Energy Consumption Analysis and Prediction based on Weather Conditions and Household Groups in London using Random Forest Regression

Angeline Mary Marchella
Computer Science and Technology
Beijing Institute of Technology
Beijing, China
angeline.marchella@binus.ac.id

Darian Elbert
Computer Science and Technology
Beijing Institute of Technology
Beijing, China
darian.elbert@binus.ac.id

Thomas Dante Wunan
Computer Science and Technology
Beijing Institute of Technology
Beijing, China
thomas.wunan@binus.ac.id

ABSTRACT

The contamination of greenhouse gases (GHGs) on Earth is increasing. One of the main contributors is the excessive amount of energy consumption. To deal with this problem, the United Kingdom (UK) governors have required households to apply smart meters to measure their daily electricity usages. In this study, the smart meters dataset for the London area is analyzed. We aim to predict energy usage according to weather conditions and household groups to help people budget their electricity consumption effectively. Through several steps of data preparation and exploration, figured that weather conditions such as temperature, humidity, and wind speed, are correlated with electricity usage behavior. Additionally, external factors like household groups and type of energy measurement also impact the total energy consumption variable. This research implemented Random Forest regression for prediction. Two experiments are created to examine whether the weather variable is the only factor that influences the changes in power usage or if other external factors also contribute. The result showed that the model that is trained with weather conditions and external factors gave more accurate predictions with 0.93 of r-squared score and 0.86 of root mean square error (RMSE). Meanwhile, by considering weather variables alone, the model returned 0.31 or r-squared and 1.12 of RMSE. Hence, we concluded that the energy consumption behavior of a household does not only depend on the current weather conditions but is also influenced by the topography and demographics of the area.

CCS CONCEPTS

• Information system → Data analytics; Decision support systems • Software and its engineering → Machine learning algorithms • Social and professional topics → Energy efficiency

KEYWORDS

Weather conditions, Households, Energy consumption, Random Forest Regression

1 BACKGROUND

Green-house gases (GHGs) concentration in the Earth's atmosphere is increasing. The factors that cause this phenomenon are human activities including excessive electricity usage, industrial processes, fossil fuel combustion, and diesel engines [1]. The global distribution of GHG emissions is shown in Figure 1. GHGs exploit far-ranging environmental and health effects. They cause air pollution, extreme weather, wildfires, and heatwaves. Hence, the need to reduce GHG emissions becomes more urgent.

According to the 2021 Glasgow Climate Pact agreement, the United Kingdom (UK) and other countries have committed to curb GHG emissions [2]. To achieve the target, governments require meaningful, accurate, and trusted information to determine how to effectively reduce GHG emissions. For almost a decade, the UK has implemented smart meters in many households and industries. Unlike standard meters, smart meters record half-hourly consumption data and automatically send meter readings to the energy provider.

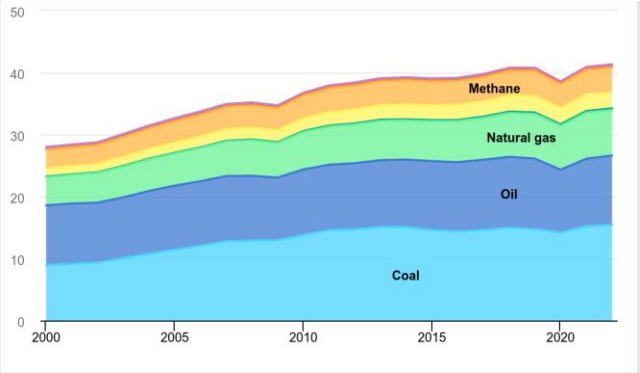

Figure 1: Global energy-related greenhouse gas emissions, 2000-2022 [3]

This innovation is highly beneficial for the nationwide [4]. For once, it gives engineers better information about the causes of power cuts. The innovative smart tariffs also enable consumers to

save money by utilizing energy outside of peak hours or when there is a surplus of clean electricity available. Moreover, customers with smart meters are simply charged for the energy they use, which allows them to budget more effectively. It has been evidently shown that consumers that maximize this information manage to save money and reduce GHG emissions.

In accordance with maximizing the power budget of customers, our research aims to make energy consumption prediction based on the weather conditions in London. A study has shown that weather variables such as temperature, wind speed, and solar radiation are strongly correlated to the amount of electricity usage [5]. Additionally, topography and demographic factors can be analyzed as other possible external factors influencing energy use. A machine learning (ML) model, namely random forest regression**,** is implemented to make the prediction.

2 METHODS

As shown in Figure 2, our proposed framework consists of five main phases: data preparation, data exploration, data preprocessing, model training, and model evaluation. Each step is detailed in the following sections.

2.1 Data Preparation

In this study, a smart meters' dataset from Kaggle is used to make the analysis and prediction of electricity usage in London [6]. The dataset consists of several files including daily energy consumption, daily weather conditions and demographic information of each household group from November 2011 until February 2014. The number of household records for different days is inconsistent as shown in Figure 3. It may happen because the adoption of smart meters in London is continuously increasing. This could lead to the incorrect perception that the energy for a specific day is high when the data was collected for more residences. To resolve this, energy consumed across households on the same day is summed up. Hence, our main focus would be to predict energy consumption collectively, rather than to predict energy consumption for individual households. Ultimately, the energy per household is becoming the target variable for prediction.

There are many attributes provided in the 'weather' dataset. However, we only need to consider numeric variables such as 'temperatureMax', 'windBearing', 'dewPoint', 'cloudCover', 'windSpeed', 'pressure', 'apparentTemperatureHigh', 'visibility'. 'humidity', 'uvIndex', 'moonPhase', 'apparentTemperatureLow', 'apparentTemperatureMax', 'temperatureLow', 'temperatureMin', 'day', and 'apparentTemperatureMin'. Then, missing values are handled to prevent bias in the machine learning model [7]. In this case, we dropped rows that contain missing values. Since we want to check other external factors that may affect energy usage, the household groups dataset (acorn) is merged with the daily energy consumption dataset. Then, 'Acorn_grouped' and 'stdorToU' ' variables are extracted. 'Acorn_grouped' contains the name of the household group, while 'stdorToU' presents the energy measurement methods, either using the standard or Time of Unit (ToU) methods.

Furthermore, based on a subset of the smart-meter data set and various additional London datasets, a dataset that contains the electric consumption of each of the London wards was created. The subset of the smart meter dataset contains the ACORN (household affluence) representation in a variety of population categories ranging from 0 for no representation to 100 for normal representation, to even larger numbers for over representation [8]. Of particular interest from the subset was the distribution of the prevalence of each ACORN in various income brackets ranging from £ 0 - £100,000+ in £20,000 intervals. This information was utilized in conjunction with data that shows the distribution of these income brackets in households throughout London to create the average weights of which each ACORN category should influence the electric consumption of each income bracket when combined with their population in the dataset, as it was designed to reflect the actual ACORN distributions in London [9]. Then, the weights were utilized to calculate the electric consumption of each income bracket based on the electric consumption of each ACORN group.

Another dataset that contains information on the median income of all households in each of the London wards is then utilized in combination with the electric consumption of each income category to create a map of the average energy consumption of each of the London Wards [10]. This was done by mapping the average income of each ward into one of the income brackets, and then assigning the electric consumption of that bracket for that ward.

2.2 Data Exploration

The dataset is explored by utilizing visualization libraries in Python, e.g. Matplotlib and Seaborn. Our focus is to search for the relationship between weather conditions and energy consumption. Hence, the data exploration mainly showed the correlation between each weather variable and the average energy consumption for different days. The variables are temperature, humidity, cloud cover, visibility, wind speed, UV index, and dew point. In addition, we also wanted to show that changes in season affect the average electricity consumption. Moreover, a correlation matrix is generated to determine the features. Ultimately, the dataset is divided into train and test sets with the ratio of 80:20.

For visualizing the energy consumption of each of the London wards, the dataset that contains information regarding the median household energy consumption that was previously created was used in combination with another dataset that mapped the boundary of each of the London wards to plot a map utilizing geopandas [11].

2.3 Data Preprocessing

The chosen independent variables are temperature, humidity, and wind speed. Since all of them have different range of values, normalization should be performed. This step would highly affect the model's performance because it helps the model assign equal weight to all variables. Min-max scaler is one of the most popular normalization methods. It maps the minimum and maximum values of a feature to 0 and 1 respectively, while all other values are transformed to a value between 0 and 1 [12]. To implement this

technique, 'MinMaxScaler' function can be simply called from ‘sklearn.preprocessing’ library. External variables, like 'Acorn_grouped', and 'stdorToU', are later used as independent variables as well. All of them are categorical data, hence label encoding is performed. Label encoding is a technique to convert categorical data into numerical format [13]. It is important because most ML models only receive numerical variables as their inputs.

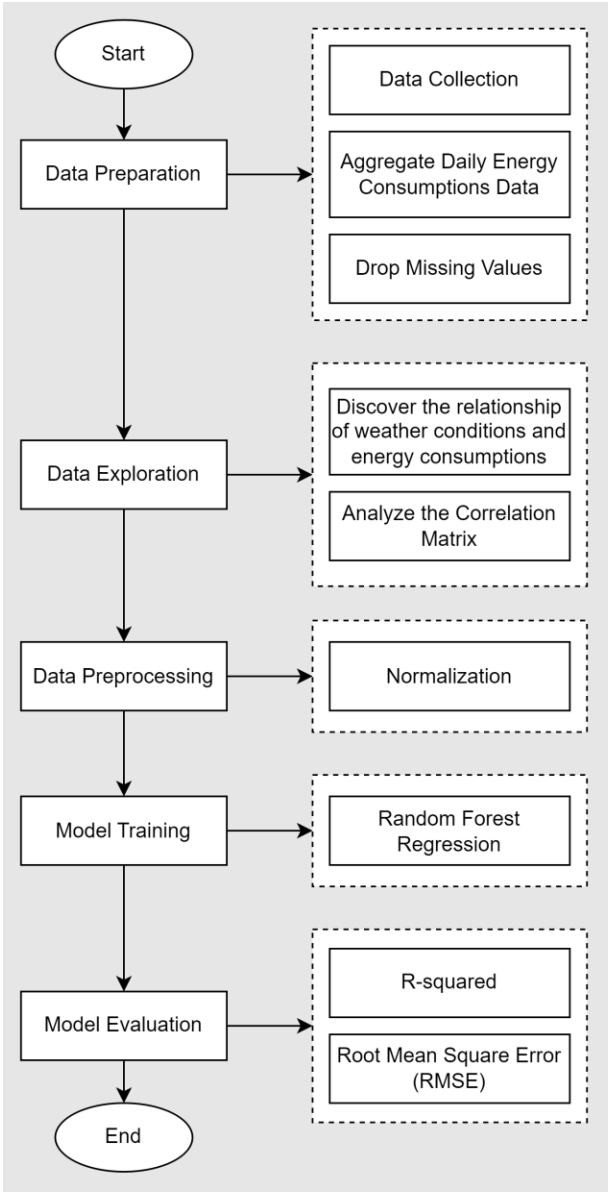

Figure 2: Research Workflow

2.4 Model Training

Random forest (RF) regression model is implemented to achieve the energy consumption prediction goal. RF regression is a supervised ML algorithm that employs an ensemble learning approach to regression [14]. In other words, it trains many models with the same data and averages their results to achieve a more accurate predictive result [15]. Bagging technique is a part of RF, meaning that the trees are running in parallel with no interaction between them. The RF model is called via ‘sklearn.ensemble’ library. To search the best combination of hyperparameters, 'RandomSearch' function is performed. Random search finds the optimal answer by mapping random combinations of hyperparameters to the model [16].

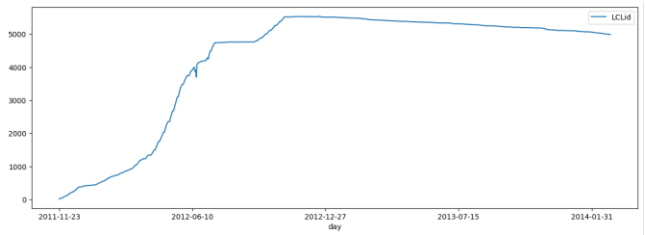

Figure 3: Inconsistent Number of Households Data Collected for Each Day

2.5 Model Evaluation

2.5.1 R-Squared. R-squared is a metric used to assess the quality of linear regression models. It represents the proportion of the variance in the dependent variable that is explained by the independent variable collectively. R-squared indicates how well the model captures the relationship between the predictor variables and the target variable, expressed as a value on a scale from 0 to 1 [17]. Typically, the higher the R-squared, the better the regression model matches the data. Nevertheless, other types of measurement should be considered as well.

2.5.2 Root Mean Square Error (RMSE). The RMSE evaluates the average disparity between the predicted values of a statistical model and the actual observed values. Essentially, it computes the standard deviation of the residuals, where residuals signify the gap between the regression line and the data points. RMSE offers insight into the dispersion of these residuals, indicating how closely the observed data aligns with the predicted values. RMSE values vary from zero to positive infinity and expressed in the same units as the dependent variable. A lower RMSE indicates a well-fitted model with more accurate predictions, while higher values indicate more error and less precise predictions.

3 RESULTS AND DISCUSSION

3.1 Exploratory Data Analysis

3.1.1 Relationships between Weather Variables and Energy Consumption

While exploring the data, we found there would be several correlations between each variable. We aim to construct a random forest regression model, it’s imperative to identify the correlations between average energy consumption and other variables. Therefore, we visualize the relationships between them through plotting.

The figures reveal that the relationship between each variable can vary from one to another. Beginning with temperature and UV Index, we observe an inverse correlation between temperature and UV Index with energy consumption. This is evidenced by the observations shown in figures 4 and 5, where peaks in energy consumption align with troughs in both temperature and UV Index.

Continuing the examination with humidity and cloud cover, as shown in figure 6 and 7, we observed that both variables follow a similar trend as the average energy consumption, with peaks aligning with other peaks and troughs aligning with other troughs. However, we can't say the same as the dew point variable, although it is a function of humidity and temperature, it doesn't follow the trend as humidity does with energy consumption. Instead, it inversely correlated with energy consumption resembling the relationship with temperature, as shown in figure 8.

On the other hand, we turn to visibility and wind speed. As shown in figures 9 and 10, both figures showed a minimal relationship between them and average energy consumption. From this observation, we can infer that neither visibility nor wind speed has a significant direct effect on energy consumption. Consequently, changes in both visibility and wind speed are unlikely to affect energy consumption substantially.

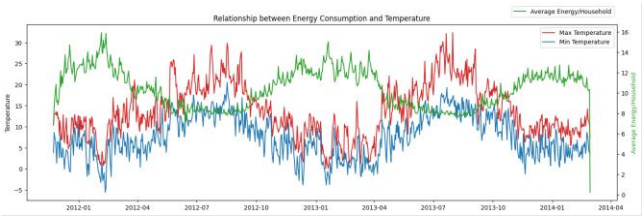

Figure 4: Relationship between Energy Consumption and Temperature

With these variables displaying different kinds of correlations, we constructed a correlation matrix, as shown in figure 11. The matrix shows us that humidity is positively correlated with energy consumption, whereas temperature exhibits a negative correlation. This alignment is also evident in figure 4 for temperature and figure 6 for humidity. Further analysis reveals other variables like dew point and UV Index shows multicollinearity with temperature, suggesting they can be excluded from our analysis. Similarly, they also disregard cloud cover and visibility due to their multicollinearity with humidity. Pressure and moon phase are also disregarded due to their low correlation with energy consumption. However, despite its low correlation, wind speed did not exhibit multicollinearity and thus included for further analysis.

3.1.2 Relationships between Seasons and Energy Consumption

Figure 12 describes the fluctuation of energy consumption related to the difference of seasons. The visualization showed that people consumed electricity most in winter. It may be due to the high usage of heaters. Meanwhile, the least electricity consumption is in summer because the weather is warmer. Spring season is the transition between winter and summer thus the energy usage is dwindled. On the other hand, the power consumption is rising during autumn.

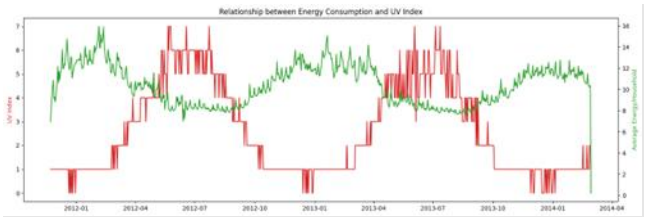

Figure 5: Relationship between Energy Consumption and UV Index

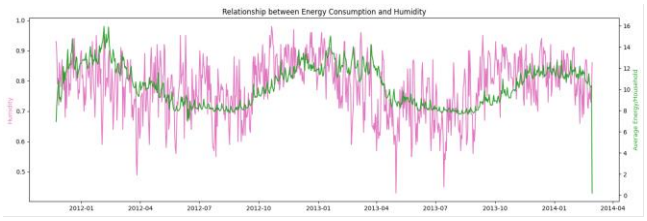

Figure 6: Relationship between Energy Consumption and Humidity

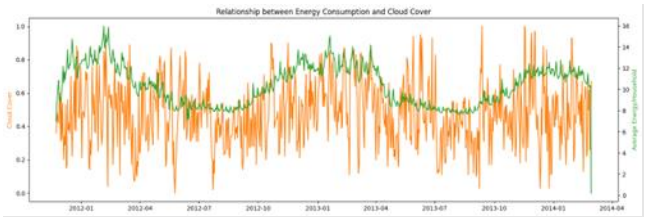

Figure 7: Relationship between Energy Consumption and Cloud Cover

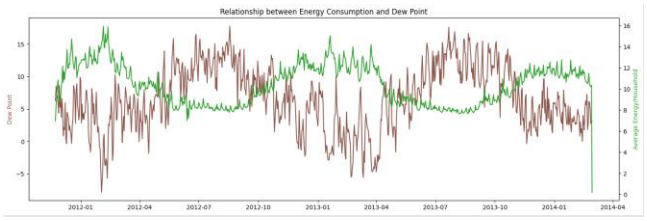

Figure 8: Relationship between Energy Consumption and Dew Point

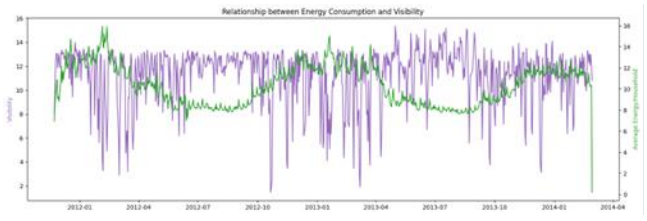

Figure 9: Relationship between Energy Consumption and Visibility

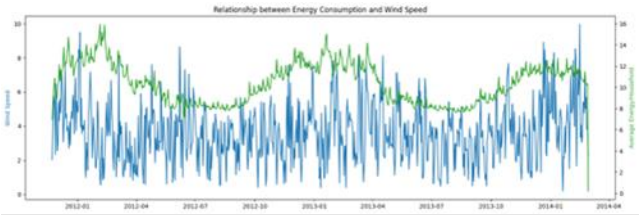

Figure 10: Relationship between Energy Consumption and Wind Speed

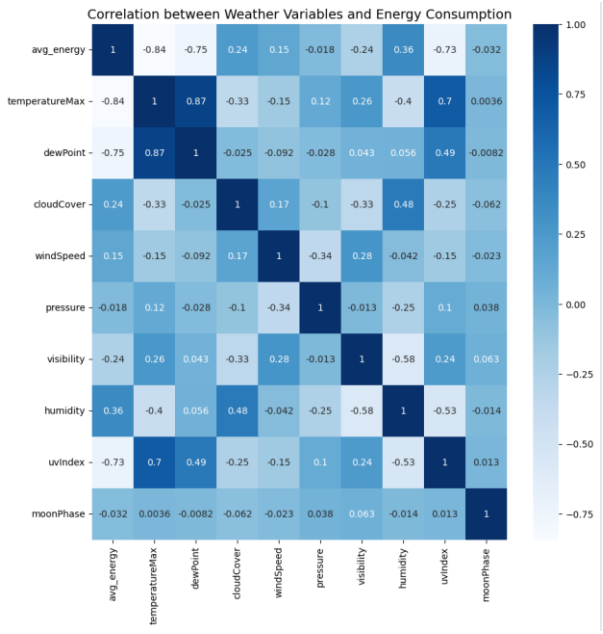

Figure 11: Correlation Matrix

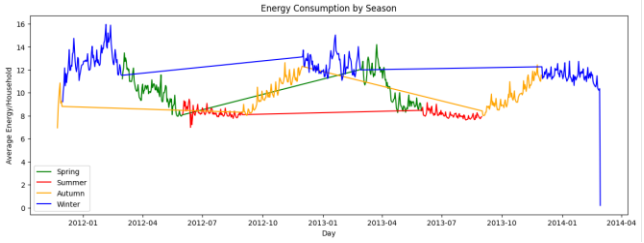

Figure 12: Energy Consumption by Season

3.1.2 Relationships between Household Variables

We also find some relationship between energy consumption and household variables. As shown in figure 13, shows the relationship between average energy and "stdorToU", the energy consumption varies from "std" and "ToU". With higher average energy in "std" and lower average in "ToU". Also from figure 14, shows the relationship between average energy and acorn group. The average energy is also distributed unevenly from each group, with the highest average energy in acorn-U.

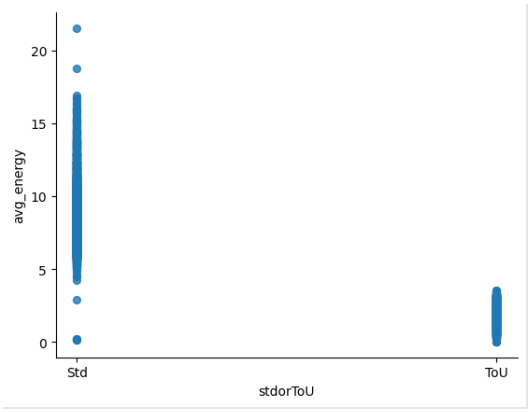

Figure 13: Relationship between Energy Consumption and stdorToU

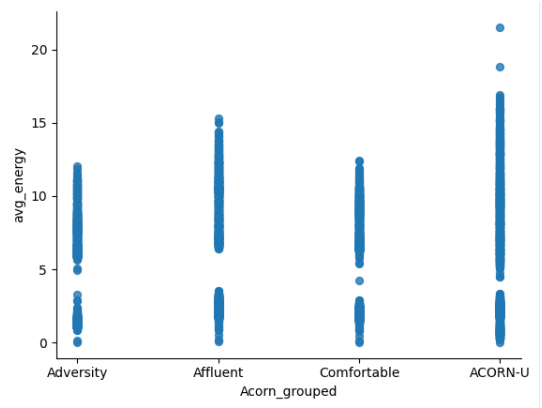

Figure 14: Relationship between Energy Consumption and Acorn Group

3.1.3 Energy Consumption in Greater London

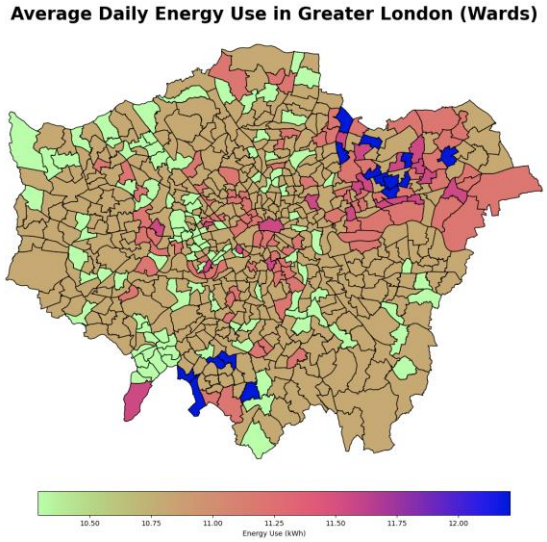

Figure 15: Average Daily Energy Use in Greater London (Wards)

As can be seen in Figure 15, the vast majority of the Greater London has a daily energy consumption that falls between 10 kwh and 11.25 kwh, but notable exceptions to this are areas in the north east and south west of London where a greater concentration of wealthier residents can be found, and where energy use falls between 11.5 and 12+ kwh.

3.2 Model Training and Evaluation

In this study, the model training is performed twice with two different combinations of features. The first experiment focuses on predicting the energy consumption based on weather variables only, namely 'temperatureMax', 'humidity', and 'windSpeed'. On the other hand, the second experiment looked on to the household groups ('Acorn_grouped'), type of measurement ('stdorToU'), and weather variables for prediction.

3.2.1 First Experiment

Random search is executed to find the best hyperparameter for the first experiment. The random search result suggested the 'max_depth' is set to 4 and 'n_estimators' is set to 996. After training and testing the RF model with these hyperparameters, it produced 0.31 of r-squared score and 1.12 of RMSE. The evaluation result is not satisfactory. As shown in figure 16, the gap between the predicted and actual values are quite large. For that reason, we assumed that other external factors must be considered which may highly affect the energy consumption measurement.

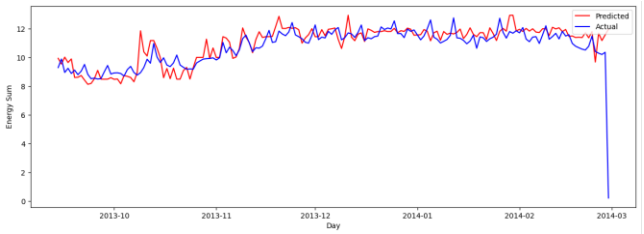

Figure 16: Predicted and Actual Values Comparison for the First Experiment

3.2.2 Second Experiment

Random searches are also executed to find the best hyperparameter for the second experiment. After doing 15 iterations, the random search has found 13 "max_depth" and 558 "n_estimators" as the best parameters. After training with the hyperparameters, we got the score of 0.94 for r-squared and 0.86 for RMSE. As shown in figure 17, we can see that it could predict good for low average energies, but we can't say the same for high average energy. This may be caused by the data distribution varies a lot between each acorn group. Overall, the model fits greatly with the data. Thus, we assume the added variables are necessary for the model to work properly.

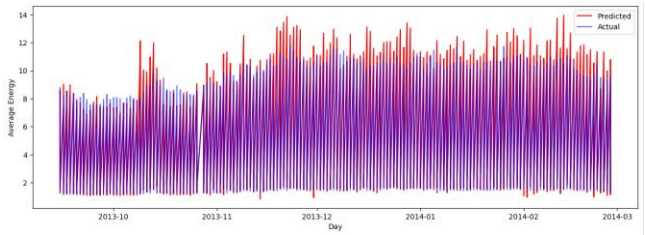

Figure 17: Predicted and Actual Values Comparison for the Second Experiment

4 CONCLUSIONS

In this paper thus far has shown the relationship between several weather conditions, and economic status on the energy use of households in Greater London. We found that the maximum temperature, dew point, and cloud cover had the strongest correlation with energy consumption, but that they were all closely related to each other, but the relatively less correlated humidity and windspeed were independent. These factors also changed with the seasons, which had a large impact on total energy use, with it peaking during the winter, and then reaching its lowest values in the summer. The Socioeconomic status of the households also influenced energy use, with more affluent households having a larger demand for energy, and destitute households having less.

The predictive model that we trained utilizing the features above to predict the daily usage of a household, as well as what type of meter each house uses. We were able to achieve an R2 Score of 0.94 and Root Mean Square Error of 0.86. Comparing this to our first experiment which only utilized weather features achieved an R2 Score of 0.31 and a Root Mean Squared Error of 1.22. Showing that the greater number of relevant features introduced to the model resulted, in a notable increase in accuracy.

With our model, individuals will now be able to judge their own electricity usage more accurately, to allow for better household budgeting. Meanwhile organizations will be able to apply a modified version of our model to the entirety of their populace to predict their total energy usage.

ACKNOWLEDGMENTS

We want to express our sincere gratitude to Asst. Prof. Ruan Sijie for his invaluable support and assistance throughout the course of this project. His expertise and guidance have been instrumental in shaping the direction of our work.

REFERENCES

[1] "Greenhouse Gas Emissions Information for Decision Making," National Academies Science, Engineering, Medicine. Accessed: Apr. 09, 2024. [Online]. Available: https://nap.nationalacademies.org/resource/26641/interactive/?gad_source=1&gclid=Cj0KCQjwiMmwBhDmARIsABeQ7xTs5wErWuupXTq2ECm6vou0WLmBc0q_c0njR0Q51ml1iaLssXR5g1EaAnG5EALw_wcB

[2] "The Glasgow Climate Pact – Key Outcomes from COP26," United Nations Climate Change. Accessed: Apr. 08, 2024. [Online]. Available: https://unfccc.int/process-

and-meetings/the-paris-agreement/the-glasgow-climate-pact-key-outcomes-from-cop26

[3] "Global energy-related greenhouse gas emissions, 2000-2022," International Energy Agency. Accessed: Apr. 09, 2024. [Online]. Available: https://www.iea.org/data-and-statistics/charts/global-energy-related-greenhouse-gas-emissions-2000-2022

[4] "Smart meters: a guide for households," Gov.UK. Accessed: Apr. 09, 2024. [Online]. Available: https://www.gov.uk/guidance/smart-meters-how-they-work

[5] S. Aisyah and A. A. Simaremare, "Correlation between Weather Variables and Electricity Demand," *IOP Conf Ser Earth Environ Sci*, vol. 927, no. 1, p. 012015, Dec. 2021, doi: 10.1088/1755-1315/927/1/012015.

[6] J. Michel, "Smart Meters in London," Kaggle. Accessed: Apr. 10, 2024. [Online]. Available: https://www.kaggle.com/datasets/jeanmidev/smart-meters-in-london/data?select=informations_households.csv

[7] N. Tamboli, "Effective Strategies for Handling Missing Values in Data Analysis (Updated 2023)," Analytics Vidhya. Accessed: Apr. 10, 2024. [Online]. Available: https://www.analyticsvidhya.com/blog/2021/10/handling-missing-value/#:~:text=You%20may%20end%20up%20building,precision%20in%20the%20statistical%20analysis.

[8] "Acorn consumer classification (CACI)," Gov.UK. Accessed: Apr 13. 10. 2024. [Online]. Available: https://www.gov.uk/government/statistics/quality-assurance-of-administrative-data-in-the-uk-house-price-index/acorn-consumer-classification-caci

[9] "Table A4 Household expenditure by gross income decile group, UK, 2014" Gov.UK. Accessed: Apr 13. 10. 2024. [Online]. Available: https://www.ons.gov.uk/peoplepopulationandcommunity/personalandhouseholdfinances/incomeandwealth/compendium/familyspending/2015/listoftablesappendixa#background-notes

[10] "Household Income Estimates for Small Areas" data.gov.uk. Accessed: Apr 13. 10. 2024. [Online]. Available: https://data.world/datagov-uk/5c4a083f-a8c6-42d8-ad40-36a9719a634c

[11] "Statistical GIS Boundary Files for London" data.london.gov.uk. Accessed: Apr 13. 10. 2024. [Online]. Available: https://data.london.gov.uk/dataset/statistical-gis-boundary-files-london

[12] S. Gupta, "How to apply Min-Max Normalization to your data," Medium. Accessed: Apr. 11, 2024. [Online]. Available: https://sourabharsh.medium.com/how-to-apply-min-max-normalization-to-your-data-f976d1633d2b

[13] "Label Encoding in Python – 2024," Great Learning. Accessed: Apr. 12, 2024. [Online]. Available: https://www.mygreatlearning.com/blog/label-encoding-in-python/#:~:text=Label%20encoding%20is%20a%20technique,only%20operate%20on%20numerical%20data.

[14] A. Chakure and B. Whitfield, "Random Forest Regression in Python Explained," Built In. Accessed: Apr. 11, 2024. [Online]. Available: https://builtin.com/data-science/random-forest-python

[15] N. Beheshti, "Random Forest Regression," Towards Data Science. Accessed: Apr. 11, 2024. [Online]. Available: https://towardsdatascience.com/random-forest-regression-5f605132d19d

[16] B. Gupta, "Random Search in Machine Learning," Scaler Topics. Accessed: Apr. 12, 2024. [Online]. Available: https://www.scaler.com/topics/machine-learning/random-search-in-machine-learning/

[17] J. Frost, "How To Interpret R-squared in Regression Analysis," Statistics by Jim. Accessed: Apr. 11, 2024. [Online]. Available: https://statisticsbyjim.com/regression/interpret-r-squared-regression/