Aalto University
School of Business

Electricity Demand Forecast

Group D

Group Report
Andriakov Nikolai
La Linh
Ta Tinh
Tran Dung

26.03.2024

1. Introduction

The objective of the assignment was to develop a model to forecast daily logarithmic electricity consumption in Finland for the upcoming day. The model was based on historical data on hourly electricity consumption in Finland and hourly temperature data for three cities: Helsinki, Tampere, and Rovaniemi. The forecasts were made in three rounds for 14.3.2024, 16.3.2024, 20.3.2024. Since the objective was to forecast logarithmic daily electricity consumption, the natural logarithm of daily electricity consumption was used in the final forecasts for each round. The output of the model was the mean and standard deviation of a forecast value distribution. During model development, AIC and visual investigation of models' fit were used as criteria for model assessments and selection of the best models. The forecasts were assessed based on the log score of the submitted electricity consumption values and the actual consumption for the day of the forecast.

2. Preliminary analysis

The analysis started with exploring the available data on electricity demand in Finland and temperatures provided for Helsinki, Tampere, and Rovaniemi. First, we plotted the consumption time series data.

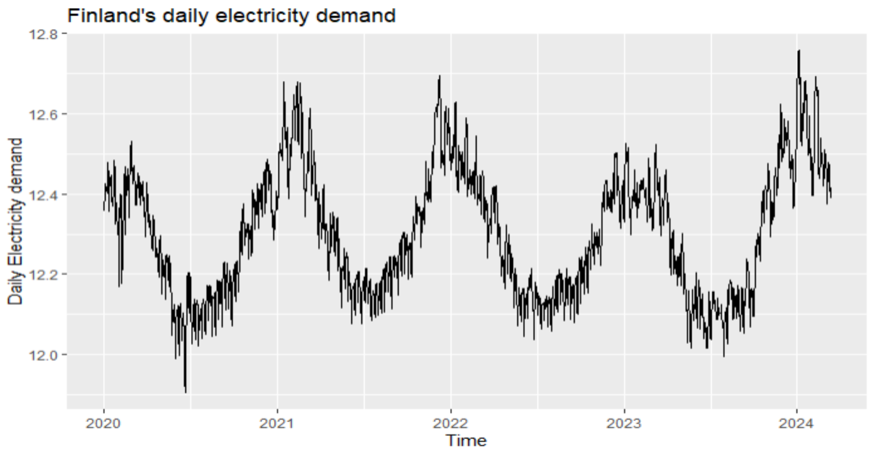

Figure 1: Log electricity consumption time series for 2020–2024.

There were significant peaks of consumption during winter months that could be related to cold weather in Finland that led to higher electricity consumption due to active usage of heating equipment. There were also local peaks that were likely to be attributed to weekly consumption patterns. It could be also noticed that peak consumption during 2023 was lower than during other years, probably due to high energy prices and the risk of an energy crisis in Europe.

In terms of data quality, there were missing data for the first half of 2022 and in the middle of 2023 which were noticeable by straight lines in the corresponding periods in Figure 1. In 2020, there were supposedly outlier consumption records represented as unusually low values of consumption for several days of 2020. These data quality issues were addressed during the data processing step of model development, and it is described below in the corresponding sections.

As for the temperature data, the figure below shows the scatterplots of electricity consumption and temperatures for the three cities. According to these scatterplots, there is a noticeable negative close-to-linear relationship between the temperatures and the consumption. All three temperatures seemed to be reasonable independent variables that could be used in a regression model of electricity consumption.

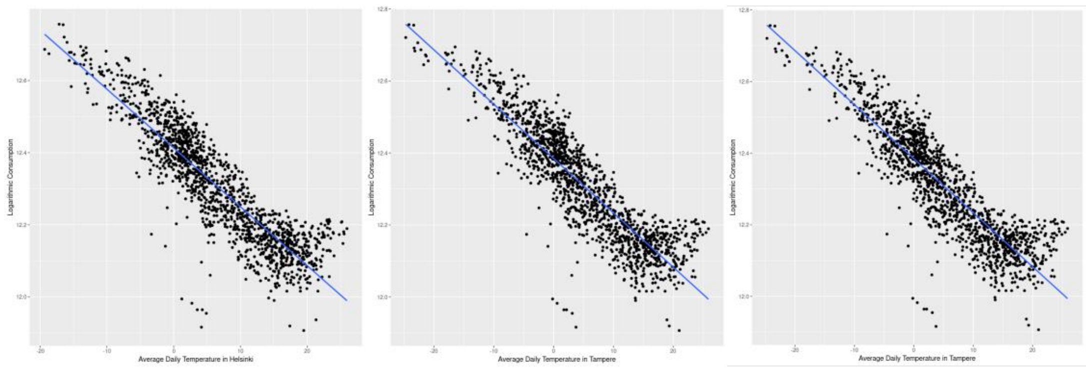

Figure 2: Scatterplots of electricity consumption and mean daily temperatures in three cities.

3. Modelling and forecasting

3.1. First forecasting round

3.1.1. Data processing

For the Helsinki, Tampere, and Rovaniemi temperature files, the average temperature per day was used for model training. For the forecast file, the average temperature per day was calculated to be used for testing the models and making forecasts.

For the consumption file, it was noticed that before 2023, some hours that ended in hh:55:00 format, which, if filtered out, would lead to data loss. Therefore, for data before 2023, 5 minutes were added to the hours that ended with 55 minutes in the End time UTC column because it was more sensible that end time be used to measure the consumption from a particular period. However, four dates that did not follow this rule and ended with 25 minutes of the hour instead, so missing consumption was added to these dates to ensure sufficient data quality. Moreover, it was noticed that from 2023 onwards, there were instances of data duplication for the same hour where consumption was recorded in 15-minute intervals instead of 1-hour intervals, so those 15-, 30-, and 45-minute intervals were removed from the dataset to ensure correct electricity consumption measurements. After, hourly consumption data was summed per day for modelling. Finally, a log consumption column was added, and the four data sets were merged.

3.1.2. Model development

The development started with time series linear regression as the benchmark and we noticed that there's a non-linear relationship between Helsinki and Tampere temperature, so another predictor of Helsinki: Tampere was added to the model with each city's temperature as well as trend() and season() predictors. The model was able to predict around 89% of the consumption variance. However, residual analysis showed strong autocorrelation, which meant that more robust methods for time series analysis were needed.

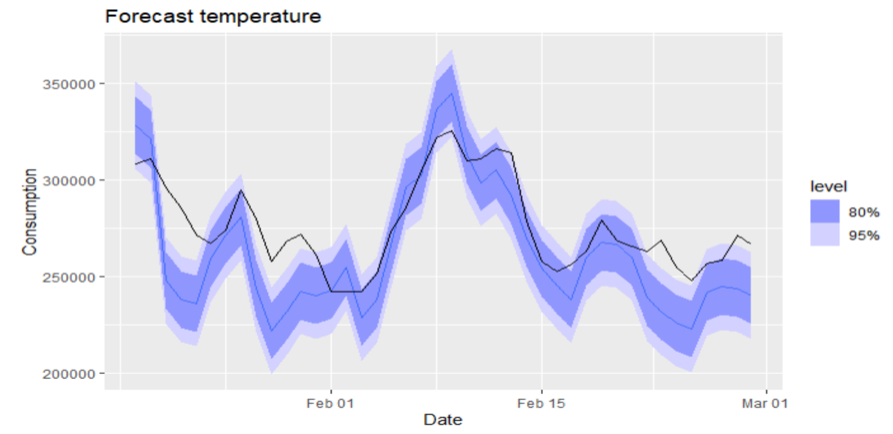

Figure 3: Time series linear regression model's fit visualization

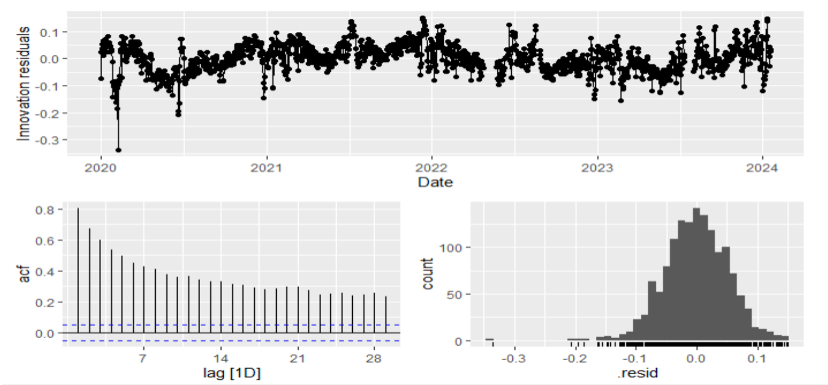

Figure 4: Time series linear regression model's residuals

Next, models with stronger seasonality assumptions of Holt's Winter and Holt's Winter Damped were used. The models' residuals seemed good, but the model did not track variation well, suggesting a combination of regression and seasonal models was needed.

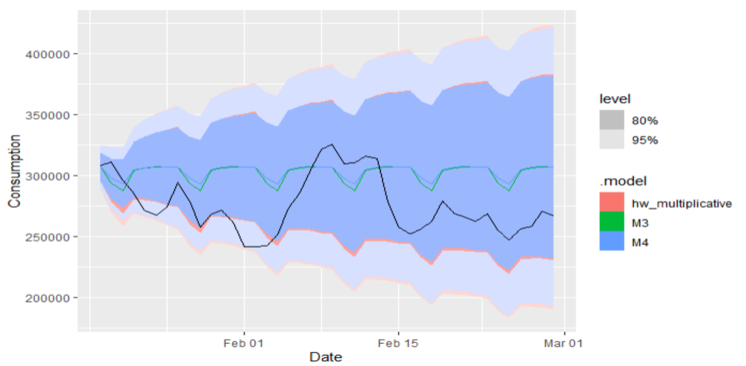

Figure 5: Holt's Winter-based models' fit visualization.

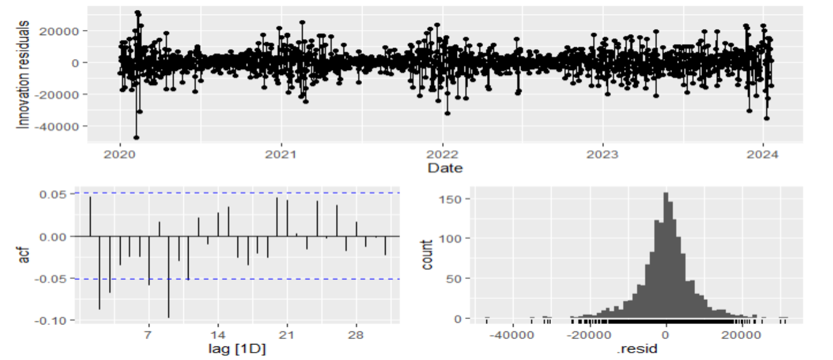

Figure 6: Holt's Winter-based models' residuals

Using the residuals and partial residuals plots of the raw consumption data, we were able to identify the seasonality of consumption which is every 7 days and with a peak on the first lagged day. Using this information, a dynamic regression model was developed for the first-round forecast.

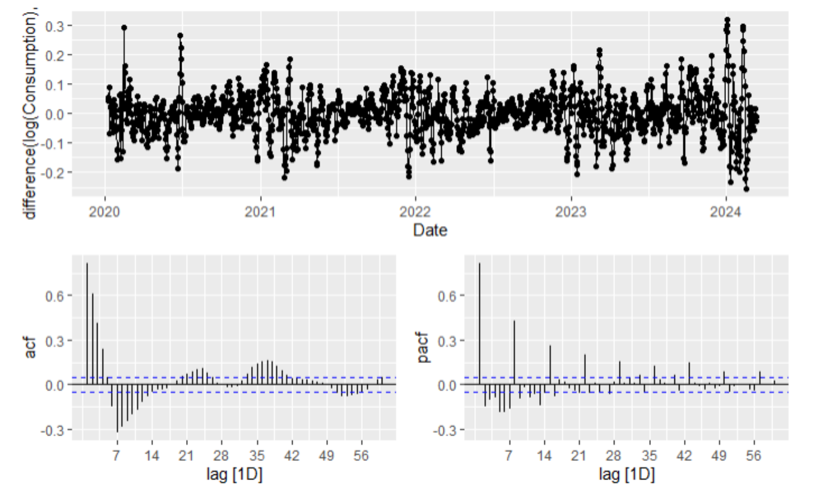

Figure 7: Partial residuals plot of log consumption.

The model was implemented as an ARIMA model with the city temperatures as predictor variables. After experimenting, the optimal ARIMA parameters were chosen as shown in Figure 10. As the model's residuals analysis showed good model quality, the AIC of this model (-6656.42) was measured and saved for future reference.

```
Series: log_consumption
Model: LM w/ ARIMA(0,1,1)(0,1,1)[7] errors

Coefficients:
          ma1     sma1  Helsinki  Tampere  Rovaniemi  Helsinki:Tampere
      -0.1644  -0.9703   -0.0041  -0.0027    -0.0012             1e-04
s.e.   0.0302   0.0077    0.0006   0.0005     0.0002             0e+00

sigma^2 estimated as 0.0005954:  log likelihood=3335.25
AIC=-6656.49   AICc=-6656.42   BIC=-6619.18
```

Figure 8: Dynamic regression parameters of the final model in the first round.

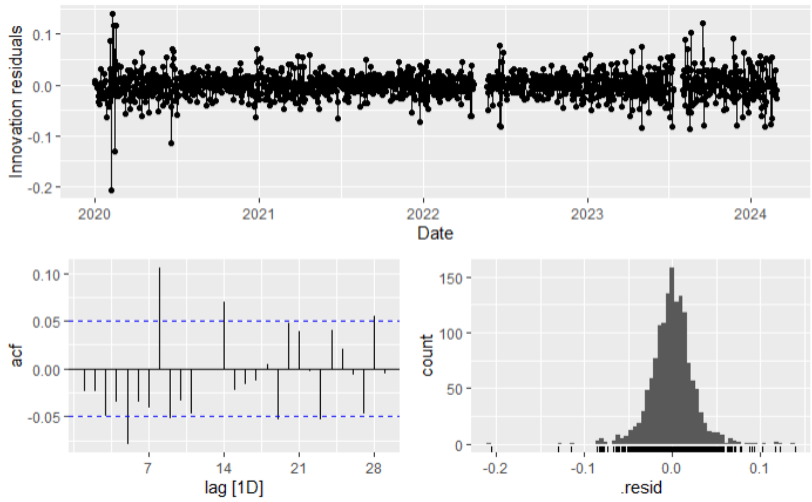

Figure 9: Dynamic regression model's residuals

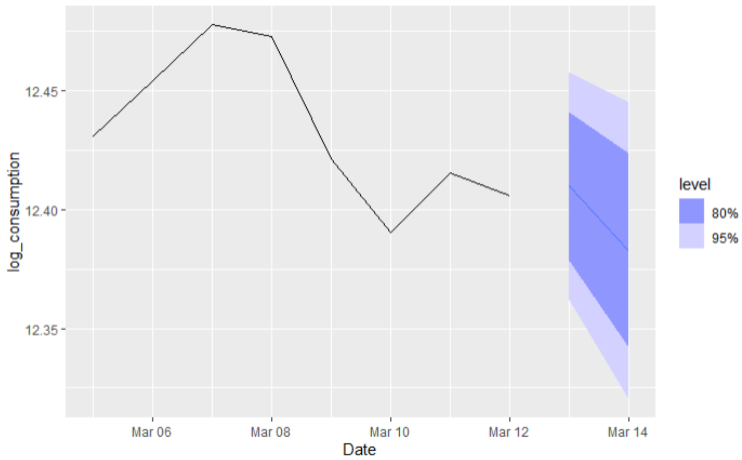

Figure 10: Forecast visualization for the first round (mean=12.383, st.dev.=0.032).

3.2. Second forecasting round

3.2.1. Data processing

In the second round, the primary focus was on addressing data processing issues. Two issues likely impacted the accuracy of the forecasts in the first round. First, certain consumption data points were omitted from the raw data because of irregular time intervals. Second, an incorrect time zone was applied during the data processing of the consumption data. Consequently, the raw data processing approach was adjusted. Instead of taking the first record for each hour, all hourly records were averaged and then summed to get daily consumption. This approach solves the first issue.

For the second issue, it was observed that the raw consumption data was in the UTC time zone, whereas the rest of the data was in the Helsinki time zone, which is in UTC+2. To improve the precision of the forecasts, the consumption records' dates and times were adjusted to UTC+2 time zone. The adjustment involves adding 2 hours to the date column.

After processing the data, the same dynamic regression model was used as in the first round. The figure below indicates that the residuals exhibit characteristics akin to white noise. While there's some autocorrelation observed in the residuals, it isn't notably significant overall. Moreover, the variance of the residuals appears relatively consistent over time. The distribution in the final chart appears to follow a normal distribution, with most values centered around zero.

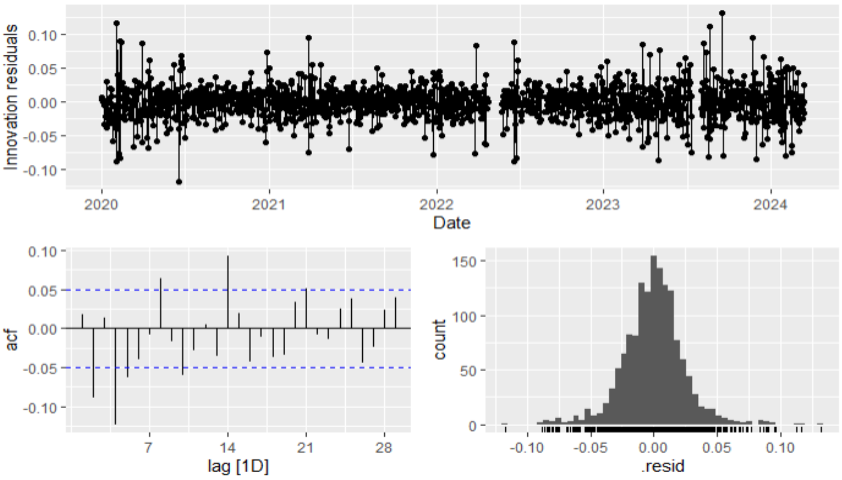

Figure 11: Dynamic regression model's residuals in the second round.

In terms of AIC, this model has a lower AIC (-6669.03) compared to the model used in the first round (-6656.42). Hence, the model for the second round was improved.

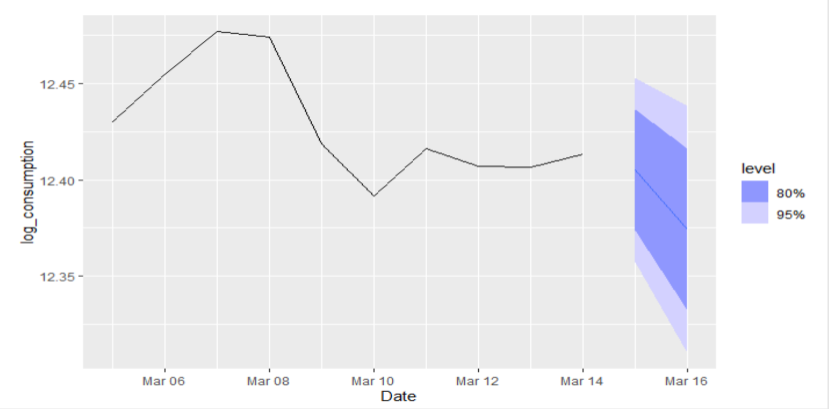

Figure 12: Forecast visualization for the second round (mean=12.372, st.dev.=0.033).

3.3. Third forecasting round

3.3.1. Data processing

In the third round, data processing was the same as in the second, but instead of using the End time UTC column in consumption records, Start time UTC was used to increase data quality due to incorrect records in End time UTC.

3.3.2. Model development

The new model for the third round integrated various COVID-19 indicators, including school and workplace closures, gathering restrictions, stay-at-home requirements, international travel limitations, public transportation closures, and movement restrictions. These indicators were represented on an ordinal scale, with values ranging from 0 to 3/4, indicating the severity or extent of each measure (where 0 indicated as no measures, 1 as the measure recommended, 2 as the measure required for some sectors/extent, 3 or 4 as the measure required for all sectors).

Initially, binary values (0, 1) from the "flag" sheets were considered, but their inclusion led to a decrease in predictive accuracy, as evidenced by an increase in AIC. Subsequently, ordinal scales were adopted for each indicator to better capture the nuances of COVID-19 measures and improve model performance. Furthermore, to account for seasonal variations, Fourier terms with K=3 were incorporated into the model. These terms enable the capture of cyclic patterns in electricity consumption data, enhancing the accuracy of the forecasting model.

The below figure suggests that residuals looked like white noise. There was some autocorrelation in the residuals, but it was not significant. Residuals had constant variance over time. Despite some outliers, the distribution seemed to be normally distributed and centered around zero.

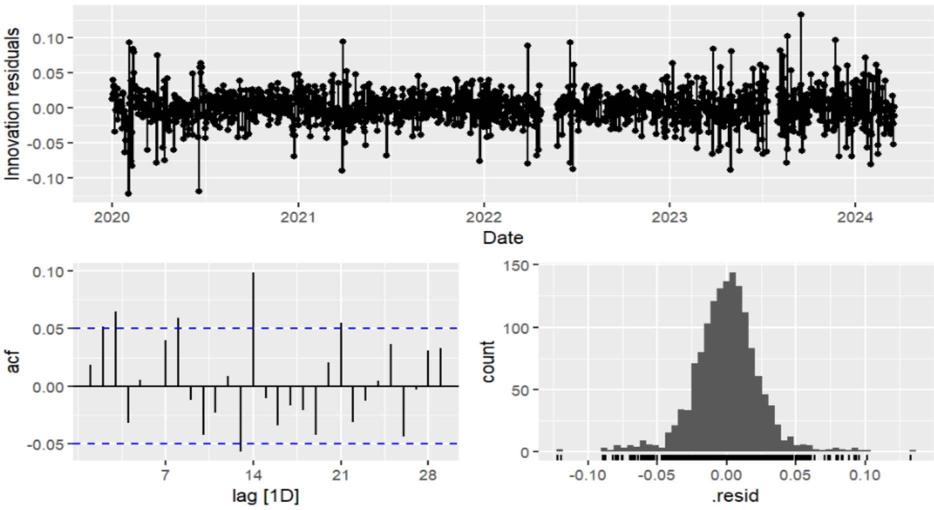

Figure 13: Dynamic regression model's residuals in the third round.

In terms of AIC, this model (M3) had the lowest AIC compared to the previous models (M1 – the model used in the second round, M2 – the model taking only the temperature of three cities and Fourier K = 3 as predictors). Hence, it was decided to adopt this model to forecast the logarithm of electricity consumption for the third round.

.model <chr>	sigma2 <dbl>	log_lik <dbl>	AIC <dbl>	AICc <dbl>	BIC <dbl>
M1	0.0005904388	3355.024	-6696.049	-6695.975	-6658.713
M2	0.0005668919	3416.173	-6800.345	-6799.988	-6714.934
M3	0.0005641561	3423.177	-6800.355	-6799.625	-6677.575

Figure 14: Comparison of different models' parameters.

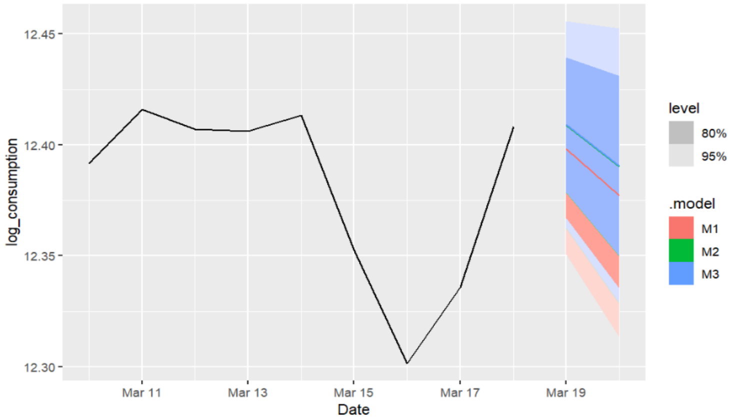

Figure 15: Forecast visualization for the third round (mean=12.390, st.dev.=0.032).

4. Discussion and conclusions

The quality of the forecasting models, as discussed above, was measured using the AIC of the model. The AICs of each round's model are provided in Table 1.

Forecasting round	Round 1	Round 2	Round 3
Final model's AIC	-6656.42	-6669.03	-6800.35

Table 1: Models' AIC comparison between the forecasting rounds.

The table shows that the AIC of the models decreased with each consecutive round, meaning that the team managed to improve its models in each round. Note that AIC does not explicitly show the accuracy of the forecasts produced by the models. However, since the models were developed before the announcement of the results from the previous round, AIC was a sufficient and consistent measure of the models' quality.

Considering the limitations of the models described above, it should be noted that outside of a university assignment, the models described above could be developed further using more relevant variables and expert knowledge. However, the assignment was limited to a specific scope of available data, thus there was a limit to how well-performing models could be developed. Another practical limitation of the model from the third round was that model M3 was chosen due to a lower value of AIC. However, model M2 was significantly lighter and required less computational resources to run despite the negligible increase in AIC. Thus, model M2 would be a preferred option for practical applications.