QBUS3830: Advanced Analytics

REDUCING POWER SUPPLY COSTS IN SOUTH AUSTRALIA USING TIME-SERIES AND MACHINE LEARNING METHODS

Rhys Kilian

November, 2018

Executive Summary

Our neural network model forecast for South Australian power demand was able to outperform the industry standard model by 11.12%. This project used very short-term load forecasting methods to improve the one step ahead (30 minute) demand forecast for the state of South Australia. In order to reach this optimal model we added Bureau of Meteorology Adelaide weather data as an additional variable on top of variables from the energy market operator. Using the 30 minute time series data for the input variables price, demand and temperature we considered traditional time series models, machine learning methods and benchmark models.

Through EDA we identified a strong daily and weekly pattern for demand which was utilised in auto regressive models and linear models. During feature engineering we created dummy variables for blackouts, heatwaves and season. The optimal neural network was identified through a randomised grid search cross validation of hyperparameters. Combination forecasts were considered but could not beat the performance of the neural network model. The expansion of input variables to include weather variables allows our neural network to be so successful that it could save millions of dollars for energy providers. By creating more accurate demand forecasts there will be less supply induced blackouts in South Australia allowing for a more reliable and less costly energy market.

Contents

1 Introduction and Business Context

The South Australian energy market has undergone a variety of challenges in recent years including extreme weather events, debate over its reliance on renewable energy and a series of power outages.

The Australian Energy Market Operator (AEMO) provides historical energy demand and price data which is used by regulators and suppliers to predict the energy demand 30 minutes in to the future, known as very short-term load forecasting (VSTLF). Our project seeks to improve the accuracy of VSTLF and ensure that shortfalls in supply occur less frequently to improve the costs of energy for consumers and providers. The industry standard loss function is the mean absolute percentage error (MAPE). It has been estimated that a reduction of 1% MAPE would lead to a US$1.6 million saving per year for a 10-gigawatt generator (Hobbs et al. 1999).

South Australia's capital city Adelaide is subject to some of the most extreme temperatures of any state in Australia. It is the driest capital city (Mack, 2013) and has a reputation for its "blistering heatwaves". These extreme weather events have put significant pressures on the state energy infrastructure (Saddler, 2013). The primary use for power is for heating and cooling (40%) houses. The secondary driver for residential energy consumption is water heating. Hence our hypothesis is that changes in temperature are a good predictor for energy use.

By combining historical climate, energy demand and price we aim to improve the MAPE of current industry models to provide cost savings to regulators and energy providers.

2 Data Understanding

Our model uses data sourced from AEMO (*Data Dashboard* 2018) and the Bureau of Meteorology (*Historical weather observations and statistics* 2017).
The key variables from the AEMO data were:

1. **Time:** The time variable was available in 30 minute intervals from 1998 to 2018.
2. **Price:** The variable 'price' is an average of the dispatch prices in South Australia over each half hour period and is known as the spot price. Industry prediction models use the spot price instead of the dispatch price because it is extremely volatile in 5 minute intervals. Predicting 30 minutes ahead is more useful for adjusting energy supply. It's units are Australian dollars.
3. **Demand:** Demand is measured in megawatts (MW) for the entire state of South Australia.

The key variables from the BOM data were:

1. **Time:** This variable was in 1h intervals for the period 01/05/2016 - 30/04/2017.
2. **Station number:** This identifies which weather station the measurements comes from
3. **Value:** For each time period there were 4 different observations within the same variable 'value'.
   (a) Current temperature (Celsius).
   (b) Maximum temperature in that period (Celsius).
   (c) Minimum temperature in that period (Celsius).
   (d) Amount of rain in that period (mm).

3 Data Processing

3.1 Time Period

A year of data was extracted from the AEMO and the BOM between 1/05/2016 to 30/04/2017. This period was selected as it contains the summer where three significant heat waves occurred. Furthermore, the more

recent weather data is extremely dirty (filled with missing values and wrong records). We considered the benefits of using complete data outweighed the benefits of using the most recent data. Hence in order to give our model the best chance of improving the industry standard we chose to focus on this period.

3.2 Weather station

The BOM has a large selection of weather stations to select from. We selected the Adelaide Airport weather station as it is considered the most reliable being located at an major airport. Furthermore the weather in Adelaide is most likely to effect the power consumption of South Australia as the capital has the highest population density. Adelaide contains over 75% of South Australia's population (ABS, 2018).

3.3 Missing data

In the weather data on the 31/05/2016 there was missing data points for 12am, 1am, 2am, 3am and 4am. This missing data was linearly interpolated.

3.4 Duplicate values

At 2pm each day the BOM data had double the amount of observations then at every other period. We concluded that this is due to the method of processing of the data. Hence we dropped the second set of values at 2pm each day.

3.5 Filtering weather data of interest

Of the four weather variables we selected current air temperature as the most relevant for our model. We excluded rainfall as this has no strong evidence for correlation with energy demand unlike temperature. In order to prevent multicollinearity issues we had to choose one of the three temperature variables. We selected current air temperature.

3.6 Matching time interval between data sources

BOM weather data intervals needed to be matched to the time intervals of AEMO data 30 minute intervals. The hourly data was expanded to 30 minute intervals via linear interpolation.

4 Exploratory Data Analysis

Our aim is to produce a prediction for the explanatory dependent variable, power demand, using the regression variables price and temperature.

4.1 Non-parametric regression

After plotting the values for air temperature and price against their index's it is clear that nonlinear time series relationship is true for both variables. This suggests that a non-linear model is likely going to perform better with this data.

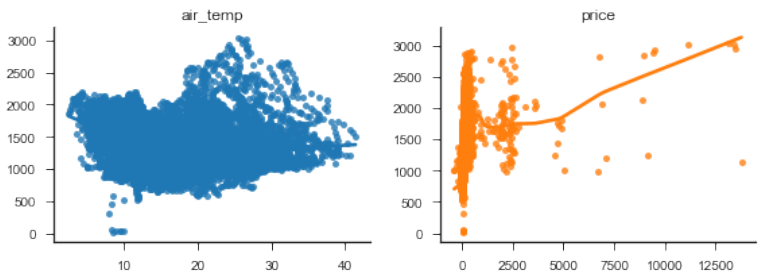

Figure 1: Local linear regression for variables air temperature and price

4.2 Time-Series Decomposition

Time series is a function of 4 components: trend component or the systematic long term trend; seasonal component which is the regular fixed period fluctuations; cyclic components representing fluctuations in the series that are not of fixed period (e.g. business cycle); and residual which is the irregular or error component. The results for this decomposition are shown in Figure 11 in Appendix A.

The time-series decomposition plot highlights a clear seasonal variation in each variable which is not proportional to the trend component. Hence, we elect to use the additive model for decomposition. Additionally, there are two clear outliers in the price variable. Further investigation found that these points occur on the 20/03/17 between 13.00-14.00 and the 21/03/17 between 11.00-13.00. During these periods the price was recorded at extreme levels from \$929-\$2651/MW, caused by an outage in Victoria.

In order to improve our modelling we decide to place a cap of \$400MW as the maximum price not considered an outlier. This was determined by considering the normal maximum price in the energy market. The outliers made up less than 1.4% of total data which we decided was reasonable. Similarly we capped demand at 1900 MW which effects less than 5% of the data. The time series decompositions before and after this cleaning are shown in the appendix (Figures 13 - 12).

4.3 Complex seasonality

With high frequency data it is common to observe multiple seasonal patterns. Complex seasonal analysis allows for identification of patterns with a non-integer period (De Livera et al. 2011). From our analysis on the demand cycle we observed a strong daily (seasonal 48) pattern and also a weekly (seasonal 336) pattern. The weekly seasonal pattern is stronger than the daily seasonal pattern. These patterns were identified using a a Box-Cox transformation of the demand data and complexity plot, as per Figure 2.

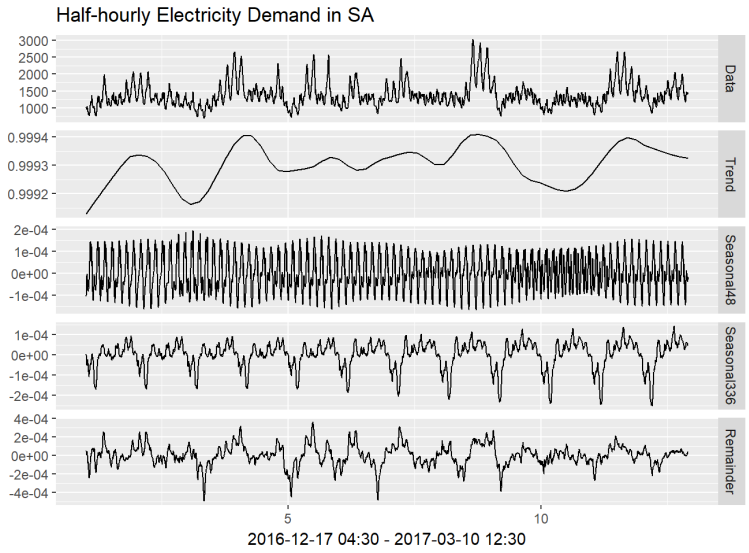

Figure 2: Complex seasonality plot for demand

However, there are some limitations of this complex seasonality analysis. Firstly these models only allow for regular seasonality. They are unable to capture seasonality associated with moving events such as Easter. Furthermore, they only show patterns the patterns present in the data. In this case we didn't provide enough data for it to show an annual pattern. At the other extreme it also struggles to capture sub daily patterns. These are likely included in the daily and weekly pattern that the above figure has identified.

5 Feature engineering

Extensive feature engineering was carried out in the project, for several reasons. Firstly, the machine learning models, including the linear and neural network models, required predictors in order to model the the response variable, South Australian power demand. In the past, this has been done by completing an autocorrelation analysis, as demonstrated by Sood et al. (2010), and then using the relevant lagged demand data as a predictor. This process is outlined in Section 5.2. Furthermore, the inclusion of several dummy variables, to model blackouts and heatwaves for example, was used to improve the model performance. The inclusion of these additional variables to predict electricity demand is not well documented and has been highlighted as an area for future work (Kotillova et al. 2012). Hence, this report aims to analyse the effects of these engineered predictors, which are outlined in this section of the report.

5.1 Dummy variables

In the study of econometrics, it is common to include dummy variables in models to improve the performance of time-series methods. For instance, Moller & Andersen (2015) made use of a number of dummy variables, model electricity demand in Denmark, from an econometrics perspective. In this investigation, dummy variables were created for: blackouts, heatwaves, seasons and public holidays. All with the purpose of increasing predictive performance, particularly for the machine learning models.

5.1.1 Blackouts

During the period over which the demand data was collected, there were several blackout events which affected the power demand during this period. The term blackout refers to the phenomena when power distribution is either bottlenecked or completely stopped, due to an invent such as a storm or a transmission fault. They are characterised by sudden decreases in demand, immediately followed by a large spike when distribution is brought online. Therefore, it was deemed critical to model these events with the use of a dummy variable. The blackout events within this dataset were defined by AMEO (*SA blackout: Why and how?* 2016):

1. **28th September, 2016** - a storm cutting the power to 1.7 million residents of South Australia until 10:00pm
2. **27th December, 2016** - blackout due to storm damage
3. **8th February, 2017** - excess load shedding during a heatwave

These events were modelled as per Equation (1).

$$dummy\_Blackout = \begin{cases} 1 & \text{if blackout during time period} \\ 0 & \text{otherwise} \end{cases} \tag{1}$$

5.1.2 Heatwaves

Heatwaves were also important events to capture using dummy variables. Our research indicated that a large component of South Australia's power demand is cooling, which is related to heatwave events (Griffiths 2018). It was expected that heatwaves would cause large upward spikes in demand, and therefore, to ensure the accuracy of predictive models, was required to be included as a predictor. The heatwaves in this dataset included:

1. **9-14th January, 2017** - temperatures exceeding 48 degrees
2. **17-21st January, 2017**
3. **31st January, 2016 - 12th February, 2017**

These were feature engineered as per Equation (2).

$$dummy\_Heatwave = \begin{cases} 1 & \text{if heatwave during time period} \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

5.1.3 Season

Given that the season affects the temperature the rationale for including dummy variables for the seasons follows on from Section 5.1.2. It is expected that in Summer the power demand would be high due to cooling devices being used. Alternatively, we expect the power demand to be high in Winter as heaters will be used. Hence, to model the four seasons, three dummy variables, as per Equations (3)-(5) were used to prevent any issues with multicollinearity (where Autumn is the reference season).

$$dummy\_Summer = \begin{cases} 1 & \text{if Summer during time period} \\ 0 & \text{otherwise} \end{cases} \tag{3}$$

$$dummy\_Winter = \begin{cases} 1 & \text{if Winter during time period} \\ 0 & \text{otherwise} \end{cases} \tag{4}$$

$$dummy\_Spring = \begin{cases} 1 & \text{if Spring during time period} \\ 0 & \text{otherwise} \end{cases} \tag{5}$$

5.2 Lagged Power Demand

As previously outlined, an autocorrelation analysis was conducted in order to determine the lagged power demand variables which would be used to model the 30 minute ahead power demand via pseudo cross sectional analysis. This type of analysis allows us to avoid using one step ahead rolling window's in ridge regression to get good prediction accuracy. The process of including these lagged features is outlined in Kotillova et al. (2012) and is summarised as follows:

1. Create an ACF plot of the demand data. This is shown in Figure 15 in Appendix A.2.
2. Rank the lags in terms of the correlation (after taking the absolute value). We took the top 6 lags (i.e. Lag 1, 48, 336, 1008, 1680, 228). These are known as the 'primary' lags
3. Using primary lags, take the surrounding lags which are called 'neighbouring' lags. The more important the primary lag (e.g. Lag 1 and 48), the more neighbours that were included as these were deemed to be the most important for prediction.

The primary lags, and their corresponding neighbours, are detailed below. Remembering that one period is equivalent to 30 minutes, therefore, allowing us to determine what time period they represent.

- **Lag 1** - 2,3,4,5,6,7 (6 neighbours from the previous lags) - period immediately before
- **Lag 48** - 45,46,47,49,50,51 (3 on either side) - 1 day before
- **Lag 336** - 334,335,337,338 (2 on either side as less important) - 1 week before
- **Lag 1008** - 1006,1007,1009,1010 (2 on either side) - 3 weeks before
- **Lag 1680** - 1679,1681 (1 on either side) - 35 days before
- **Lag 288** - 287,289 (1 on either side) - 6 days before

Furthermore, the lagging process created missing values within the dataset. These were removed to prevent any issues with the machine learning modelling. This corresponded to less than 0.5% of the data so this shouldn't have a major effect on the size of our dataset.

5.3 Data Split for Modelling

Table 1 shows how the data was split for the machine learning models (linear and neural network).

Table 1: Data Split

	Period	Approximate Percentage (%)
Train	May 2016 - Dec 2016	63.0
Validate	Jan 2017 - Feb 2017	18.5
Test	Mar 2017 - Apr 2017	18.5

This split does have some selection bias as the sets cannot be randomly shuffled before the division. However this bias was minimised by random shuffle after the split in the machine learning models. This was done to ensure that the machine learning and time-series models could be appropriately compared over the same period. For the time-series models there was no validation set as alternative model selection methods (e.g. AIC) were used by convention.

5.4 Stationary check and transformation

Times-series data is considered stationary when the joint distribution of the time series does not depend on time. In practice, the more common case is weakly stationary, which means that both mean and variance are constant over time and the covariance only depends on the time horizon. Stationary data is required for general time series model, such as ARMA. We check stationary by plotting autocorrelation function (ACF) and partial autocorrelation function (PACF) plots for each variable as shown in the appendix (Figures 26-28).

Based on these plots we conclude that demand and temperature require a first and seasonal difference and price requires a first difference. The ACF and PACF plots, after differencing, are shown in Figures 3-5.

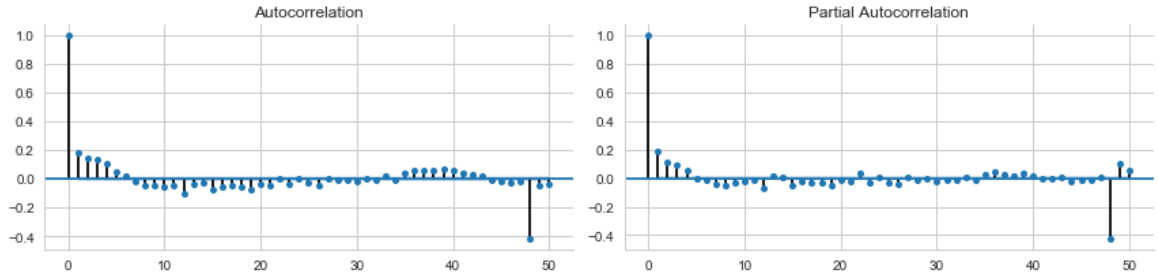

Figure 3: ACF/PACF plot for demand variable

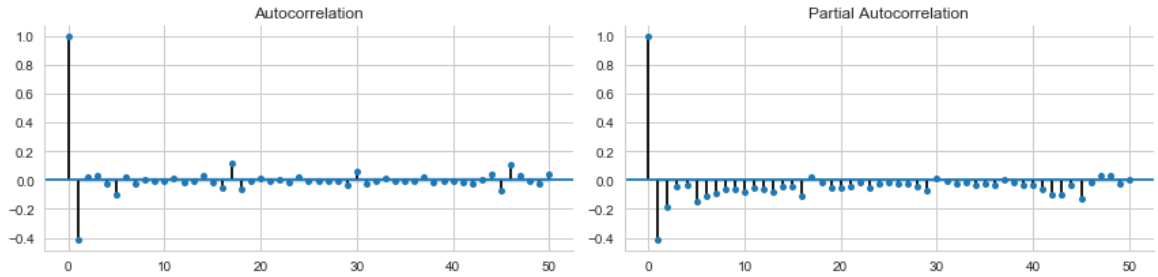

Figure 4: ACF/PACF plot for price variable

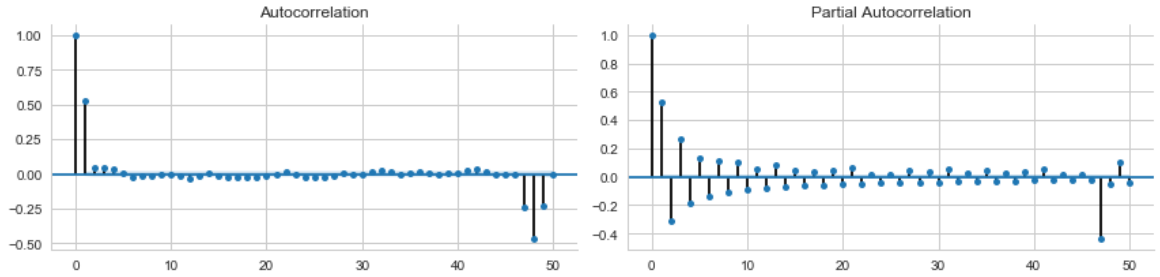

Figure 5: ACF/PACF plot for temperature variable

In order to test if the data is now stationary we used the Augmented Dickey Fuller (ADF) test. This test has a null that there is a one unit root time series so a rejection of this suggests stationary data. The results rejected the null and provided us with some evidence that we had successfully transformed the data into stationary.

5.5 Machine Learning EDA

EDA was also conducted on the lagged demand features and the dummy variables created in this section. As expected, this analysis did not reveal any interesting conclusions due to the similarity in the features. For instance, the variable, 'lag_2' has identical observations to 'lag_1', except that is has been shifted in time by a single period. The only differences in the data were related to the dummy variables. For example, the power demand in Winter was higher than the reference season of Autumn, as expected. Regardless, a small subsample of the plots, are provided in Appendix **(insert reference)**, for reference.

6 Modelling

This section outlines the various modelling techniques used to accurately predict the 30 minute ahead power demand within South Australia. Namely, several classes of models were considered, including: baselines - random walk and industry standard models for comparative purposes, time-series models - traditional forecasting methods, machine learning models - evaluating complex non-linear models such as neural networks, and simpler linear models such as linear regression.

6.1 Model selection criterion

We selected two metrics to assess the performance of our models. The Mean Absolute Percentage Error (MAPE) and the Mean Absolute Error (MAE). According to Kotillova et al. (2012) these are the two most common metrics for reporting electricity forecasting in industry and academic settings respectively. Using these metrics we aimed to find the best linear model and the best neural network model using the validation set.

On the other hand, the time-series models used either the Akaike Information Criterion (AIC) or visual identification for model selection. For instance, the order selection for ARIMA models could be achieved using ACF and PCF plots as per the previous section, or using an automatic selection process based on AIC.

Nonetheless, using the test set, the best machine learning model and the best time-series model could be compared against the industry models the the time-series benchmarks. For this particular activity, the MAPE and MAE were reported for the test period outlined in Section 5.3 in Table 1.

6.1.1 MAPE

MAPE is considered the "preferred metric by [power] industry forecasters" (Kotillova et al. 2012) and will be our primary model selection measure. It is widely used due to its simplicity and ease of comparison between models as it expresses accuracy as a percentage. The scale independence of MAPE is a advantage compared to MAE. The MAPE is defined as per Equation (6).

$$MAPE = \Big(\frac{1}{n}\sum_{j=1}^{n}\frac{|y_j - \hat{y_j}|}{y_j}\Big) \times 100 \tag{6}$$

The MAPE does have a few limitations that need to be considered when using it. To prevent division by 0 none of the demand values can be 0, we have protected against this by filling any missing values with reasonable replacements as power demand should never be zero. The other issue is that the measure puts a higher penalty on overestimated values compared to under estimated values. This can lead to a tendency to select models that under forecast values over a model that over forecasts. In our circumstance under estimation of demand is problematic so we consider MAE alongside MAPE.

6.1.2 MAE

MAE is a widely used metric by researches in time series analysis that estimates accuracy by averaging the absolute errors of prediction. It is defined as per the following equation,

$$MAE = \frac{1}{n}\sum_{j=1}^{n}|y_j - \bar{y_j}|. \tag{7}$$

MAE is a scale dependent metric meaning that it produces results in the same scale as the data being measured. Therefore, caution should be used when comparing models with transformations applied.

6.2 Justification for models selected

Based on the paper by (Kotillova et al. 2012), we have tried four baseline models for comparison. These are naive prediction methods. The best performing baseline model was the random walk one step ahead. This makes sense as most time series processes are dependent on prior outcomes.

6.3 Industry standard model

Based on the paper by Kotillova et al. (2012), we use their definition of the industry model. This model predicted one step ahead t+1 demand using:

1. The previous 5 lags (t, t–1, t–2, t–3, t–4)
2. The same time (t+1) in the previous week
3. The previous 5 lags in the previous week

These features have a logarithmic transformation applied then the differences between the successive values are calculated. This is leads to a model with 9 features (Sharmsollahi et al., 2002).

This industry model is used to predict 5 minute ahead demand for the New England area (located in America). Different data sources are used to train this model so it does have some limitations when comparing against our model. Despite this model being used used in a different context to the Adelaide demand 30 minute ahead prediction modelling we still believe it acts as a good industry benchmark.

6.4 Time series models

We considered a wide range of time series models including Seasonal Exponential Smoothing (ES), Seasonal ARIMA, Bayesian structural time series (BSTS) and Auto regressive (AR) models. We discarded ES as it performed very poorly in all measures of accuracy suggesting that ES models are not appropriate for this data. ARIMA and BSTS models could not beat AR model performance and were discarded due to their complexity In ARIMA models our seasonality is 48 periods (one day) which makes the rolling window method very slow and BSTS modelling takes even longer to train without much improvement in performance. Due to these limitations we discard them and focus on the best performing model, AR.

6.4.1 Auto Regressive models

AR models use a linear combination of the output variable's previous values to forecast future values. In this case we use the previous values of demand in order to predict the next value of demand. AR models can be expanded to include seasonality. In order to select the number of lags (p) to include in the model we use the partial autocorrelation plot to select the maximum lag as the one beyond where the PACF goes to zero. An AR(p) model is of the form shown in Equation 8 where the prediction is based on a weighted value of the previous demand value.

$$X_t = c + \sum_{i=1}^{p} \phi_i X_{t-i} + \epsilon_t \tag{8}$$

AR models can be expanded to include seasonality. Hence when there is seasonality the stationary assumption is violated as the average demand during summer is higher then in winter. In order to address this seasonal differencing is required which removes any trend to produce stationarity.

6.5 Machine Learning Methods

6.5.1 Linear regression

Multiple linear regression (MLR) is basic model useful as a comparison between other models. The key assumptions for the model to be valid as as follows.

1. Linear: the time series follows a model which is linear in its parameters
2. Zero conditional mean: expected value for error term is zero if we have all the data.
3. No perfect collinearity: no independant variable is constant or a perfect linear combination of the others.
4. Homoskedasticity: errors have constant variance across time
5. No correlated errors: errors in two different time periods are uncorrelated
6. Independent and identically distributed errors

These assumptions are rarely satisfied in practice, particularity the linear assumption. It is important to check these assumptions to give context to the MLR model results. Therefore we consider methods such as variable selection and shrinkage methods to improve MLR.

Shrinkage methods: Ridge regression
Ridge regression is a shrinkage method that introduces the tuning parameter λ into the least squares fitting procedure. This parameter adds a l_2 norm penalty term, $\lambda \sum_{j=1}^{p} \beta_j^2$, to the coefficients to drag them towards zero. The purpose of this penalty is to reduce the variance of the prediction with a small increase in bias (James et al., 2015). Ridge regression works best when the least squares estimates have high variance. However the tuning parameter will never decrease a parameter to 0 and therefore cannot perform variable selection. When these variables are not all important Ridge will retain them and produce a less interpretable model. The tuning parameter is chosen using cross validation to select the best value.

Shrinkage methods: LASSO
The Least Absolute Shrinkage and Selection Operator (LASSO) was developed to address the limitation of Ridge regression that all p predictions are included in the model. LASSO is able to perform variable selection by forcing some coefficients to 0 if the tuning parameter is large enough. To do this it adds an l_1 norm penalty term $\lambda \sum_{j=1}^{p} |\beta_j|$ to the residual sum of squares. The result is a model that is easier to interpret with a reduction in variance and slight increase in bias. The tuning parameter is also chosen via cross validation. A disadvantage of this model is that in situations where none of the true parameters are equal to zero LASSO can exclude these variable leading to worse prediction.

Shrinkage methods: Elastic net
Elastic net attempts to combine the advantages of Ridge and LASSO by including both a l_1 and l_2 norm penalty term. The result is a shrinkage method that can still perform variable selection like LASSO and reduces coefficients of correlated predictors like Ridge regression.

6.5.2 Neural networks

Neural networks are a machine learning method inspired by the interconnected node structure of human neurons. In their most basic form neurons receive information from input nodes, process the information in some way and produce an output (Hippert et al., 2001). They have the advantage of being incredibly adaptive to complex patterns due to their non-parametric nature. They make no assumptions on the structure of the input data and are usually able to 'learn' complex patterns much better than traditional methods. The main disadvantage of neural networks is they act as a black box which makes it hard to interpret the model.

6.6 Time series: Model selection and diagnostics

The classical linear model assumptions for time-series regression are what AR and linear models are based on. See section 6.5.1 for the assumptions.

6.6.1 Auto regressive models

Based on the results from our complex seasonality plot in section 4.3, as the weekly seasonality seems stronger than daily seasonality, we try three AR candidates. AR(1), AR(48) (in-day seasonality) and AR(336) (in-week seasonality).

AIC
With larger lags AIC always decreases in the training data so that when lag reaches 336 AIC approaches 0, shown in Figure 31 (appendix). Therefore selecting the model with the lowest AIC will just select the most complex model and may result in overfitting. Hence we also considered AR(48) as we observed a large decrease in the AIC plot at p=48 (Figure 31).

Despite these concerns of overfitting the fact that our data is such high frequency makes AR(336) reasonable despite containing over 300 parameters.

Table 2: AR Models AIC: training data

	AR(1)	**AR(48)**	**AR(336)**
AIC	37689.72	164	0.003

ACF
After these 3 models were chosen for consideration we performed a residual check using an ACF plot. The AR model that satisfies the residual check was AR(336) shown in figure 6. All of the remaining ACF plots are available in the appendix (Figures 29 and 30).

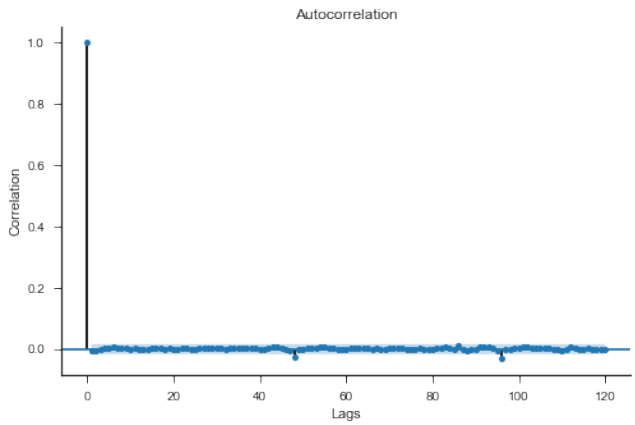

Figure 6: ACF plot for AR(336)

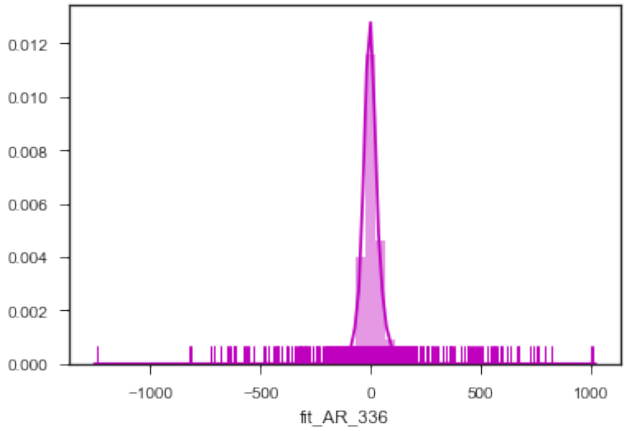

Figure 7: Histogram and kernel density estimate for AR(336)

Based on these plots we select AR(336) as this model best satisfies the assumptions of stationarity and normal distribution. There is some evidence that there is still some outliers based on the long tail potentially not completely satisfying the white noise assumptions (see Figure 7) .However this model satisfies our assumptions the best. The overfitting might be a problem in the AR(336) model due to too many estimated parameters, but considering that our dataset is high frequency and the simplest in-day seasonality includes 48 observations already, 336 order should be reasonable here.

Our model selection and diagnostic processes considered a variety of AR models but we focus our further testing on the AR(48) and AR(336). AR(1) has very poor AIC performance and ACF plot so we drop it from consideration.

6.7 Machine learning: Model selection and diagnostics

6.7.1 Linear Models

The hyperparameters for the linear models were selected using cross-validation (CV) for our three regularised linear models. CV provides an estimate of the out-of-sample MSE. We used 5-fold CV on the training data to select the complexity parameters. 5-fold CV randomly partitions the data into 5 groups using 1 fold as the proxy validation set and the remaining 4 as the training set. The MSE is calculated on the observations in the left out fold and the procedure is repeated for each fold. The final CV MSE is an average of the results for each fold. We used k fold CV as it is less computationally expensive compared to leave one out cross validation.

The hyperparameters selected for the three regularised linear models are shown in Table 3.

Table 3: Linear Models: Hyperparameter Selection

	Hyperparameter/s
Ridge	$\alpha = 1.400$
LASSO	$\alpha = 0.165$
Elastic Net	$\alpha = 0.057$, l_1 ratio $= 1.000$

What is interesting about Table 3 is the choice of l_1 ratio for the elastic net model. By selecting a value of 1, the elastic net model essentially moves towards the LASSO model. Therefore, we expect these two models to have similar results.

These complexity parameters shrink the coefficients for the ridge and elastic net models, and may also have the effect of removing coefficients all together for the LASSO and elastic net cases. Nonetheless, as

will be described in the next section, the ridge model has the highest performance on the validation data, therefore only its coefficients are interpreted here. Figure 8 shows the ridge model coefficients.

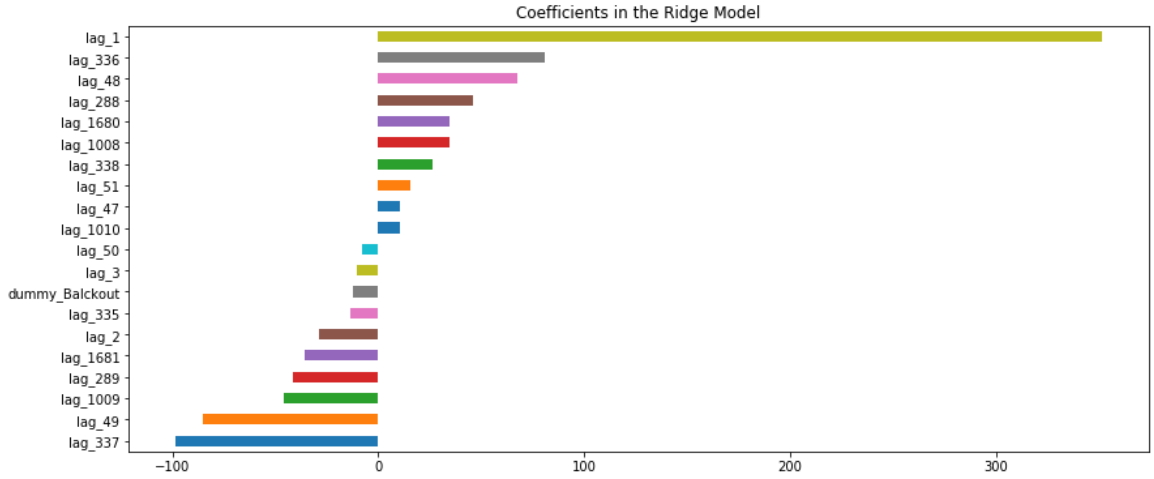

Figure 8: Ridge Regression Coefficient Plot

In Figure 8 larger coefficient, in terms of magnitude, indicates a larger marginal effect on the response. It should also be noted that our model includes both continuous and dummy variables. Where, the continuous variables can take on a larger range of values (not 0 or 1), therefore, they can have a larger impact on the demand prediction. The most important variables selected by ridge were:

- **Lag 1**: demand 30 minutes before
- **Lag 337**: demand 7 days and 30 minutes before
- **Lag 49**: demand 1 day and 30 minutes before

The effect of seasonality when predicting power demand is clear in the ridge regression model. In particular, this model picks up on the weekly and daily seasonality when predicting the 30 minute ahead power. These results are similar to the time-series models, which also used the same seasonality, to make these predictions as well. Despite their differences, it is interesting to note how these types of forecasting approaches have converged to a similar conclusion.

6.7.2 Neural networks

Similarly to shrinkage methods neural networks have a range of hyperparameters that need to be selected. These include selecting the number of hidden layers and number of neurons in each layer. These add further complexity to the model and may lead to overfitting. Dropout is used to combat this by randomly removing some neurons from the model. Each neuron needs to have a weight which is applied to the output depending on its importance. These are updated but need to be initialised. Once initialised the weights are updated using an optimiser algorithm. These are updated by running the data through the network a certain number of times known as epochs. Normally the data set is too large to be sent through in a single run so a batch size is also set. All these parameters need to be tunned to minimise the test error.

The hyperparameters were selected using a process of cross-validation randomised grid search. This process follows these steps:

1. The hyperparameter space is constructed. This means that all the permutations using the above hyperparameters are found which is a very large set of options
2. A random number models are selected from this hyperparameter space (75 in this case)
3. Each model undergoes 3-fold cross validation to find the MSE for each fold on the training data.

4. Take the mean of the 3-folds for each of the 75 models

5. Select the 3 models which have the lowest mean CV score based on the MSE

The parameters of the 3 best models are outlined in Table 10 in Appendix D.3. These are the three models which will be tested alongside the linear and time-series models on the validation data.

7 Model Validation

The goal of this section is to identify the best performing models on the validation data to identify the best: linear model, NN model and time-series model.

7.1 Validation results

7.1.1 Linear models

The model validation results for linear models are shown in Table 4.

Table 4: Model Validation Results for the Linear Models

	MLR	**Ridge**	**LASSO**	**Elastic Net**
MAPE (%)	1.80	**1.79**	28.65	28.66
MAE (MW)	**24.55**	24.65	25.60	24.89

In terms of our primary metric, Ridge models performs the best on the MAPE. Therefore, this will be selected for further model evaluation.

Surprisingly, the MLR model performs well on the validation data (even outperforming Ridge in terms of the MAE). However, this model is expected to have multicollinearity issues as our input variables are the lagged time-series values. Therefore, we dont expect this model to uphold its assumptions (which Ridge regression overcomes).

LASSO (and Elastic net which is very similar in this case) both perform really poorly in terms of the MAPE. This is due to the definition of MAPE in Section 5.1 Model Selection Criterion. Namely, these models are probably mispredicting values which are very close to 0. This blows up the MAPE making it get such a large result.

Nonetheless, we select Ridge regression for further evaluation.

7.1.2 Neural network models

When training and validating the neural network it became apparent that the choice of dummy variables was inappropriate for this problem. This is in particular reference to the seasonal dummy variables introduced in Section 5.1.3. Since this problem required the machine learning models, including the neural networks, to be evaluated against the time-series models, a consistent test was required as introduced in Section 5.3. However, as outlined, this induces sampling bias within the data.

The particular periods which were selected for validation and evaluation were the warmer months. Therefore, power demand was likely to be higher due to increased cooling requirements. However, since the training data did not include many observations where 'dummy_Summer' and 'dummy_Autumn' were equal to 1, the predictions in the out-of-sample set were systematically upwards biased as per Figure 36 in Appendix F. It is clear from Figure 36 that for any given predictions, *all* the residuals are positive, by a significant amount. This is a clear indication that the predictions of power demand are systematically inaccurate by approximately 550MW. This is the effect of not including a dummy variable for the warmer months within the training set. Thus, it is concluded that the use of seasonal dummy variables inappropriate for neural

network modeling.

Nonetheless, the validation scores for neural network models, after excluding the seasonal dummies, is shown in Table 5.

Table 5: Model Validation Results for the Neural Network Models

	NN 1	**NN 2**	**NN 3**
MAPE (%)	**1.72**	1.81	1.96
MAE (MW)	**23.65**	25.60	26.32

NN 1 was our best NN model on the validation data set so we select this model for evaluation on the test data.

7.1.3 Time-series model

As previously outlined in 6.6.1 the best time series model was selected based on AIC not on validation data as is convention in time series models. This lead us to select AR(48) and AR(336) for further evaluation.

8 Model Evaluation

Everything in this section is using the test data. Here, we want an idea of how well our model will perform on unseen data and how generalisable it is. We also want to compare it to the relevant benchmarks too.

8.1 Time series models evaluation

Table 6 shows the performance of the time series models on the test data. From these results the best performing model was AR(336). In order to calculate the MAPE and MAE for time series models we used rolling window, with the fixed window length of training data, to calculate the forecast.

Table 6: Model Evaluation for the time series models by Rolling Window one step ahead

	AR(48)	**AR(336)**	**RW one step ahead**	**RW Seasonal (In day)**	**RW Seasonal (In week)**
MAPE (%)	2.56	**1.83**	3.00	12.12	14.64
MAE (MW)	33.58	**23.60**	39.64	161.10	194.07

8.1.1 Hypothesis Tests for model comparison

Granger Causality test

The Granger Causality Test acts as a variable selection hypothesis test. The test checks if adding price or air temperature improves demand forecasting when we are already using demand's history for prediction. From this test we reject the null hypothesis that these variables don't help demand prediction (Appendix Table 11 and Table 12). This suggests that price and temperature have contributed something to the model.

Diebold-Mariano test

The Diebold-Mariano test compares the forecast accuracy of two forecast methods and acts like model selection. The null hypothesis is that the two methods have the same forecast accuracy.

From this test we can conclude

- AR(336) gives a better forecast accuracy compared to the benchmark model RW one step ahead
- AR(336) gives a better forecast accuracy compared to AR(48)
- NN model gives a better forecast accuracy compared to Ridge
- NN model gives a better forecast accuracy compared to AR(336)

These results confirm NN as our best time series model.

8.1.2 Final model comparison

We selected the best linear, best NN and best time series model. They were compared against the relevant benchmarks and the results are shown in the following table.

Table 7: Test performance for selected models

	Ridge	**NN Model**	**Industry**	**AR(336)**	**AR+Ridge**	**Rand Walk**
MAPE (%)	2.89	**1.48**	12.60	1.83	1.61	3.00
MAE (MW)	38.73	**19.72**	174.11	23.60	20.73	39.64

Therefore we conclude that the best model based on both the MAPE and MAE is the NN model. The one step ahead forecast is shown in Figure 9.

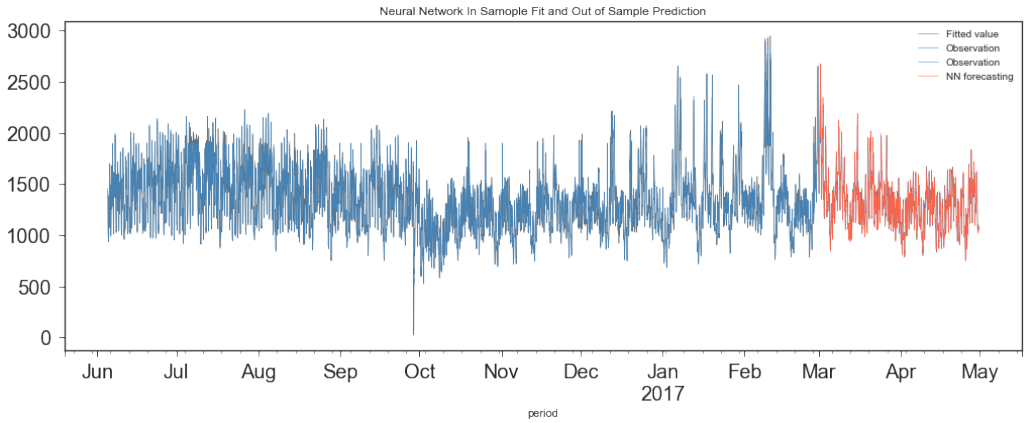

Figure 9: NN one step ahead forecast

Combination Forecasts Mothod

Combining forecasts have been shown in the literature to greatly improve forecasts and also intuitively makes sense as well. If we have two forecasts for the same time period from 2 different models we denote this as $\hat{y}_{t+1}^{(1)}$ and $\hat{y}_{t+1}^{(2)}$. The respective forecast errors are denoted as $e_{t+1}^{(1)}$ and $e_{t+1}^{(2)}$. From this, letting λ be the weight parameter, we denote the combined forecast as:

$$\hat{y}_{t+1}^{c} = (1-\lambda)\hat{y}_{t+1}^{(1)} + \lambda\hat{y}_{t+1}^{(2)} \tag{9}$$

From this, the variance of the combined forecast error is given in Equation 10.

$$Var(e_{t+1}^{c}) = (1-\lambda)^2\sigma_1^2 + \lambda^2\sigma_2^2 + 2\lambda(1-\lambda)\rho\sigma_1\sigma_2 \tag{10}$$

We can then optimise λ to minimise the variance, which gives us Equation 11

$$\lambda^* = \frac{\sigma_1^2 - \rho\sigma_1\sigma_2}{\sigma_1^2 + \sigma_2^2 - 2\rho\sigma_1\sigma_2} \tag{11}$$

However, we need to use estimates so we have Equation 10 where we use $\hat{\sigma}$ as an estimator for σ,to be the residuals in our estimated model.

$$\hat{\lambda^*} = \frac{\hat{\sigma}_1^2 - \hat{\sigma}_{12}}{\hat{\sigma}_1^2 + \hat{\sigma}_2^2 - 2\hat{\sigma}_{12}} \tag{12}$$

We attempted to combine the best performing models to improve this result further. We considered combinations of our three best performing models; NN with AR, NN with ridge, NN with Ridge and AR. All of these combinations could not improve the MAPE any further than NN alone. Combining AR(336) and ridge regression did improve the results of either model along which makes our combined AR and Ridge model the second best overall model as shown in Table 7.

8.2 Statistical inference

In this section of the report, we are aiming to quantify the uncertainty associated with the results in Section 8.1.2. Statistical inference, using a computation resampling method known as bootstrapping, is used to generate a 95% confidence interval on the test metrics. We perform this on the best performing model, which was the neural network model. Additionally, bootstrapping is used to determine whether the added complexity of the neural network, provides a statistically different result than the simpler ridge model.

8.2.1 Performance Metrics

Bootstrapping is used to determine a 95% confidence interval on the performance metrics (MAPE and MAE). We are going to bootstrap the observations, which is a non-parametric bootstrapping approach, since we are not interested in the model coefficients. In this process, we randomly draw $10,000$ bootstrap samples from the test data and use our best machine learning model to generate predictions. We then compare our predictions to the true power demand values, allowing the MAPE and MAE bootstrap statistics to be calculated.

From this, we can calculate the pivotal and percentile 95% confidence intervals for the neural network regression, as shown in Table 8.

Table 8: NN 95% CI

	Pivotal	**Percentile**
MAPE (%)	(1.53, 1.57)	(1.36, 1.40)
MAE (MW)	(19.99, 20.39)	(17.59, 17.98)

If we take the pivotal confidence interval for the MAPE, we can interpret this result by imaging if this investigation was repeated 100 times, then for 95 of those investigations we would find the true population value for the MAPE to be between 1.53% and 1.57%. The other intervals can be interpreted in a similar fashion.

Each of the confidence intervals in Table 8 demonstrate variation in the ranges. However, it is a positive sign that each of the four confidence intervals in the table are relatively narrow. The tightness of the intervals is evidence that the neural network produces fairly reliable results, suggesting that it has not been overfit in the training stage.

Moreover, it is interesting to note that the test statistics calculated for the NN in Table 7 (MAPE=1.48% and MAE=19.72MW), are not contained in any of the confidence intervals. While unusual, this is not an issue, because using our analogy of the repeated trials from before, this investigation may be one of 5 investigations (out of 100), which do not contain the true population value.

Finally, the high performance of the neural network model is also highlighted in Table 8. Even with confidence bands applied to the test scores, the upper interval for both the MAPE and MAE are still substantially less than the second highest performing model which was the combined AR Ridge model. Clearly, the ability of the NN to capture the non-linear relationships within the demand data is contributing to the success of its result.

Appendix H contains the distributions for the bootstrapped MAPE and MAE for the neural network model is shown in Figures 41 and 42.

8.2.2 Comparison Between Ridge and NN

We can also extend this bootstrapping process in the previous section to provide a comparison between the ridge regression model and the neural network model. While it would be beneficial to compare the best neural network developed in the study with the industry prediction model, it is clear that the industry model developed is simply inappropriate for 30 minute ahead prediction. Therefore, this section aims to determine whether the added complexity of a neural network model, provides a sufficient enough improvement in terms of predictive accuracy over the simpler ridge model (which would be less prone to overfitting). This time, our bootstrapped statistic is defined as:

$$MAPE_{difference} = MAPE_{ridge} - MAPE_{nn}$$

Since we are trying to minimise the MAPE, we want the statistic to be as positive as possible. Again, we can construct a pivotal and percentile 95% confidence interval shown in Table 9.

Table 9: 95% CI Comparison Between Ridge and NN

	Pivotal	**Percentile**
$MAPE_{difference}$	(1.52, 1.57)	(1.29, 1.34)

Using these intervals, it is also possible to construct a hypothesis test. Where we define our hypotheses as follows:

$$H_0 : MAPE_{ridge} \leq MAPE_{model}$$
$$H_1 : MAPE_{ridge} > MAPE_{nn}$$

Based on the results in Table 9, we observe that the 95% confidence intervals are both positive (and do not contain 0). Hence, we reject the null-hypothesis at the 5% significance level, and conclude that we have enough evidence to suggest that the neural network model proposed is statistically significant (i.e. lower MAPE) than the ridge model. This is clear validation that the non-linearity in the South Australian power demand is better modelled using the neural network, and the added complexity is an appropriate trade-off.

The bootstrap distribution of this statistic is contained within Figure 43 in Appendix H.3. A similar distribution of the difference has been created for the MAE in Figure 44. The resulting confidence interval has a similar conclusion.

8.3 Diagnostics

Neural network models are not constrained by the entire list often prohibitive assumptions imposed on Generalised Linear Models. However, it is still important to complete residual diagnostics to understand deficiencies which may be hold back the predictive accuracy. For instance, Figure 10 shows the residuals vs. fitted values, using the training data. The horizontal axis, at the top of the plot, shows the distribution of the power demand values, whereas the vertical axis plot shows the distribution of the residuals.

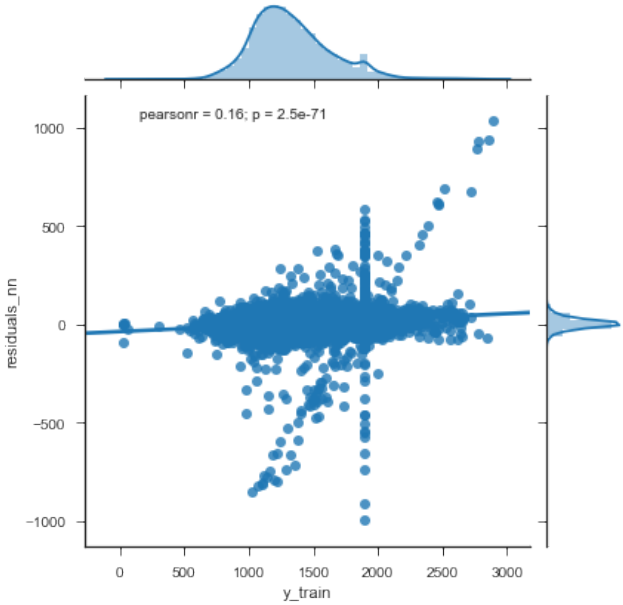

Figure 10: Residuals vs. Fitted - Neural Network

It is clear from Figure 10 that there are issues with the specification of the model, particularly with the two lines of residuals (vertical and diagonal). These systematic patterns suggest that there are linear relationships in the training data which have not been captured by the neural network. A similar plot is shown for the ridge regression case, in Figure 45, in Appendix H.4. From this plot, no clear systematic patterns are observed, suggesting that the ridge model is correctly capturing these linear effects. Hence, the neural network could be improved in the future by combining the predictions with the ridge model to ensure that these linear effects are correctly modelled. These systematic deviations aside, the neural network demonstrates a fairly small range of residuals, with no clear outliers.

9 Deployment

9.1 Trade off between ML model and Time-series model

Machine learning methods such as NN have the best accuracy. It could be implemented by updating the NN each month or quarter which is avoids having to continually retrain models. This is convienent as retraining the NN model takes significant time and computational costs. However NN models are act as a black box making them hard to interpret due to the complex concepts behind them. This makes them susceptible to overfitting to training data if not implemented successfully. They also require significant feature engineering into to function at the peak potential. This contrasts with Ridge Regression which is easy to understand and interpret however it produces less accurate forecasts.

Time series models also have their strengths in their easy interpretation and low computational cost. The AR(336) model took less than 15 minutes to produce 2 months of predictions. They also are less susceptible to the overfitting problem that machine learning methods have. However they are not as practical for implementation as machine learning methods as they would require frequent updating every half hour.

Hence in order to maximise the accuracy of our forecasts we select neural networks despite their drawbacks in interpretation.

10 Literature Discussion

It became clear that the results from similar studies were overstated, in particular with the predictive accuracy of neural networks for power demand. For example, the work conducted by Kotillova et al. (2012) and Bunn (2000) both concluded and praised the performance gains which could be obtained by using this type of algorithm. However, these studies failed to correctly partition their data as to prevent overfitting of this algorithm. In both investigations, the data was split into two: training and test data. While this is not uncommon, the paper's alluded to the fact that the test data was used for *both* hyperparameter optimisation and model evaluation. Because of this, their results are likely overfit to the test data, and if their models were to be tested on another out-of-sample set, then they would likely exhibit poor generalisability due to overfitting.

Moreover, it has been mentioned throughout this report, that modelling the effects of the seasons has been a large challenge to the group. For instance, Section 8 highlights that the use of seasonal dummies failed to improve predictive accuracy for the machine learning models, due to the way in which the data was partitioned. However, several authors overcame this problem by only analysing the the power demand within a single season (3-month period). For instance, Kotillova et al. (2012) only looked at the data from Winter, 2010. This meant that the authors did not have to model the complexities associated with the changing seasons, which in turn affects the power demand. Thus, is a major limitation of their work, creating overly optimistic results.

11 Conclusion

11.1 Future work

Panel data: The demand forecast for South Australia could be improved by considering the demand and price of other states. This technique involves combining time series data with cross sectional data to produce a more powerful multidimensional approach to forecasting.

Interval forecasting: Instead of forecasting a point estimate and the accuracy of that point estimate with measures such as MAPE and MAE future work could consider forecasting Bootstrap intervals. This could provide another dimension of information to energy operators. Also, the bootstrap interval forecasting could be implemented within a rolling window method.

11.2 Project outcome

Our best neural network model was able to improve on the industry model by 11.12%. This equates to savings of about $25 million for a 10-gigawatt generator (Hobbs et al. 1999). Improvements of this scale have the potential to have a real life impact for business's and people in Australia. These kind of savings will assist operators to reduce their costs and consumers could benefit with lower power bills in the competitive power market.

References

Bunn, D. W. (2000), 'Forecasting loads and prices in competitive power markets', *Proceedings of the IEEE* **88**(2), 163–169.

Data Dashboard (2018).
URL: *https://www.aemo.com.au/Electricity/National-Electricity-Market-NEM/Data-dashboard#aggregated-data*

De Livera, A. M., Hyndman, R. J. & Snyder, R. D. (2011), 'Forecasting Time Series With Complex Seasonal Patterns Using Exponential Smoothing', *Journal of the American Statistical Association* **106**(496), 1513–1527.
URL: *http://www.tandfonline.com/doi/abs/10.1198/jasa.2011.tm09771*

Griffiths, L. (2018), 'Heatwave triggers Code Yellow', *The Australian* .

Historical weather observations and statistics (2017).
URL: *http://www.bom.gov.au/climate/data-services/station-data.shtml*

Hobbs, B. F., Jitprapaikulsarn, S., Konda, S., Chankong, V., Loparo, K. A. & Maratukulam, D. J. (1999), 'Analysis of the Value for Unit Commitment of Improved Load Forecasts', *IEEE Transactions on Powcr System* **14**(4), 1342–1348.

Kotillova, A., Koprinska, I. & Rana, M. (2012), Statistical and Machine Learning Methods for Electricity Demand Prediction, *in* T. Huang, Z. Zeng, C. Li & C. S. Leung, eds, 'Neural Information Processing', Lecture Notes in Computer Science, Springer Berlin Heidelberg, pp. 535–542.

Moller, N. F. & Andersen, F. M. (2015), 'An econometric analysis of electricity demand response to price changes at the intra-day horizon: The case of manufacturing industry in West Denmark', *International Journal of Sustainable Energy Planning and Management, Vol 7 (2015)* .

SA blackout: Why and how? (2016), *ABC News* .

Sood, R., Koprinska, I. & Agelidis, V. G. (2010), Electricity load forecasting based on autocorrelation analysis, *in* 'The 2010 International Joint Conference on Neural Networks (IJCNN)', pp. 1–8.

A EDA

A.1 Time-Series Decomposition

Figure 11 shows how the time-series for the power demand can be separated into the four components.

Figure 11: Time series decomposition: Demand before outliers removed

Figure 12: Time series decomposition: Demand after outliers removed

Figure 13: Time series decomposition: Price before outliers removed

Figure 14: Time series decomposition: Price after outliers removed

A.2 Machine Learning EDA

When creating the lagged features in the machine learning model, an ACF plot was used to inform which lags were the most correlated with the demand. This is shown in the figure below:

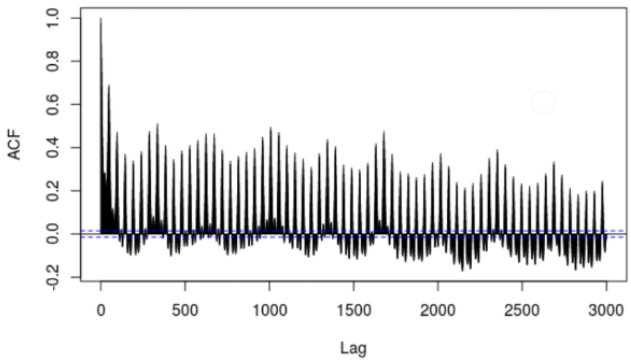

Figure 15: ACF Plot of South Australian Power Demand

After engineering the features, the following figures show the EDA conducted on the machine learning features. As many of the features contain the same values, they are excluded for brevity.

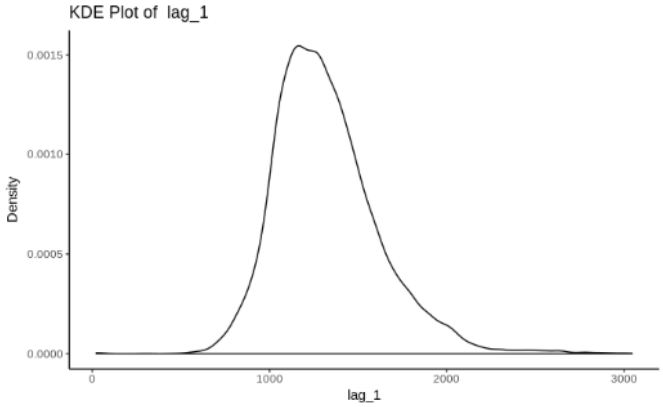

Figure 16: KDE of 'lag_1'

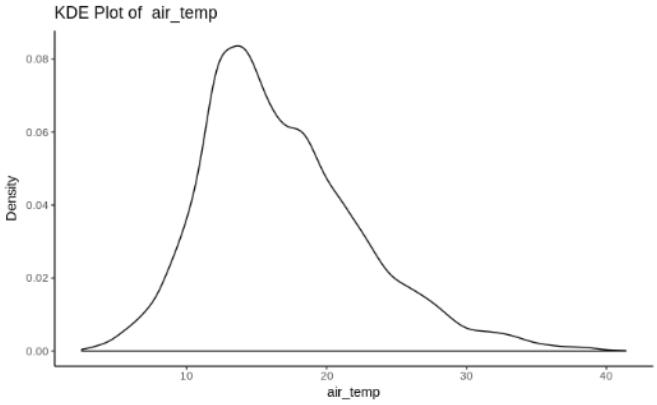

Figure 17: KDE of 'air_temp'

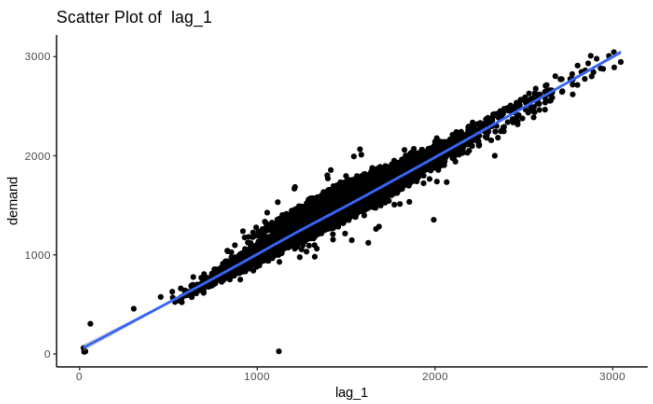

Figure 18: Scatter of 'lag_1'

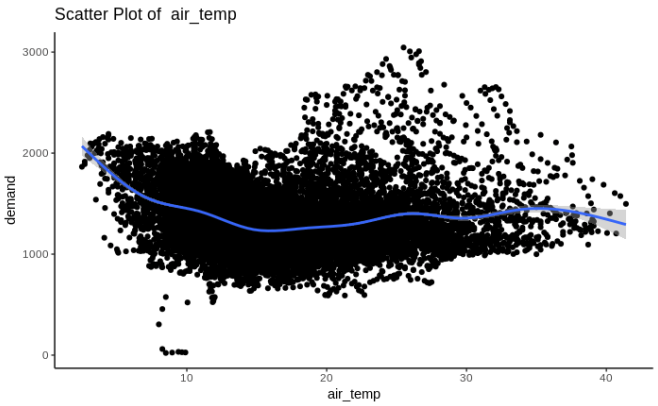

Figure 19: Scatter of 'air_temp'

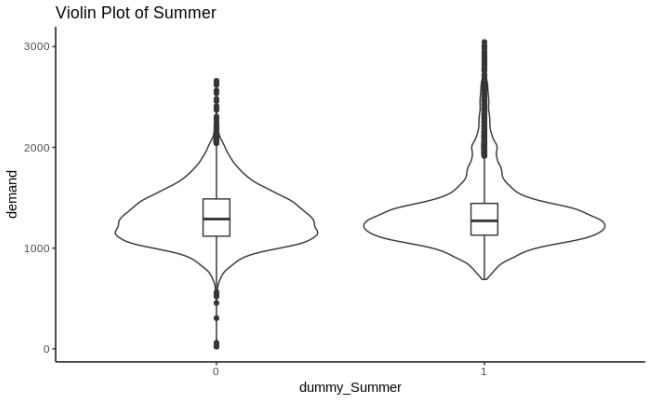

Figure 20: Violin of 'dummy_Summer'

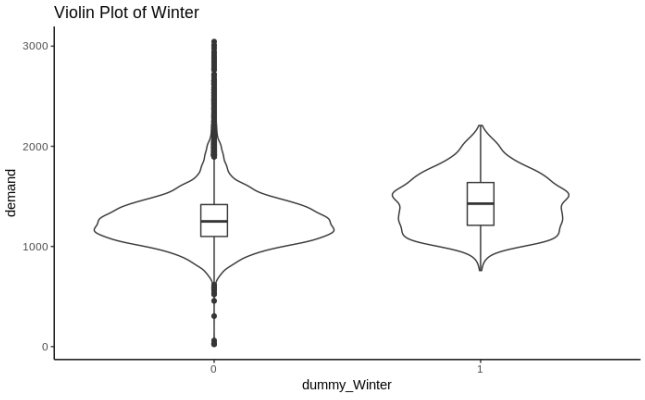

Figure 21: Violin of 'dummy_Winter'

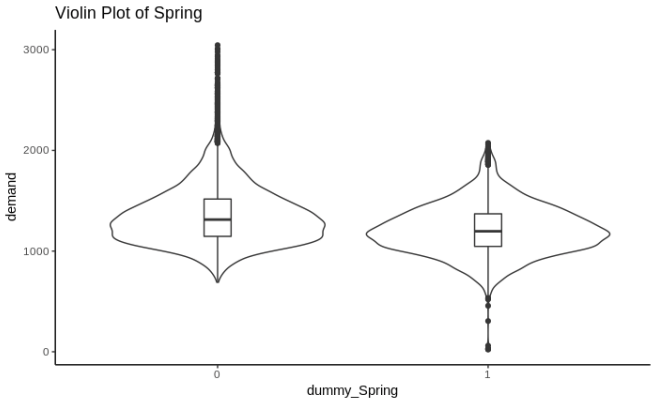

Figure 22: Violin of 'dummy_Spring'

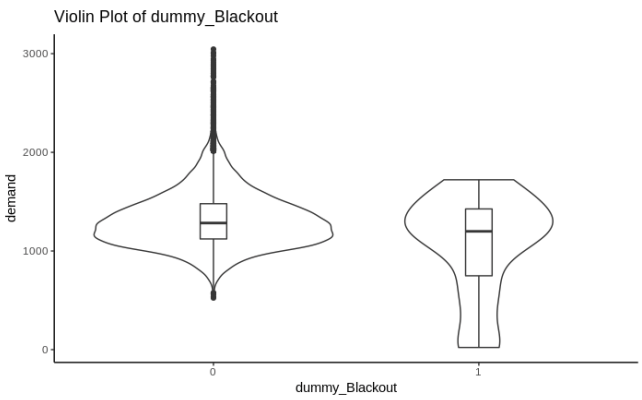

Figure 23: Violin of 'dummy_Blackout'

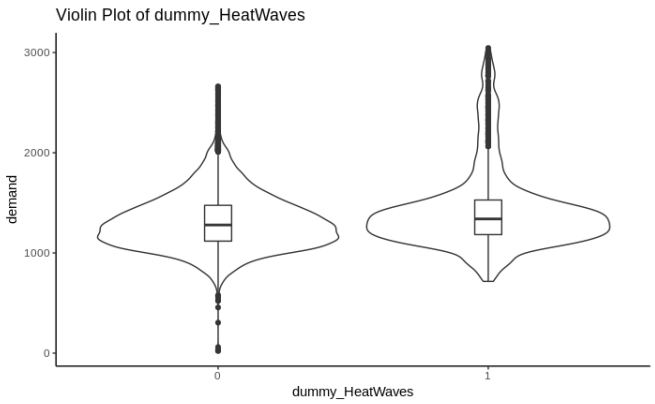

Figure 24: Violin of 'dummy_Heatwave'

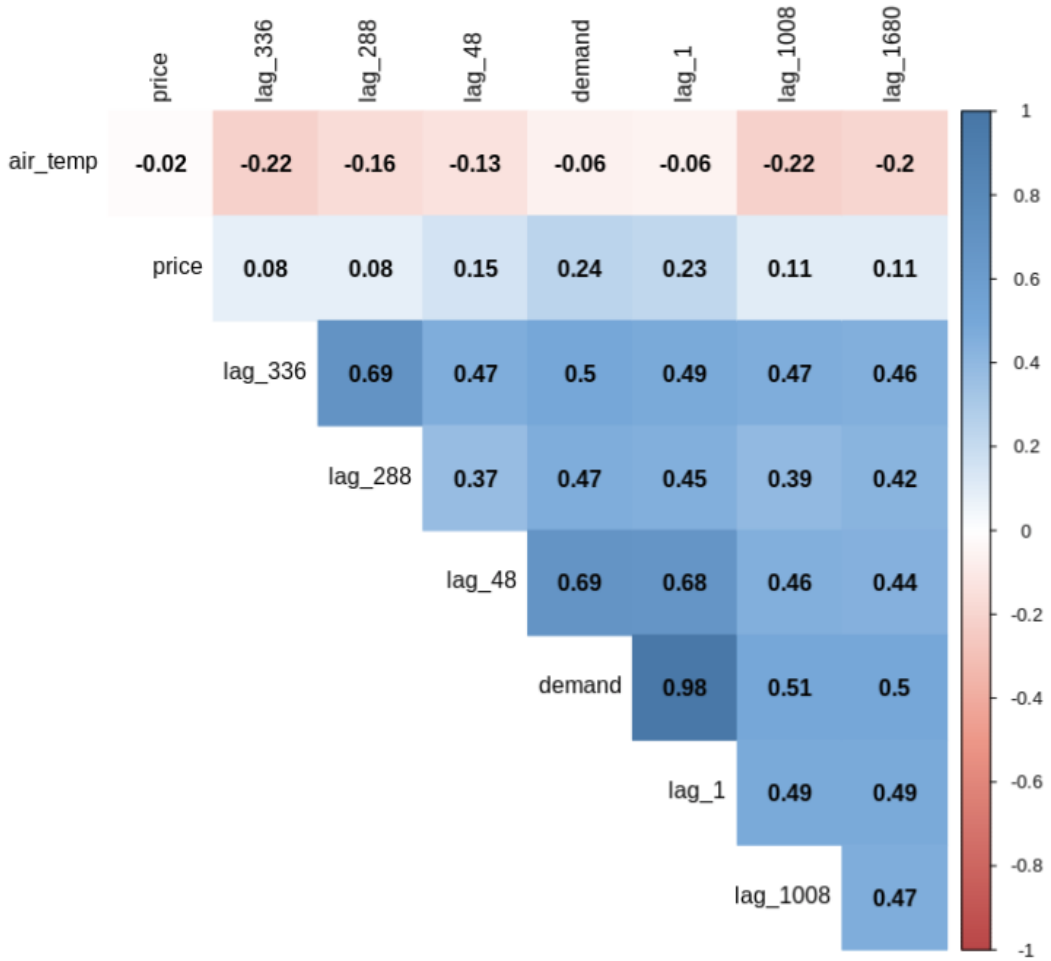

Figure 25: Correlation Heatmap of Select Variables

B Stationary check and transformation

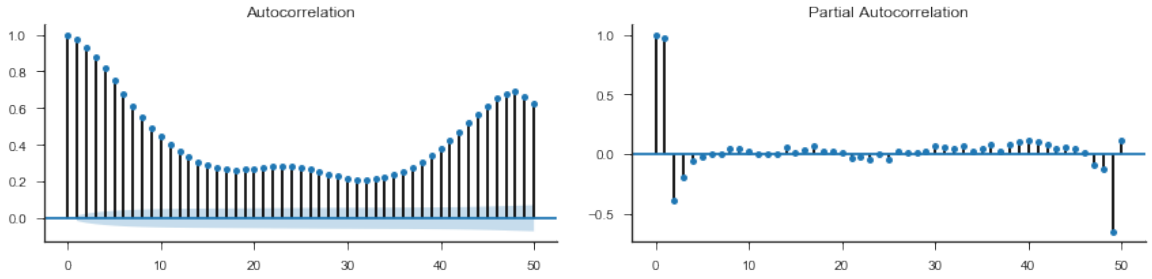

Figure 26: ACF/PACF plot for demand variable

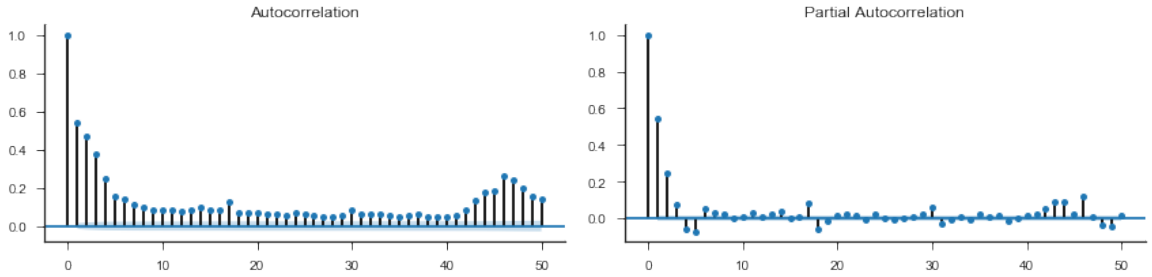

Figure 27: ACF/PACF plot for price variable

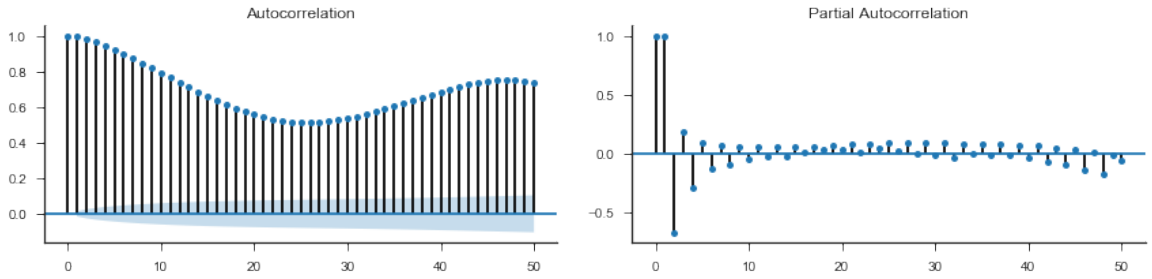

Figure 28: ACF/PACF plot for temperature variable

C Auto regressive models

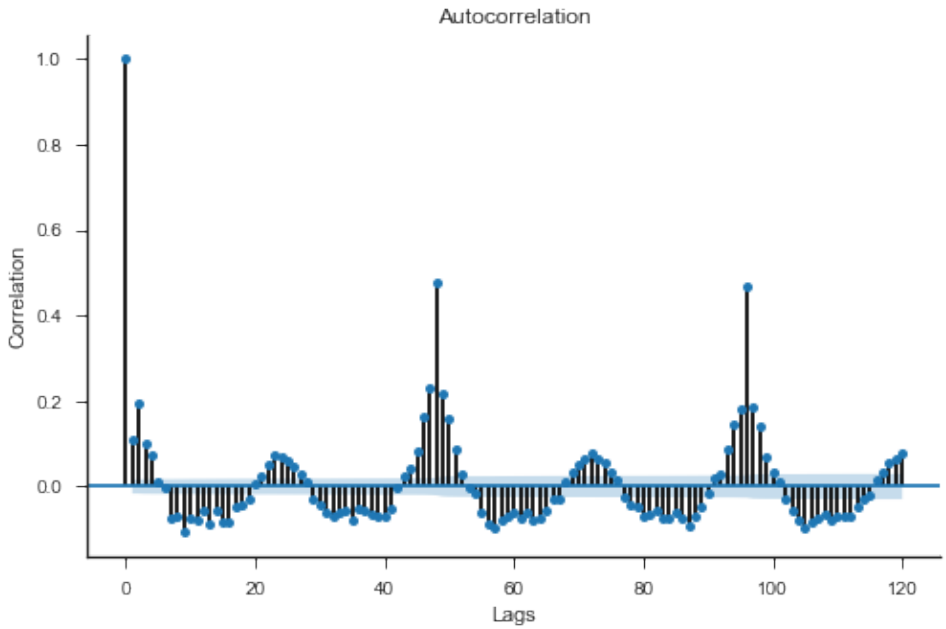

Figure 29: ACF AR(1)

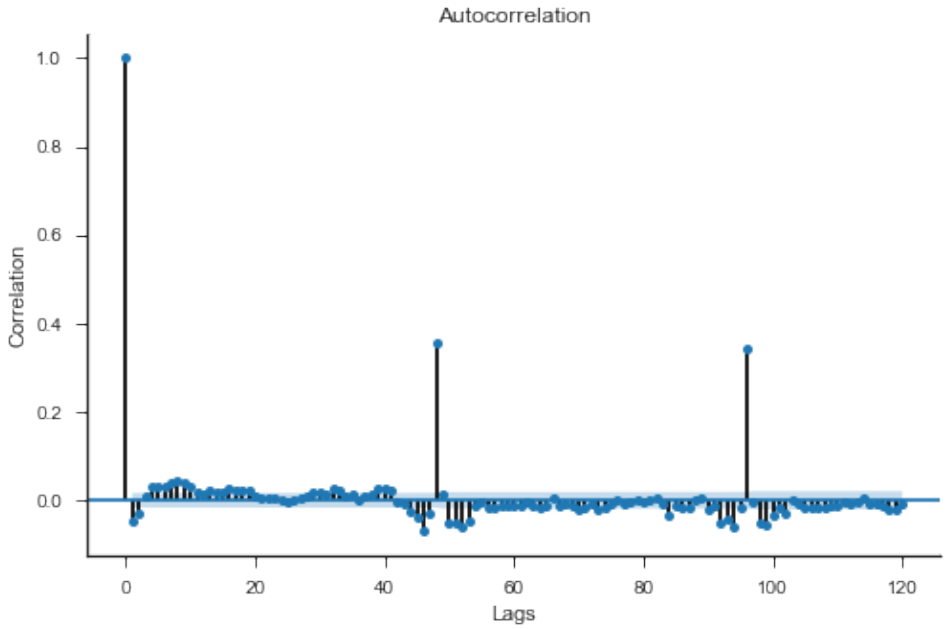

Figure 30: ACF AR(48)

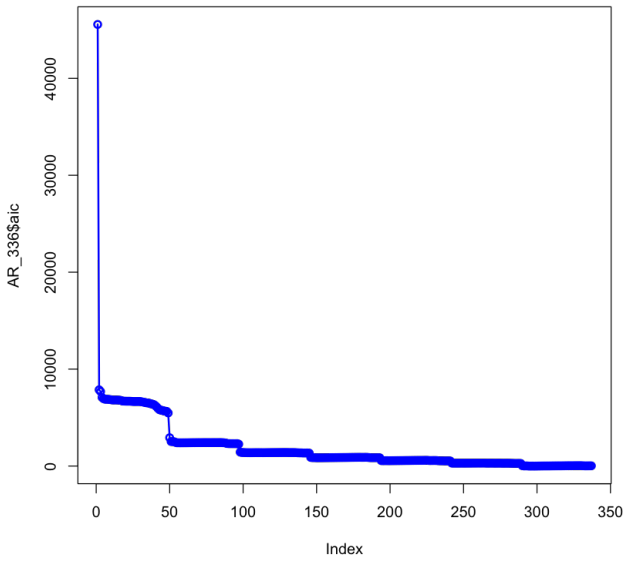

Figure 31: AIC plot by lag number

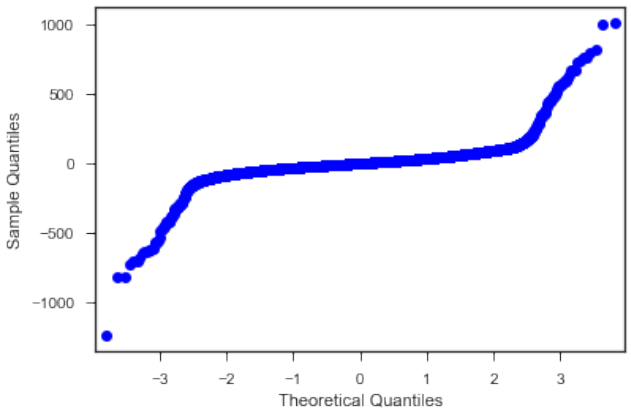

Figure 32: QQ Plot of AR(336)

D Model Validation

This Appendix outlines the further model validation that was completed in this investigation.

D.1 LASSO

The best alpha, based on cross-validation on the training data, was $\alpha = 0.165$. The coefficients are shown in the following plot:

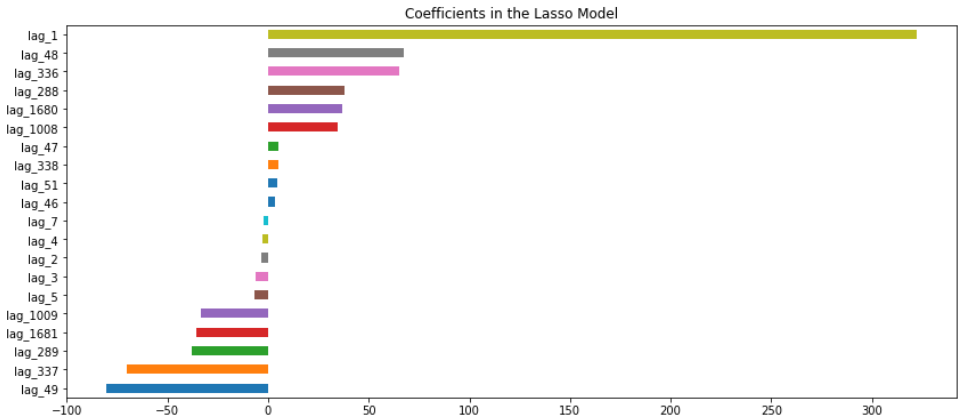

Figure 33: LASSO Coefficient Plot

The most important variables were the same as the Ridge case however the LASSO model shrunk 12 coefficients to zero. Hence the LASSO model has reduced variance and is more interpretable compared the ridge model.

D.2 Elastic net

The best alpha from our analysis was $\alpha = 0.057$ with the best l_1 ratio $= 1.0$. This ratio suggests that the elastic net has essentially picked LASSO regression and we expect the LASSO model and elastic net to have similar results.

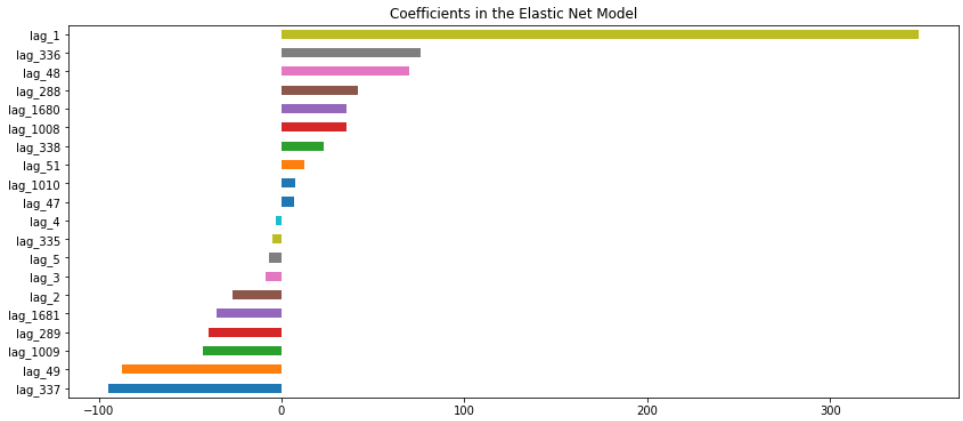

Figure 34: Elastic Net Coefficient Plot

The elastic net eliminated 8 variables in contrast to LASSO's 12.

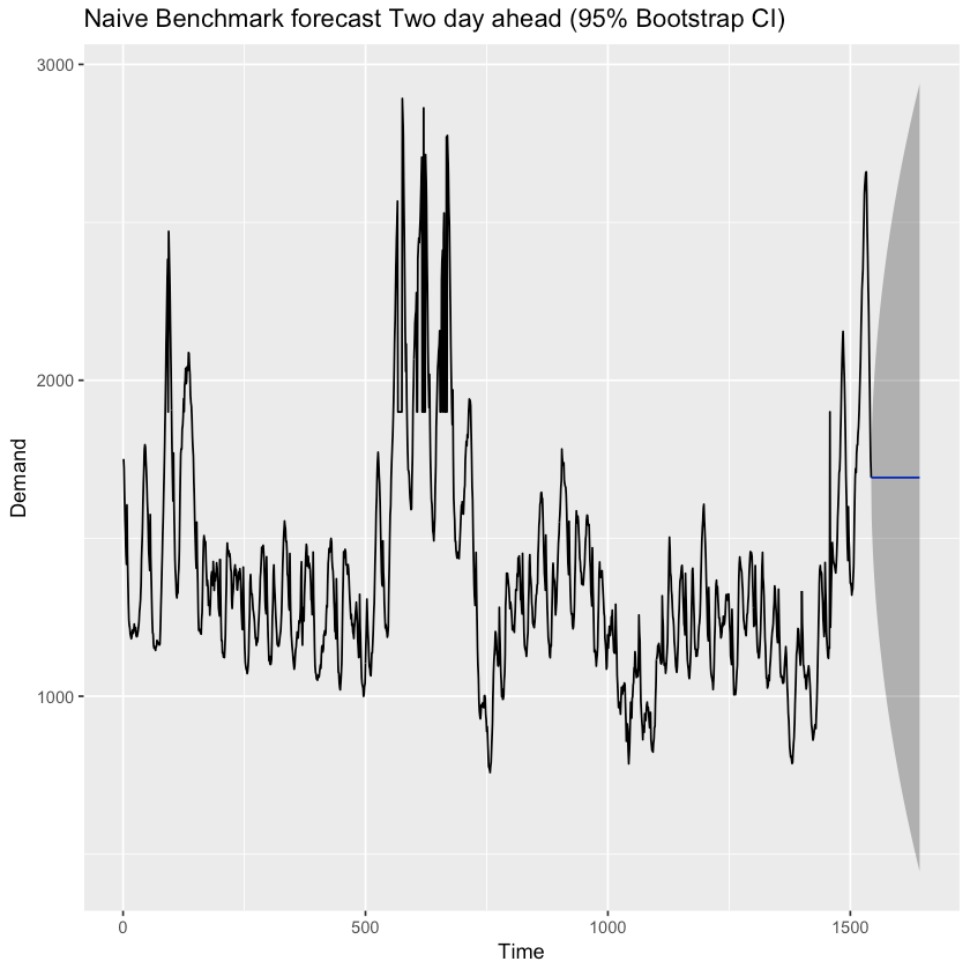

Figure 35: Forecast two days ahead Naive Benchmark 95% Bootstrap CI

D.3 Neural Network

The 3 best neural network models, based on their cross-validation randomised grid search scores, are shown in the following table:

Table 10: Hyperparameter Optimisation Results

	NN 1	**NN 2**	**NN 3**
Layer 1 Neurons	25	25	40
Layer 2 Neurons	25	40	30
Layer 3 Neurons	40	30	30
Weight initilisation	he_normal	he_normal	glorot_normal
Epochs	1200	900	1200
Batch size	50	10	10
Optimisation algorithm	nadam	nadam	adam
Dropout	0.2	0.2	0.2
Max. weight constraint	3	1	2

E Granger Causality Test Results

Model 1 is the unrestricted model that includes the Granger-causal terms. Model 2 is the restricted model where the Granger-causal terms are omitted. The test is a Wald test that assesses whether using the restricted Model 2 in place of Model 1 makes statistical sense (roughly speaking).

H0: No Granger causality

H1: Granger causality

As Pr(<F)< 0.01, we reject H0 for both price and temperature.

Table 11: Granger Causality Test: Temperature

Res. Df	**Df**	**F**	**Pr(>F)**
17306	NA	NA	NA
17354	-48	6.58	1.14e-40

Table 12: Granger Causality Test: Price

Res. Df	**Df**	**F**	**Pr(>F)**
16442	NA	NA	NA
16778	-336	2.47	3.20e-42

F Neural Network Residuals

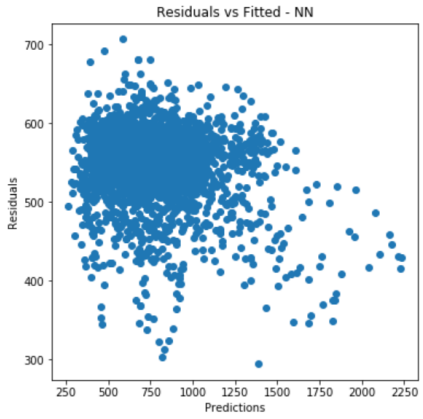

Figure 36: Residuals vs. Fitted - NN with Seasonal Dummies

G One step ahead forecast plots

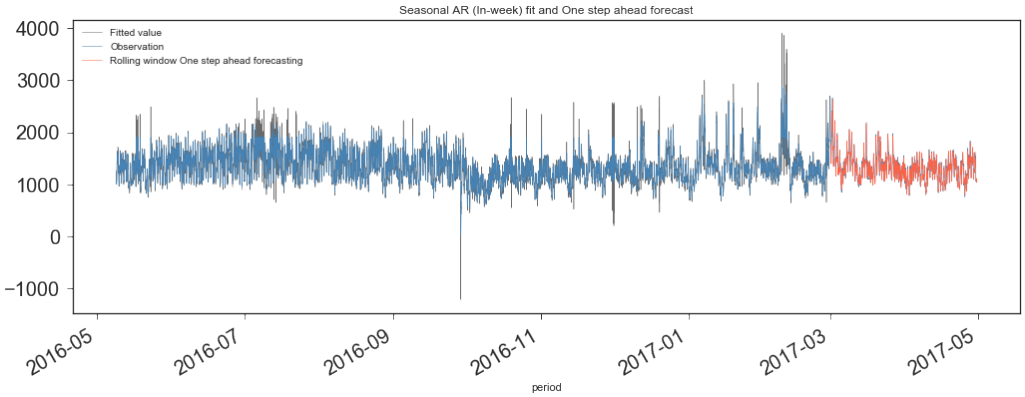

Figure 37: AR(336) one step ahead forecast

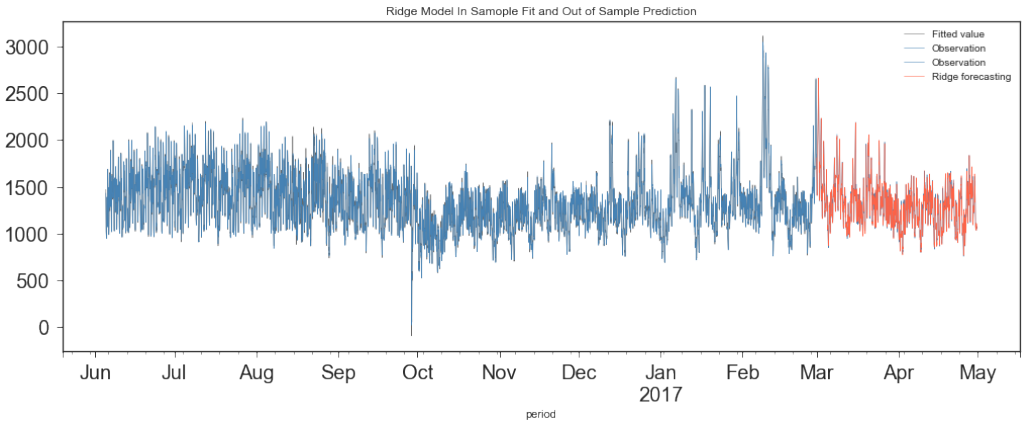

Figure 38: Ridge Regression one step ahead forecast

H Bootstrap

H.1 Ridge

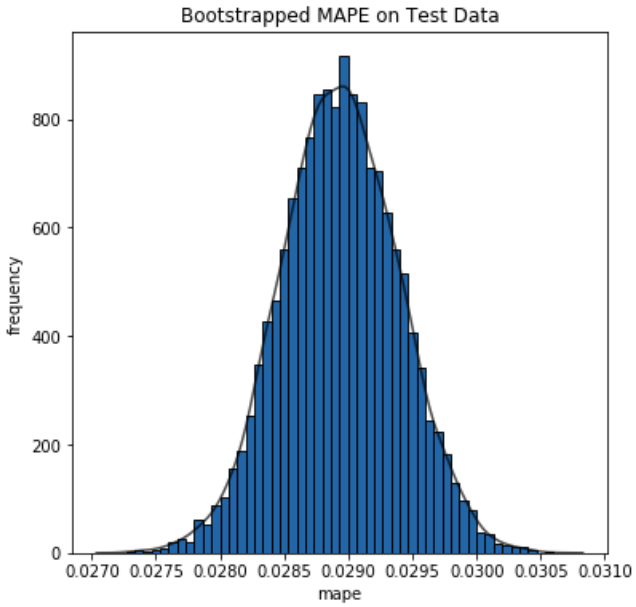

Figure 39: Ridge: Bootstrap MAPE Distribution

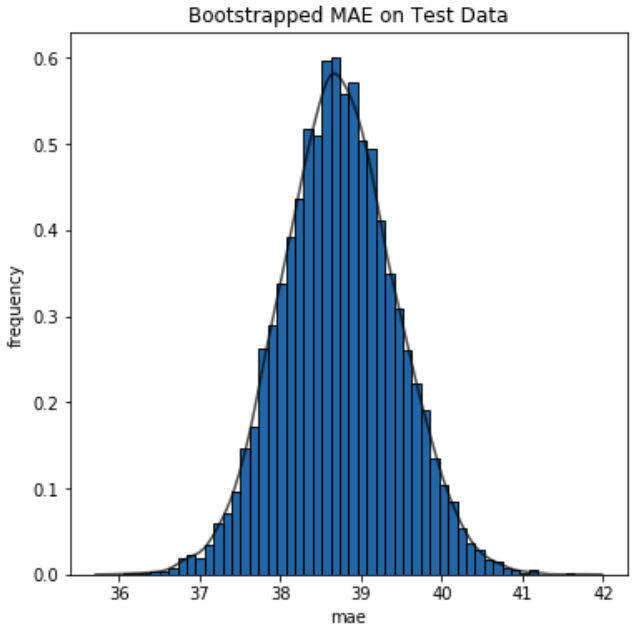

Figure 40: Ridge: Bootstrap MAE Distribution

H.2 Neural Network

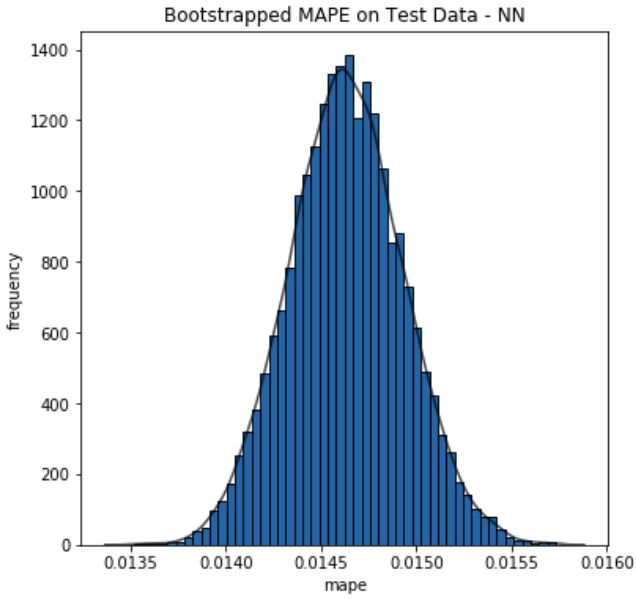

Figure 41: NN: Bootstrap MAPE Distribution

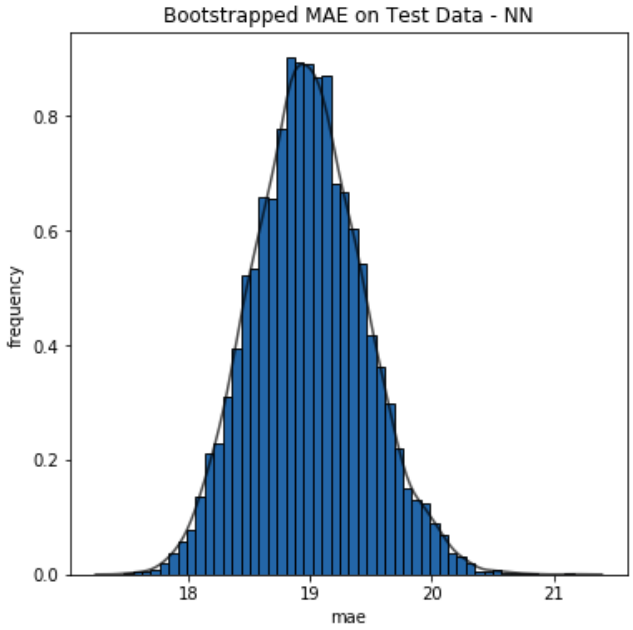

Figure 42: NN: Bootstrap MAE Distribution

H.3 Difference

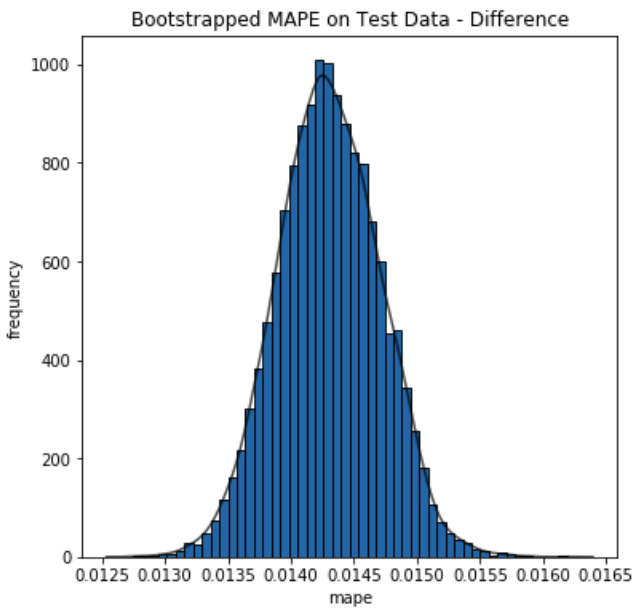

Figure 43: Difference: Bootstrap MAPE Distribution

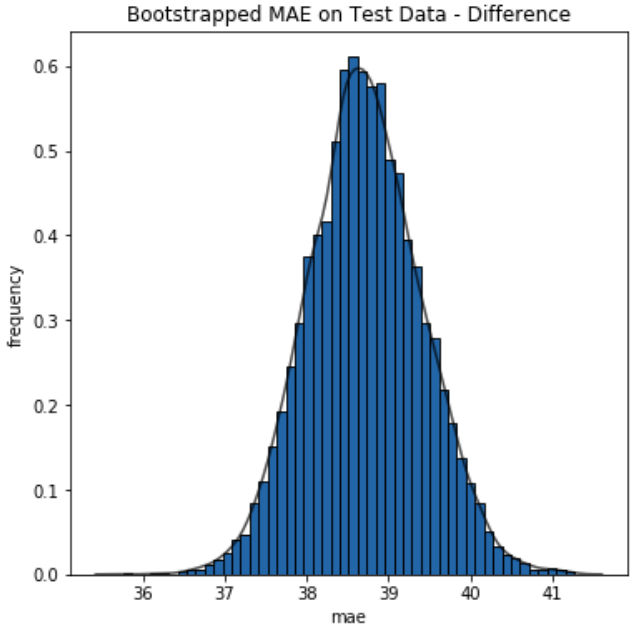

Figure 44: Difference: Bootstrap MAE Distribution

H.4 Diagnostics

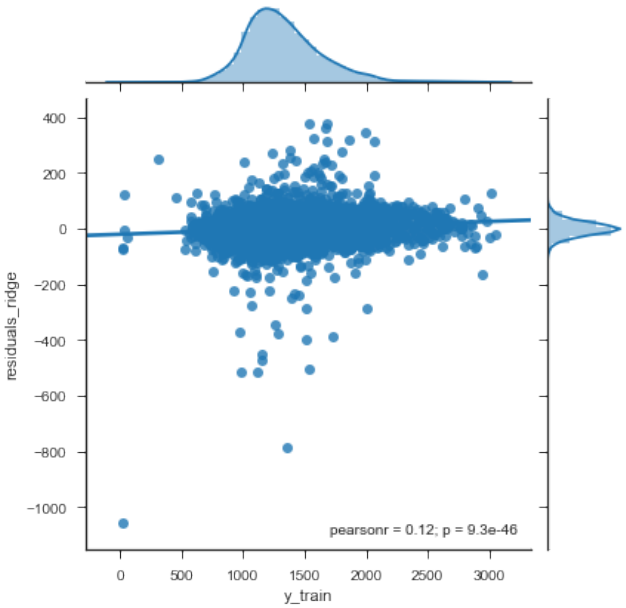

Figure 45: Residuals vs. Fitted - Ridge Regression