CAPSTONE PROJECT BY TEAM B

THE VALUE OF DIVERSE WEATHER DATA IN FORECASTING ELECTRICITY CONSUMPTION IN AUSTRALIA

Alexander Gould (z3062201), Alexa Kelly (z5276338), Kelvin Tang (z5287799), Mark Tsui (z3397409).

School of Mathematics and Statistics
UNSW Sydney

June 2021

Submitted in partial fulfilment of the requirements of the capstone course ZZSC9020

Plagiarism statement

I declare that this thesis is my own work, except where acknowledged, and has not been submitted for academic credit elsewhere.

I acknowledge that the assessor of this thesis may, for the purpose of assessing it:

- Reproduce it and provide a copy to another member of the University; and/or,
- Communicate a copy of it to a plagiarism checking service (which may then retain a copy of it on its database for the purpose of future plagiarism checking).

I certify that I have read and understood the University Rules in respect of Student Academic Misconduct, and am aware of any potential plagiarism penalties which may apply.

By signing this declaration I am agreeing to the statements and conditions above.

Signed: (Alexander Gould) Date: 14 June 2021

Signed: (Alexa Kelly) Date: 14 June 2021

Signed: (Kelvin Tang) Date: 14 June 2021

Signed: (Mark Tsui) Date: 14 June 2021

Abstract

Short term load forecasts ("SLTF") are an important element in energy control systems. As demand for energy fluctuates, supply must adjust in response. However, many power generators have a start-up time, which means forecasts of up to 24 hours are required to ensure power is available on demand. Fluctuations in energy demand are driven in part by changes in weather conditions, which influence consumption by individuals and small businesses. Electrical grid demand may also change in response to changes in the amount of wind or solar energy available off the grid. We explored the extent to which STLF may be improved by utilising a diverse set of weather data, including both capturing weather readings from a larger number of weather stations and also capturing more varied data from each station. Using the state of NSW Australia as a case study, we demonstrate that the use of more diverse weather data can improve STLF accuracy. In particular, several contrasting methods found an improvement in accuracy (up to 7% reduction in RMSE) when more granular weather data was incorporated. Importantly, dimensionality reduction techniques such as aggregate statistics and PCA were necessary for the modelling methods considered to handle the large amount of highly correlated data generated by many weather stations.

Contents

Chapter 1

Introduction and Motivation

Endgame Economics ("Endgame") creates solutions for the energy sector and has asked us to apply our data science skills to help them build better short term load forecasts ("STLF") for their clients. STLF with forecast windows of between 1 and 24 hours are needed to ensure energy providers can make operational decisions to ensure electricity supply matches demand, as unlike other goods it is difficult to efficiently store electricity in bulk. For a gas or coal plant, a warm start-up would take three hours (Gonzalez-Salazar, Kirsten, and Prchlik 2018), and this is the forecast window we use in our study.

Weather is an important determinant of short term variability in energy consumption. For example, weather can influence the concurrent use of appliances such as heaters or air conditioning by both residential and commercial consumers (Australian Energy Market Operator 2020). Weather varies not only over time but also across different geographical regions; a rainy day in a region with high residential solar panel penetration could result in additional demand from the grid, and this may not be captured by a model which only has access to the temperature at a single weather station.

Our group examined the question, how much better are STLF when they are based on a diverse set of weather data? Diversity in weather data can mean both capturing weather readings from a larger number of weather stations and also capturing more varied data from each station. We developed STLF for the state of NSW in Australia and found evidence to support the use of weather data collected from multiple weather stations, and using more weather metrics than just temperature. We found that our models were generally unable to handle the granular and highly correlated raw data from a large number of weather stations and some sort of dimensionality reduction, such as the use of aggregate statistics or applying principle components analysis ("PCA") was necessary.

Capturing and processing a larger volume of data has an associated cost in terms of model complexity and analytical expertise. Our analysis may help Endgame and its clients to make a more informed decision about whether such a cost is justified in terms of predictive capability. A relatively small improvement in STLF accuracy of 1-3% can result in substantial cost savings, measured in the millions of dollars (Islam, Baharudin, and Nallagownden 2017). Based on our results, we believe there would be clients who would see a positive return on investment from seeking to use more diverse weather data.

Chapter 2
Literature Review

2.1 STLF currently used in practice

The Australian Energy Market Operator ("AEMO") generates forecasts each day for the day ahead, known as the pre-dispatch forecast. A range of predictors are used in their statistical models (Australian Energy Market Operator 2021a), including:

- Time: type of day (weekday or weekend), school holidays, public holidays, daylight davings
- Prior demand: historical and real-time metered loads
- Weather: historical temperature and humidity, forecast temperature and humidity, forecast rooftop solar generation, forecast non-scheduled solar and wind generation

For weather forecasts, AEMO use 13 weather stations: Armidale, Bankstown Airport, Campbelltown, Canberra Airport, Cessnock Airport, Coolangata, Gosford, Orange, Penrith, Sydney Airport, Terrey Hills, Wagga Wagga and Wollongong Airport.

AEMO also provide details on their models used for forecasting for annual consumption and maximum and minimum demand (Australian Energy Market Operator 2020). In these models only temperature at Bankstown Airport weather station is considered, and in particular, AEMO use the concept of Heating Degree Days and Cooling Degree Days, which are the number of degrees the average daily temperature is above 19.5°C and below 17.0°C respectively.

Our project may help inform best practice in this area, as we consider a broader range of stations and weather metrics than those currently in use.

2.2 STLF academic research studies

Zhang et al. (2019) applied STLF to the California market, comparing three methods and considering a range of weather factors. They note that many methods can be applied to STLF and that machine learning methods are particularly applicable. They found gradient boosting outperformed simpler methods. While temperature was found to be by far the most significant weather-related factor, other factors were found to be significant.

Islam, Baharudin, and Nallagownden (2017) used a genetic algorithm to optimise a neural network for STLF for NSW Australia. They note that non-parametric models such as neural networks tend to produce better STLF than parametric models, as they can capture non-linear relationships that drive electricity demand; however, the performance of such models is dependent on decisions about architecture and hyperparameters, and they are more likely to have issues converging or avoiding local minima.

Our research follows a similar path, considering multiple modelling approaches, including neural networks, and with a focus on the NSW Australia energy market.

2.3 Drivers of electricity demand

Li (2020) outlined common time-based factors used in STLF include time of day, day of the week and month of the year. Prior demand data has been used in simple models such as ARMA and exponential smoothing, and is a relevant feature in more complex models. Temperature is the most well known and commonly used weather-related factor, but humidity, wind speed, solar irradiance and precipitation are also linked to electricity consumption.

We include time, prior demand and weather related factors in our study. While our research is focused primarily on the impact of using more diverse weather related data, it is important to include commonly used time and prior demand related factors in our model, in order to assess the potential gain that a market player might expect if they incorporate more diverse weather related data into an existing STLF framework.

2.4 Combining data from multiple weather stations

Sobhani et al. (2019) considered the task of combining weather station forecasts, noting that measurements from multiple stations are needed to reflect the geographic spread of the load. They considered seven alternative methods to the traditional approach of using a simple average, and found that while none of the methods considered consistently outperforms, model accuracy may be increased by using more sophisticated approaches than just the simple average.

Li (2020) took a different approach and considered using an optimal subset of weather stations. They noted that in practice, a small number of weather stations are often used for STLF in part due to practical concerns such as requiring domain knowledge and expertise to select the most relevant stations and concerns about data quality and data history. However, the winning entries in the Global Energy Forecasting Competition in 2012 and 2014 demonstrated that better predictions are possible when readings are combined from multiple weather stations. They explore simpler alternatives to genetic algorithms which are being used but which can be computationally intensive.

Our research builds on these studies by considering whether dimensionality reduction is necessary, or whether a suitably complex model, such as a deep neural network or gradient boosted ensemble, would be able to dynamically weight the highly correlated data taken from a large number of weather stations.

Chapter 3

Data and Assumptions

3.1 Data sources

Endgame provided us with three sets of data:

- Total electricity demand for NSW at 30-minute intervals between 1 January 2010 and 17 March 2021 ("Actual demand")
- Forecast electricity demand for NSW at 30-minute intervals for the corresponding period, with multiple lead times up to 40 hours in advance, based on the AEMO modelling discussed in section 2.1 ("Benchmark")
- Actual temperature at Bankstown Airport in NSW at short intervals (around 6 minutes apart) for the corresponding period

We were able to source daily weather data free of charge from the Bureau of Meteorology ("BOM") FTP site for the corresponding period as the data provided by Endgame. This data includes temperature, solar exposure, rainfall and wind speed data, which have been used elsewhere in the literature (Apadula et al. 2012; Sailor and Muñoz 1997; Čupeljić 2016). Humidity data is also available, but not incorporated into our dataset as we did not appreciate its availability and relevance until late in the project. In section 7 we note possible areas for future work, and including humidity data is one such area.

More granular 30-minute interval weather data is also available at a cost, but was not used for our research.

3.2 Data processing

3.2.1 Missing values

The historical demand data provided by Endgame was relatively complete, with no significant concerns due to missing data. A handful of periods had no data; these were excluded from our analysis and this is not expected to have any material impact on the conclusions of our study. These records represent an extremely small proportion of the overall data used for training and performance assessment.

The weather station data obtained from the BOM did have material amounts of missing data. The data is collected from 104 different weather stations, some of which have been inactive at points in time. The figure below shows the proportion of days with no readings for each weather station and year between 2010 and 2021.

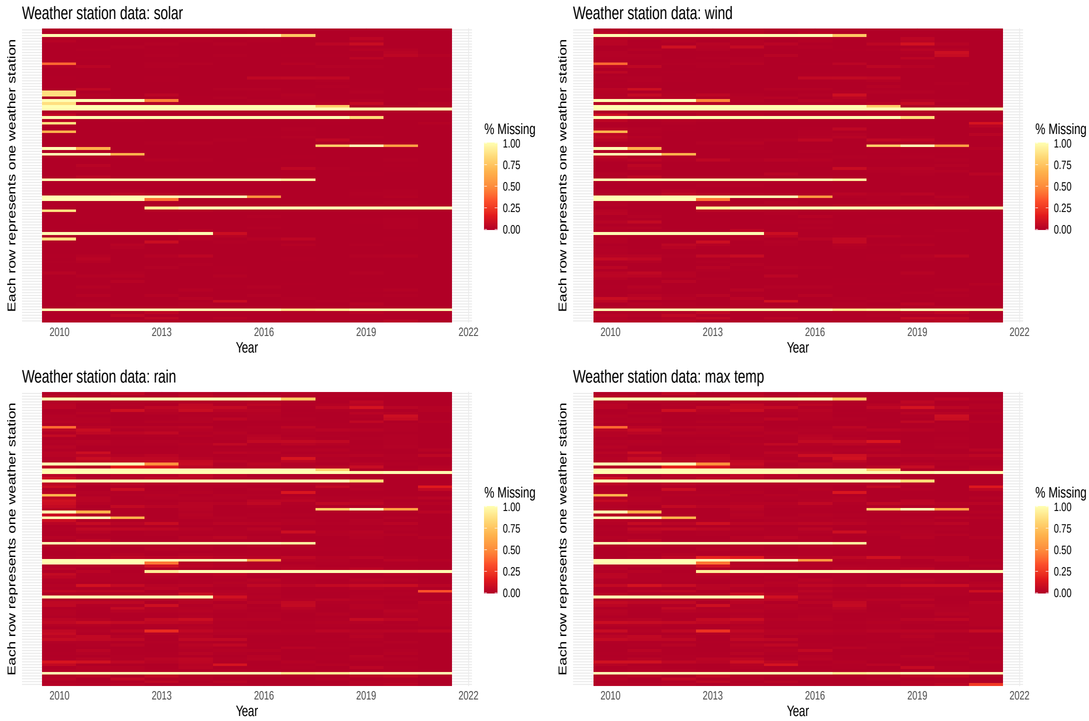

Figure 3.1: Overall missing proportion by weather station; solar, wind, rain and temperature

From this exercise we observed:

- The majority of weather stations have relatively complete data
- Stations with missing values for one type of weather data generally have missing values for all types of weather data. One exception to this is solar in 2010 which had a higher proportion of missing values.
- There are a number of weather stations with periods of missing data; we observe stations coming "online" (missing in earlier years), going "offline" (missing data in later years), and also stations going into "hibernation" (missing values for one or two years).
- There are two weather stations with no data at all.

The figure below shows the overall proportion of missing data for each of the weather stations:

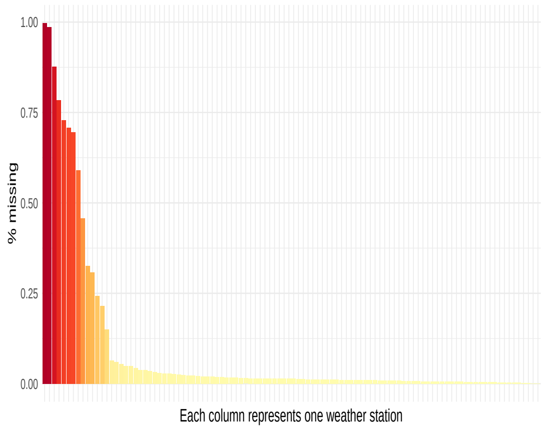

Figure 3.2: Overall missing proportion by weather station

Based on the analysis above, we decided to remove 14 out of 104 weather stations, being those stations with >10% missing overall. This left 90 stations for modelling. For these stations, two steps were taken to address the remaining missing values:

1. Splines were used to interpolate missing values for a handful of dates where solar radiation data was missing across all weather stations. While considering various univariate strategies for imputing missing data, Lee et al. (2020) mentioned that spline interpolations produce smoother interpolants with smaller errors for missing values.
2. Mean imputation was used for the remaining values. Given that the proportion of values impacted was small, and the data was at a daily level of granularity, this approach was considered reasonable for the purpose of our exercise.

We experimented with a more sophisticated approach, a random forest imputation algorithm for missing data implemented in the missForest package (Morgan 2020), however found that with over 500 columns it was not feasible to run the algorithm. It is also questionable whether this approach would translate into practice; if a weather station fails to provide a reading, using the historical average reading for the station is more feasible.

An alternative approach would have been to use the latest historical reading, which may have provided a better estimate for days missing completely at random as it would have captured the weather conditions of the season. However, our missing data analysis suggests that stations with missing values tend to have long periods without data, so this may not have been a reasonable approach to take. It would also not work for stations with missing data at the start of the dataset, which we observed in our analysis.

3.2.2 Historical periods included in model training data

While Endgame provided us with ten years of historical demand data, and we were able to obtain daily weather data over the same timeframe, we restricted our training data to the past four years of data (2017 onwards). This timeframe is consistent with models used in practice, with AEMO using two years for short term seasonality patterns and 4-6 years for longer term trend patterns. While our models are largely focused on short term variation, it is also important to capture longer term trends to avoid them being interpreted as weather impacts.

Using the full history of data is inappropriate if the data generating process has changed over that timeframe. The relationship between demand and its drivers can change over time and using older periods which are less representative of the current data generating process is not always helpful (Fan and Hyndman 2012).

A key trend in the industry over the past decade is the proliferation of solar photovoltaic (PV) panels on rooftops and in solar farms, which reduce the amount of electricity demanded from the grid in general, but also increase the sensitivity of demand from the grid to weather factors such as the amount of solar radiation. The figure below shows the increase in solar PV capacity over the past decade. We consider 2017 to be an appropriate starting point for our modelling data, as by this point solar capacity had reached a material threshold and the rate of increase has been relatively constant from this point onwards.

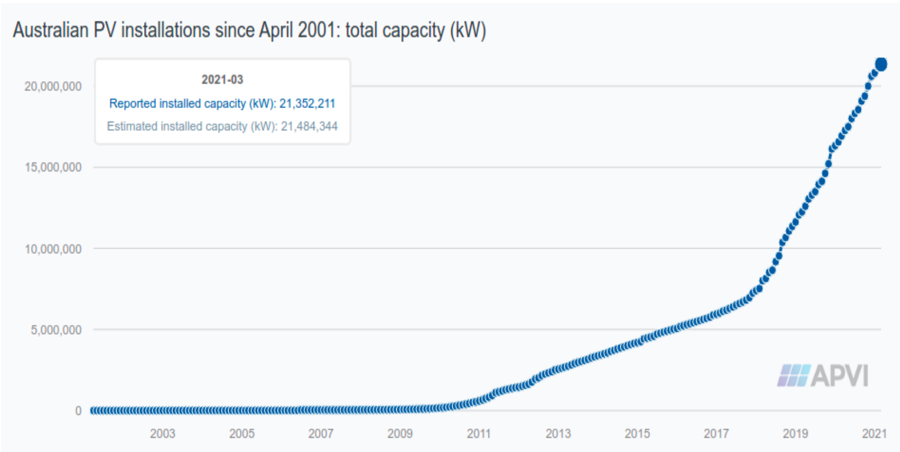

Figure 3.3: Solar PV capacity (source: https://pv-map.apvi.org.au/analyses)

The trend towards more solar PV capacity may have contributed to the reduction in demand for electricity from the grid over time, particularly between 2010 and 2017, as shown in the figure below. This provides further support of the use of 2017 as a cut of point, as the level of demand has only reduced slightly since 2017, compared to a more significant reduction between 2010 and 2017.

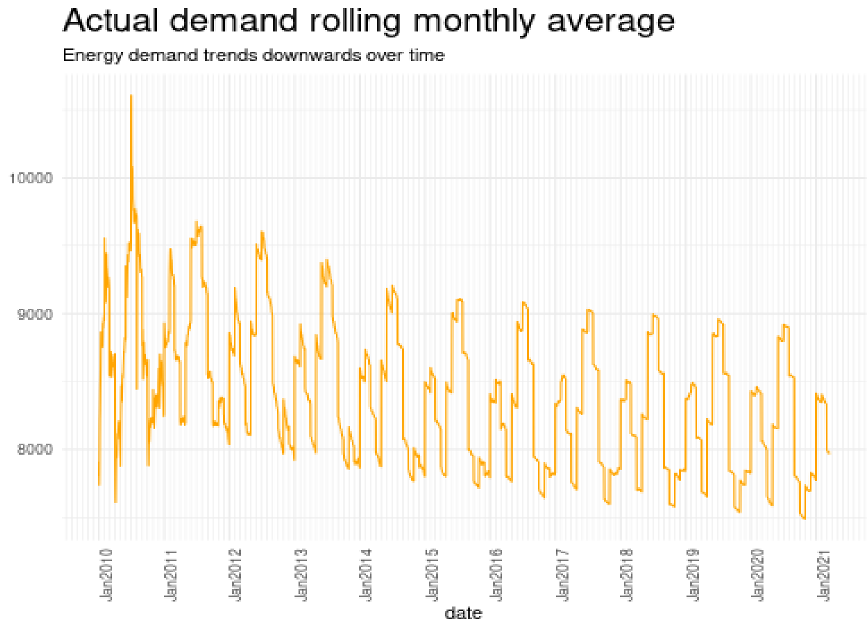

Figure 3.4: Rolling demand by month

3.2.3 Aggregate weather metrics

We used the simple mean and standard deviation of each metric across all weather stations in our models. A weighted mean was also derived using measures of population density, under the assumption that weather stations located near higher population density areas would exhibit greater correlation with total electricity demand. This approach was informed by Pardo, Meneu, and Valor (2002) who used population-weighted temperature because energy use is highly correlated with population (DeLong and Burger 2015).

3.3 Modelling Approach

3.3.1 Base, Aggregate and Granular models

We measure the impact of using more diverse weather data by comparing the accuracy attained by three models, which use different levels of weather related data. These are:

1. Base: Bankstown min/max daily temperature
2. Aggregate: mean, standard deviation and weighted mean (Pardo et al. 2002) of daily weather metrics (min/max temp, rain, wind and solar) across 90 weather stations
3. Granular: daily individual station data for the same weather metrics as the Aggregate model

In addition to weather data, the models include demand and time based predictors. These are included identically in each of the models, so that the difference between the models only relates to the treatment of weather data only. The non-weather related predictors include:

- Time-based factors: year, month of the year, day of the year, day of the week, hour of the day, public holiday

- Demand-based factors: recent half hourly demand (3-24 hours), demand in the prior day (24 hours leading up to 3 hours prior) and prior week (7 days leading up to 3 hours prior)

By comparing Base and Aggregate, or Base and Granular, we obtain an estimate of the impact of using more diverse weather data. The Aggregate data will have far fewer columns, which may result in faster training and more robust models. Some domain expertise is necessary to understand the best way to aggregate the individual station data, and there will be some loss of information in aggregating the data. The Granular data could potentially lead to greater accuracy with less human intervention, but only if one of the modelling approaches considered is able to dynamically combine the raw data.

3.3.2 30 Minute and Daily models

We decided to use a three hour forecast window for our research, as this reflects the real world scenario of a warm start-up lead time for a gas or goal plant. Having greater confidence that demand will ramp up in the coming three hours would reduce the risk of being unable to sell electricity at the anticipated price and making less profit or even a loss on the decision to start up the power plant.

As discussed in section 3.1, we were able to obtain daily weather data for our experiments. Ideally, we would use intra-day weather data, as this would better capture features such as rapid changes in weather which would impact daily demand but may not be apparent from daily min/max metrics. This would link in better with the benchmark intra-day predictions mentioned above.

Given the daily granularity of weather data available, we ran some models using data aggregated to the daily level. Intra-day predictors such as hour of the day and prior demand in the 3-24 hours leading up to the forecast time were removed. Our hypothesis was that we might observe a more pronounced effect of weather related data when the intra-day non-weather factors were removed from the picture.

3.3.3 Model types

We created STLF models using several contrasting methods, including linear models, tree-based ensembles and neural networks. The same datasets were used for all models. In comparing the results from these models, we get a better sense of whether our conclusions are robust. We can also exploit the potential of different model types, with some models being more explainable or robust, while others have have greater potential for accuracy if applied well.

3.3.4 Training and validation data

To assess model results fairly, data was split into train and test sets in an 80:20 ratio, with the former being used to train models and the latter being used to assess the models. In splitting the data, all records relating to a given week were allocated to either the train set or the test set. This approach was taken to reduce the risk of models exploting features in the train set, such as demand in the 3-24 hours prior to the forecast window, to infer the target values in the test set. We understand from discussions with Endgame that this is also a standard approach in practice. The same test set was used between different model types to make results more comparable.

3.3.5 Performance measures

Models were generally tasked with minimising root mean squared error (RMSE). This is a standard metric for regression tasks. We also considered mean absolute error (MAE) as an alternative. Our rationale for optimising for RMSE was that Endgame have indicated that it is more important to have accurate predictions for days when demand is much greater than normal, and RMSE will tend to give greater weight to larger errors which would occur on such days. In contrast, MAE might provide better accuracy on typical days when it is less relevant.

When comparing models, we tended to rely on mean absolute percentage error (MAPE) and R-squared (R2). These metrics have more intuitive interpretations in terms of relative error. For example, RMSE is much higher for daily models than for 30 minute models, whereas MAPE and R2 are more comparable.

3.4 Assumptions

3.4.1 Daily weather data

We assume that daily weather data is still a reasonable predictor to include in modelling of 30 minute demand windows. This is partly due to the potential for models to capture the interaction between time-based predictors such as hour of the day and the daily weather-based predictors. Most models should be able to predict, for example, that demand will be higher on a hot day, and that within that day demand will be relatively high at certain points during the day.

We also assume that the impact of using more diverse weather data on an intra-day basis would be at least as significant as it is on a daily basis, and probably more so. Intra-day weather data will be more predictive as it can capture changes in the middle of the day, such as a cool change that leads to a hot day becoming pleasant, reducing demand for electricity, where on other days it might remain hot and demand would remain high.

3.4.2 Weather on the day

We have used actual daily weather data for the day being forecast. In practice, this data would not be available at the time that the forecast is made, and would only be available with some delay for measurement and processing. However, in practice, reasonable forecasts would be available, and actual data may be available if making forecasts later in the day. We use actual weather on the day as a proxy for having such weather forecasts, noting that some forecast error is expected to occur and so the results we obtain represent an upper bound on the impact of using daily weather data. The impact of this is likely to be relatively small as temperature is the most significant weather-related predictor and temperature at least can be forecast quite accurately up to five days ahead (Li 2020). AEMO provides an analysis of the accuracy of the 7 hour and 24 hour temperature forecasts used in their modelling which show gaussian errors which are generally less than a few °C (Australian Energy Market Operator 2021b).

3.4.3 Impact of model hyperparameter selection

A core assumption in our approach is that we can create models using datasets that differ only with respect to the granularity of weather data available, and that differences in model accuracy is driven by the difference in input data provided to the model. In reality, model performance will be impacted by choice of hyper-parameters and architecture. A poorly fit base model should not be compared to a well fit alternative model. We have attempted to minimise this risk by adopting a similar approach to hyper-parameter selection, having the same or similar hyper-parameters where appropriate to do so, and testing neighbouring hyper-parameter selections to ensure models are, at least locally, fit fairly optimally.

3.4.4 Geographic spread of electricity demand

We assume that the geographical distribution of weather stations used in our modelling capture differences in weather conditions that may drive short term fluctions in demand. There are some regions with particularly high concentration of residents (and small businesses), such as the region surrounding Bankstown airport, and others with much lower population density. In using metrics such as the simple average across all stations, we are assuming that different stations contribute equally which may not be accurate, however if more stations are located in regions with dense population this may be a reasonable approximation. In using the weighted average based on population density, we assume that population is a strong driver of fluctuations in energy demand, both from residential consumption and small businesses which serve nearby residents.

The figure below shows the spread of weather stations from which data was obtained. The stations span the full geography of NSW/ACT and have a higher concentration in densely populated regions, supporting the assumptions above.

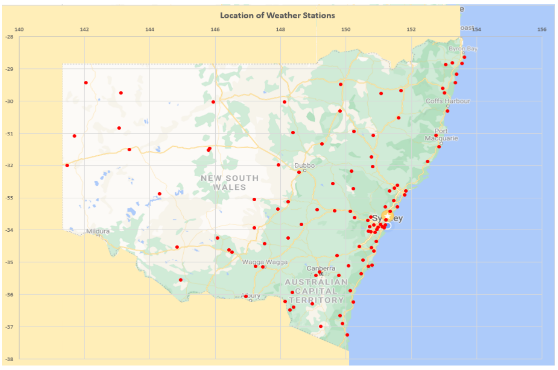

Figure 3.5: Location of Weather Stations in NSW

The figure below shows the population density weights for the top 15 weather stations. Endgame provided us with temperature data from Bankstown Airport, which is ranked third on the list, behind Sydney and Canterbury Racecourse.

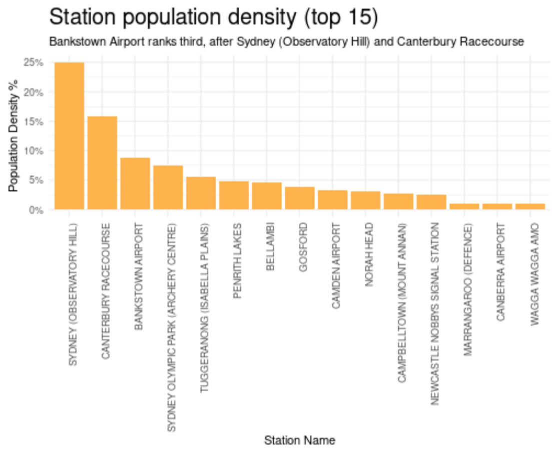

Figure 3.6: Weather station population weights

We assume that the stations with a high proportion of missing values are located in remote areas with low population, or are located in populated regions but are close to other stations whose readings can be used instead.

3.5 Software and source code

Our group used R and Python as these languages are well suited to data processing and model building.

The complete source code may be found in the project Github repository, which is located at `https://github.com/alexa-malexa/capstone`. As this is a private repository, for access please contact Alexa Kelly at alexandra.kelly@student.unsw.edu.au. Fixed seeds were used where possible to ensure reproducibility. The input datasets used for modelling have been saved in the Rds format.

Chapter 4

Exploratory Data Analysis

Explanatory variables used in our model generally fall under time-based, demand-based and weather based models. This section of the report explores the raw data and provides evidence to suggest that the factors used in our models are likely to be useful in predicting electricity demand.

The figure below shows scatterplots of several weather metrics against demand. The most pronounced relationship is between demand and temperature, with higher demand at more extreme (high or low) temperatures. There is a lesser relationship between demand and low levels of solar and wind, which may reflect the reduction in energy from solar and wind sources off the grid. There is a slight negative relationship between demand and rainfall, although it is unclear what might be driving this.

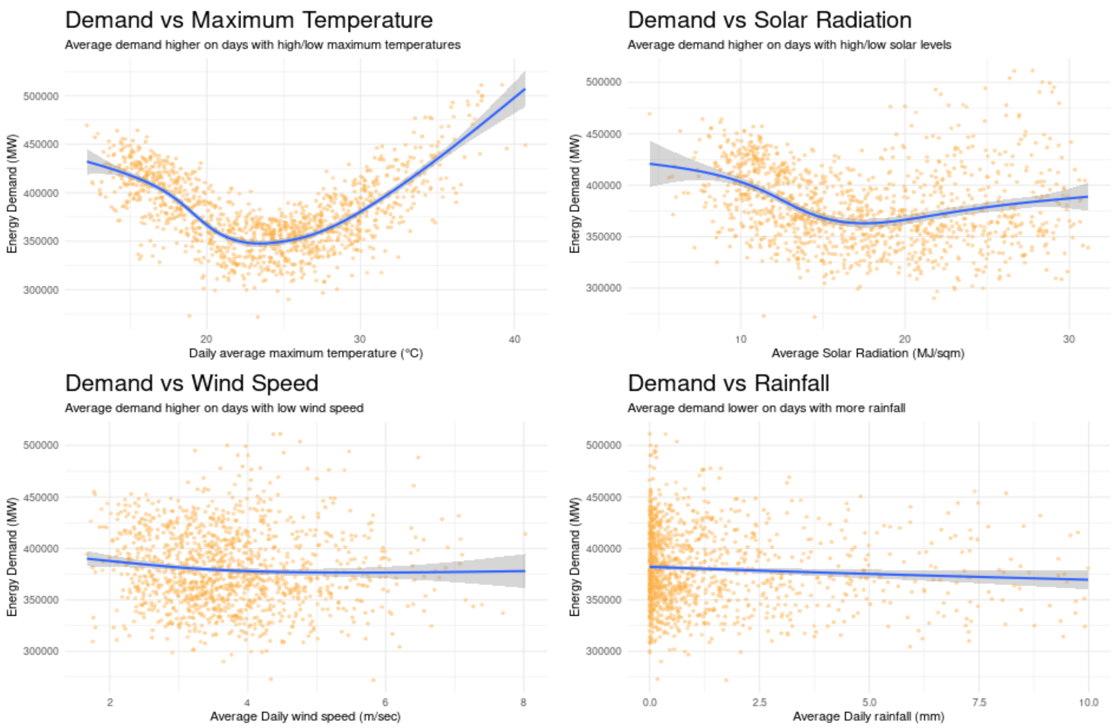

Figure 4.1: Weather-related drivers of demand

The impact of working days is broken down in the figure below, which indicates that demand is higher for weekdays than for weekends. There are small differences between days within these two categories, such as higher demand on Saturdays compared to Saturdays.

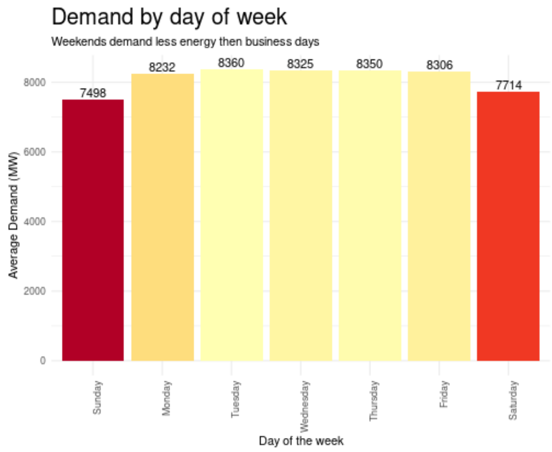

Figure 4.2: Average demand by day of the week

In addition, public holidays may fall on any given day of the week. The box plots in the figure below show that public holidays have lower demand and variability than other days.

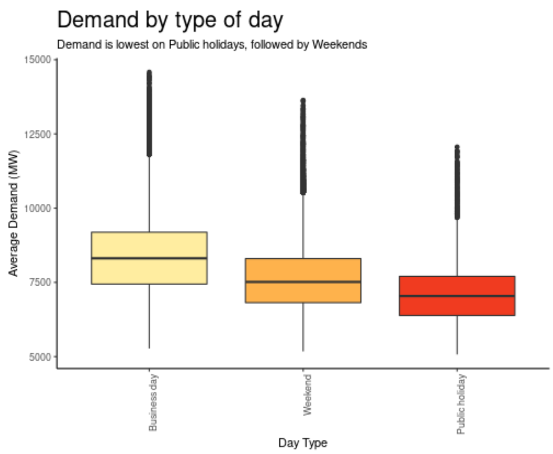

Figure 4.3: Forecast model accuracy over time

The figures below support the use of prior demand measures, such as demand for the same period the day prior, and demand over the 24 hours leading up to the projection. There is a strong correlation between demand today and prior demand measures. We note this likely reflects other factors in the model, such as the tendency for weekdays to follow weekdays or for there to be sustained periods with similar weather. It seems likely that models without such autocorrelative demand features would see an increased effect from such other factors in the model.

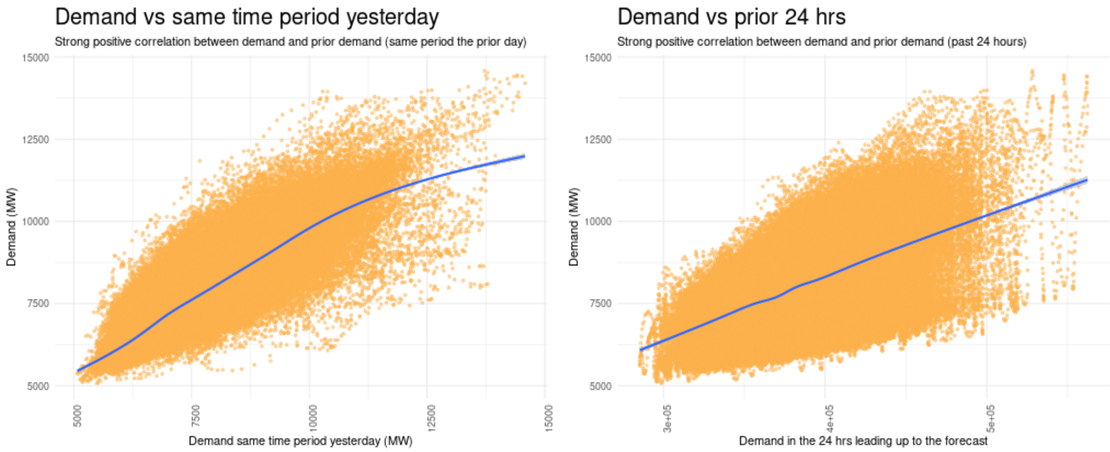

Figure 4.4: Forecast model accuracy over time

Endgame provided us with benchmark forecasts. The figure below examines the benchmark forecast accuracy over time. We observe a significant improvement in forecast accuracy around 2012, and a possible shift since 2017 with the model tending to overestimate demand. This suggests comparisons to the benchmark should ideally be based on more recent years.

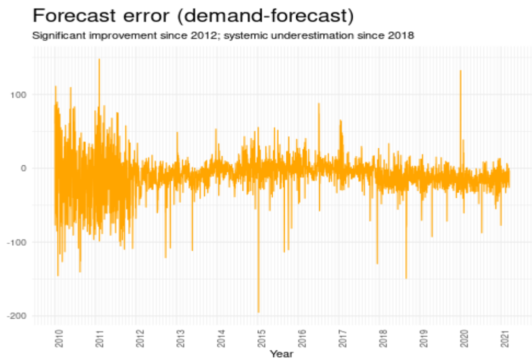

Figure 4.5: Forecast model accuracy over time

The figure below shows a few examples of the pattern of demand occuring in the actual and benchmark forecast over the course of a week. We note that on some days there is a bimodal pattern with high demand in the morning and evening, but a dip in the middle of the day, while on other days the pattern is more unimodal with a further increase in demand during the middle of the day. While there is a high amount of seasonality in the data, differences such as this may undermine some time series models that are often applied to STLF.

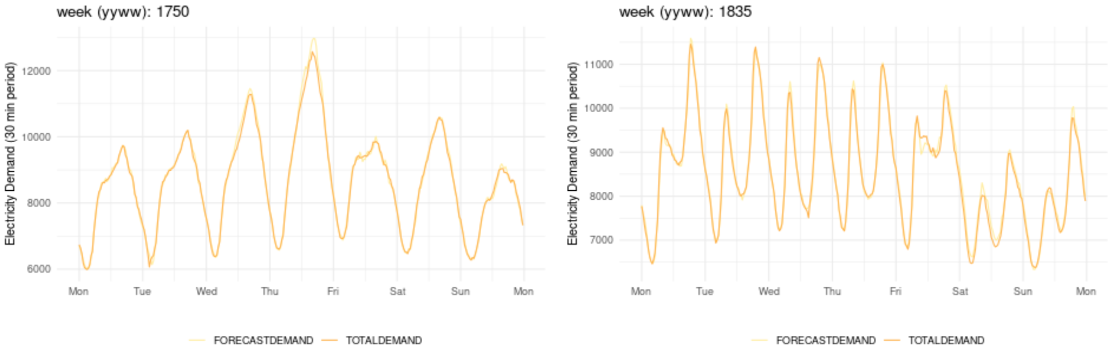

Figure 4.6: Examples of demand patterns throughout the week

Chapter 5

Analysis and Results

5.1 Neural networks

5.1.1 Background

Feed forward neural networks were considered due to their ability to handle non-linear interactions, given the large number of relevant predictors and the potential for interactions between them. For example, demand may be different for a hot day with wind compared to a hot day without wind, and the difference due to wind may be more evident on hot days than on cool days. Neural networks also have the capacity to model complex functions, and we hypothesised that this might allow for the dynamic combination of the granular weather station data. The model was applied to both the 30 minute and Daily forecast tasks.

5.1.2 Data preparation

In addition to the data preparation steps covered in section 3.2, the data used to train the model was scaled using a standard normal scaler applied to all columns (including boolean indicators). This approach makes the network more likely to converge and to converge faster than without any scaling (Geron 2019).

5.1.3 Selected hyperparameters

For each model (Base, Aggregate and Granular), a range of alternative hyperparameters were considered. This was important to ensure that the different models were fit appropriately and any differences observed were driven by the different granularity of weather data being used, rather than by hyperparameter choices. The selected hyperparameters were chosen as test set performance was at a low point relative to neighbouring alternative choices or performance was relatively similar over the range of alternatives considered. In the end, the same hyperparameters were used across all models. These are shown in the table below.

Hyperparameter	Selected Value	Alternatives Considered
Epochs	100	10, 40, 100
Learning rate	0.005	0.0005, 0.001, 0.005, 0.01, 0.05
Hidden layers	3	1, 2, 3, 4, 5
Neurons per layer	64	8, 16, 32, 64, 128
L1 regularisation	0.01	0, 0.001, 0.01, 0.1
L2 regularisation	0.01	0, 0.001, 0.01, 0.1
Dropout	None	None, Dropout, Alpha Dropout

Early stopping was used, so the number of epochs actually used was generally less than 100. When using smaller values such as 10, model accuracy was adversely impacted. Adam optimizer was used for all experiments as it is a reasonable default choice which is robust to choice of other hyperparameters (Goodfellow, Bengio, and Courville 2016). Relu activations paired with He initializations were used for all experiments; this is a popular approach (Geron 2019). The results for experiments with Alpha Dropout were particularly poor, and so Selu activations were used, as recommended by the Tensorflow documentation, paired with Lecun Normal initialisation, as suggested by Geron (2019), however this did not lead to Alpha Dropout being effective.

A small amount of regularization was used as in early experiments there was some evidence to suggest this led to a slight improvement in test performance and it is a good idea to include some regularization by default (Geron 2019). In addition to testing L1 and L2 regularisation in isolation, Elastic Net regularisation (i.e. both L1 and L2) was also considered. However, given the limited impact of both L1 and L2 regularisation, this also had little impact on the results, so Sensitivity 6 (Elastic Net Regularisation) has not been shown in the sensitivity results shown in this report.

For each of the alternatives considered, the model was run five times to give a 95% confidence interval. This approach reduced the risk of conclusions being drawn on the basis of randomness in the model training process, rather than a genuine impact of the change in hyperparameter.

The figure below shows two examples of this sort of analysis, the first for the Base 30 Min model and the second for the Granular 30 Min model.

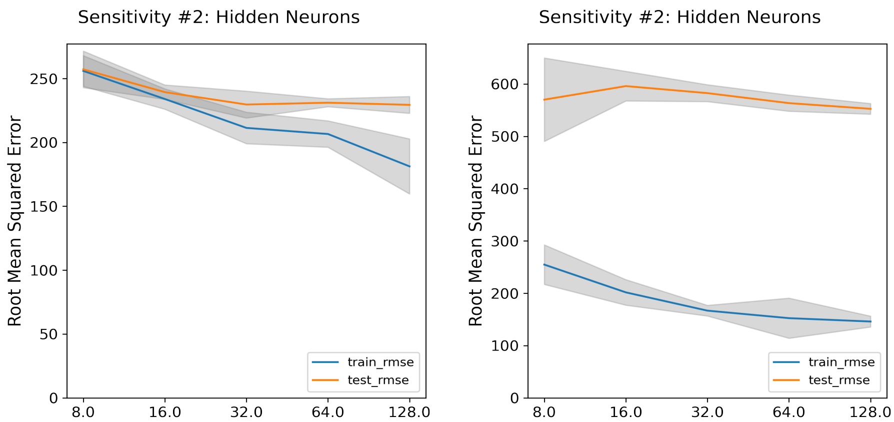

Figure 5.1: Example model hyperparameter sensitivities

The Appendix includes hyperparameter sensitivities for each of the Base, Aggregate and Granular models, for both the 30 minute and Daily granularity.

We found clear evidence of overfitting in the Granular models, with training set RMSE much lower than test set RMSE. This prompted the exploration of L1 and L2 regularisation and different dropout layers as part of the hyperparameter selection process, however these generally proved ineffective. Some more extreme parameters were also considered (not shown in the sensitivity figures included in this report), such as very low and very high learning rates, and sensitivities were also examined for a smaller network (i.e. fewer hidden layers and neurons per layer), however results were generally unchanged and/or training error increased without a commensurate reduction in test error. We concluded that neural networks are unable to handle hundreds of highly correlated predictors, and some sort of dimensionality reduction is necessary.

We applied PCA to reduce the dimensionality of the individual weather columns. The figure below shows the amount of variance explained at each dimension. 10 dimensions were selected as this appeared to be an elbow inflexion point on the graph, and subsequent testing indicated test set accuracy fell for lower dimensions (e.g. 5, 8) and higher dimensions (e.g. 15, 20). The explained variance with 10 dimensions was the same for the 30 minute and Daily data.

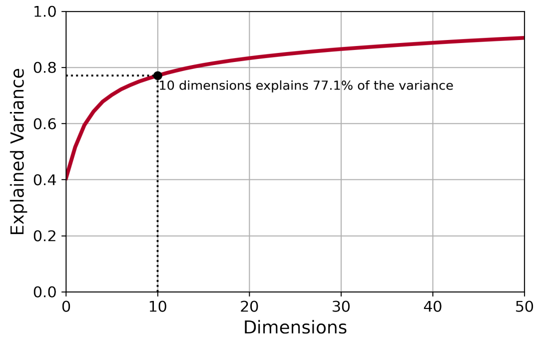

Figure 5.2: PCA explained variance for selecting dimension

We also examined learning curves to ensure the models were given adequate time to train, given the use of early stopping. The figure below shows an examples for the 30 min Base model and 30 min Granular model.

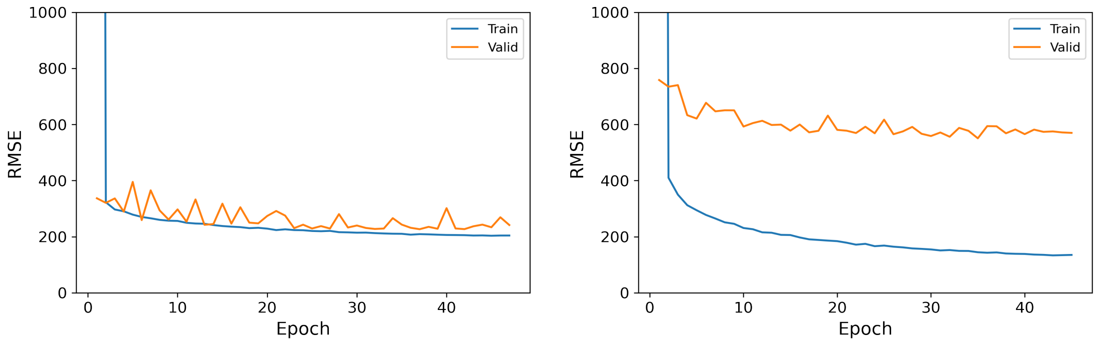

Figure 5.3: Example learning curves

Learning curves for each of the models are included in the Appendix.

5.1.4 Results

The figure below compares the results for the Base, Aggregate, Granular and Granular-with-PCA models, for neural networks predicting demand in a 30 minute window 3 hours prior. A small improvement can be observed for the Aggregate model compared to the Base model (1.1% reduction in RMSE). The Granular model exhibited a strong tendency to overfit, so while the training error was relatively low, the test performance was quite poor. However, when PCA was used to reduce the dimensionality of the Granular weather data, the model achieved superior accuracy to the Aggregate model (7.0% reduction in RMSE).

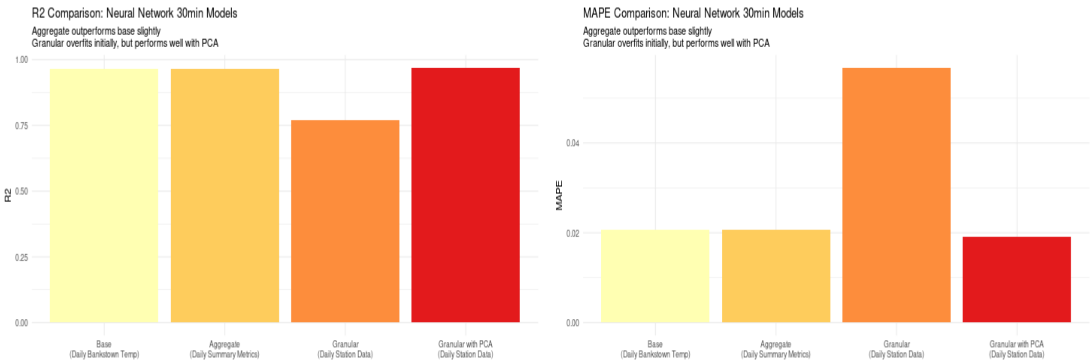

Figure 5.4: Model results for 30 Minute Models

The figure below shows the equivalent results for neural networks predicting Daily demand. We considered this additional level of granularity as we hypothesised that the effect of daily weather data may be more obvious when predicting daily demand, whereas for the 30 minute model, features such as demand 3 hours prior to the 30 minute window would be more predictive. The results were fairly consistent between the 30 minute and daily models, however.

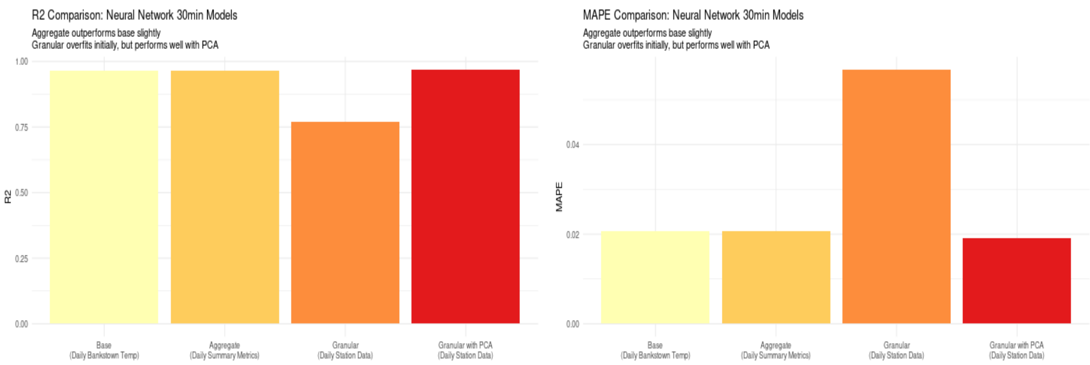

Figure 5.5: Model Results for Daily Models

Using the Daily model, an additional Aggregate model was run with only the temperature related predictors included and rain, solar and wind predictors omitted. The results of this exercise suggest that around 85% of the improvement in accuracy for the Aggregate model is driven by temperature. This results is consistent with the literature (Zhang et al. 2019) and the exploratory data analysis shown in section 4.

5.2 Ensemble (XGBoost)

5.2.1 Background

XGBoost was considered due to its ability to handle highly correlated features, high dimensionality data and non parametic data. Since XGBoost is a correlation robust algorithm there was less need to apply PCA, normalisation or scaling. (Chen et al. n.d.). The model was applied to the 30 minute forecast task.

5.2.2 Data preparation

In addition to the data preparation steps covered in section 3.2, year was removed as a predictor. Some methods, such as linear regression, are able to extrapolate trends. For example, if demand increases each year due to population growth, the linear regression model will be able to predict a higher level of demand in future years that the model has not been trained on. However, XGBoost is a tree-based method that will at best segment the data and pick the level of demand corresponding to the most recent year(s). Due to this feature, we do not consider that year would not be a robust factor to use in practice, unless the model was regularly retrained.

5.2.3 Selected hyperparameters

For each model (Base, Aggregate and Granular), a range of alternative hyperparameters were considered. This was important to ensure that the different models were fit appropriately and any differences observed were driven by the different granularity of weather data being used, rather than by hyperparameter choices. The selected values were those which provided better or similar test set performance to the alternatives considered. In the end, the same hyperparameters were used across all models. These are shown in the table below.

Hyperparameter	Selected Value	Alternatives Considered
nrounds	200	100, 500, 1000
min_child_weight	4	3, 5, 6, 7
gamma	0.3	0:0.9
max_depth	5	3, 4, 6
eta	0.3	0.1, 0.2, 0.4, 0.5

The Base, Aggregate and Granular models showed evidence of overfitting, with a noticible gap between train RMSE and test RMSE. The model was more prone to overfitting for number of rounds (nrounds) above 200. Model performance declined for learning rates above 0.3. The gamma hyperparameter was adjusted as higher values make the model more conservative which can reduce overfitting. Max depth and min child weight were also adjusted, as blocking feature interactions can reduce overfitting. Overall, attempts to rectify the overfitting issue by adjusting minimum child weight, gamma, maxdepth and eta were unsuccessful. Increasing the data history from 4 years to 10 years had no significant impact.

5.2.4 Significant features

The figure below shows the most significant features in the model. Most models relied heavily on demand for the same period yesterday (demand_lag_48 represents demand with a lag of 48 half hourly periods, equivalent to 24 hours), demand 3 hours prior (demand_lag_7 is the half hour immediately prior to the 3 hour forecast window), and monday/weekday (to adjust for the days where the same period yesterday swaps from a weekday to a weekend). A relatively small amount of importance was placed on the weather features (features beginning with X represent different weather stations by BOM ID, with weather metrics appended).

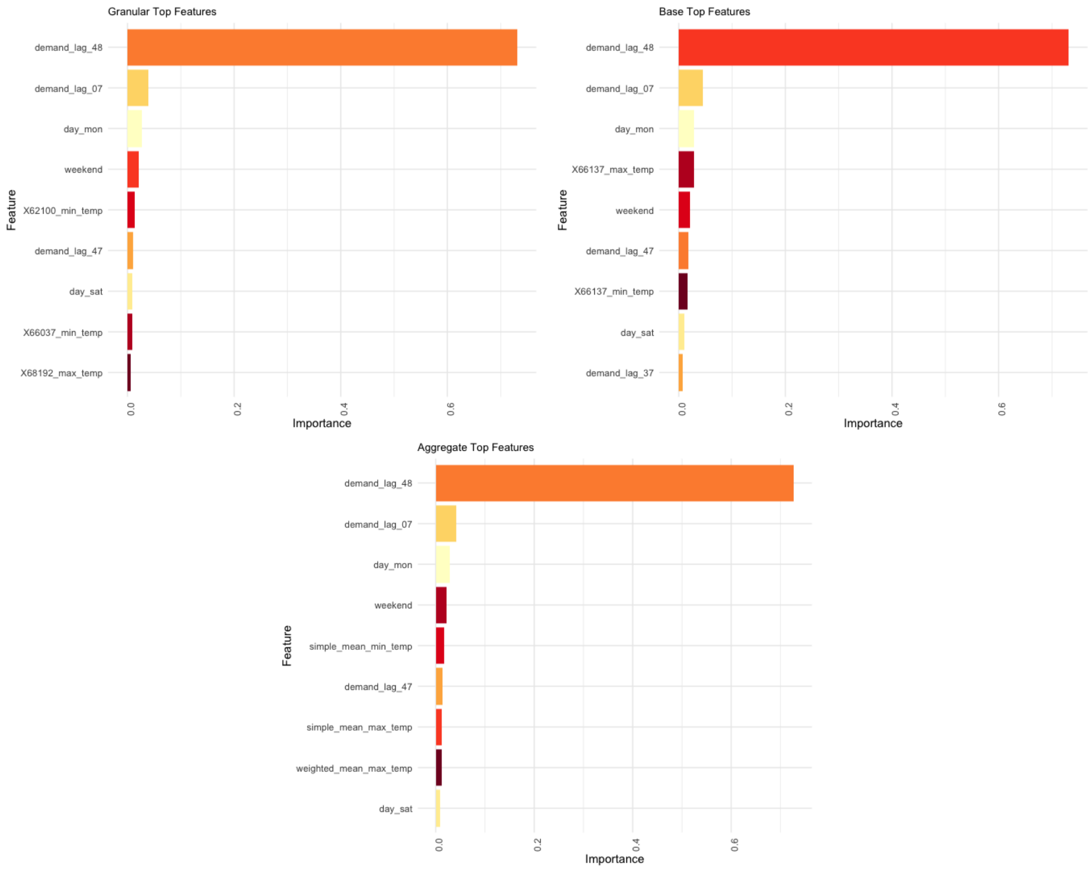

Figure 5.6: XGBoost feature significance

5.2.5 Results

The figure below compares the results for the Base, Aggregate and Granular models, for the XGBoost models. The results are relatively similar across the board, with little evidence to support the use of either Aggregate or Granular weather data.

Figure 5.7: Model results for XGBoost models

5.3 Random forest

5.3.1 Background

Random forest was considered as it has been successfully applied to STLF previously and requires less tuning than methods such as neural networks (Lahouar and Ben Hadj Slama 2015).

5.3.2 Selected hyperparameters

For the 30min model, max_depth was set to 6 to prevent overfitting. This was a simple initial model and no attempt was made to adjust the hyperparameters further.

For the daily model, the number of estimators was set to 1000 and subsample per tree was set to 10,000. This model was built independently by a different team member, in part to see whether approaching the task with different hyperparameters would lead to a similar or different conclusion.

5.3.3 Significant features

The figure below provides examples of the feature importance plots for the 30min Base and Granular models. There is little difference between the two and the results closely resemble the XGBoost results, with the model primarily relying on demand for the same period yesterday (demand_lag_48), demand for the latest half hour before the forecast is made (demand_lag_07), and flags for monday, saturday and weekend, to capture changes in patterns between working and non-working days. Some weather features are included further down the list (starting with X and a BOM ID number). Feature significance was similar for the Daily models.

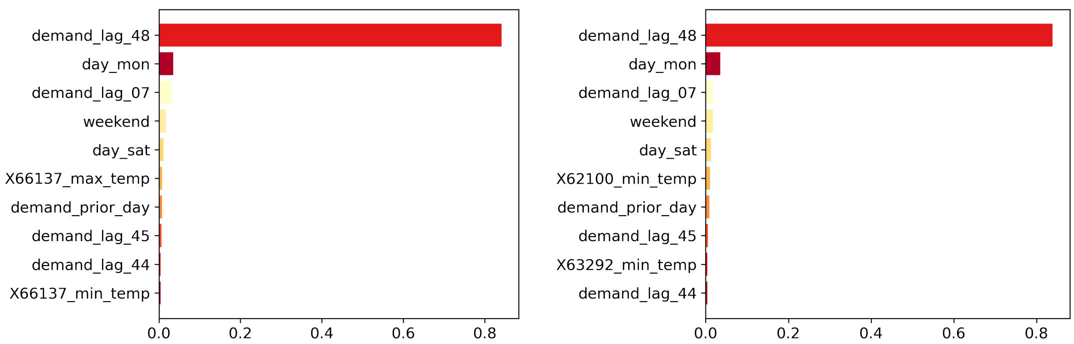

Figure 5.8: Feature importance for Random Forest 30min models

5.3.4 Results

The figure below compares the results for the Base, Aggregate and Granular models, for the Random Forest 30min models. There was little improvement due to incorporating the Aggregate and Granular weather data.

Figure 5.9: Model results for Random Forest daily models

The figure below compares the results for the Base, Aggregate and Granular models, for the Random Forest Daily models. There was a more noticible improvement from incorporating the Aggregate and Granular weather data, compared to the 30 minute models, with the Aggregate data performing better than the Granular data.

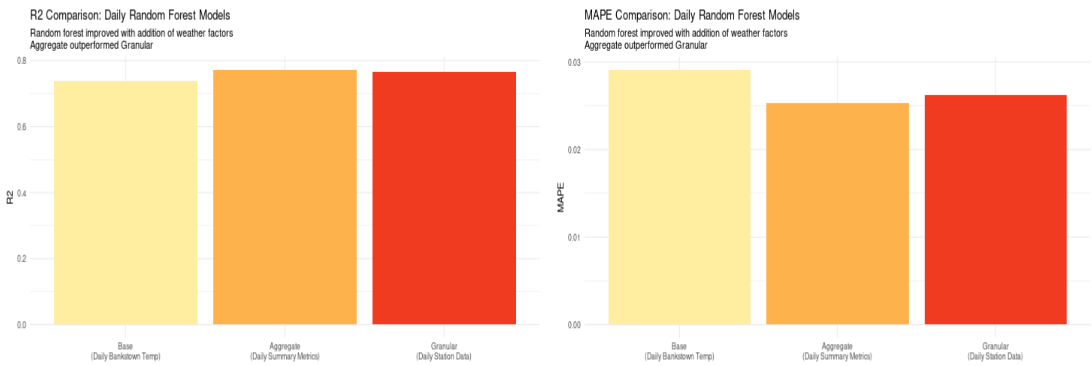

Figure 5.10: Model results for Random Forest daily models

5.4 Linear model

5.4.1 Background

The linear model methodology used takes guidance from Shah et al. (2019) in using component-wise approach to model the logarithm of the demand against a deterministic component and a stochastic component. Daily demand D_t is modelled as:

$log(D_t) = F_t + R_t$

Conceptually applying a similar approach, the F_t comprises a running week index (long term trend), annual, seasonal, weekly cycles and calendar effects. It is modelled as:

$F_t = l_t + a_t + s_t + w_t + h_t$

In the data, l_t is the weekly index numbering 1 to 585 which is the total number of weeks in the data set. As the data set was restricted to start from 2017, the starting week number was changed to 365. a_t is the day of the year (between 1 and 365) and unlike the rest of the variables in F_t which are modelled using OLS, a_t is modelled using a smoothing spline. s_t is a set of dummy variables representing the seasons. Seasons is preferred in the linear model instead of a dummy month variable in order to minimise predictor variables and reduce overfitting. With reference to common understanding of months in seasons, September to November is mapped to Spring, December to February is mapped to Summer, March to May is mapped to Autumn and June to August is mapped to Winter. w_t is a set of dummy variables for Monday to Sunday to indicate the day of the week. h_t is a dummy variable to indicate holidays.

l_t is modelled separately from periodic variables a_t, s_t, w_t and h_t (Lisi and Nan 2014). This is done by modelling l_t on $log(D_t)$ and then the periodic variables a_t, s_t, w_t and h_t on the residuals of $log(D_t) - \hat{l}_t$. Subsequently the stochastic component R_t is modelled against the residuals of $log(D_t) - \hat{l}_t - \hat{a}_t - \hat{s}_t - \hat{w}_t - \hat{h}_t$.

In various literature, the stochastic component is usually modelled using Autoregressive Moving Average (ARMA) type models. ARMA models require the training and testing data to be split as two successive periods, however our modelling data was split randomly by week. Hence, R_t is modelled as:

$R_t = log(d_{t-1}) + log(\sum_{i=1}^{7} d_{t-i})$

$log(d_{t-1})$ is the logarithm of the demand of the prior day and $log(\sum_{i=1}^{7} d_{t-i})$ is the logarithm of the demand of the prior week.

The above approach was followed to derive an initial base model with no weather data. This was compared to the following alternatives in order to separately test the impact of using more diverse weather metrics, and the impact of using aggregate metrics over more weather stations.

1. Adding daily temperature variables taken solely from Bankstown Airport
2. Adding daily temperature, wind speed and solar exposure variables taken solely from Bankstown Airport
3. Adding daily population density weighted mean of temperature, wind speed and solar exposure taken across all weather stations in NSW

5.4.2 Results

The figure below compares the results for the Base, Aggregate and Granular models, for the Linear models. Using only maximum temperature led to an improvement in model accuracy, and adding more diverse metrics led to a further improvement. However, using aggregated weather measures across all stations was less effective than solely using weather variables from Bankstown.

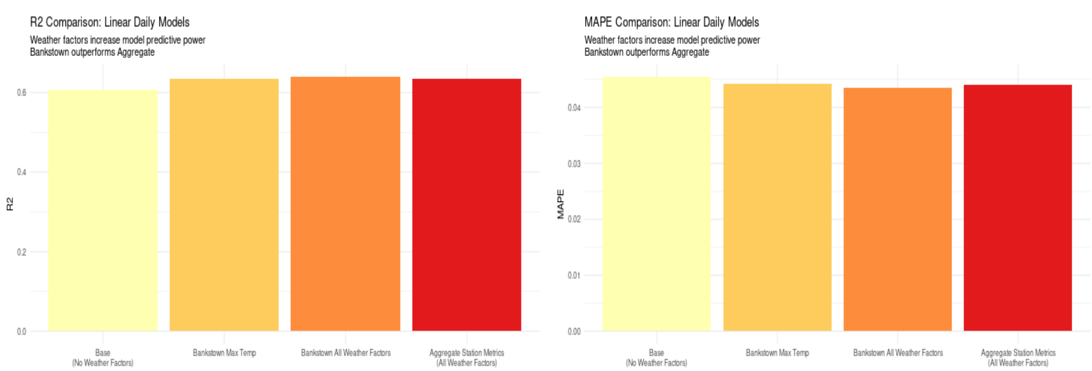

Figure 5.11: Model results for Linear models

Chapter 6

Discussion

We were provided with benchmark forecasts, based in part on temperature readings taken every 6 minutes throughout the day at Bankstown airport. The following figure compares the accuracy of this model with the six base case models that only use the minimum and maximum temperature for the day at Bankstown airport. The benchmark model performance is quite high, and this suggests a significant benefit to using intra-day weather data.

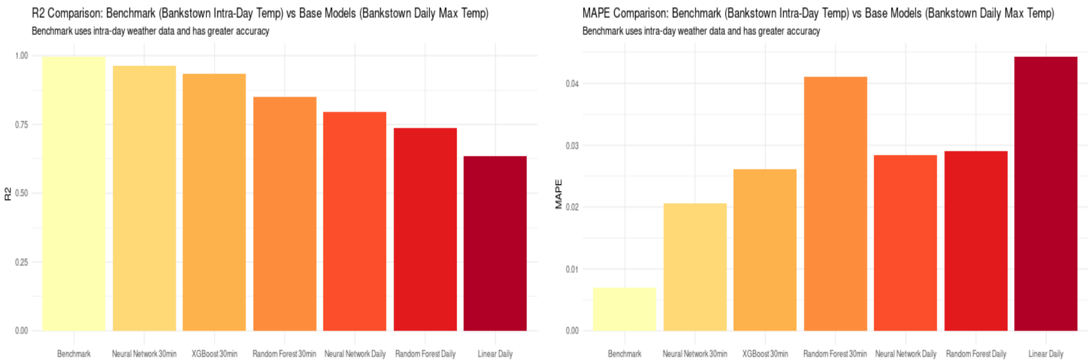

Figure 6.1: Comparison of Benchmark Forecast and Base Model

The 30 minute models have higher R2 and lower MAPE than the Daily models. Between the different model frameworks considered, the Neural Network models tend to have the best performance, although they were more difficult to tune and require more effort to interpet. Of the remaining models, the XGBoost performed relatively well, followed by the random forest and linear modelling approaches.

To summarise the findings of each of the models, relative to the Base models:

Model	Findings
Neural Networks: 30min	7.0% reduction in RMSE for Granular model when using PCA. 1.1% reduction for Aggregate model.
XGBoost: 30min	Negligible impact due to additional weather variables
Random Forest: 30min	Negligible impact due to additional weather variables
Neural Network: Daily	5.0% reduction in RMSE for Granular model when using PCA. 4.2% reduction for Aggregate model.

Model	Findings
Random Forest: Daily	5.6% reduction in RMSE for Granular model. 6.7% reduction for Aggregate model.
Linear: Daily	0.7% reduction in RMSE when adding additional weather variables (Bankstown station only)

Based on the above results, and giving some weight to the relative accuracy of the Base models, we conclude there is sufficient evidence to argue that using more diverse weather data may improve STLF accuracy by up to 7%. The reduction attained may depend on the model used; in our experiments tree-based methods seemed to give little weight to the weather data where neural networks relied on it more. Some sort of dimensionality reduction seemed necessary to handle the large amount of highly correlated data from so many weather stations. We experimented with both aggregate statistics, such as the population weighted mean, and the use of PCA, and found both approaches successful.

In general, the impact of weather data on the models was much less significant than the impact of time and prior demand factors. However, the benchmark forecasts suggest that intra-day weather data is more predictive than daily weather data and so these results may represent a conservative estimate of the potential gain of using more diverse weather data in practice.

Chapter 7

Conclusion and Further Issues

Our group examined the question, how much better are STLF when they are based on a diverse set of weather data? We developed STLF for the state of NSW in Australia and found evidence to suggest that using more diverse weather data can improve accuracy by up to 7%. We found that our models were unable to handle the granular and highly correlated raw data from a large number of weather stations. Some sort of dimensionality reduction, such as the use of aggregate statistics, or PCA, was required for the models to avoid overfitting.

Capturing and processing a large volume of weather data has an associated cost in terms of model complexity and analytical expertise. Our analysis may help Endgame and its clients to make a more informed decision about whether such a cost is justified in terms of predictive capability. A relatively small improvement in STLF accuracy of 1-3% can result in substantial cost savings, measured in the millions of dollars (Islam, Baharudin, and Nallagownden 2017). Based on our results, we believe there would be clients who would see a positive return on investment from seeking to use more diverse weather data.

Future areas of work might focus on:

- We neglected to include humidity in our modelling data, despite the data being available and it being used in practice. Models such as neural networks are likely to benefit from capturing the interaction between temperature and humidity and we would expect to see a further improvement in accuracy if this were included.
- Our models used daily weather data; our study could be replicated using intra-day weather data, and we would expect to see a much greater impact.
- Our models are all point estimate models; probabilistic models may be more useful for decison making purposes (Li 2020)

References

Apadula, Francesco, Alessandra Bassini, Alberto Elli, and Simone Scapin. 2012. "Relationships Between Meteorological Variables and Monthly Electricity Demand." *Applied Energy* 98 (October): 346–56. https://doi.org/10.1016/j.apenergy.2012.03.053.

Australian Energy Market Operator. 2020. "Electricity Demand Forecasting Methodology Information Paper." Australian Energy Market Operator. https://aemo.com.au/-/media/files/electricity/nem/planning_and_forecasting/inputs-assu 2020/2020-electricity-demand-forecasting-methodology-information-paper.pdf?la=en.

———. 2021a. "Power System Operating Procedure - Load Forecasting." Australian Energy Market Operator. https://aemo.com.au/en/energy-systems/electricity/national-electricity-market-nem/nem-forecasting-and-planning/operational-forecasting/load-forecasting-in-pre-dispatch-and-stpasa.

———. 2021b. "Temperature Forecast Analysis for Winter 2020." Australian Energy Market Operator. https://aemo.com.au/-/media/files/electricity/nem/planning_and_forecasting/load-forecasting/temperature-forecast-analysis-fo pdf?la=en.

Chen, Tianqi, Tong He, Michael Benesty, and Yuan Tang. n.d. "Understand Your Dataset with Xgboost." cran.r-project.org. Accessed June 8, 2021. https://cran.r-project.org/web/packages/xgboost/vignettes/discoverYourData.html#numeric-v.s.-categorical-variables.

Čupeljić, B. 2016. "A Short Term Forecasting of Electricity Demand: The Case of Slovenia." PhD thesis. cek.ef.uni-lj.si/magister/cupeljic2227-B.pdf.

DeLong, John P., and Oskar Burger. 2015. "Socio-Economic Instability and the Scaling of Energy Use with Population Size." Edited by Rodrigo Huerta-Quintanilla. *PLOS ONE* 10 (June): e0130547. https://doi.org/10.1371/journal.pone.0130547.

Fan, Shu, and Rob J. Hyndman. 2012. "Short-Term Load Forecasting Based on a Semi-Parametric Additive Model." *IEEE Transactions on Power Systems* 27 (February): 134–41. https://doi.org/10.1109/tpwrs.2011.2162082.

Geron, Aurelien. 2019. *Hands-on Machine Learning with Scikit-Learn and Tensorflow Concepts, Tools, and Techniques to Build Intelligent Systems.* O'Reilly Media, Inc.

Gonzalez-Salazar, Miguel Angel, Trevor Kirsten, and Lubos Prchlik. 2018. "Review of the Operational Flexibility and Emissions of Gas- and Coal-Fired Power Plants in a Future with Growing Renewables." *Renewable and Sustainable Energy Reviews* 82 (February): 1497–1513. https://doi.org/10.1016/j.rser.2017.05.278.

Goodfellow, Ian, Yoshua Bengio, and Aaron Courville. 2016. *Deep Learning.* The Mit Press.

Islam, Badar, Zuhairi Baharudin, and Perumal Nallagownden. 2017. "Modified Meta Heuristics and Improved Backpropagation Neural Network-Based Electrical Load Demand Prediction Technique for Smart Grid." *IEEJ Transactions on Electrical and Electronic Engineering* 12 (June): S20–S32. https://doi.org/10.1002/tee.22420.

Lahouar, A., and J. Ben Hadj Slama. 2015. "Day-Ahead Load Forecast Using Random Forest and Expert Input Selection." *Energy Conversion and Management* 103 (October): 1040–51. https://doi.org/10.1016/j.enconman.2015.07.041.

Li, Yike. 2020. "Optimal Weather Station Selection for Electric Load Forecasting." PhD thesis.

Lisi, Francesco, and Fany Nan. 2014. "Component Estimation for Electricity Prices: Procedures and Comparisons." *Energy Economics* 44 (July): 143–59. https://doi.org/10.1016/j.eneco.2014.03.018.

Morgan, Liam. 2020. "RPubs - Missforest - Missing Data Imputation." rpubs.com. https://rpubs.com/lmorgan95/MissForest.

Pardo, Angel, Vicente Meneu, and Enric Valor. 2002. "Temperature and Seasonality Influences on Spanish Electricity Load." *Energy Economics* 24 (January): 55–70. https://doi.org/10.1016/s0140-9883(01)00082-2.

Sailor, David J., and J. Ricardo Muñoz. 1997. "Sensitivity of Electricity and Natural Gas Consumption to Climate in the U.s.a. - Methodology and Results for Eight States." *Energy* 22 (October): 987–98. https://doi.org/10.1016/S0360-5442(97)00034-0.

Shah, Ismail, Hasnain Iftikhar, Sajid Ali, and Depeng Wang. 2019. "Short-Term Electricity Demand Forecasting Using Componentsestimation Technique." *Energies* 12 (July): 2532. https://doi.org/10.3390/en12132532.

Sobhani, Masoud, Allison Campbell, Saurabh Sangamwar, Changlin Li, and Tao Hong. 2019. "Combining Weather Stations for Electric Load Forecasting." *Energies* 12 (April): 1510. https://doi.org/10.3390/en12081510.

Zhang, Ning, Zhiying Li, Xun Zou, and Steven M. Quiring. 2019. "Comparison of Three Short-Term Load Forecast Models in Southern California." *Energy* 189 (October): 116358. https://doi.org/10.1016/j.energy.2019.116358.

Appendix

Neural Network 30min Model Hyperparameter Sensitivities

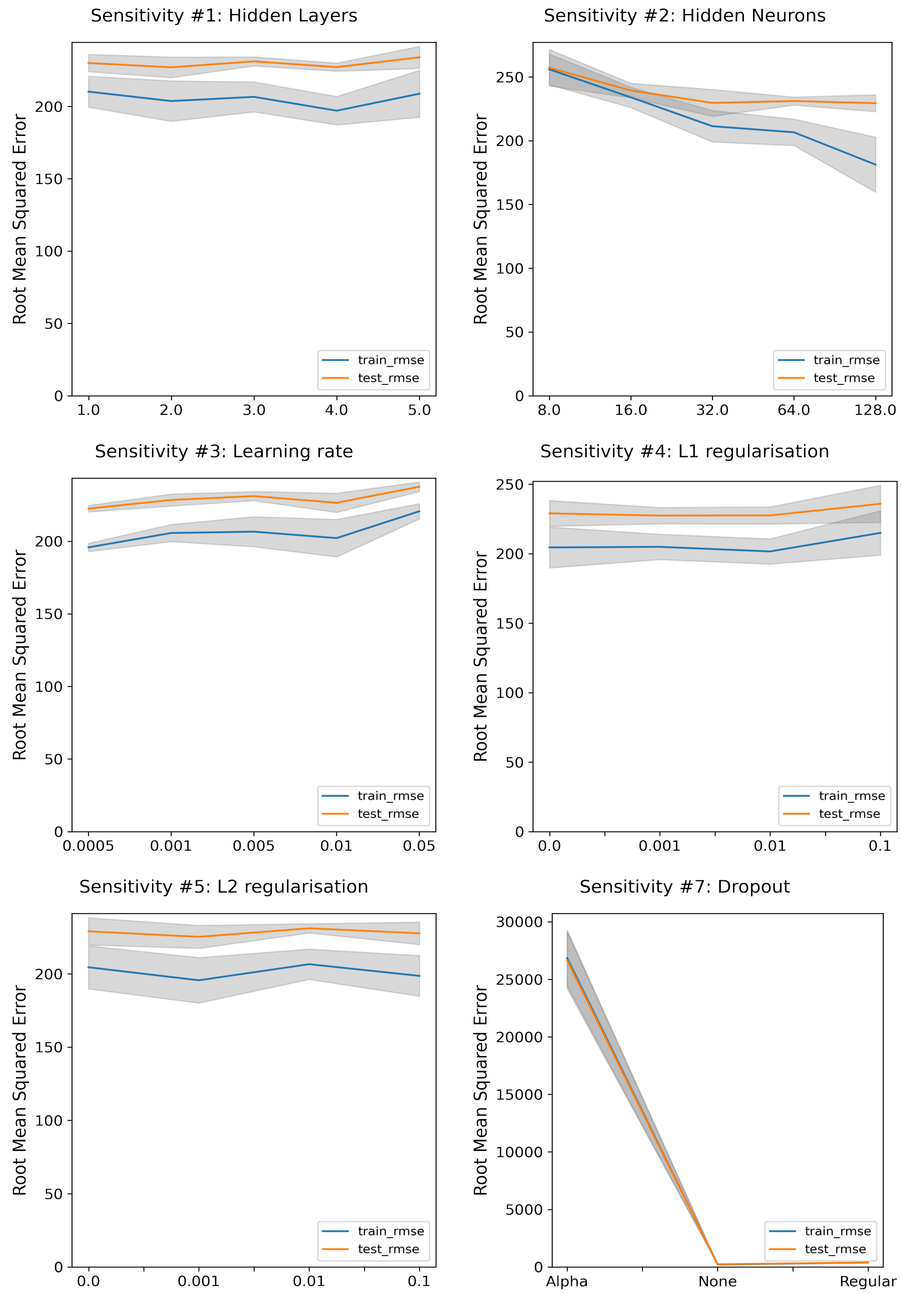

Figure 7.1: Base model hyperparameter sensitivities

The most sensitive hyperparameter was the number of hidden neurons; with fewer neurons the network was unable to capture patterns in the data as well (higher training set error) while with greater neurons the model tended to overfit (greater gap between training and set set error).

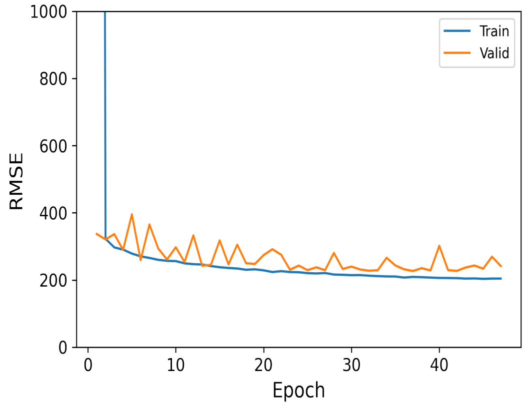

Figure 7.2: Base model learning curve

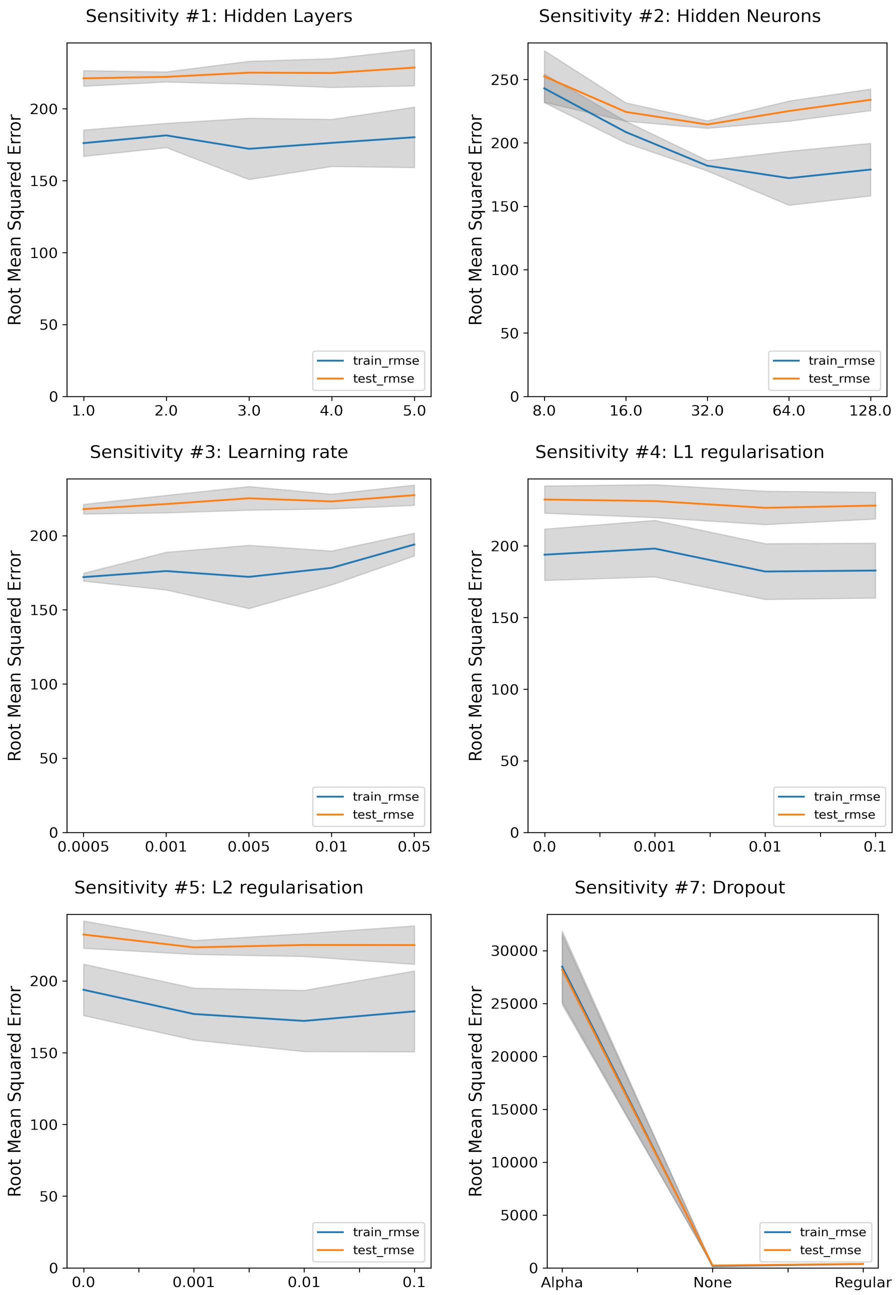

Figure 7.3: Aggregate model hyperparameter sensitivities

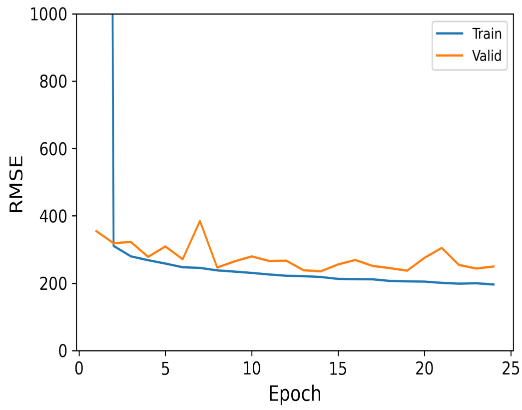

Figure 7.4: Aggregate model learning curve

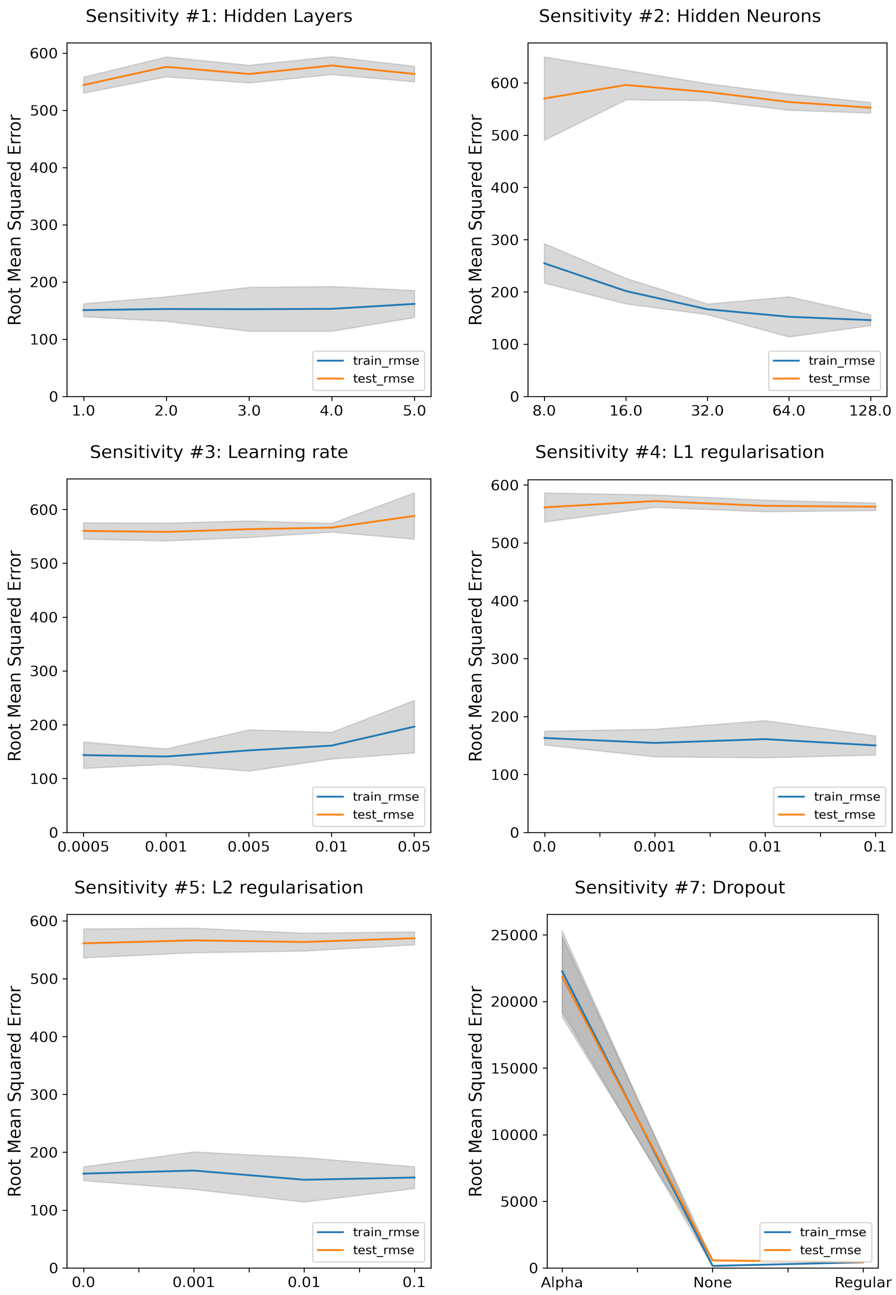

Figure 7.5: Granular model hyperparameter sensitivities

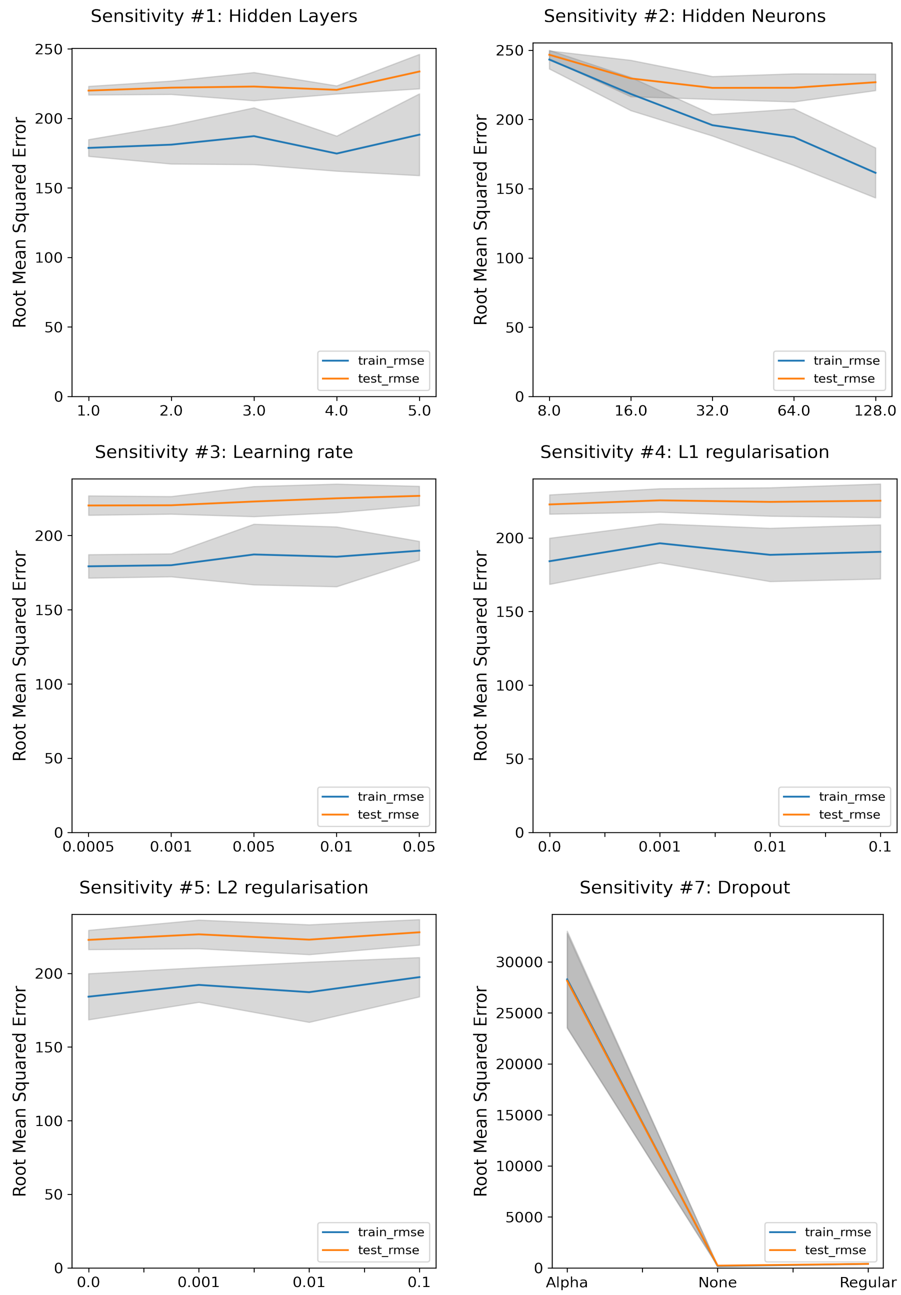

Figure 7.6: Granular model with PCA hyperparameter sensitivities

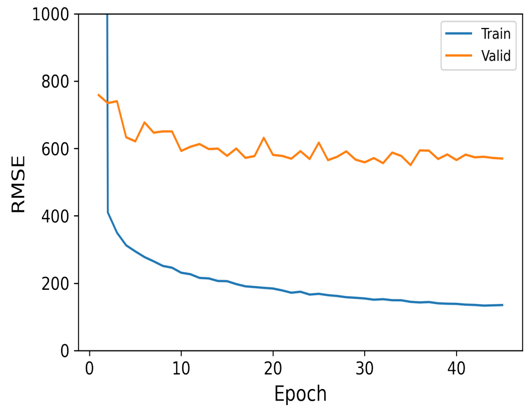

Figure 7.7: Granular model learning curve

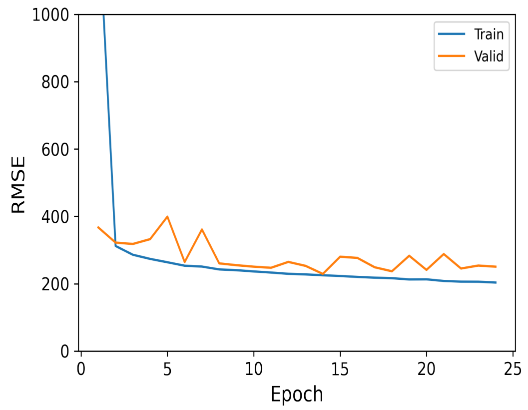

Figure 7.8: Granular model with PCA learning curve

Neural Network Daily Model Hyperparameter Sensitivities

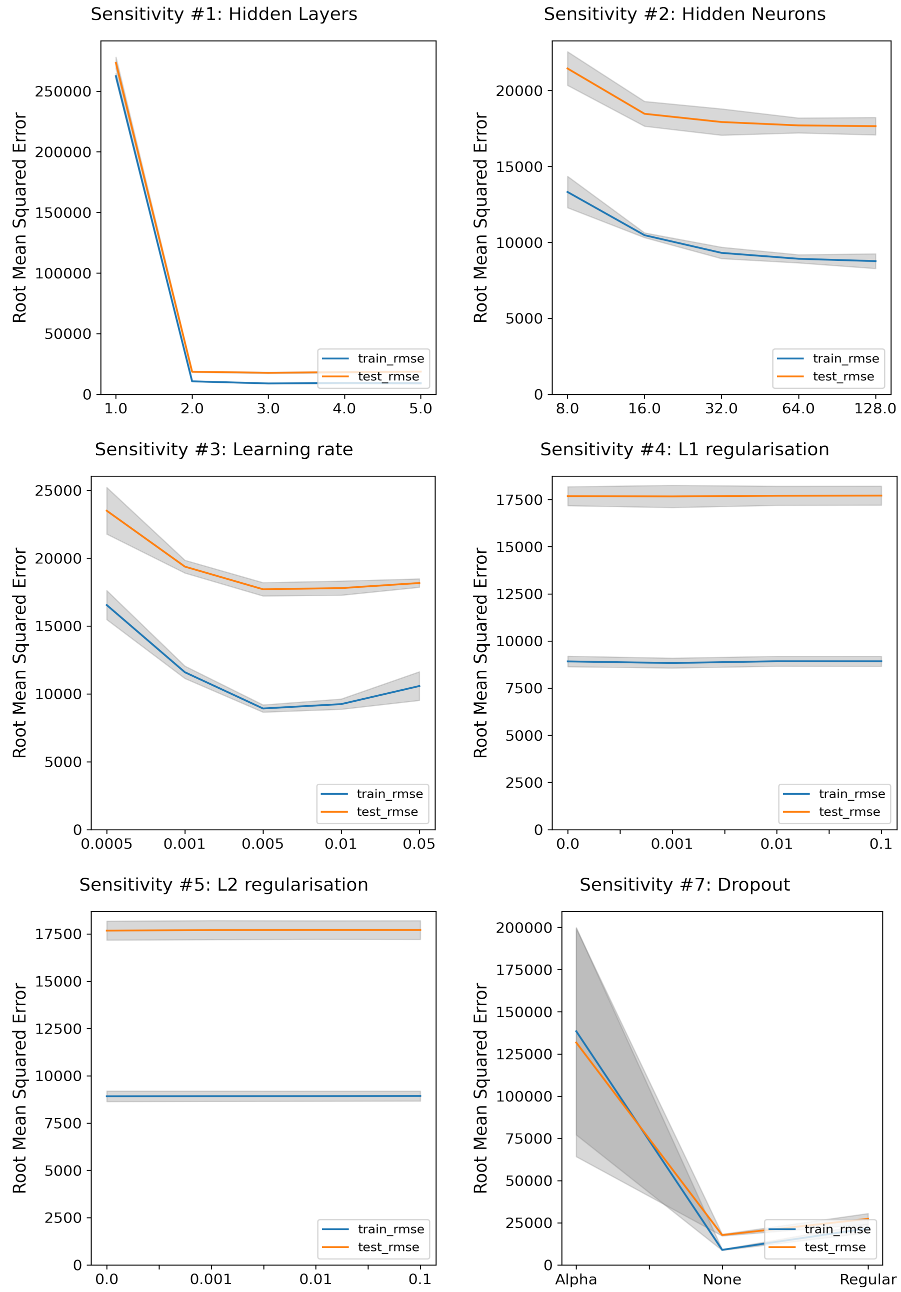

Figure 7.9: Base model hyperparameter sensitivities

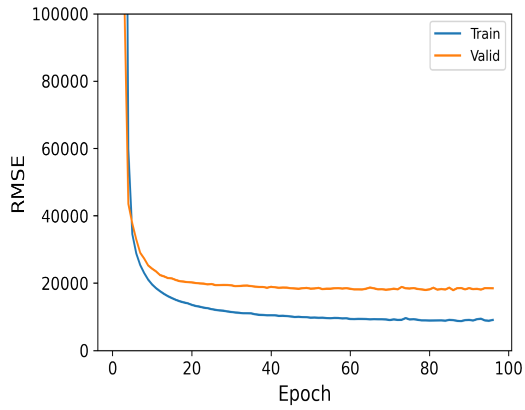

Figure 7.10: Base model learning curve

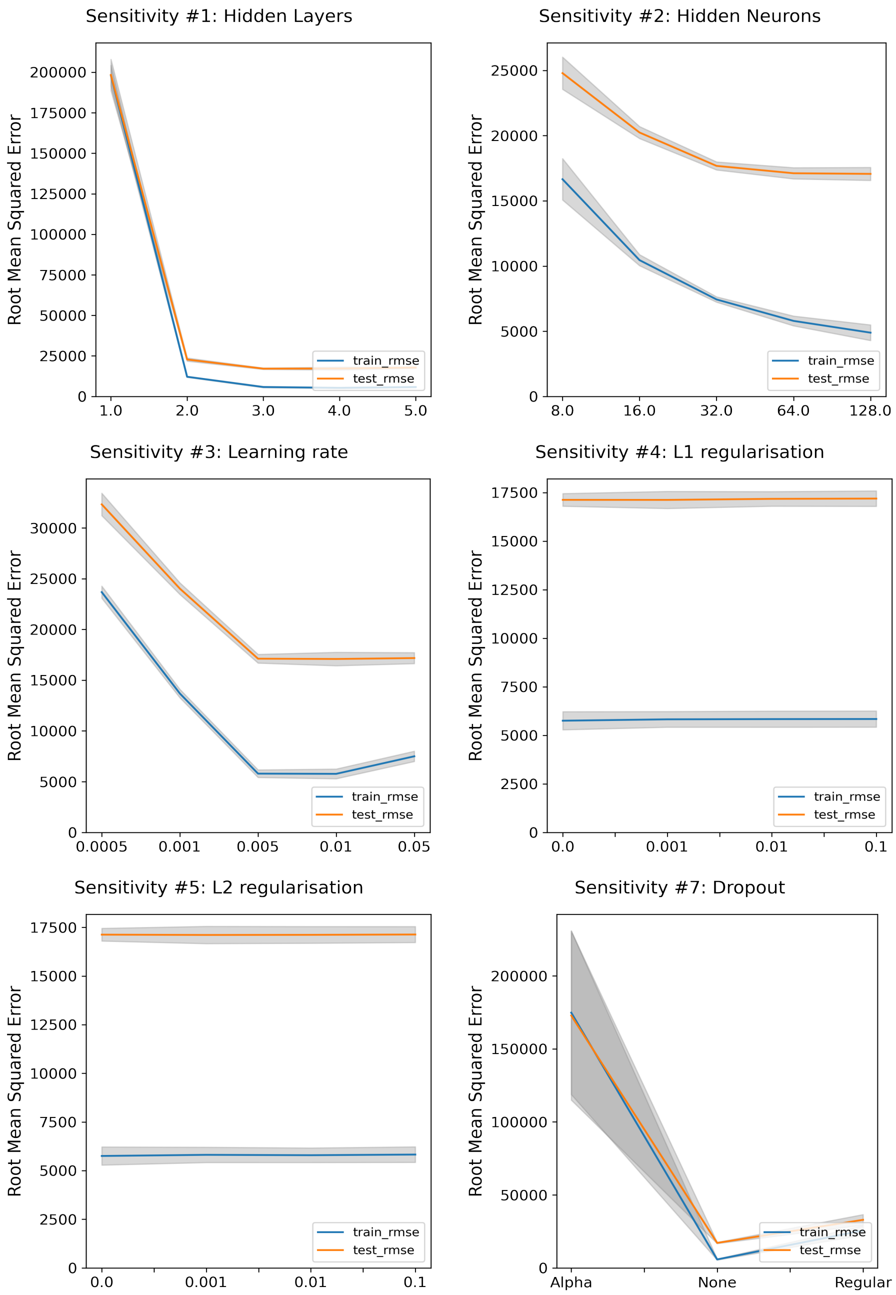

Figure 7.11: Aggregate model hyperparameter sensitivities

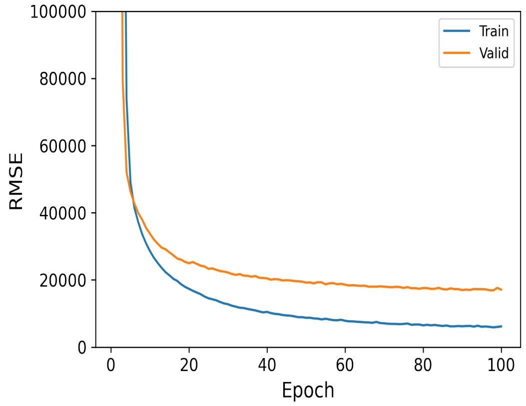

Figure 7.12: Aggregate model learning curve

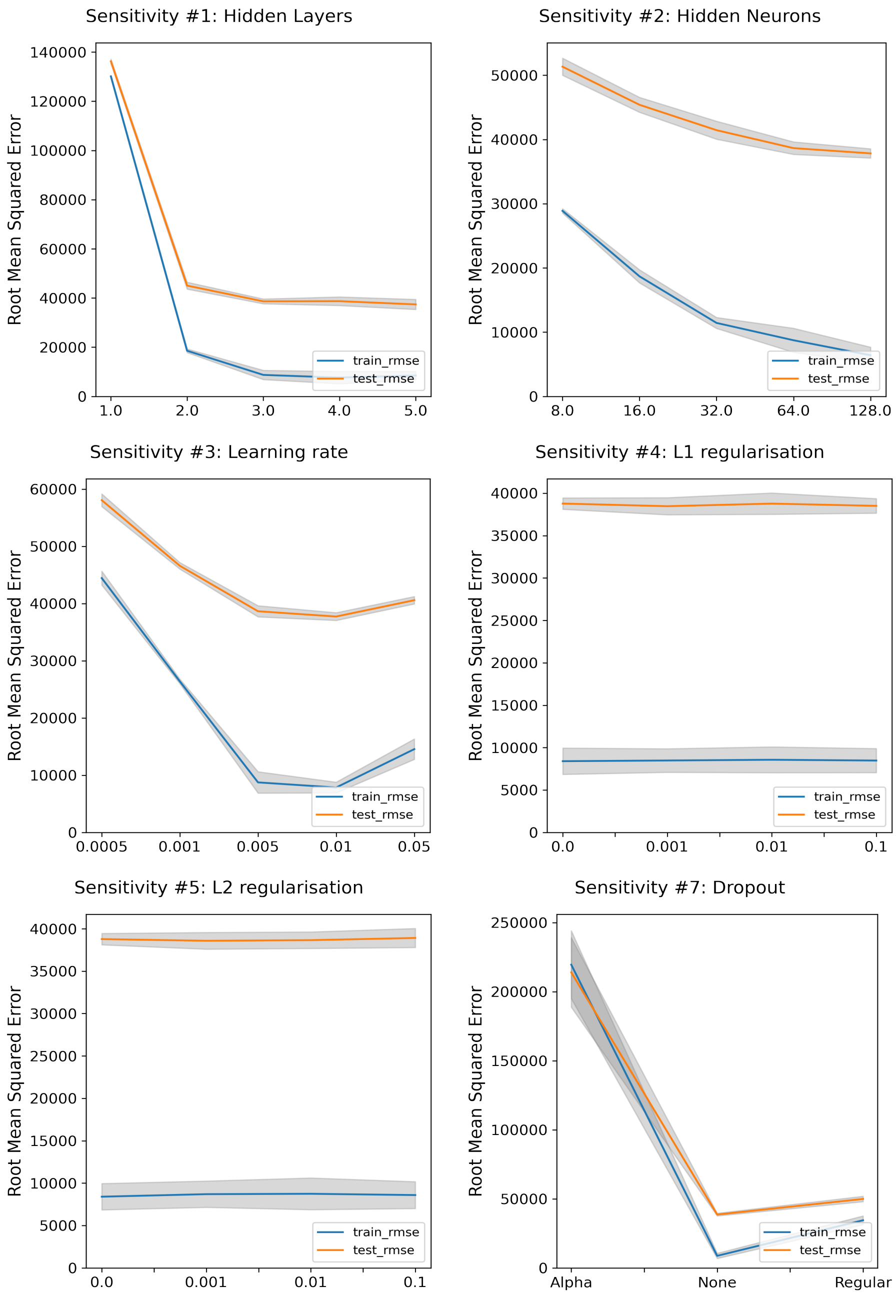

Figure 7.13: Granular model hyperparameter sensitivities

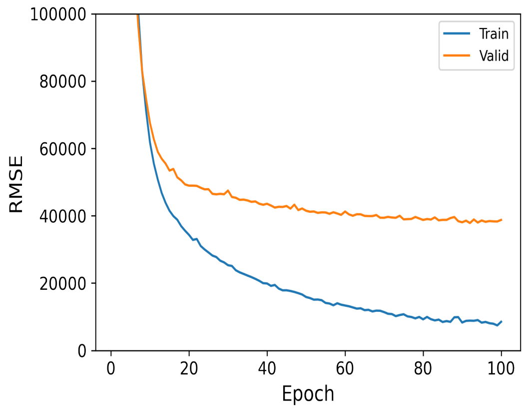

Figure 7.14: Granular model learning curve

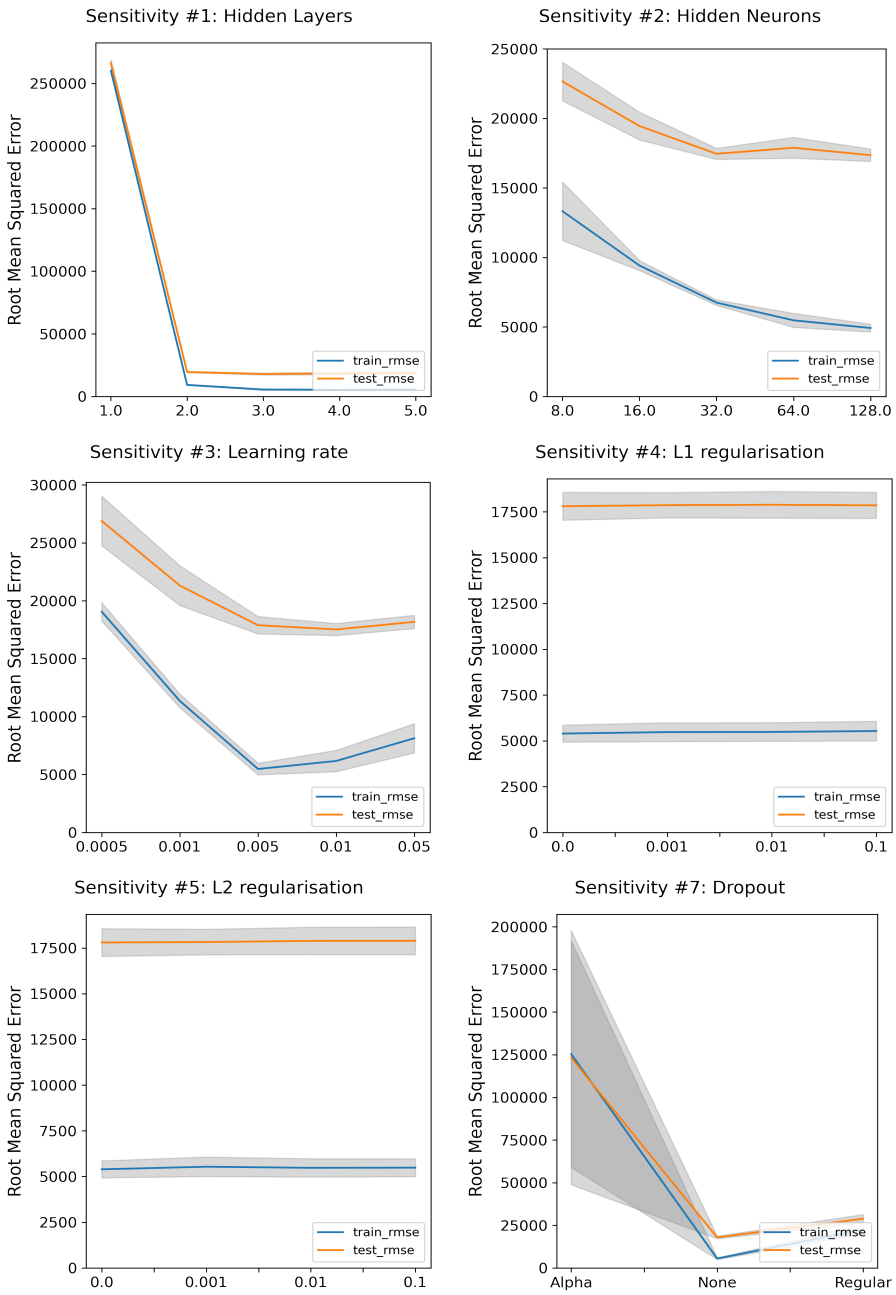

Figure 7.15: Granular model with PCA hyperparameter sensitivities

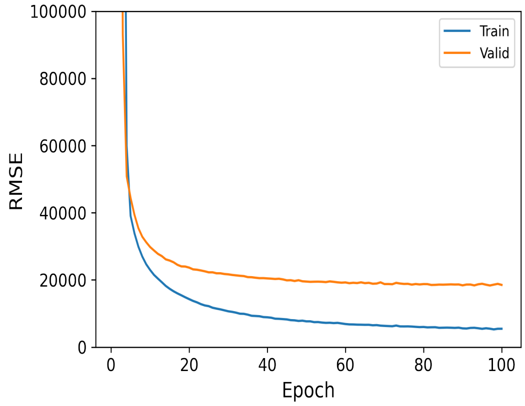

Figure 7.16: Granular model with PCA learning curve