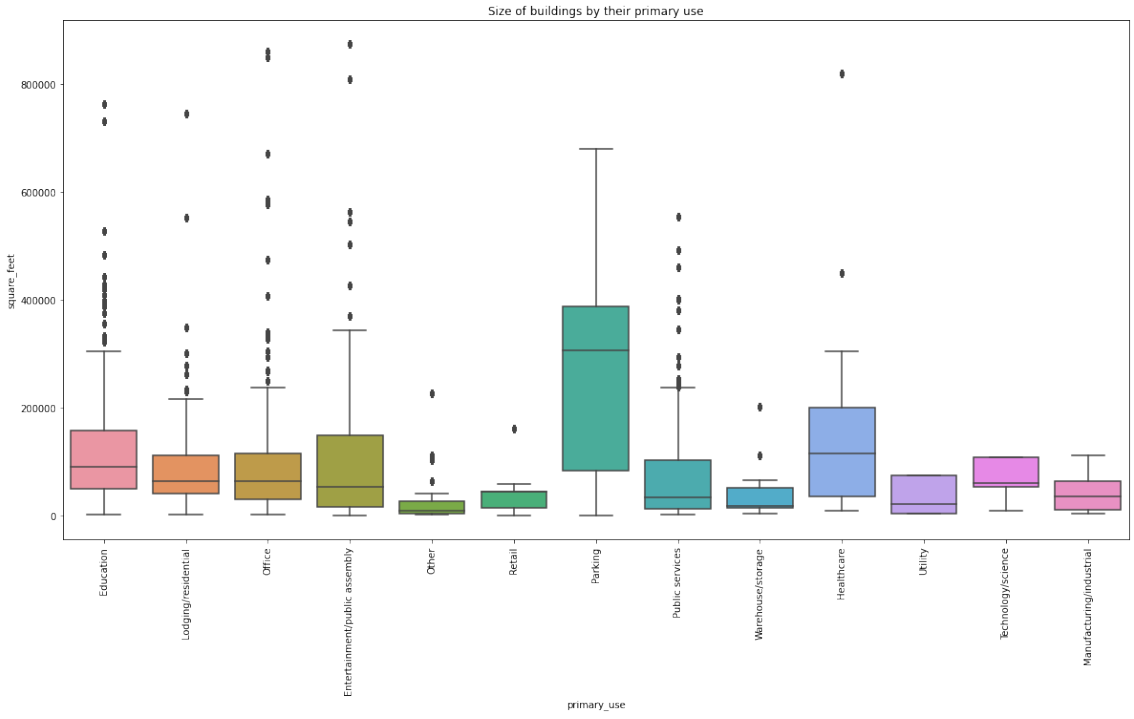

Figure 5: Correlation between pairs of weather variables

Next thing that needs addressing is correlation of weather variables. Multicollinearity of a pair, or multiple variables can negatively impact models' performance, that's why it needs to be addressed in data cleansing phase, and before moving on to the modelling

phase. Figure 6 shows correlation between pairs of weather variables, as can be seen on the chart, there is very little linear correlation between pairs of features in this dataframe. The only exception from this is the pair of dew temperature and air temperature. They will be monitored at a modelling stage, and should the variables' coefficients go in the opposite directions remediating steps will be taken.

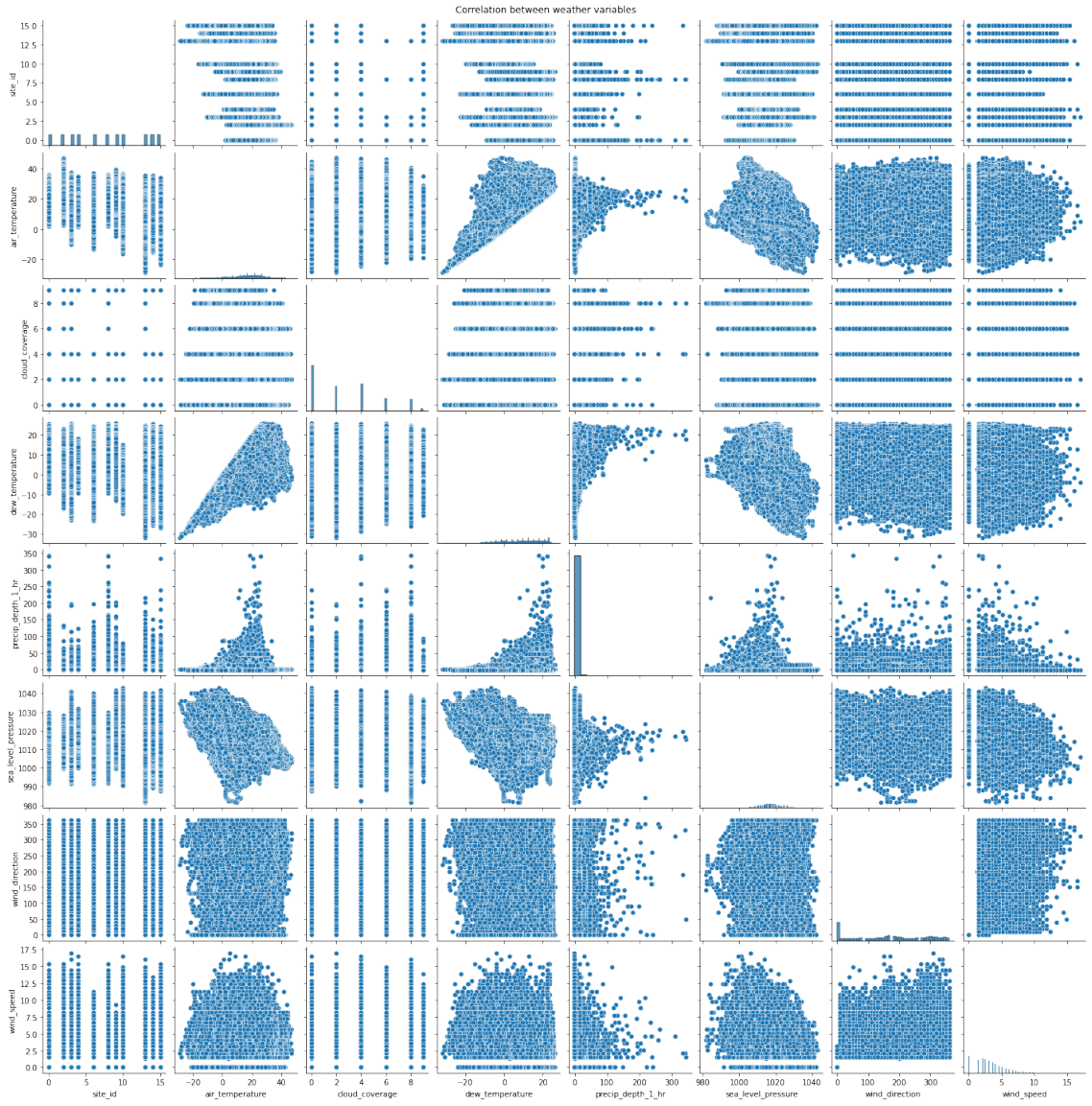

Figure 6: Pairplot of weather variables

To proceed to the modelling phase all data needs to be contained within one dataframe. This can be done by using the keys to merge all three of them. At this stage categorical variables can also be converted to discrete variables using one hot encoding. Since part of the project is also to check for any synergy effects variables that will measure that synergy can also be created (one of the variables will check if there is a combined impact of cloud coverage with air temperature, another one will check the synergy impact between wind speed and wind direction).

The electricity meter had the most readings in the dataset therefore it can guarantee the most consistent data (even after data cleansing). The electricity meter was also the one that was showing the strongest correlation with independent variables and that is the

reason for which it was selected for the remainder of this analysis. Since there are number of locations to choose sample data from, the data selection will focus on the sample with the least datapoints to be the most representative. The locations chosen: location 6, 8, and 10 to move on to the following step on CRISP-DM methodology: Data Modelling. These 3 sample datasets will be split pseudo-randomly into train, test, and validate parts in the 70:20:10 proportion. Figure 7 shows correlation matrix of features with electricity meter reading. As can be seen on it, some variables show really strong correlation with the electricity meter reading which will be the dependent variable variable, the strongest correlation with meter reading has the area variable (square_feet). While the chart below shows site 6, the correlation between area of the building and the meter reading was also true for other sites. There are also number of independent variables that are correlated with each other. There are also variables correlation of which does not make sense from the logical point of view such as a negative correlation between height of the building and electricity consumption (it would mean that taller buildings have smaller electricity consumption).

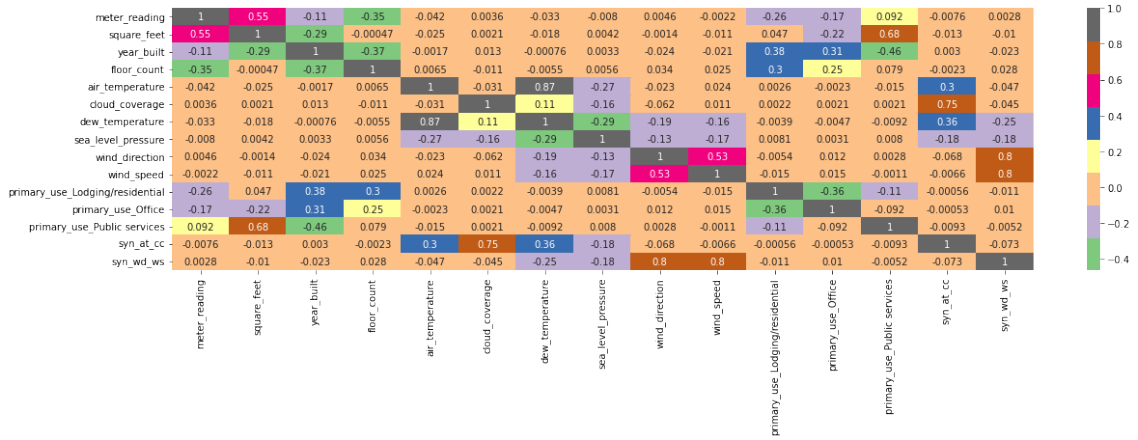

Figure 7: Correlation matrix of various features for the electricity meter for the site 6 train dataset

Bearing this correlation table mind, the analysis moves on to the next stage, as correlation does not mean that these variables will have impact on the modelled variable as the features might not be statistically significant. There will be number of techniques tested and compared:

The first technique used will be **(OLS) Ordinary Least Squares**, it will also serve as a base for comparison of results. The objective of this method is to minimise the distance of predicted values from the actual dependent variable. This is achieved by measuring a distance at each actual datapoint and measuring the distance to the predicted value for this datapoint and fitting a function that minimises the euclidean distance between the two. This method, due to its ease of interpretability of the results, will also be used to determine statistically significant variables.

Second technique used for comparison will be **(KNN) K Nearest Neighbours**. It takes K nearest neighbours and calculates the average of their values to find the numerical target. Similarly to OLS, it has been used for over half a decade. Typically this technique is used for classification purposes, but it can also be successfully applied to regression tasks as well.

The next method applied will be **(PCR) Principal Component Regression**. The

technique uses Principal Component Analysis and reduces the number of independent variables by drawing the curve with number of eigenvalues. This curve can be used for determining number principal components using a visual "elbow" method to determine their number, and building model using those components. Another method that uses PCA is **(PLS) Partial Least Squares** method. It is similar to PCR, however this method focuses on latent-variables. It identifies a new set of linear features, that are combinations of the source variables.

A slight variation of OLS technique is **Ridge** regression. Ridge is enhancing OLS technique by altering the cost function slightly and thus preventing overfitting of models. This is done by a technique called L2 regularisation which adds a penalty function to the parameters in models.

A very similar technique to Ridge is **Lasso**. It also a modified OLS technique, with altered cost function however, the cost function is added using L1 regularisation rather than L2 as in case of Ridge.

A combination of the two previous techniques is a method called **ElasticNet**. It uses both L1 and L2 regularisations and applies both cost functions to make the model more robust.

Similarly to K Nearest Neighbours **Decision Tree** is mostly known for its application to classification problems, but it can also be used for prediction of continuous variables. The algorithm splits the dataset alongside decision leafs based on the variability of the data. Depending on the complexity of the data a desion tree might split the dataset in only two subsets (so called "tree stump") or it may consist thousands of decisions (leafs).

Multiple Decision Trees can be build randomly forming a technique called **Random Forest**. The dataset is split pseudo-randomly into multiple subsets, decision tree is built for each of the subsets, results are combined and final output is the output of this technique.

Another ensemble method where multiple models are built is called **(AdaBoost) Adaptive Boosting**. The technique samples data, builds stump trees, assigns lower weights to weaker models. The tree stumps are built sequentially therefore past inputs improve accuracy.

Similar to AdaBoost is a technique called **Gradient Boosting**. It is also a sequential ensemble technique, where previous models improve future ones. The main differences to AdaBoost are that Gradient Boost introduces learning rate rather than lower weight to previous models, and that the trees built are full size versus AdaBoost's stumps.

(XGBoost) Extreme Gradient Boosting is a technique that combines Gradient Boosting with Elastic Net. It uses models sequentially, and adds L1 and L2 regularisation to prevent overfitting of the gradient boosting technique.

The last technique used for comparison will be **(ANN) Artificial Neural Network**. The method mimics a brain of a living animal. The literature suggested that using 3 layers (input, hidden and output) is the most optimal set-up as the addition of extra layers was only slowing the prediction. The method was praised in literature for the accuracy to predict non-linear relationships.

4 Results and Discussion

The metric cited most in the literature to evaluate the models was coefficient of determination (R^2). It is also a metric that allows comparison of results from different models

as it takes values between 0 and 1 while other metrics such as MAE, MAPE have ranges depending on the value of dependant variable, which makes it difficult to compare results with those in the literature. While there are no perfect models, the value of R^2 should be as close to 1 as possible.

4.1 Results

The R^2 achieved was similar, and in number of cases higher than the one quoted in literature. It is also worth noting, that results depended on the site that they were measured for, e.g. site 6 produced the highest results, while site 10 had the lowest. The charts and tables below show results for site 6, since the results were closest to the ones cited in the literature.

Figure 8 shows comparison of R^2 of models and their processing time for site 6. The top part shows the accuracy of the models while the bottom part shows processing time for various techniques.

Figure 8: comparison of R^2 of models and their processing time (site 6)

As can be seen, majority of the techniques applied produced comparable results. The only exceptions were **ElasticNet**, and **Neural Network** which had R^2 of 0.48 and 0.56 respectively.

What is interesting to see is the processing time of different methods. Table 2 provides information about the results for each model, processing time and weighted time for each of the modelling techniques for site 6. All times are in seconds, and the weighted processing time is included to bring all models to the common denominator for the perfect model (R^2 of 1). The results are ordered from the model which would produce perfect result fastest to the one that was the slowest.

Table 2: R^2, processing time, and weighted time per model (site 6)

Method	**R^2**	**Training time [s]econds**	**Weighted time[s]**
PLS	0.882	0.01	0.012
PCR	0.964	0.03	0.03
OLS	0.957	0.19	0.196
KNN	0.959	1.58	1.646
Ridge	0.963	1.7	1.762
Lasso	0.945	2.3	2.435
DecisionTree	0.931	3.31	3.551
ElasticNet	0.48	2.62	5.456
RandomForest	0.958	44.33	46.263
AdaBoost	0.901	47.33	52.542
Neural Network	0.56	34.52	61.609
GradientBoost	0.962	60.28	62.651
XGBoost	0.962	73.44	76.327

The fastest method was **Partial Least Squares** closely followed by **Principal Component Regression**, while the slowest were **Boosting** techniques and **Neural Network**. Using **PLS**, to achieve perfect results could be reached in as low as 0.01 second (rounded to two decimal places) while it would take **XGBoost** over 76 seconds (over 6,000 times slower).

It is worth noting that the highest overall R^2 of 0.964, was that of the second fastest technique: **PCR**. **XGBoost** quoted in literature, as a method which was becoming a new standard, was able to achieve similar results, however it were marginally lower at 0.962. The time it took to achieve perfect result for XGBoost vs PCR was over 2500 times slower.

Another method that was recommended in the literature was **Artificial Neural Network**. The one used for the purpose of this project contained 3 layers: input, hidden, and output. While this is a simple design, it was quoted that designing more complex Neural Network would slow the training process without gains Wang and Dong (2009). Neural Network was able to achieve sub-par R^2 of 0.56. Comparing it to PCR, it would take over 2000 twice as long to achieve a perfect R^2.

Figure 9 shows output of Ordinary Least Squares model for site 6.

	coef	std err	t	P>|t|
site_id	-7.7527	0.847	-9.148	0.000
square_feet	0.0010	4.17e-05	23.718	0.000
year_built	0.0557	0.002	34.730	0.000
air_temperature	-0.5886	0.051	-11.493	0.000
wind_direction	-0.0134	0.005	-2.669	0.008
wind_speed	1.7242	0.326	5.290	0.000
primary_use_Education	143.9231	1.162	123.854	0.000
primary_use_Lodging/residential	-64.4066	1.697	-37.948	0.000
primary_use_Office	-10.7647	4.113	-2.617	0.009
primary_use_Public services	-77.1383	6.600	-11.687	0.000
building_id_745	-201.6801	4.350	-46.359	0.000
building_id_747	-201.7126	3.815	-52.874	0.000
building_id_748	-41.9711	4.332	-9.690	0.000
building_id_749	14.1767	4.048	3.502	0.000
building_id_750	-47.8973	3.651	-13.117	0.000
building_id_753	-198.6073	3.482	-57.040	0.000
building_id_755	-61.3855	4.342	-14.137	0.000
building_id_759	-19.6092	3.577	-5.483	0.000
building_id_760	62.1414	4.421	14.057	0.000
building_id_761	-97.4340	3.273	-29.772	0.000
building_id_762	-139.8252	3.239	-43.165	0.000
building_id_764	-207.1846	3.142	-65.933	0.000
building_id_765	24.8498	3.022	8.222	0.000
building_id_766	-26.4854	4.773	-5.549	0.000
building_id_767	94.8305	4.788	19.806	0.000
building_id_768	192.4149	2.964	64.910	0.000
building_id_769	-186.4936	2.995	-62.259	0.000
building_id_770	-232.5456	3.043	-76.432	0.000
building_id_771	-8.7250	3.611	-2.416	0.016
building_id_773	-25.4508	3.463	-7.349	0.000
building_id_774	-23.7214	3.536	-6.708	0.000
building_id_775	16.8283	5.559	3.027	0.002
building_id_776	-62.1981	5.560	-11.188	0.000
building_id_777	79.1436	3.467	22.825	0.000
building_id_778	79.6786	2.859	27.871	0.000
building_id_780	363.4622	3.254	111.695	0.000
building_id_782	139.4238	3.376	41.300	0.000
building_id_784	303.1468	5.221	58.060	0.000
building_id_785	561.4219	7.317	76.731	0.000
building_id_787	-39.7847	8.874	-4.483	0.000
building_id_788	-77.1383	6.600	-11.687	0.000

Figure 9: Model summary OLS (site 6)

Coming back to Figure 7 it is worth noting that while "square_feet" variable was the variable showing the strongest positive correlation with the meter reading, on the other side "floor_count" variable had the strongest negative one. Looking at t-stat output of the model in Figure 8, while there is a strong impact of that variable, the impact of "primary_use_Education" variable had statistically stronger impact. When it comes to the interpretation of "floor_count" variable, it turned out that it did not have statistically significant impact on energy consumption. Primary_use_Education was the strongest

positively impacting variable in model for site 6, had a relatively weaker impact (although statistically significant) for site 10, and was not statistically significant at all for site 8, potentially that could mean that schools and universities are using proportionally more energy in comparison to other types of buildings. On the opposite side was Air_temperature had a negative impact on energy consumption for site 6, and 10 however, which makes logical sense, since the higher the temperature, the lower the electricity consumption, however it was not relevant for site 8, this could mean that use of electricity for heating was not necessary. As shown on Figure 8 the variables that were meant to show synergy effects between different variables are missing. This means that there was no impact of the combined variables. The only model where they combination of wind speed and wind direction had statistically significant impact was model built for site 10. This however can not prove that there was synergy effect between these variables since in the other two models, this variable was not statistically significant.

4.2 Discussion

While in general, the results were comparable to the ones found in literature, there were number of suggestions that could be applied to further improve the outcomes of models.

It is possible to design more complex Neural Network to improve the results. Lu et al. (2022) mentioned that wider or deeper neural networks can be built to achieve greater accuracy, it is also possible to implement other other Neural Networks technique (Feed Forward NN, different types of Recurrent NN, or Convolutional NN) that can be used for the purpose of energy prediction. Blaszczyk (2020b) tested wider and deeper Neural Networks for credit score prediction, however this did not improve the prediction, the only improvement that was the use of Sigmoid function instead of RELU. This however improved the R^2 from 0.84 to 0.85, since both credit score and energy consumption predictions are both regression problems, it is possible that the impact would be similar.

When it comes to the time it took to train the model, it is possible to develop models that will use parallel computing and therefore reducing the time required to train the model.

5 Conclusion and Future Work

This section completes the thesis by reiterating the key findings and future work in relation to the research question, sub-research questions and objectives from the section 1.1:

Research Question:"Which machine learning method can be used to predict energy consumption most accurately?"

To answer this question, the following tasks were completed. Data was gathered from Kaggle website, the raw data was explored, and where gaps were identified a relevant techniques were applied to cleanse it. Following this the tables were merged and three samples were selected. Each of the samples were split into train, test, and validation parts. Number of machine learning techniques were applied to all of them to evaluate if the findings from one model were replicated in another.

It is worth reiterating that in this project, the training dataset contained a sample of only 7000 observations. Typical datasets used to train machine learning models are significantly bigger and it is not uncommon for the datasets to contain millions of obser-

vations. Needless to say this would significantly increase the time required to train such a model.

Due to reduction in number of features used and relative simplicity of the method, PCR was able to outperform other methods.

With the progress in processing power and cloud computing it is relatively easy to design more complex methods (e.g. adding extra layers to Neural Network), however as demonstrated, some of the simple techniques can outperform more complex ones. They did not only run faster, but were able to achieve more accurate results. There are still many areas that need to be explored in order to achieve better results in the area of energy usage forecasting, hopefully this project contributed to the state of general knowledge, and techniques that might be useful.

It was suggested in the literature that introduction of number of independent variables (such as behaviour of inhabitants), it might be especially important now, when companies are adopting hybrid models with "in office" days. Majority of offices are not used fully throughout the week so it would be interesting to measure the impact of lack of occupation on energy consumption.

Acknowledgements

I wish to thank Dr. Vladimir Milosavljevic who through his suggestions and support greatly contributed to the completion of this project.

I am grateful to the lecturers and staff of National College of Ireland I have met throughout the two years of study in my Postgraduate Diploma and later Masters in Data Analytics.

Figure 10: Thank you

References

Abanda, F. and Byers, L. (2016). An investigation of the impact of building orientation on energy consumption in a domestic building using emerging bim (building information modelling), *Energy* **97**: 517–527.
URL: *https://www.sciencedirect.com/science/article/pii/S0360544216000037*

Amin, S. and Khan, F. (2020). Modelling energy demand in bangladesh: An empirical analysis, *The Journal of Developing Areas* **54**.

Blaszczyk, G. (2020a). Data mining and machine learning i project, kaggle competition: Ashrae - great energy predictor iii, *Postgraduate Diploma, NCI, Dublin* .

Blaszczyk, G. (2020b). Domain applications of predictive analytics, credit score prediction using ensemble methods of genetic algorithms and artificial neural networks, *Postgraduate Diploma, NCI, Dublin* .

Ferrarini, L., Fathi, E., Disegna, S. and Rastegarpour, S. (2019). Energy consumption models for residential buildings: a case study, *2019 24th IEEE International Conference on Emerging Technologies and Factory Automation (ETFA)*, pp. 673–678.

Konstantinou, N. and Paton, N. W. (2020). Feedback driven improvement of data preparation pipelines, *Information Systems* **92**: 101480.
URL: *https://www.sciencedirect.com/science/article/pii/S0306437919305320*

Lu, C., Li, S. and Lu, Z. (2022). Building energy prediction using artificial neural networks: A literature survey, *Energy and Buildings* **262**: 111718.
URL: *https://www.sciencedirect.com/science/article/pii/S0378778821010021*

Lü, X., Lu, T., Kibert, C. J. and Viljanen, M. (2015). Modeling and forecasting energy consumption for heterogeneous buildings using a physical–statistical approach, *Applied Energy* **144**: 261–275.
URL: *https://www.sciencedirect.com/science/article/pii/S0306261914012689*

Merlin, M. L. and Chen, Y. (2021). Analysis of the factors affecting electricity consumption in dr congo using fully modified ordinary least square (fmols), dynamic ordinary least square (dols) and canonical cointegrating regression (ccr) estimation approach, *Energy* **232**: 121025.
URL: *https://www.sciencedirect.com/science/article/pii/S0360544221012731*

Olu-Ajayi, R., Alaka, H., Sulaimon, I., Sunmola, F. and Ajayi, S. (2022). Machine learning for energy performance prediction at the design stage of buildings, *Energy for Sustainable Development* **66**: 12–25.
URL: *https://www.sciencedirect.com/science/article/pii/S0973082621001307*

Pham, A.-D., Ngo, N.-T., Ha Truong, T. T., Huynh, N.-T. and Truong, N.-S. (2020). Predicting energy consumption in multiple buildings using machine learning for improving energy efficiency and sustainability, *Journal of Cleaner Production* **260**: 121082.
URL: *https://www.sciencedirect.com/science/article/pii/S095965262031129X*

Purbasari, A., Rinawan, F. R., Zulianto, A., Susanti, A. I. and Komara, H. (2021). Crisp-dm for data quality improvement to support machine learning of stunting prediction in infants and toddlers, *2021 8th International Conference on Advanced Informatics: Concepts, Theory and Applications (ICAICTA)*, pp. 1–6.

Robinson, C., Dilkina, B., Hubbs, J., Zhang, W., Guhathakurta, S., Brown, M. A. and Pendyala, R. M. (2017). Machine learning approaches for estimating commercial building energy consumption, *Applied Energy* **208**: 889–904.
URL: *https://www.sciencedirect.com/science/article/pii/S0306261917313429*

Roden, B., Lusher, D., Spurling, T. H., Simpson, G. W., Klein, T., Brailly, J. and Hogan, B. (2022). Avoiding gigo: Learnings from data collection in innovation research, *Social Networks* **69**: 3–13. DATA COLLECTION FOR SOCIAL NETWORKS RESEARCH.
URL: *https://www.sciencedirect.com/science/article/pii/S0378873320300332*

Shen, M., Lu, Y., Wei, K. H. and Cui, Q. (2020). Prediction of household electricity consumption and effectiveness of concerted intervention strategies based on occupant behaviour and personality traits, *Renewable and Sustainable Energy Reviews* **127**: 109839.
URL: *https://www.sciencedirect.com/science/article/pii/S1364032120301337*

Wang, S. and Dong, X. (2009). Predicting china's energy consumption using artificial neural networks and genetic algorithms, *2009 International Conference on Business Intelligence and Financial Engineering*, pp. 8–11.

xiang Zhao, H. and Magoulès, F. (2012). A review on the prediction of building energy consumption, *Renewable and Sustainable Energy Reviews* **16**(6): 3586–3592.
URL: *https://www.sciencedirect.com/science/article/pii/S1364032112001438*