Energy Analytics Forcasting

Group 7

2022

Contents

Project Description **2**
Variables 2

Exploratory Data Analysis (EDA) **2**

Statistical Analysis **4**
ARIMA 4
Regression Models 5

Conclusion **11**

Project Description

In this project, we aim to forecast the total electricity demand of England based on past available data. The forecast takes 6 rounds in total and we predicted the electricity demand for the following day in each round. The data set we used for forecasting contains historical demand of electricity, past data on temperature and sunshine duration, as well as ahead weather forecasts. The data is recorded separately for three regions in England, which are Bristol, London and Leeds. Through out the forecasting process, we tried out different approaches such as ARIMA and regression models, and continuously refined the models according to relevant statistical measures. In this report, we will step by step discuss the data analyzing and model fitting processes in each round and the refinements we have made with the models, also, we will address the difficulties encountered in the project.

Variables

The response variable is the demand of electricity, which refers to total electricity demand in England recorded on a daily basis. We used log scoring method in this study by transforming daily demand on log scale.

The explanatory variables include temperature, sunshine duration, time, indicator of weekend and lag of log demand. Temperature and sunshine duration include both historical data and day ahead forecasts data, the weather information is recorded since the start of 2018. Temperature data is collected on a 6-hour basis, which is recorded at midnight, 6 am, 12 noon and 6 pm; sunshine duration is given on a daily basis. Moreover, we constructed a variable to account for non-linear time trends in data, an indicator variable for weekend, and a variable to indicate one day lag of log demand.

Exploratory Data Analysis (EDA)

We have taken several steps to prepare the data for model fitting. Firstly, since the temperature is recorded four times a day while other variables are given on a daily basis, we need to transform temperature to a compatible scale. Thus, we grouped the temperature data by date, and calculated the average temperature in each day, and used daily average temperature for further analysis. Then we transformed the electricity demand to log scale, to avoid the case when the demand variations change over time.

Then we used ts() function to convert historical demand data into time series data frame with an annual cycle, we set frequency to be 365 and starts at 2018 Jan 1st. After this step, we used stl() function to decompose time series components. The plot below shows the result of time series decomposition. The uppermost plot shows the raw data and subsequent plots show the trend component, seasonal component and remainder respectively. From the seasonal component plot, we observe that the seasonal variation of demand seems to be constant over time, which means the seasonal pattern is similar among different years; the demand of electricity tends to peak in winter while drops to lowest point during middle of year. Besides, we found that the demand overally exibits a downward trend and the data has random residuals. The grey bars at right of each panel shows the importance of each component, we observe that the variation in trend is less important compared to variations in data, seasonal component and remainder.

To prepare the data for regression model, we computed a lagged version of demand by shifting time

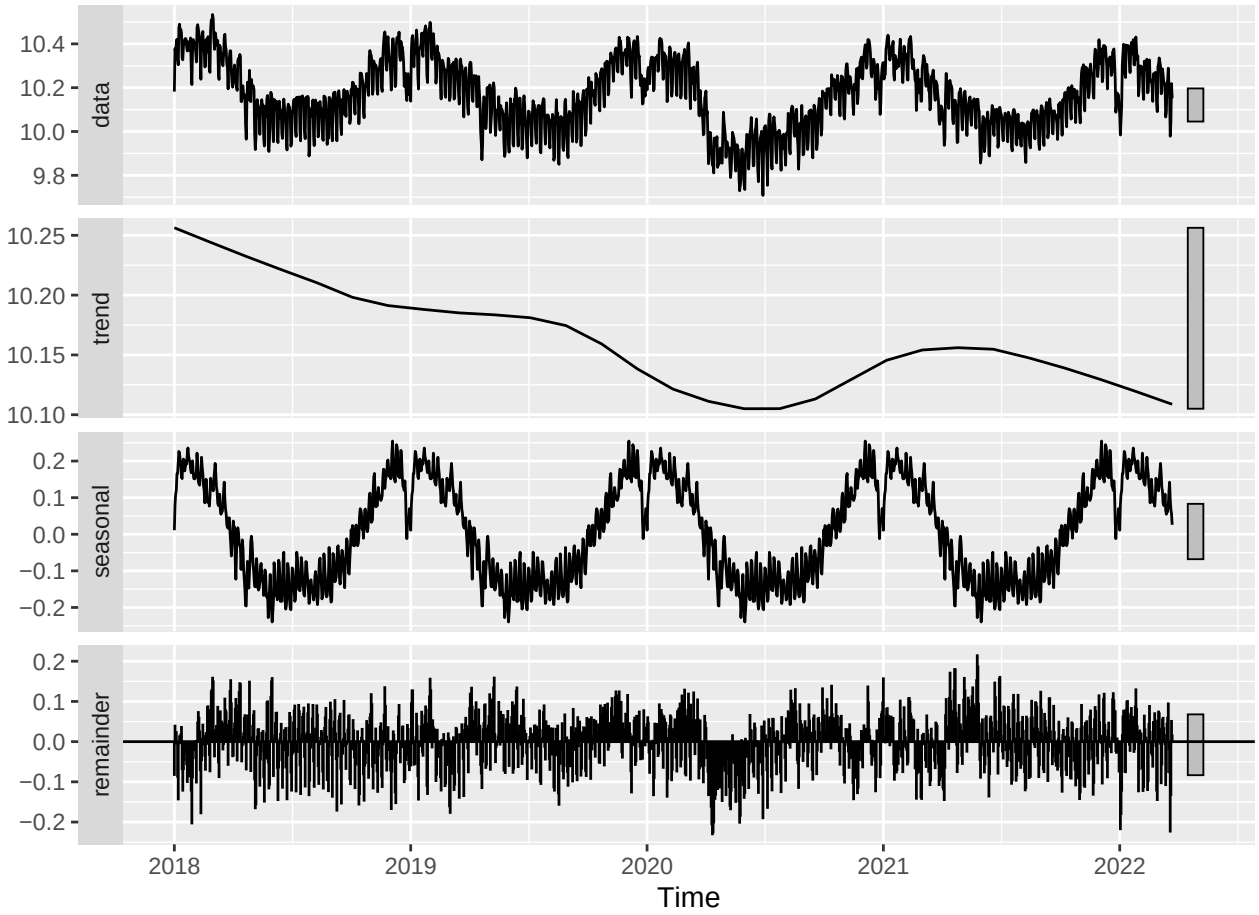


Figure 1: Plot of time series decomposition

base back by one day and obtained the variable d_{lag}. Besides, we created an indicator variable *weekend* to indicate whether a particular day is in weekend. Lastly, we converted date into variable *time* to account for non-linear time trends in data. Variable temperature and sunshine duration will also be used to forecast future demand in regression models.

Statistical Analysis

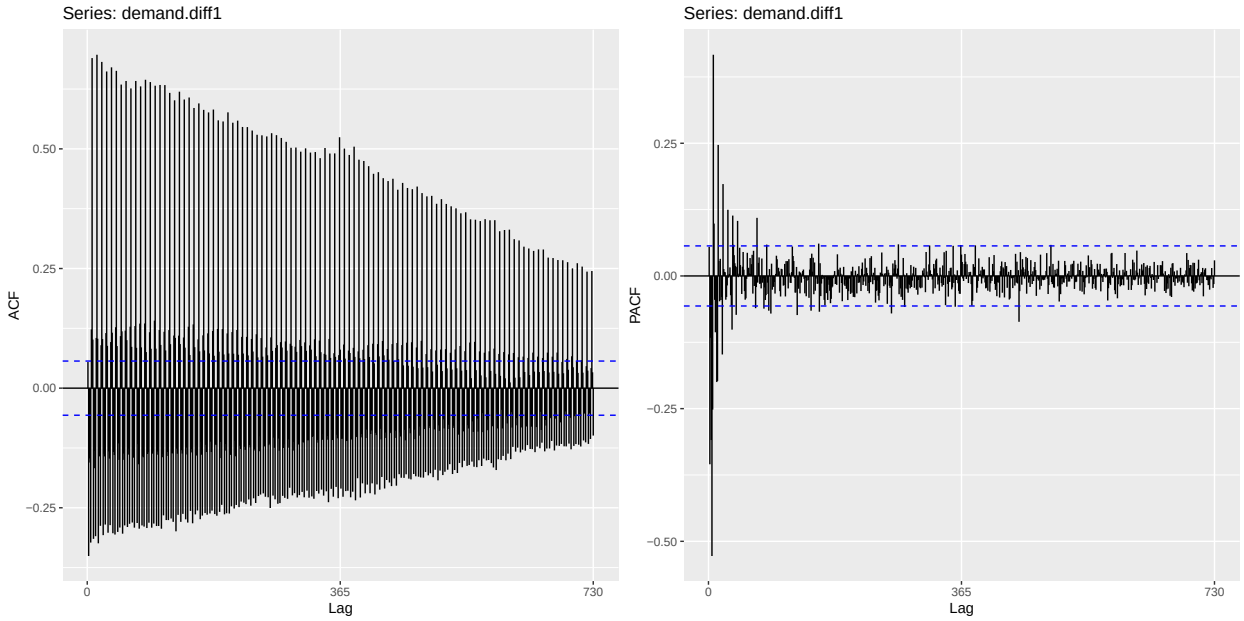
ARIMA

ARIMA is a statistical analysis model that uses time series data to predict future trends. ARIMA models predict future values based on past value.

```
##
##  Augmented Dickey-Fuller Test
##
## data:  demand.diff1
## Dickey-Fuller = -13.209, Lag order = 10, p-value = 0.01
## alternative hypothesis: stationary
```

```
##
##  Phillips-Perron Unit Root Test
##
## data:  demand.diff1
## Dickey-Fuller Z(alpha) = -661.24, Truncation lag parameter = 7, p-value
## = 0.01
## alternative hypothesis: stationary
```

```
##
##  KPSS Test for Level Stationarity
##
## data:  demand.diff1
## KPSS Level = 0.016938, Truncation lag parameter = 7, p-value = 0.1
```



The series in the figure keeps rising and falling in the long run, showing very obvious non-stationarity. Therefore, we first-differenced the data, and figure shows the difference data. They appear to be stationary, so we don't do further differencing. Then, we used the ACF plot to decide which one of these terms we would use for our time series. After plotting the ACF plot we move to Partial Auto-correlation Function plots (PACF). PACF is a summary of the relationship between an observation in a time series with observations at prior time steps with the relationships of intervening observations removed. Results showed that the PACF plot has a significant spike only at lag 1, meaning that all the higher-order auto-correlations are effectively explained by the lag-1 auto-correlation, so we could use the AR model. After analyzed by the auto.arima() function, the best model would be ARIMA(5,1,2).

The residual auto-correlation plot of the ARIMA(5,1,2) model shows that all the auto-correlation coefficients are within the confidence domain, which reflects that the residuals resemble white noise. The univariate mixture test yielded a large p-value, which also means that the residuals are similar to white noise. So we chose this model to forecast the value at 8th March 2022.

Since we need a criteria to determine which model is better. The scoring rule we choose is to use the log likelihood of our predicted distribution at the observed value. We use log since it would obtain a higher score if our predicted distribution is closer to the true distribution.

Learning from the result of auto.arima(), we could find that the best model would be ARIMA(5,1,2)(0,1,0)[365] . Then we calculate the score of our predicted distribution as stated above.

The score for ARIMA model is as below. The result getting from the ARIMA model is unreliable. Since the ARIMA model can prove inaccurate under certain market conditions, such as financial crises or periods of rapid technological change and there are many other factors that could effect the total electricity demand for England and Wales, we decided to try the regression model and compare the scores between those two to find out the best model to forecast the electricity demand.

```
##                          ME       RMSE        MAE          MPE      MAPE
## Training set  0.0005960959 0.06077732 0.04160206  0.004871483 0.4097782
## Test set     -0.2558897374 0.28021607 0.25783055 -2.521763891 2.5409319
##                   MASE        ACF1 Theil's U
## Training set 0.442728 -0.02621296        NA
## Test set     2.743826  0.67157442   3.85717

## The test score for ARIMA model is: -903.455683676824
```

Regression Models

Regression model is an intuitive way to include other variables like temperature and sunshine duration. The variable we chose is as below.

- t: The temperature of cities, we will experiment and try different weights to get a weighted value. It make sense to include temperature since the electricity demand for heating should increase when the temperature is low. We expect to see a negative coefficient for this variable.
- s: The sunshine duration of cities, we will get a weighted value as temperature. When the sunshine duration is long, the electricity generated by solar should be higher. Thus, the electricity demand would be lower (since solar power is not included in the demand). We expect to see a negative coefficient for this variable.

- *time*: There might be a trend when the time passed. We include *time* and $time^2$ do deal with nonlinear trends. If there is no trend, the coefficient in regression model will be insignificant.
- *weekend*: A binary variable indicates that whether this day is weekend or not. In the data exploration, it is clear that the demand in weekends are substantially lower than the demand in weekdays. We expect to see a negative coefficient for this variable.
- d_{lag}: The one day lag of log demand. It makes sense to include this variable since the weather for today and tomorrow would be similar, and weather is highly correlated with electricity demand. We expect to see a positive coefficient for this variable.

The regression formula is as below.

$$d = \alpha + \beta_1 t + \beta_2 s + \beta_3 time + \beta_4 time^2 + \beta_5 weekend + \beta_6 d_{lag} + \epsilon$$

where d is the log of electricity demand and ϵ is the residual.

Due to the uncertainty in forecasts, we forecast both the expected value and the standard deviation. The standard deviation is calculated through prediction interval. The prediction interval is calculated by

$$\hat{y}_f \pm z_{\alpha/2} \times [\hat{\sigma}^2 + se(\hat{y}_f)^2]^{1/2}$$

where $\hat{\sigma}^2$ is the estimated variance of ϵ and $se(\hat{y}_f)^2$ is the estimation variance of the OLS estimates. $[\hat{\sigma}^2 + se(\hat{y}_f)^2]^{1/2}$ is standard deviation of the prediction error. Since the perdition interval is available in R, we can calculate the standard deviation backward.

$$sd = \frac{fitted - lower}{z_{\alpha/2}}$$

where $fitted$ is the predicted value and *lower* is the lower bound of prediction interval.

Since we are comparing models with different weights, we need a criteria to determine which model is better. The scoring rule we choose is to use the log likelihood of our predicted distribution at the observed value. We use log since it would obtain a higher score if our predicted distribution is closer to the true distribution.

Model 1: Use London data only

The first model we choose is only using the temperature and sunshine duration in London. This is intuitive since the electricity demand of London is probably higher than the demand of Bristol and Leeds.

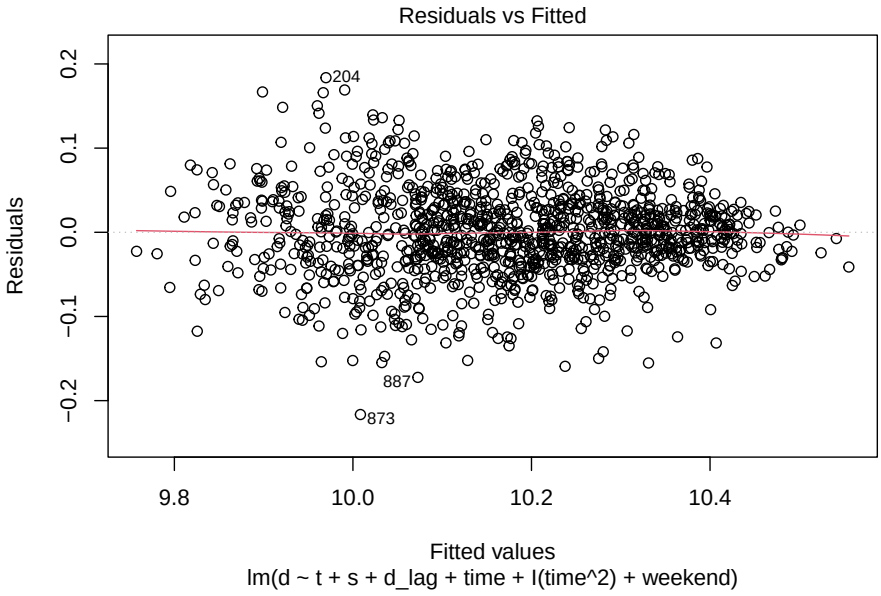
The summary of regression model is as below.

```
##
## Call:
## lm(formula = d ~ t + s + d_lag + time + I(time^2) + weekend,
##     data = data_train)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.216516 -0.029888 -0.001209  0.031021  0.183680
```

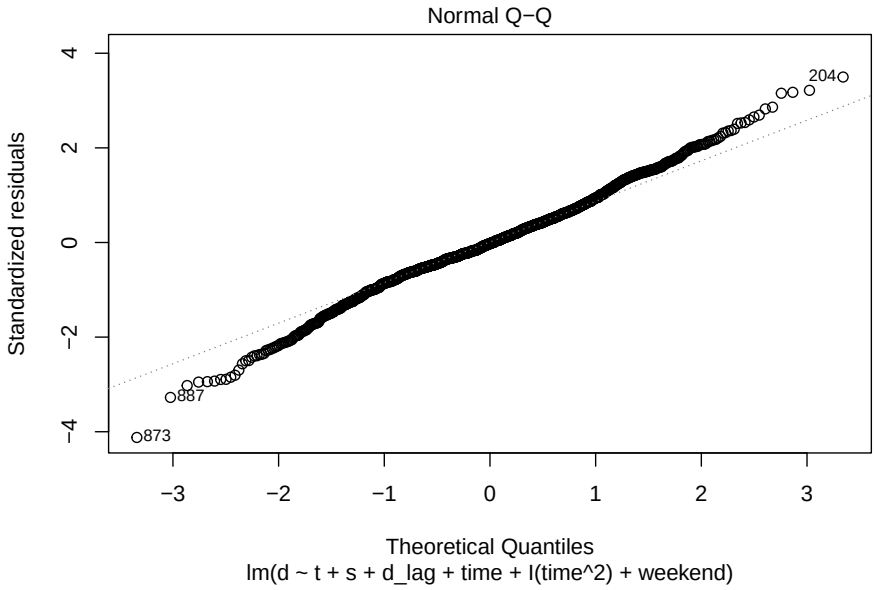
```
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.080e+00  1.547e-01  26.375  < 2e-16 ***
## t           -6.535e-03  3.939e-04 -16.590  < 2e-16 ***
## s           -6.963e-05  6.334e-06 -10.995  < 2e-16 ***
## d_lag        6.150e-01  1.470e-02  41.826  < 2e-16 ***
## time        -8.926e-05  1.834e-05  -4.866 1.29e-06 ***
## I(time^2)    3.683e-08  1.469e-08   2.507   0.0123 *
## weekend     -1.051e-01  3.432e-03 -30.606  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.05274 on 1193 degrees of freedom
## Multiple R-squared:  0.893,  Adjusted R-squared:  0.8925
## F-statistic:  1660 on 6 and 1193 DF,  p-value: < 2.2e-16
```

As our expectation, the coefficient for t, s and $weekend$ is negative; the coefficient for d_{lag} is positive. All the variables are fairly significant and the overall R-squared is 0.893, which means 89.3% of variance is explained in this model.

We further check if the regression model satisfied the OLS assumption. First, we look at the residual plot.



In the residual plot, the mean of residual is about 0, and the variance of residual is fairly consist. Next, we check if the normality assumption is satisfied by normal Q-Q plot.



In the normal Q-Q plot, the closer the point and the dashed line, the more normal it is. In this plot, the points and the line are very close, so the assumption is also satisfied.

Finally, calculate the score of our predicted distribution as stated above. The score for model 1 is as below.

```
## The test score for model 1 is: 556.047902758597
```

Model 2: Average all three cities

Though the electricity demand of London is the highest, it may not be representative to the whole electricity demand in England and Wales. Therefore, using the data from other cities might improve the prediction. The most intuitive may is averaging the temperature and sunshine duration of all three cities.

The regression model is:

$$d = 3.413 - (5.962e-03)t + (-6.802e-05)s + (-7.289e-05)time + (2.888e-08)time^2 + 0.676d_{lag} + \epsilon$$

The result is very similar to the first model, the R-squared increased slightly to 0.8976. The score of our predicted distribution is as below.

```
## The test score for model 2 is: 558.243363530912
```

We can see that the overall score also increase slightly, so taking average may be a better idea than only using the data of London.

Model 3-5: Test different weights

Since the electricity demand in London is still the highest, a higher weight for data in London make sense. In the following models, we try different weights of data.

In model 3, we use a weighted average with 50% weight for London and 25% weight for other cities. The score is as below.

```
## The test score for model 3 is: 558.937437389715
```

Since the score is slightly higher than averaging all three cities, adding weights for London seems better. Therefore, in model 4, we use a weighted average with 60% weight for London and 20% weight for other cities. The score is as below.

```
## The test score for model 4 is: 558.895586491351
```

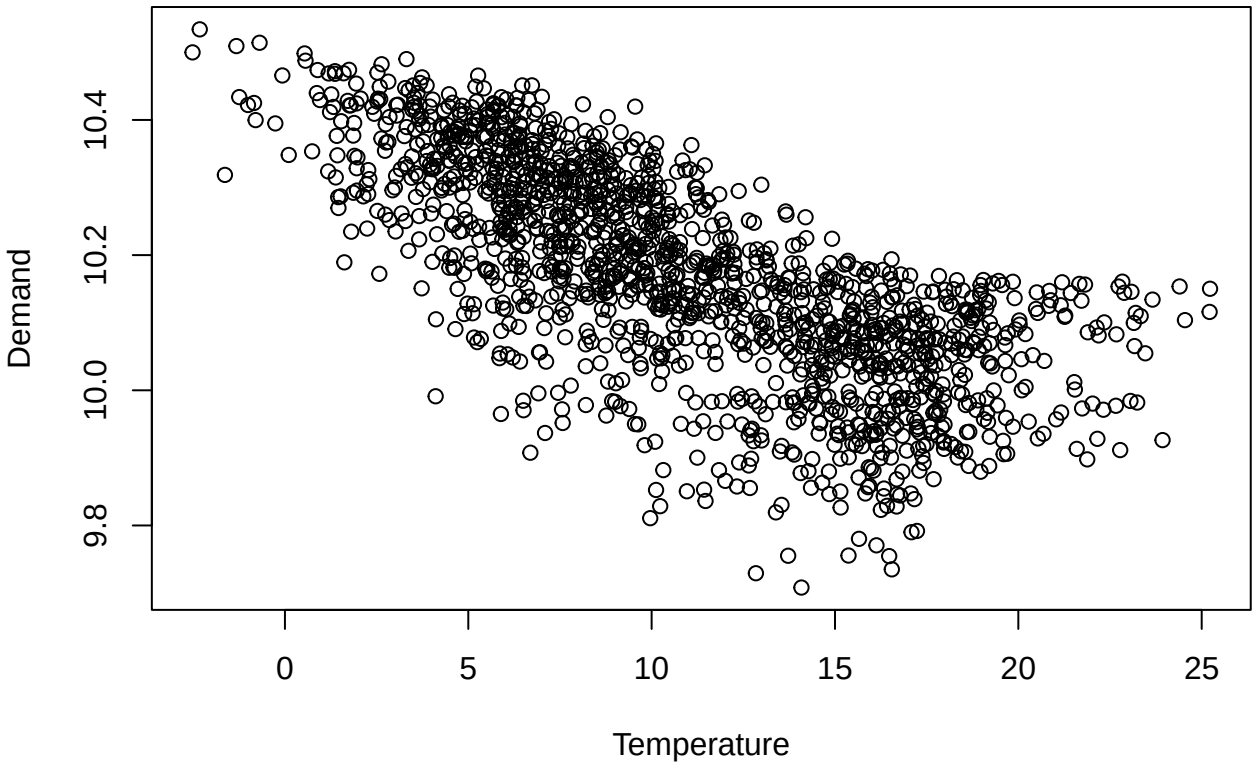
The performance is worse than 50% weight for London, so we try to use a lower value between 50% and 33%. In model 5, we use a weighted average with 40% weight for London and 30% weight for other cities. The score is as below.

```
## The test score for model 5 is: 558.638017175989
```

Still, it does not outperform the third model. Therefore, we decide to use 50% weight for London and 25% weight for other cities.

Model 6: Add a knot in temperature

The relationship between temperature and electricity demand may not be linear. To be more specific, there might exist a critical point (knot) that the linear relationship changes. To see if this exist in our data, we plot a scatter plot using electricity demand and temperature.



It seems like that there are two different linear relationships. The critical point is about 11. When the temperature is lower than 11, the lower the temperature, the higher the electricity demand. When the temperature is higher than 11, the demand does not change much if the temperature changes. This make intuition sense since when the temperature is low, people will turn on their heater. When the temperature is high, few household have air conditioner to turn on.

We turn the temperature t in the previous model into two variables w_t and w_0.

- w_t: If the temperature t is greater than or equal to 11, the value of w_t is the same as t, otherwise it is 0.
- w_0: If the temperature t is less than 11, the value of w_0 is the same as t, otherwise it is 0.

The regression formula thus become:

$$d = \alpha + \beta_1 w_t + + \beta_2 w_0 + \beta_3 s + \beta_4 time + \beta_5 time^2 + \beta_6 weekend + \beta_7 d_{lag} + \epsilon$$

Running the regression, we get the result as below.

```
##
## Call:
## lm(formula = d ~ wt + w0 + s + d_lag + time + I(time^2) + weekend,
##     data = data_train)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.202221 -0.029572 -0.001502  0.029810  0.184450
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.325e+00  1.567e-01  27.598  < 2e-16 ***
## wt          -7.188e-03  4.306e-04 -16.691  < 2e-16 ***
## w0          -7.319e-03  7.205e-04 -10.157  < 2e-16 ***
## s           -8.902e-05  7.215e-06 -12.338  < 2e-16 ***
## d_lag        5.922e-01  1.492e-02  39.687  < 2e-16 ***
## time        -9.155e-05  1.809e-05  -5.061 4.82e-07 ***
## I(time^2)    3.642e-08  1.446e-08   2.519   0.0119 *
## weekend     -1.060e-01  3.370e-03 -31.446  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.05172 on 1192 degrees of freedom
## Multiple R-squared:  0.8972, Adjusted R-squared:  0.8966
## F-statistic:  1487 on 7 and 1192 DF,  p-value: < 2.2e-16
```

In the result, the coefficient difference between w_t and w_0 seems trivial. Therefore, the result may be very close to the model without knot. The score of the predicted distribution is as below.

```
## The test score for model 6 is: 558.83418593604
```

The test score is indeed very close to the model without knot. The table below are the summary of regression models we have used.

Table 1: Regression Models Used

Model_Name	London_Weight	Bristol_Weight	Leeds_Weight	With_Knot	Score
Model 1	100%	0%	0%	N	556.0479
Model 2	33%	33%	33%	N	558.2434
Model 3	50%	25%	25%	N	558.9374
Model 4	60%	20%	20%	N	558.8956
Model 5	40%	30%	30%	N	558.6380
Model 6	50%	25%	25%	Y	558.8342

Conclusion

In this report, we work through data selection, model selection and refining with six rounds. With the final fitted regression model, we forecast the total electricity demand of England based on past data. The final result is produced based on a regression fit model 6, with weights of 50% on London, 25% on Bristol and Leeds, and with the prove based on a temperature turning point indicator. The final result of the demand has a mean of 10.13 and a standard derivative of 0.0516.

The final result is yielded from the model below:

$$d = 3.369 - (5.917e{-}03)w_t - (6.397e{-}03)w_0 - (6.800e{-}05)s - (6.800e{-}05)time + (2.821e{-}08)time^2 + 0.6804_{lag} + \epsilon$$

The report has mainly three parts to improve the forecasting. The first step involves the model selection, we select the time series with the first glace of the data, then choose the regression model as our target model. Finally, we keep adjusting parameters on the regression model and stepping towards our final prediction.

Firstly, we do the log data transformation by transferring electricity demand. With the assumption check of time series, we select the first-difference method to apply to the model. With the indication from the PACF plot, we fitted the dataset with ARIMA(5,1,2). Further analysis rejects the performance of the time series model, so we aim to move to regression to forecast.

Regression is what we mainly determined in this report. Six models are developed with different factors and model components selected. Including more cities acts an increase on the performance when moving from model 1 to model 2, and changing the weights of those three cities to improve the accuracy throughout model 3 to model 5. Then, by adding a knot and including indicators to clarify the linearity, we further improve the accuracy and come up with the final model. This model differs from others since it not only includes the influence of the cities, and also considers the electricity swifts of temperature.