**Topic: Energy Consumption Prediction based on Polynomial Regression and Support Vector Machines ECS784P**

**Abstract:** This research contains findings of energy consumption prediction related to various weather factors. Polynomial Regression and Support Vector Regression (SVR) models have been used as data analytics methodologies. This dataset consists of 8 dimensions that are correlated with weather measurements from 36 locations across the UK. Data management, standardization and exploratory data analysis have also been discussed and displayed in this report. Five literature reviews about energy consumption and our approaches are included in this report. An explanation on selecting these two data analytics models and their performance evaluation were included. Finally, a conclusion is made based on the analysis result.

**1. Introduction**

This project will study the prediction of energy consumption using Polynomial Regression and SVR models. This is a real-world problem, and the findings in this report can be used for future applications and analysis, ideally with more industrial level research and data analytics models approaches. The underlining reason for the project topic will be explained and discussed below.

It is known that climate change is causing significant warming every year worldwide including UK. Therefore, it is expected that UK will become warmer and wetter in the future. In which, the Met Office (2022) forecasted several changes in 2070 as below:

* Winter will be between 1 and 4.5°C warmer and up to 30% wetter
* Summer will be between 1 and 6°C warmer and up to 60% drier
* Heavy rainfall is more likely to happen.

Despite there are multiple factors interacting with the demand for energy, such as the population expansion, economic growth, behaviour of individuals and organizations, and the pace of technological development (Bas J. van Ruijven, 2019), it is no doubt that climate change is one of the most critical factors having an impact on global energy use. The link between global climate change and emissions generated from non-renewable energy resources was even proved by Khan and Arsalan (2016). Meanwhile, a study has revealed that the changes in climate 2050 will have a moderate impact on energy consumption of 7–17% globally, depending on the degree of warming (Enrica De Cian, 2019). Therefore, it is assumed that there is a certain context of the relationship between weather factors and energy consumption, and it is worth exploring more in this research.

**Objective:**

* Use at least two data analytic models on historic weather factors to produce a prediction outcome of total energy consumption.
* Conduct exploratory data analysis to understand the trend and seasonality along with the historic data.
* Explore the correlation between independent features and target feature which determines the actual value of total energy consumption.
* Highlight possible improvements for future data analysis.

**2. Literature Review**

In this report, related research and literature reviews will be presented and discussed.

The first report to be discussed is called “Exploring the Relationship between Electricity Consumption and Drivers of Climate Change: A Functional Data Analysis Approach” by *Amira Elayouty and Hala Abou-Ali*. It is a fundamental data analysis report which studied the relationships between energy consumption and CO2 emission across the globe. The purpose of this report was primarily in response to the sustainable development goals declared by the United Nations in 2016, which included the commitments to the Paris Agreement on climate change and SDG13 which concerned to “take urgent action to combat climate and its impacts.‘ This report highlighted a growing trend of CO2 emission particular in the higher-income country and concluded that a functional data analysis approach can further help quantify the gap between electricity demand and supply, as a large gap between the supply and demand in the energy sector may introduce an energy crisis. This establishes a strong foundation for our report and highlights that with enough knowledge of the features related to energy consumption, it can help to predict the electricity demand more precisely and hence improve its energy efficiency as well as carbon efficiency.

The second report that will be discussed is titled “On the Impact of Climate Change on Building Energy Consumptions: A Meta-Analysis” by *Ludovica Maria Campagnaand Francesco Fiorito.* The main aims of this report are to investigate the relationship between global warming and energy demand and to evaluate the impact of climate change on building energy consumption. As their prediction model was solely based on temperature values, they did not consider other weather variables such as humidity or solar radiation in their approach. Hence this could be a limitation in their approach as it assumed a simple linear relationship between energy demand and temperature, but the relationship is often more complicated. This report provided further confidence for our dataset as it contained different weather variables such as days of air frost, precipitation, sunshine duration, which should also be considered in our project.

The third report to be discussed is titled “Linear Model to Predict Energy Consumption using Historical Data from Cold Stores” by Majeed Safe, Birendra KC and Mahdi Safa. In this study, a multiple linear regression model (MLR) was used to predict energy consumption. It showed that the MLR model could be fitted to energy usage data and accounted for around 79% of the variance. It also discussed that the accuracy of the historical data would be a critical factor in developing a model with a minimum margin of error. This study provided strong evidence to select linear (polynomial) regression as one of the machine learning models in our project, given that it is a similar project to ours.

The fourth report that will be discussed is titled “Support vector regression (SVR) for energy consumption prediction” by Tarannom Parhizkar. This academic journal provided a piece of fundamental knowledge to us as it discussed the main difference between linear regression and Support vector regression (SVR). It highlighted that SVR can be deemed as a modified regression algorithm, in which we try to use regression to minimize the error rate, while with SVR we try to fit the error within a specific threshold.

The final report that will be discussed is titled “Using regression analysis to predict the future energy consumption of a supermarket in the UK” by M.R. Braun, H. Altan, S.B.M. Beck. It highlighted that the energy consumption of supermarkets is highly related to temperature instead of humidity. This study used multiple regression analysis based on gas and electricity data for 2012 to estimate the respective consumption in the period 2030-2059. It displayed a comparative trend where the gas use dropped dramatically compared to a modest increase in electricity use in the UK. This report proved that temperature, one of the independent attributes in our dataset, is an impactful factor affecting energy consumption in the UK. Also, it strengthened the reason why energy consumption usage in electricity generation should be chosen as a target feature, given the increasing usage of electricity in the UK.

**3. Data Management**

**3.1 Data Source and Description**

Two datasets have been merged and facilitated in this project. The first dataset, which is purely about UK weather data, was found from Kaggle where the data was sourced from the MET Office website. All data were aggregated and summarized in the average value among all UK cities on monthly basis. An attribute monthly mean temperature (tmean) was added and calculated from the average of the mean daily maximum and mean daily minimum temperature (°C) i.e. (tmax+tmin)/2. af is the days of air frost recorded that month (days); rain is the total rainfall (mm); sun is the total sunshine duration (hours).

The second dataset, which is about the consumption of electricity, was found from the UK government website (Department for Business, Energy & Industrial Strategy). The total energy consumption (includes coal, oil, gas, nuclear and hydro), which is the target attribute, was retrieved from the worksheet “Table 5.3 Fuel used in electricity generation by major producers” and concatenated into the first dataset above. The energy consumption was measured in the unit of million tonnes of oil equivalent (Mtoe). As for the merged dataset, all data were collected and displayed on monthly basis. All data are numerical data, and no categorical variable is included in this data set. The shape of the dataset is 240x9.

Table, Excel

Description automatically generated

*Table 1 Data Sample after merged two datasets*

**3.2 Dealing with Missing Data**

It is given that some sectors included a significant number of missing values. Those missing values do not denote 0 and therefore they are displayed in NA values. In the dataset used for this report (before merging), there were no duplicated values but an unexpectedly high number of missing data for independent attributes (2532 missing values out of 8116 in sun variable), and they can cause significant effect in the following analysis. Although there are numerous ways to deal with the missing data such as using the drop.duplicate() function or establishing a predictive model to refill the missing values. Given the significant large proportion of missing data, the former method can result in biased data and thus affect the accuracy of analysis, while the latter one is too complicated in the scope of this project. Therefore, fillna.mean() function was simply used, which is to replace the missing value with the mean for continuous data, as it is the simplest way to deal with this problem.

**3.3 Feature Selection**

There are multiple worksheets and features in the second dataset, but only the variable of energy consumption (total fuel used in electricity) was chosen to be the target attribute as it is the most representative data. Both attributes year and month were reserved because they contain the features in terms of trend and seasonality in our data respectively. Thus, they can be used as X variables in the following supervised model.

**3.4 Times Series Decomposition**

Time series decomposition is a technique to extract multiple types of variation from our dataset. The chart below shows a downward trend and a strong seasonality. Details are also discussed in the exploratory data analysis session. Graphical user interface

Description automatically generated with medium confidence

**3.5 Data Standardization**

Many techniques can be used to make the features follow a normal distribution, such as scaling, normalization, and logarithmic transformers. According to the official Sklearn guide (2021), “if a feature has a variance that is orders of magnitude larger than others, it might dominate the objective function and make the estimator unable to learn from other features correctly as expected.” To achieve higher performances in the following data analysis, the scaling method was used for all attributes in the dataset, where it can transform numerical features to have a mean of 0 and a variance of 1. Meanwhile, log transformation was used for attribute af and rain, as both followed a skewed distribution by observation of histogram.

**4. Methodology**

Given that our dependent variable is continuous data, polynomial regression and sector support regression have been used as the regression model to forecast the amount of energy consumption. They are a set of supervised learning methods used for the machine learning task. Although they may not be the best models when dealing with time-series data, it is known that supervised models can still be used for time series, if the seasonality can be extracted and put it into a variable (Joos Korstanje, 2021). The explanation and reasons behind selecting these models are explained as follows:

**Polynomial Regression Justification**

Polynomial Regression is a variation of linear regression model which can be deemed as a curved line to estimates the relationship as an nth degree polynomial. Since our data has many fluctuations between each month and year, in which several features of trend and seasonality were observed in the exploratory data analysis (next session), we should not solely use a simple linear regression model to predict the target values in a straight line. To prove this justification, the model performance of both simple linear regression and polynomial regression were tested and compared, in terms of RMSE (root mean square error) and accuracy of the model (*R*2). As the results, RMSE and *R*2 for linear regression were 0.896 and -1.349, while they were 0.36 and 0.621 for polynomial regression respectively. Simply say, a smaller value of RMSE and a higher value of R2 accuracy implies a better model. Hence, we could conclude that the polynomial regression model demonstrated higher goodness of fit and better performance in our project.

A picture containing text

Description automatically generatedText

Description automatically generated with low confidence

**SVR Justification**

Support vector machine (SVM) is a popular supervised learning model developed by Valadimir Vapnik used for data classification and regression (IBM Cloud Education, 2020), where SVR can be regarded as a combination of SVM and Regression. Given that we don’t need classification in our data, SVR is used to predict continuous numerical outcomes. As the algorithm, SVR tries to minimize the error within a certain threshold between two boundary lines. One of the advantages of using SVR is that it can predict values within a nonlinear threshold, given that our dataset is in a non-linear structure (Drew Scatterday, 2019).

Diagram

Description automatically generated

*Figure 1 Illustration of how SVR works*

**5. Exploratory Data Analysis**

Time Series decomposition Is conducted below. Once plotting the values of total energy consumption into multiple line graphs by years, an obvious seasonality and trend could be found. Both graphs showed similar seasonality, in which the energy consumption often reached the peak from December to March (winter) and maintained a low from April to August (summer). That is probably because people use more heaters in the winter season. On top of the summer/winter seasonality, both graphs showed a downward trend in UK energy consumption particularly from 2003 to 2019. This trend could be due to global warming, which forces people to use less energy to keep warm in a warmer climate. Also, it was found that the energy change among each month was decreasing, when comparing different years in the two graphs below. It can give us some insight overall, that is the energy consumption efficiency, or the related technology may be improving in the UK. Chart, line chart

Description automatically generatedChart, line chart

Description automatically generated

After making bins for the mean temperature and total energy consumption based on the values in it and analyzing the corresponding energy consumption level for each bin, a stacked bar chart was produced. It showed that the mean temperature has a negative association/correlation with energy consumption. In another word, when the mean temperature is lower, energy consumption is higher. This result also aligned with the previous conclusion above.

Chart, bar chart

Description automatically generated

A correlation matrix was visualized to find out the correlation among all attributes in the dataset.

Pandas corr() function was used and only the correlation between target attribute and other attributes has been shown as below. The heatmap was labelled with the intensified colour and sorted in descending order. It is shown that ‘af’

days of air frost has the highest correlation score (r) when compared to others. Also, most of the attributes displayed a negative correlation.

A picture containing table

Description automatically generated

To further dig deep into the correlation among attributes, some methods such as kbest from the sklearn library have been studied and compared. As chi-squared test can only be used for computing statistics between each non-negative feature and class, it is only valid for classification but our target value (energy consumption) contains continuous values. Therefore, two methods namely f\_regression in kbest and mutual information regression were used separately to find the highest predictive features of 'y'. As shown in the tables below, both tables indicated that the feature ‘rain’ has the lowest score and thus it could be deemed as an unimportant attribute. However, this feature is still retained as it is believed that more weather variables can better predict energy consumption according to the literature review discussed previously.

**Table

Description automatically generated Table

Description automatically generated**

*Table 2 Correlation matrix using kbest (left) and mutual information regression (right)*

**6. Analysis, testing, results**

This section will explain how to implement the models, and how the model performances were evaluated. However, it cannot be guaranteed that both models perform perfect accuracy due to the complex structure of time series data.

**Polynomial Regression Results**

The implementation of both models is relatively simple. The algorithm and coding are based on the python library sklearn. It is a library that contains multiple regression and classification models, which particular includes polynomial regression (varied from linear regression) and SVR (varied from SVM) in our project.

In detail, the first step is determining the degree of the polynomial and creating the new features. Here the degree is set to 2, as we want to work with a 2nd degree polynomial. A typical 2nd degree polynomial formula is y = ax2 + bx + c, while we have 8 independent variables so the formula will probably be much longer and more complex. After that, we can fit and transform the independent variables (X variable) into polynomial features.

The second step is using the train\_test\_split() method to split the data into train and test data. It is noted that as our data is time series in which data are autocorrelated and contain trend and seasonality that discussed previously, we should split the data by time instead of shuffling the data randomly. Here we split the train and test data in a proportion of 8:2.

Once the data are transformed into polynomial features and split into training and test data, we can initialize linear regression and use it to fit the data with the given function. Linear regression is imported because polynomial regression is a variation of linear regression, which was discussed previously in the methodology session.

Graphical user interface, text, application

Description automatically generated

After the model is executed, we try to evaluate the performance of model by several indicators. The first performance measure is called root mean squared error (RMSE), a commonly used metric. RMSE is the square root of the average squared errors, which is the difference between observed and predicted data. In general, a lower RMSE means better performance, as zero RMSE indicates no error.

The second performance measure is the score() method. It is also a frequently used metric to evaluate the performance of a regression-based machine learning model about the coefficient of determination – *R*2. Here the RMSE and R2 are 0.36 and 0.621 respectively, which implies a good performance without overfitting.

As an advanced evaluation matric, cross-validation can avoid overfitting problem, and evaluate model performance in a more robust way than a simple train-test. There are various techniques in cross-validation, such as k-fold, LeaveOneOut (LOO), Shuffle Split. However, using any of them to evaluate the training model can be problematic as they assume each observation is independent, given that it is not true on our time series data. It was observed that our dataset apparently has both trend and seasonality in those 20 years, whereas k-fold cross-validation does not consider these time structural changes. For time-series data, it is important to evaluate our model on the “future” observation least like those that are used to train the model (Scikit-learn, 2021). Therefore, time series split cross-validation was introduced as the third indicator.

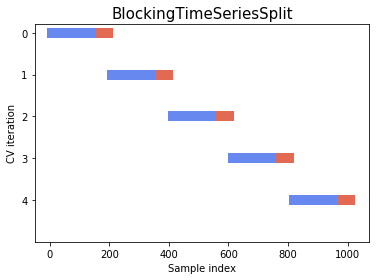
Since the data have already been sorted by date, we can simply use TimeSeriesSplit() method to split the training and testing data into multiple segments on a rolling basis. It is called ‘rolling basis’ because of the “origin” at which the forecast is rolling forward in time. By doing so, the temporal dependency between observations can be preserved during testing. In our testing, we set the number of rolling (k- fold) to be 10 for both time series cross validation and blocked cross validation. It means the cross-validation would perform the fitting procedure a total of 10 times, which also is standard practice.

**Chart, bar chart

Description automatically generated**

*Figure 2 Visualization of time series split cross validation*

As for the fourth indicator, blocked cross-validation was introduced for better comparison. This method has a different approach and is used to avoid leakage from future data by adding margins at two positions.



*Figure 3 Visualization of blocked split cross validation*

The test result showed a contractual difference between the first two and last two indicators. As discussed in the methodology session, the lower RMSE (0.36) and higher R2 accuracy (0.62) could indicate better performance. However, both time series cross validation and blocking time series cross validation showed a negative score (-11 and -15). It means the polynomial regression may not work fit on time series cross-validation.

Graphical user interface, text

Description automatically generatedChart

Description automatically generated

*Figure 3 Comparison of observed data and predicted data with a line chart (blue = observed, red = predicted)*

**Support Vector Regression (SVR) Results**

For our second machine learning model SVR, all implementation steps are almost the same as the polynomial regression. After we imported the SVR model from the sklearn library, we can simply train the model with the training data.

Similarly, to access the model performance, the model was tested with the four indicators again. Both kernels in ‘linear’ and ‘rbf’ showed similar results, but only the ‘linear’ result is demonstrated below for simplicity. Simply say, the test result showed a contractual difference between the first two and last two indicators again. Although SVR did not perform as well as polynomial regression in terms of RMSE and R2 accuracy, it has a relatively higher score in both cross-validation metrics. It implied that the prediction of SVR does not work as well as polynomial regression in our test data, but it may work well in time series cross-validation. In other words, it may work better on the past data.

Graphical user interface, text

Description automatically generated

Chart

Description automatically generated with medium confidence

*Figure 3 Comparison of observed data and predicted data with a line chart (blue = observed, red = predicted)*

**7. Conclusion**

Overall, the aims of the project have been achieved.

Regarding the limitations, it is known that time series forecasting is a complex subject topic in machine learning. For instance, as there is an implicit dependence on previous observations, we cannot simply split the data randomly and hence classical machine learning approach may not work well on time series. Likewise, we cannot use typical k-fold cross validation to access the models due to the unique properties of time series data. Otherwise, it will create biased and irrelevant results.

Due to the scope of our project, only supervised machine learning models were considered. However, there are various fantastic models that we didn’t consider in this project, like the classical time series models or even the deep learning architectures. For instance, the ARIMA family is a popular set of time series models that focuses on relations between the past and the present data. It is believed that using more machine learning models can greatly improve the results and improve the overall quality. Moreover, if the source of daily-basis data is available, it may even create more insights for us as it probably contains more seasonality factors like the difference between daytime and night, given that we often use more energy in the daytime and less at night.

**References**

Amira Elayouty and Hala Abou-Ali (2021) Exploring the Relationship between Electricity Consumption and Drivers of Climate Change: A Functional Data Analysis Approach”.

Met Office (No Date) Effects of climate change. Available at: https://www.metoffice.gov.uk/weather/climate-change/effects-of-climate-change (Accessed: 2 Mar 2022).

Barten, Anton P. (1987). "The Coefficient of Determination for Regression without a Constant Term". In Heijmans, Risto; Neudecker, Heinz.

Bas J. van Ruijven, Enrica De Cian & Ian Sue Wing (2019) Amplification of future energy demand growth due to climate change. Available at: [https://www.nature.com/articles/s41467-019-10399-3](https://www.nature.com/a) (Accessed: 2 Mar 2022).

Drew Scatterday (2019) Walking through Support Vector Regression and LSTMs with stock price prediction. Available at: https://towardsdatascience.com/walking-through-support-vector-regression-and-lstms-with-stock-price-prediction-45e11b620650

(Accessed: 11 Mar 2022)

Enrica De Cian & Ian Sue Wing (2019) Global Energy Consumption in a Warming Climate. Available at: [https://link.springer.com/article/10.1007/s10640-017-0198-4](https://link.springer.com/article/10.1) (Accessed: 2 Mar 2022).

GOV.UK (2021) Energy Trends electricity tables (ODS). Available at: <https://www.gov.uk/government/statistics/electricity-section-5-energy-trends> (Accessed: 2 Feb 2022).

IBM Cloud Education (2020) What is Supervised Learning? Available at: <https://www.ibm.com/cloud/learn/supervised-learning#toc-supervised-QVA1W1YW> (Accessed: 11 Mar 2022)

Joos Korstanje (2021) How to Select a Model For Your Time Series Prediction Task. Available at: <https://neptune.ai/blog/select-model-for-time-series-prediction-task> (Accessed: 10 Mar 2022).

Kaggle (2020) UK MET Office Weather Data. Available at: <https://www.kaggle.com/josephw20/uk-met-office-weather-data> (Accessed: 4 Mar 2022).

Khan, J. and Arsalan, M.H. (2016) Solar power technologies for sustainable electricity generation-A review, Renewable and Sustainable Energy Reviews, 55, 414425.

Ludovica Maria Campagnaand Francesco Fiorito (2022) On the Impact of Climate Change on Building EnergyConsumptions: A Meta-Analysis.

Majeed Safe, Birendra KC and Mahdi Safa (2015) Linear Model to Predict Energy Consumption using Historical Data from Cold Stores”.

M.R. Braun, H. Altan, S.B.M. (2014) Using regression analysis to predict the future energy consumption of a supermarket in the UK”.

Packt Editorial Staff (2019) Cross-Validation strategies for Time Series forecasting [Tutorial]. Available at: <https://hub.packtpub.com/cross-validation-strategies-for-time-series-forecasting-tutorial/> (Accessed: 11 Mar 2022)

Scikit-Learn (2021a) sklearn.feature\_selection.mutual\_info\_regression. Available at: [https://scikit-learn.org/stable/modules/generated/sklearn.feature\_selection.mutual\_info\_regression.html#sklearn.feature\_selection.mutual\_info\_regression](https://scikit-learn.org/stable/modules/generated/sklearn.feature_selection.mutual_info_regression.html%22%20%5Cl%20%22sklearn.feature_selection.mutual_info_regression) (Accessed: 2 Mar 2022)

Scikit-Learn (2021b) sklearn.feature\_selection.f\_regression.

Available at: [https://scikit-learn.org/stable/modules/generated/sklearn.feature\_selection.f\_regression.html#sklearn.feature\_selection.f\_regression](https://scikit-learn.org/stable/modules/generated/sklearn.feature_selection.f_regression.html%22%20%5Cl%20%22sklearn.feature_selection.f_regression) (Accessed: 2 Mar 2022).

Scikit-Learn (2021c) sklearn.model\_selection.cross-validation. Available at: <https://scikit-learn.org/stable/modules/cross_validation.html> (Accessed: 11 Mar 2022)

Tarannom Parhizkar (2021) Support vector regression (SVR) for energy consumption prediction.

Tamas Ujhelyi (2021) Polynomial Regression in Python using scikit-learn (with a practical example). Available at: <https://data36.com/polynomial-regression-python-scikit-learn/>

(Accessed: 11 Mar 2022)