ECEN 489-503: Final Project Report

Addison Maupin and Yunfei Xie

Due: April 30, 2020

ELECTRICAL & COMPUTER ENGINEERING
TEXAS A&M UNIVERSITY

Addison Maupin
Yunfei Xie

Table of Contents

Overview and Purpose:

This project will analyze the data from smart meters from 5,567 London households that took part in the Low Carbon London project led by the UK Power Network [1] between November 2011 - February 2014, for a total of 27 months. The data "seems associated only to the electrical consumption" [1]; however, this report will aim to find relationships between the daily energy usage and the demographics and tariff type of the household and predict future energy consumption for certain households.

Specifically the following two questions will be answered: 1) Can the average energy consumption of a household be predicted from their Acorn group and tariff type, and 2) Can the future energy consumption of a household be predicted from their past consumption?

The economic condition of a household will be measured by the provided Acorn type which classifies neighborhoods into 16 different categories, from A to R [2]. This data set also contains Acorn type U and Acorn type -, neither of these types have an explanation in the dataset and thus have been taken to mean unclassified and will be filtered out during all analysis.

The tariff type is either a standard tariff or a time of use tariff. A standard tariff has a default price that can vary with the market [3] while a time of use tariff charges cheaper rates at off-peak times and higher rates at peak times [4]. A time of use tariff is often cited as being more environmentally friendly and cheaper for the end consumers [4].

In our midterm project we established there was a relationship between Acorn groups and the energy consumption of a household but there was very little correlation between tariff types and energy consumption. Since tariff types did not have a strong enough correlation with energy consumption to be used on its own we used both Acorn groups and tariff types to predict the average energy consumption. For our second question we decided to do a time series analysis to see if we could predict the future energy consumption for individual households.

The project uses the previously collected data from the midterm project but assigns Acorn groups and tariff types integer values to be used in the models. Each Acorn group has a value from 1 to 18 with group A being 1 and group U being 18; however, as mentioned previously, group U will eventually be dropped as it is not a recognized Acorn group. The standard tariff was assigned a value of 0 and time of use tariff is represented by 1.

After this data was prepared, two different models were built, one was a simple regression model and one was an ARIMA model for time series. The results from these models and shown and discussed in the following sections.

Technical Instructions:

The intOutput.csv, which contains aggregated data including the daily household energy average and Acorn groups, and the Jupyter Notebooks for the two models have been provided for convenience. The results can be obtained by running the Jupyter Notebooks which contain detailed comments explaining the code and how to recreate the results. The libraries used are pandas, matplotlib, Scikit-Learn, seaborn, statsmodels, and numpy. Please note that the data set must be saved in the same folder to run the Jupyter Notebooks. All 19 files from the data set are needed.

Outcomes and Results:

The project's outcomes can all be viewed in the submitted Jupyter Notebooks. The intOutput.csv file contains all the aggregated data including the daily household average, the total household average, and the Acorn group's daily energy usage as well as the Acorn group and tariff type in integer values. There are two different models, Model1.ipynb and Model2.ipynb. Their specific outcomes are shown below.

Model 1:

Our first model is a multiple linear regression model which takes in two inputs: Acorn group and tariff type and produces one output: the daily average energy consumption for a household. To validate this model we split our dataset into 80% training and 20% testing subsets. We used Scikit-Learn to perform multiple linear regression and then predict the average energy consumption based on the Acorn group and tariff type [5].

We first plotted a histogram of the data to help visualise the data.

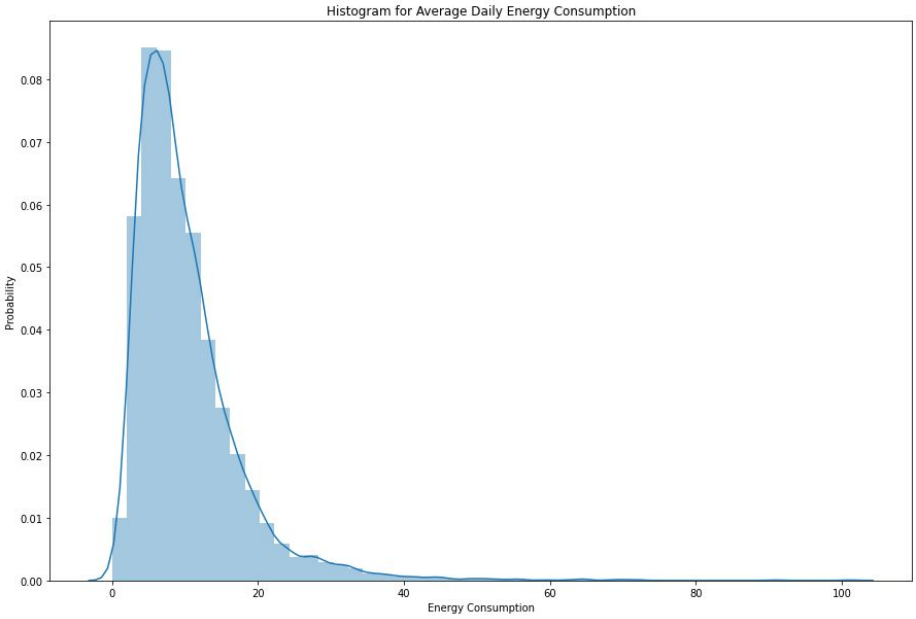

Figure 1: Histogram for Average Daily Energy Consumption

From this graph we can see that most households consume around 5-15 kW of energy per day. Furthermore we see that this data follows a fairly normal distribution which means it is a good dataset for linear regression [5].

After splitting the dataset into the training and testing subsets the model was trained and obtained the following results for the regression equation.

```
Intercept =  12.813098166961058
Coefficient for Acorn Types =  -0.28282077218214713
Coefficient for Tariff Types =  -0.9957565929732806
```

Figure 2: Parameters for Regression Equation

After the regression model was trained it was simple to begin predicting values using the test set. The results of the predicted daily average energy consumption are shown below as a table and a graph. Note that the table only shows a few readings.

	Actual	Predicted
0	12.187364	12.530277
1	12.747934	10.120417
2	11.448270	11.681815
3	3.330599	9.136428
4	5.736595	8.287966
...	...	...
1103	3.402794	11.398994
1104	7.579990	9.702070
1105	13.057939	11.398994
1106	1.962596	8.005145
1107	7.468176	11.681815

1108 rows × 2 columns

Table 1: Actual vs Predicted Results for Daily Average Energy Consumption

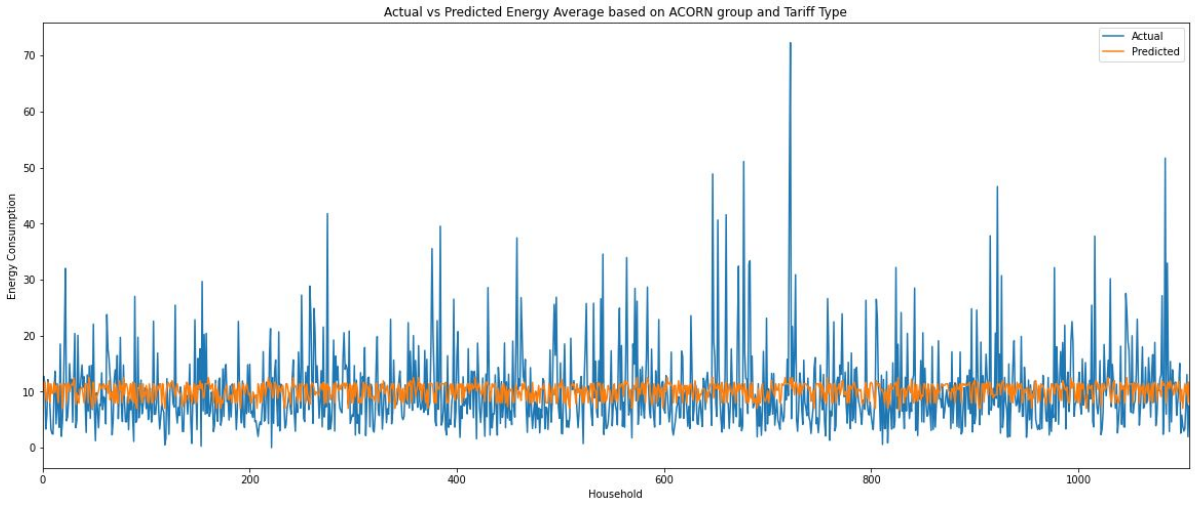

Figure 3: Actual vs Predicted results for Average Daily Energy Consumption

As we can see from this graph, the program is predicting a fairly constant range of values between 8 to 13 kW roughly. Initially this was assumed to be a poor result as the predicted values did not follow closely

enough with the observed trend; however, to check the accuracy of this model three statistical tests were performed.

```
R2 coefficient:  0.05522760600452681
Mean Absolute Error:  4.943243057490394
Mean Squared Error:  49.97461952840958
Root Mean Squared Error:  7.069272913702624
NRMSE:  0.06991319699058125
```

Figure 4: Statistical test for Regression Model

The R^2 coefficient is a measure of how well the model fits the data. The R^2 value can range from 0 to 1 and the higher the value the better as more variance is explained by the model [6]. From the figure above we can see we have a very low R^2 coefficient which explains why our model is predicting a relatively small range of numbers since it will not respond to variance as well. However, our RMSE is fairly low considering the original dataset has a range of 0 to 101.115 kW. This indicates our model is fairly accurate. We can normalize this value using the difference of the range to obtain a Normalised RMSE of 0.0699 which again indicates high accuracy [6]. Therefore even though our model is not very reactive to variances, it can still be considered accurate.

Model 2:

Our second model is a time series analysis of individual households. In this report only one household will be discussed: MAC000239. This household was chosen as it was in a “Comfortable” acorn group and had enough daily readings to properly analyse. In the submitted Jupyter Notebook there are two other households which can be uncommented and run.

A time series analysis has three main components: visualising the data to find any trends and seasonality, stationarizing the dataset, and fitting an ARIMA model to learn and predict the pattern of the series [7].

First the dataset was plotted to find any trends and seasonality. From the graph below we can see that energy usage tends to spike in the winter and then decrease in the summer.

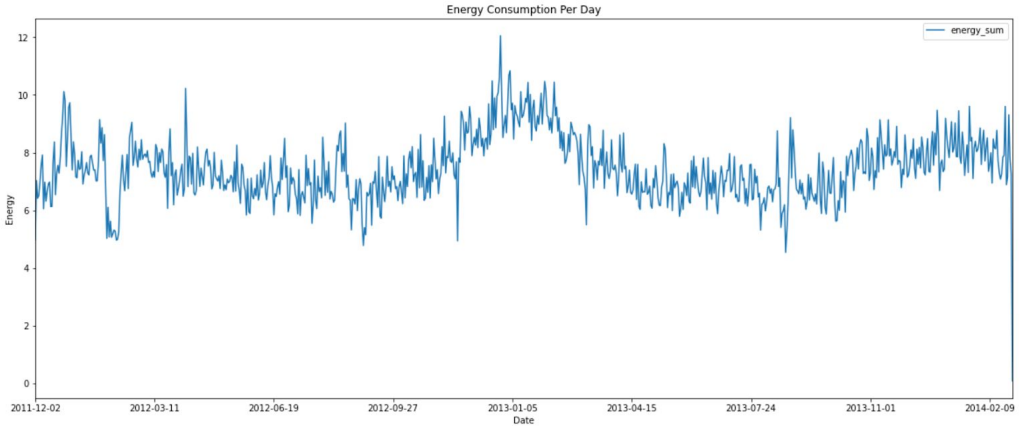

Figure 5: Daily Energy Consumption

To explore the seasonality and trends more, the seasonal_decompose method was used. This analyzed seasonality using a frequency of 365 days to find the yearly trends. The results are shown below.

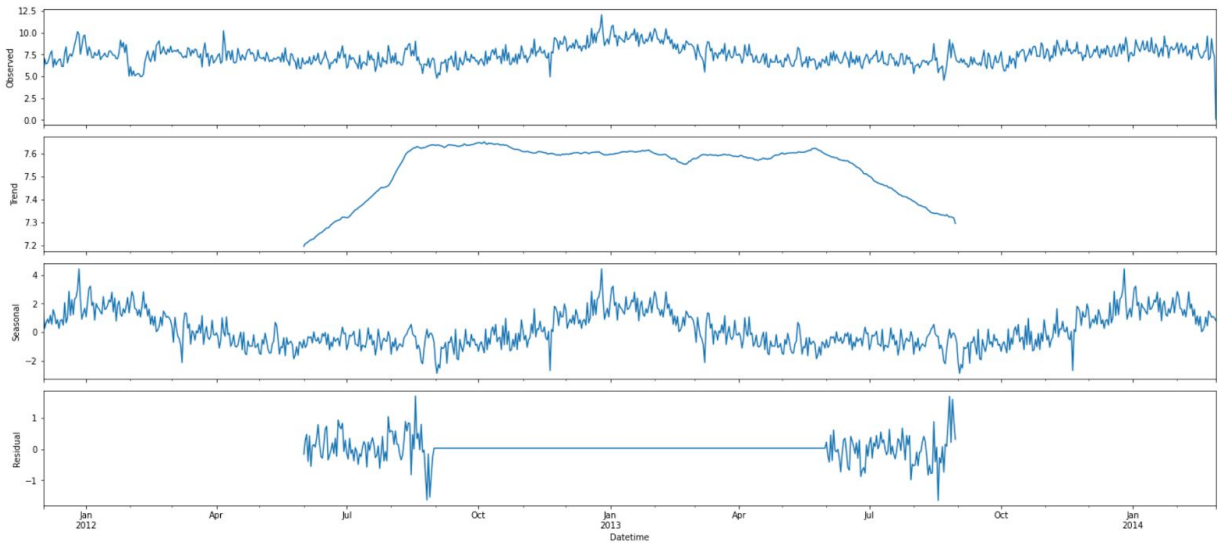

Figure 6: Seasonal and Trend Data

We can also look at the lag plot for energy consumption which yields important information about the autocorrelation of the dataset. From the graph below we can see that the dataset is weakly autocorrelated [8]. Since this dataset only has weak autocorrelation our time-series analysis might not yield as highly accurate results as possible but we should still be able to perform a decent analysis [7].

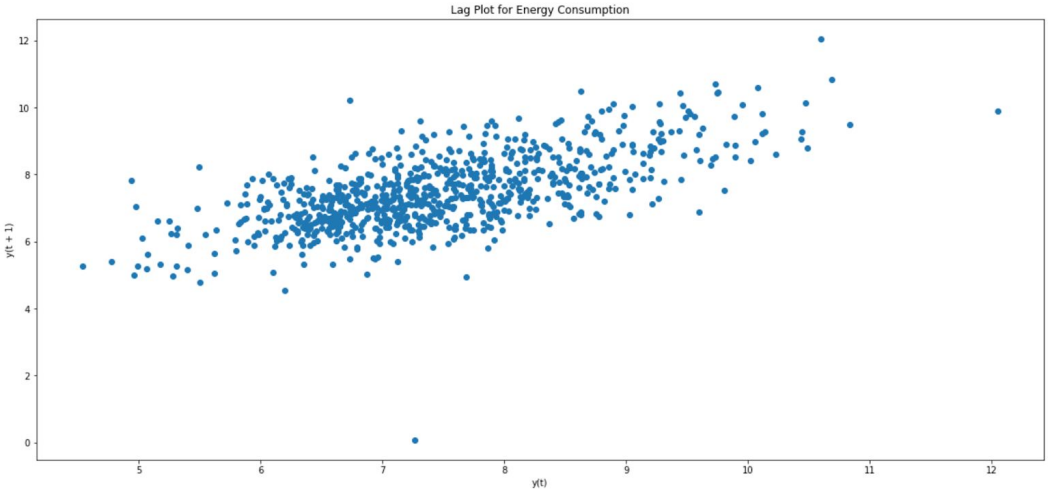

Figure 7: Lag Plot for Energy Consumption

After understanding the seasonality of this data we can then begin to test whether it is stationary or not. Stationary data is not dependent on time and therefore has constant mean and variance. There is a simple test to determine if a dataset is stationary called the Augmented Dickey-Fuller test [9]. This is a hypothesis test with the null hypothesis being that the series is not stationary. If the p-value is greater than 0.05 this null hypothesis will be rejected and the series will be assumed to be stationary [9]. The results for this dataset after the Augmented Dickey-Fuller test are shown below.

```
ADF Statistic:  -4.090461768550885
p-value:  0.0010042880689532883
Critical Values:
        1%: -3.438
        5%: -2.865
        10%: -2.569

Reject Null Hypothesis: Series is stationary
```

Figure 7: Results from Augmented Dicky-Fuller Test

Since this dataset is considered stationary, we can then begin to fit an ARIMA model to predict future values. An ARIMA model consists of autoregression terms represented by p, integrated terms represented by d and moving average terms represented by q [10]. The best fit for this model was determined to be (p, d, q) = (19, 1, 0) . The model was then predicted approximately 5 months into the future. The results are shown below.

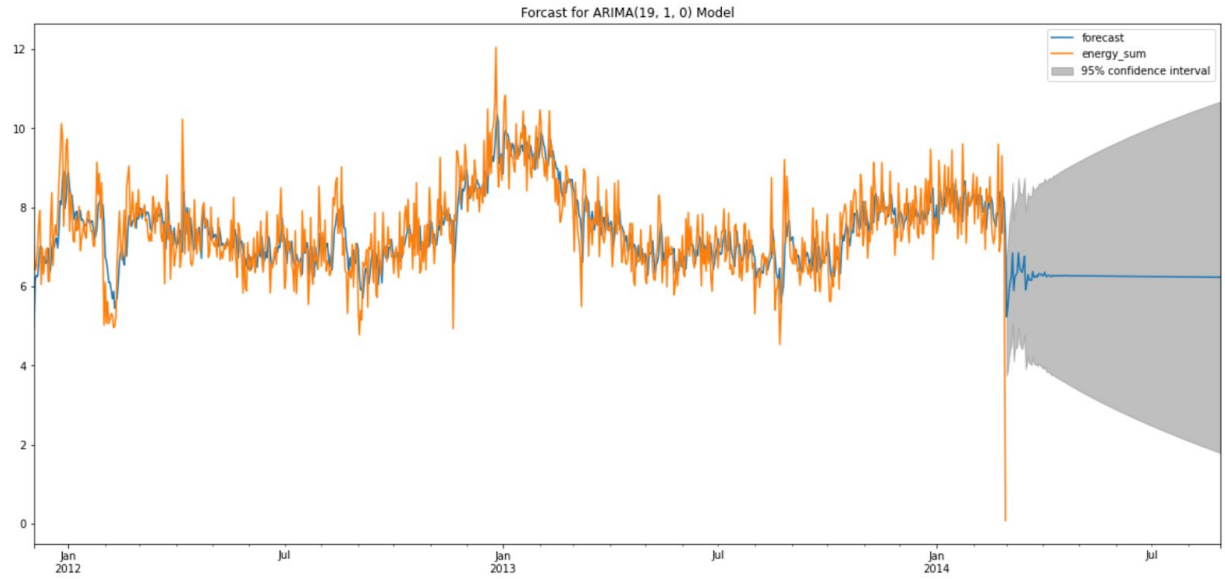

Figure 8: Forecast vs Actual results for Energy Consumption

As you can see, the model tracks very well with the actual results. Even though the forecast drops off to a straight line for future data, the model can still predict the energy sum up to the 95% confidence interval. The 95% confidence interval becomes so large that it is hard for the model to give out precise prediction, thus the predicted result is just at the middle of the confidence interval. Therefore, we can deduct that with a credible interval, the model works well for predicting the energy sum.

Results:

In general, our results obtained from the two models are accurate and correct. We were considering using the data from the Acorndetails.csv, however, due to the fact that the dataset is not concise or organized enough, we could not put those data as parameters into our model. We also had no way to track the data stored in this file back to specific households. If this data could be utilized it would vastly improve our first model as we would have many more parameters which would overall improve our accuracy. Overall, we used most data from the original datasets and translated those data into the standardized form to build two different models. With various graphs and statistical information, we can conclude that our two models work very well both in accuracy and correctness.

Weekly Reports:

Week 1: 23 March - 30 March

Progress Made:

This week we focused on figuring out our non-trivial questions. For the midterm project we looked at 1) the relationship between the daily energy usage and the economic condition of a household and 2) the relationship between tariff types and energy consumption. For the final project, we have currently decided to build a model that will predict two things: the economic condition from the total energy usage and the tariff type from the total energy usage. This is subject to change if we feel like we have a better question in mind.

Issues Encountered:

We are trying to figure out exactly how to train the model since we mainly have classified images before and not just data contained in csv files. So far, we have had issues using Keras with Pandas and getting our data into readable formats.

Solutions Tried:

We began by adapting our Homework 2 model to work with the data we have for this project. This was eventually successful, but more testing is needed as we are not sure if this model is correct since so far it has always predicted the same result.

Plans for Next Week:

For next week we will clean up our dataset and ensure it is readable by our current model based off of Homework 2. Our current model needs to be tested more and made more accurate. We are also debating changing our questions as predicting the tariff does not seem to be accurate due to a lack of correlation in the data. We will hopefully have our final questions settled by next week.

Week 2: 30 March - 6 April

Progress Made:

We built a model to classify households into ACORN groups based on their daily average energy usage. We also ensured all of our data was readable with the Keras and Pandas library and assigned integer values to both the tariff type and ACORN group.

Issues Encountered:

We ran this model with 20 epochs and only had an accuracy of around 0.0003%. This is obviously an issue although we are not sure if this is an error with our model, a lack of data, or a lack of correlation between the data sets. Furthermore this model always predicts the same result which happens to be the first ACORN type in the ouput.csv file.

Solutions Tried:

We are a little lost on how to fix this but plan to ask Dr. Tao the following questions: 1) Is there a way to fix our model so it is not alway predicting the same result and 2) is there a way to use the acorn_details.csv to influence our model? With acorn_details.csv, we need to ensure there is a way to tie the data back to the household.

Plans for Next Week:

We will continue to work on this model to improve its accuracy. We will do this through further research and reaching out to Dr. Tao. We also need to finalize our non-trivial questions as we are worried our original two questions will not be predictable with our model.

Week 3: 6 April - 13 April

Progress Made:

After discussing with Dr. Tao we decided to restart our model to ensure our input dimensions were correct. We changed the model to take in 4 inputs and produce 1 output. Our network is currently structured as shown below:

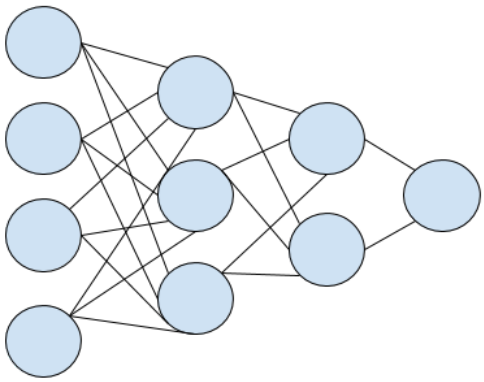

Figure 9: Network Architecture

We have chosen to evaluate our model loss with Mean Square Error and used ADAM as an optimizer. The dataset has been split into training data and validation data. The dataset consists of: energy sum for

the household, daily energy average for the household, the daily energy avergy for the ACORN group of the household, and the tariff type for the household. The output is the household's ACORN group.

Issues Encountered:

We are still getting very low accuracy rates ranging from 0.06% to 0.04%.

Solutions Tried:

I have tried to take out layers, reducing the neural network to be just an input and output later or an input, hidden, and output layer. I have also tried increasing my training epochs; however, neither of these changes increased my accuracy. I also reduced the number of inputs into my model limiting it to two: daily energy average for the group and tariff type. This again had an accuracy of about 0.04%.

Plans for Next Week:

We plan to do more research on improving the accuracy of models as we are unsure of how to continue from here.

Week 4: 13 April - 20 April

Progress Made:

After discussing with Dr. Tao, we decided to build a linear regression model using the SciKit library instead of Keras. Currently we have a model comparing the ACORN type and tariff type against the average energy consumption of households.

Issues Encountered:

We have measured the accuracy of our model through the mean absolute error, the mean squared error, and the root mean squared error. We are little unsure of how to interpret these results and need to clarify how these metrics are used to understand our model's accuracy. We also need to find a second question to discuss as we have combined tariff types and ACORN types into one model.

Solutions Tried:

We are considering using time series analysis to predict future energy consumption for each household (or a subsection of households) for our second question.

Plans for Next Week:

Next week we plan to have our second question finalized and a simple model built so the final three days we have remaining can be spent on debugging and writing the final report.

Week 5: 20 April - 27 April

Progress Made:

The model for our first question was improved and finalized. We also finalized our second question by using time series analysis to predict future energy consumption for a subsection of households. We decided to use the ARIMA model for our second model to make predictions. As this is the last week for this final project, we prepared for the final presentation and wrote the final report as a conclusion to this project.

Issues Encountered:

For the first model, from the statistics results, we noticed that we had a very low R^2 score of 0.0552% which means the model does not fit our data well. However, we have a fairly low RMSE of 7.069 which is also pretty low after normalization. These two results seem to contradict each other but since our RMSE indicates accuracy we believe this model is indeed accurate. For the second question, we need to stabilize our model more so it can make better predictions.

Solutions Tried:

We talked about the issues we are still having during the presentation and received suggestions from Dr. Tao.

Plans for Next Week:

Our final project will be done! Thank you for offering such an interesting and instrumental course to us!

Bibliography:

[1] J.-M. D., “Smart meters in London,” *Kaggle*, 22-Feb-2019. [Online]. Available: https://www.kaggle.com/jeanmidev/smart-meters-in-london. [Accessed: 18-Feb-2020].

[2] “The Acorn User Guide ,” *The Acorn User Guide* . CACI, London, 2014.

[3] “Energy tariffs explained,” *Energy tariffs explained - Which? Switch*. [Online]. Available: https://switch.which.co.uk/energy-advice/energy-tariffs-explained.html. [Accessed: 18-Feb-2020].

[4] “Time of use tariffs,” TheGreenAge. [Online]. Available: https://www.thegreenage.co.uk/tech/time-of-use-tariffs/. [Accessed: 18-Feb-2020].

[5] N. S. Chauhan, “A beginner's guide to Linear Regression in Python with Scikit-Learn,” *Medium*, 25-Feb-2019. [Online]. Available: https://towardsdatascience.com/a-beginners-guide-to-linear-regression-in-python-with-scikit-learn-83a8f7ae2b4f. [Accessed: 30-Apr-2020].

[6] R. Ng, "Evaluating a Linear Regression Model", ritchieng.github.io, 2020. [Online]. Available: https://www.ritchieng.com/machine-learning-evaluate-linear-regression-model/. [Accessed: 30- Apr- 2020].

[7] “Predict Electricity Consumption Using Time Series Analysis,” KDnuggets. [Online]. Available: https://www.kdnuggets.com/2020/01/predict-electricity-consumption-time-series-analysis.html. [Accessed: 30-Apr-2020].

[8] “Lag Plots,” *NCSS Statistical Software*. [Online]. Available: https://ncss-wpengine.netdna-ssl.com/wp-content/themes/ncss/pdf/Procedures/NCSS/Lag_Plots.pdf. [Accessed: 30-Apr-2020].

[9] J. Brownlee, “How to Check if Time Series Data is Stationary with Python,” *Machine Learning Mastery*, 23-Nov-2019. [Online]. Available: https://machinelearningmastery.com/time-series-data-stationary-python/. [Accessed: 20-Apr-2020].

[10] J. Brownlee, "How to Create an ARIMA Model for Time Series Forecasting in Python", Machine Learning Mastery, 2020. [Online]. Available: https://machinelearningmastery.com/arima-for-time-series-forecasting-with-python/. [Accessed: 30- Apr- 2020].