**2 Analysis**

**2.1 Introduction**

An analysis of the dataset was undertaken concurrently with the development of the system to appreciate the load forecasting challenges identified in **chapter 1.** The *Jupyter Notebook* analysis environment enabled the creation of load forecasting models and evaluation of their performance statistically in a sandbox environment. Complex data manipulation functionality was written in *Python* in the analysis and was integrated into the solution. Models found to have the best forecasting performance were included in the system. The analysis of model performance was used as a source of proof to cross verify the system’s outputted performance results. The dataset was also assessed on its viability to facilitate proposed solution features that leverage specific characteristics of the data.

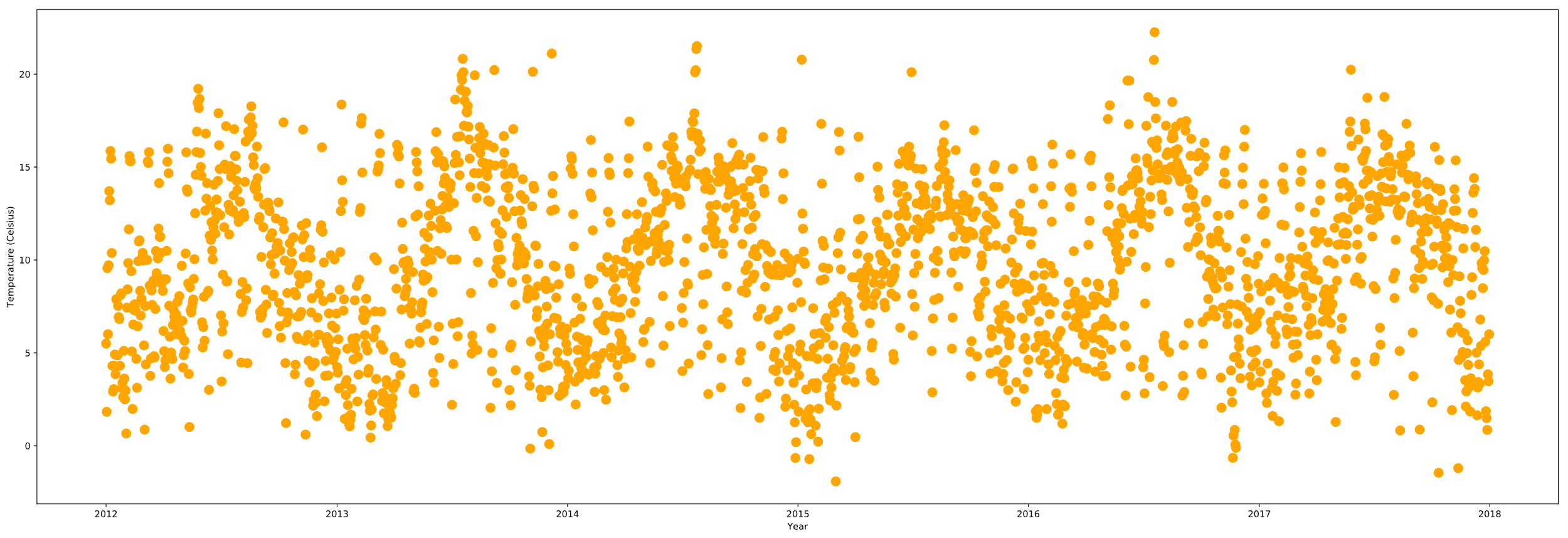
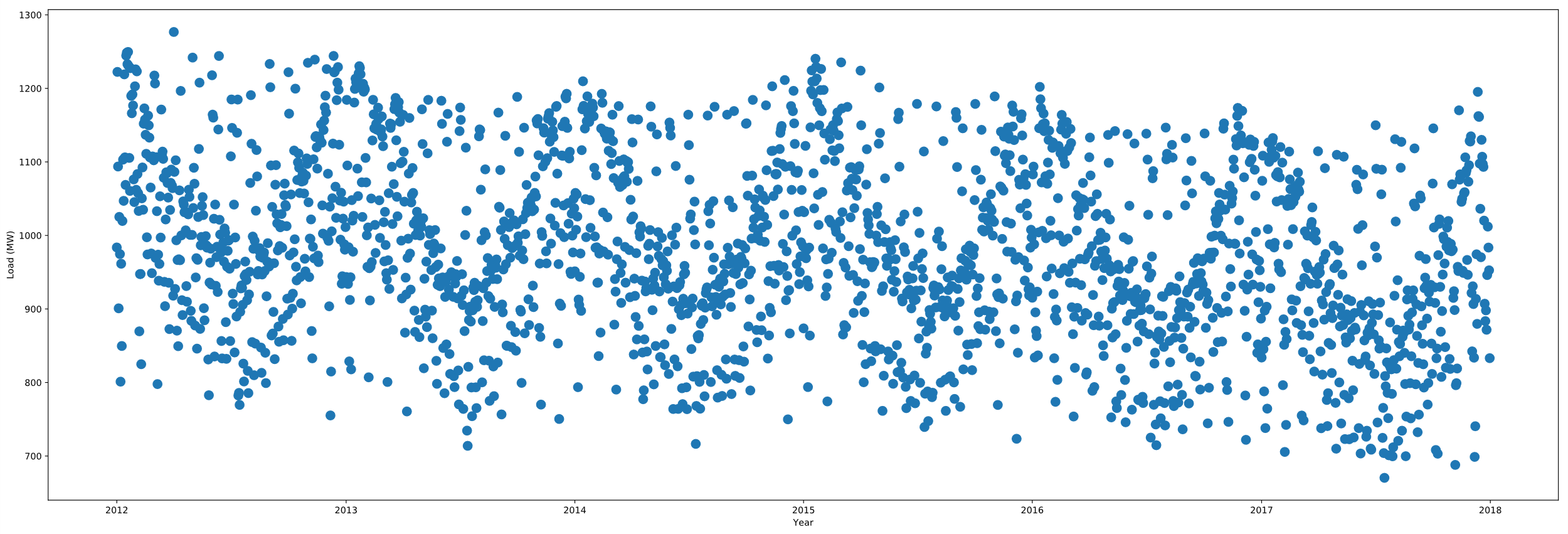
**2.2 The Dataset**

The dataset used for analysis was provided by SONI (System Operator for Northern Ireland). The data was provided in MATLAB data format (.mat). To use the dataset in the analysis environment, its data was converted into Excel workbook format (.xlsx) using MATLAB and then saved as a comma-separated values file (.csv) using Microsoft Excel. The dataset was indexed by date ascending ranging from 1st February 2010 to 1st July 2018, with an interruption in the load value entries between 7th February 2011 and 10th November 2011. The time series interval was 30 minutes.

The main column of interest for model performance analysis was the ‘Load’ column. This is the megawatt (MW) load system demand of the Northern Ireland’s power grid for a date and time. The column was used in analysis to assess a model’s forecasting performance by comparing the column’s value with a model’s predicted load. The dataset was also augmented with columns that can be used as candidate explanatory variables in load forecasting regression models. Meteorological columns sourced from the MET office from one location in Northern Ireland (Aldergrove) are included in the dataset. Furthermore, the dataset includes columns derived from the dataset’s ‘Date’ column. Further date derived variables not present in the dataset were added as part of data cleaning process. **See Appendix A** for a table containing all the columns of the dataset.

**2.4 Load-Temperature Analysis**

Studies on short term electricity load forecasting identify weather as being influential in deviations of load demand [1]. It has been observed that people’s activity patterns change between abnormally warm or cold periods. Hence, temperature was chosen for analysis to assess whether it was an influential weather variable that has a strong linear relationship with load.



***Figure 2.1******(a)*** *24-hour system load average 2012-2018* ***(b)*** *Temperature 24-hour average 2012-2018*

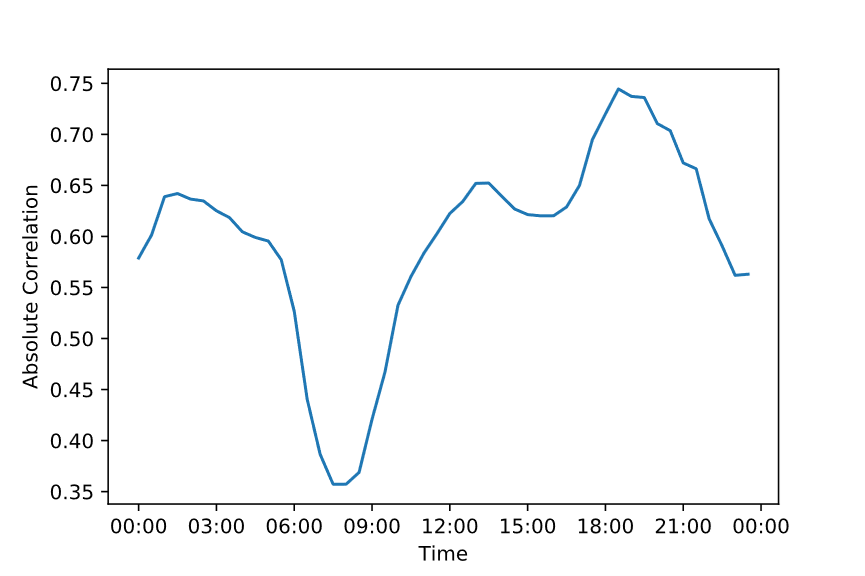
(a)

(b)

There is a sine wave like cyclic pattern in both the load and temperature graphs, **see Figure 2.1.** Load follows an inverse cycle to Temperature, with the troughs and rises in **(a)** visually contrasting **(b).** From the graphs, it can be determined the different seasons of the year influence the temperature each year, causing a cyclic pattern across the year. Therefore, an observation is that in warmer periods, less electricity load is used while in colder periods, more electricity load is used.

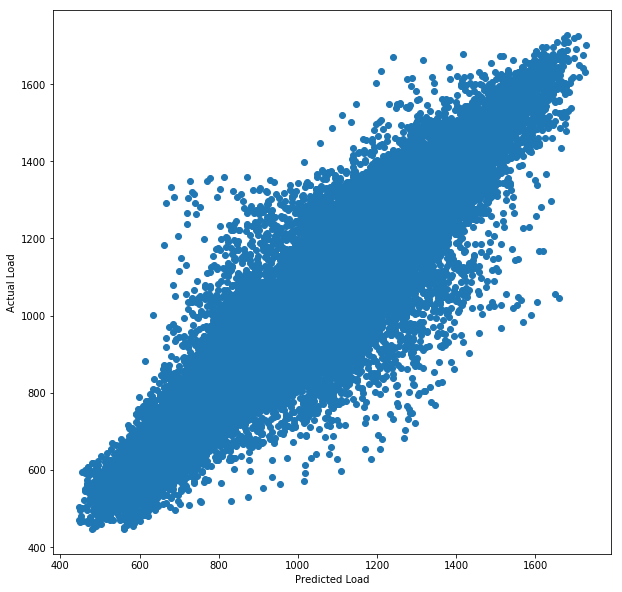
An investigation into whether temperature and load had a linear relationship in a more specific split of the data than yearly trends was undertaken. This was important to determine the value in building forecasting models which include a data entry’s datetime characteristics for model input variables. The data was split by the time of the day, **see figure 2.2**. Pearson’s correlation coefficient is a metric used to measure the strength of a positive or negative linear relationship between load and temperature. The highest correlation is in the evening and the lowest in the morning. This can be attributed to people being inactive at night and hence less reactive to changes in temperature. During day hours there is sunshine, and this this could be influencing the weaker temperature correlation with load in the day with photovoltaic electricity generation reducing the system demand. regression models.

***Figure 2.2*** *Absolute correlation between load and temperature for each half hour interval****.***



**2.5 Displacement Model Analysis**

|  |  |
| --- | --- |
| **Model** | **Correlation** |
| SDLD | 0.928988 |
| SDLW | 0.961703 |
| SDLY | 0.959353 |

Different displacement models were considered for analysis: Last Day, Last Week and Last Year. The models were built on the dataset 2013-2018 to ensure a full range of displaced load entries were present in the data to use as the prediction load value. These approaches to creating a displacement model have their advantages and unique limitations. However, the displacement models all have the advantage of using the same time entry of the day, meaning the different pattern observedin the previous chapter specific to a time of day was considered in the load forecast.

**(a)**

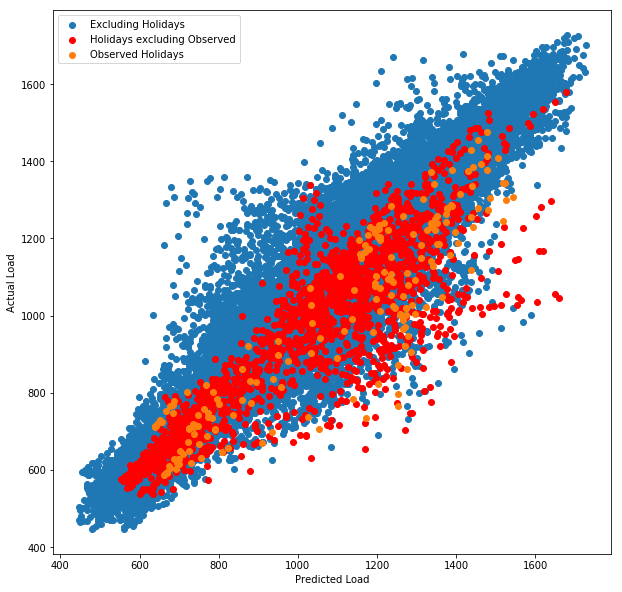
**(b)**

***Figure 2.3******(a)*** *Correlation between actual load and predicted load for each displacement model* ***(b)*** *predicted load-actual load for SDLW Model using half hour load entries between 2013-2018*

The model with the strongest linear relationship of its forecasted load output to the actual load entry is Same Day Last Week, and weakest is Same Day Last Day, **see Figure 2.3 (a)**. For building linear regression models in further analysis SDLW model forecasts will be included as a variable in the dataset. This will be for selection as an explanatory variable, or as a variable that can be corrected by other explanatory variables to produce more accurate load forecasts.

**2.6 Holiday Analysis**

|  |  |
| --- | --- |
| **Day type** | **Correlation** |
| Excluding Holidays | 0.968 |
| Holidays excluding Observed | 0.875 |
| Observed Holidays | 0.898 |

Holiday days that have lower system demand load entries than normal days impact the load forecasting performance of displacement models. The dataset contains an abnormal day binary column that records data entries that are holidays or affected by an observed ‘bleeding’ effect of holidays on surrounding non-holiday days. Taking advantage of Python, the ‘*holidays’* library was used to generate a list of all the holiday days in Northern Ireland. The list includes observed holiday days - days which follow a weekend the holiday occurred on.

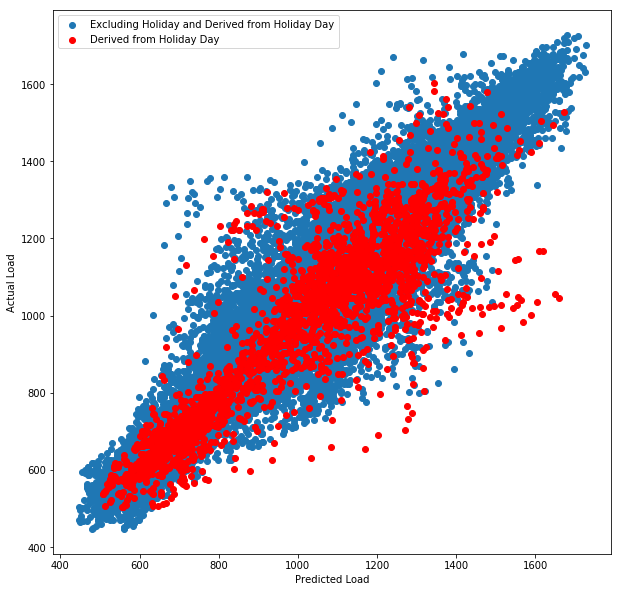
**(a)**

**(b)**

***Figure 2.4******(a)*** *Correlation between predicted load using the sdlw model and actual load for each day type* ***(b)*** *predicted load-actual load for SDLW Model using half hour load entries between 2013-2018*

By mapping the ‘date’ column of the dataset with the generated list of Northern Ireland holiday dates, the dataset entries were labelled if they were a holiday and/or an observed holiday. The SDLW displacement model was used for performance analysis with holiday days. ‘Excluding Holidays’ days have the strongest positive correlation with actual load using the SDLW model forecasts, and ‘Holiday excluding observed’ holidays entries have the weakest, see **Figure 2.4 (a).** Visually in **Figure 2.4 (b)**, there is load over forecasting for holiday day entries with the model forecasted load having a larger MW value than the actual load.An explanation for this is that holiday days are not using holiday days to predict the load. This supports the assumption that people in Northern Ireland do not follow the same social pattern on holidays. There is a marginally stronger model performance for observed holiday days than holidays days, but they still underperform compared to excluding holiday day entries.

|  |  |
| --- | --- |
| **Day type** | **Correlation** |
| Excluding Holiday and Derived from Holiday Day | 0.970 |
| Derived from Holiday Day | 0.885 |



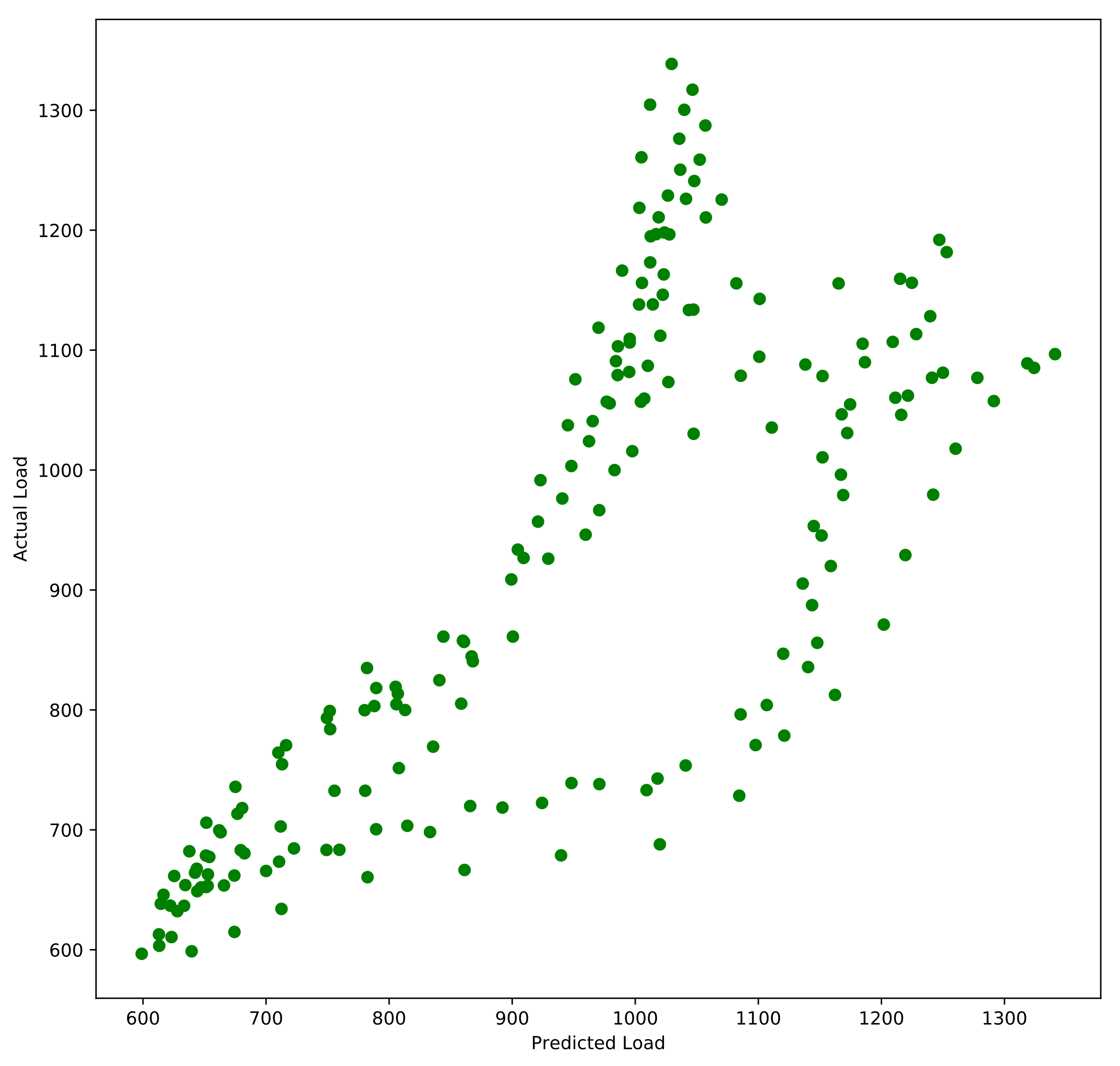
**(b)**

***Figure 2.5******(a)*** *Correlation between predicted load using the SDLW model and actual load for each day type* ***(b)*** *predicted load-actual load for SDLW Model using half hour load entries between 2013-2018*

**(a)**

There are non-holidays entries that use holiday day load entries as their predicted load entry in the SDLW model. The performance of these entries was considered and are referred to as ‘Derived from Holiday Day’. The ‘Excluding Holiday and Derived from Holiday’ days have a stronger linear relationship with actual load than excluding only holiday days, **see Figure 2.5 (a).**  The derived from holiday day entries do not perform well in the SDLW model, with a weak linear relationship with actual load. Visually, there is a concentration of entries both under forecasting load and over forecasting load, **see (b).** The under forecasting of the actual load can be attributed to, as identified in the previous analysis, holiday days having abnormally low load entries for the period of the year [2].

|  |  |
| --- | --- |
| **Holiday** | **Correlation** |
| Holiday New Year's Day | 0.739, **0.833** |
| Derived Christmas Day | 0.748, **0.788** |
| Holiday Easter Monday | 0.755 |
| Holiday Christmas Day | 0.820, **0.941** |
| Holiday Boxing Day | 0.841, **0.985** |

The holiday library enables identification of which holiday the data entry is and which holiday entry the load entry is derived from. Therefore, the performance of the SDLW model for different individual holidays was compared.

**(a)**

**(b)**

***Figure 2.6******(a)*** *Top 5 weakest actual load-predicted SDLW load correlated holidays. Correlation using SDLY model for Christmas period days is in bold – see* ***Appendix B*** *for**full list.* ***(b)*** *Predicted load-actual load for SDLW Model for data entries on New Year’s Day 2013-2018.*

The worst performing holiday using the SDLW model is New Year’s Day, **see Figure 2.6 (a)**. Visually, the SDLW under forecasts the load on New Year’s Day, **see (b)**. The Christmas period holiday days all have a weaker linear relationship with Load. The SDLW model for these days and days derived from these days does not produce an accurate load forecast. A correction strategy was created to create an accurate load forecast for these days.

The correction strategy chosen was using the SDLY (365 days) model for days derived from the Christmas holidays and the actual holiday days. A displacement of 365 days ensures it is the same holiday day in non-leap years and hence the holiday load pattern is considered. This improved the performance of all the Christmas period days **see (a).**

**2.7 Linear Regression Model Analysis**

Linear regression models are a popular choice for load forecasting [2]. Analysis in other research on the SONI dataset has observed accurate forecasts using linear regression models [2]. For analysis the Python machine learning library **scikit-learn** was used which uses the least squares fitting technique to fit the model. A full year of data was used as training data to train the model and the ascending years were used as the test data to evaluate the forecasting performance of the trained model. Mean absolute percentage error (MAPE) was the metric chosen to measure the prediction accuracy of the linear regression model forecasts. It is the most widely used metric in load forecasting [3].

Temperature Only Linear Regression Model

Temperature has been identified as having a linear relationship with load. A linear regression model was created that included ‘Base Temperature’ as an explanatory variable to predict load :

***TABLE 2.1*** *The MAPE of the temperature only linear regression model fitted with a training year predicting the load of the test year. White cells: Year ahead load forecast performance. Blue cell (right): The average MAPE of the training year predicting test years.*

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Test** | 2012 | 2013 | 2014 | 2015 | 2016 | 2017 | **Average** |
| **Training** |  |  |  |  |  |  |  |
| 2012 | 23.334 | 24.265 | 24.410 | 24.096 | 25.201 | 27.894 | 24.867 |
| 2013 | 23.402 | 24.077 | 24.145 | 23.906 | 24.744 | 27.000 | 24.546 |
| 2014 | 23.172 | 23.808 | 23.825 | 23.543 | 24.307 | 26.401 | 24.176 |
| 2015 | 23.082 | 23.705 | 23.700 | 23.402 | 24.137 | 26.165 | 24.032 |
| 2016 | 23.081 | 23.564 | 23.504 | 23.170 | 23.757 | 25.463 | 23.757 |
| 2017 | 23.239 | 23.584 | 23.464 | 23.046 | 23.435 | 24.653 | 23.570 |

The average MAPE of the model forecasting a year ahead: 24.311%.

SDLW Linear Regression Model

In prior analysis the SDLW model forecasts have been identified as having the strongest linear relationship with Load. Hence, a linear regression model was created that included ‘Load Last Week’ as an explanatory variable to predict load :

The average MAPE of the model forecasting a year ahead: 24.311%. - **see Appendix C** for a table of results. This is a considerable improvement in forecasting performance over the Temperature Only Linear Regression Model.

SDLW Temperature Corrected Linear Regression Model

There are forecasting errors in using only SDLW as a dependent variable because of the difference in temperature between the forecasted day and the day’s load used for the prediction which influences load. A linear regression model was created that corrected the Load Last Week prediction with the displaced temperature difference :

The average MAPE of the model forecasting a year ahead: 4.183% - **see Appendix C** for a table of results.

SDLW Weather Corrected Linear Regression Model

Strong performing linear regression models in other research include a wider range of weather variables as explanatory variables than only temperature [1]. A linear regression model was constructed with variables: the displaced Temperature-48 hours difference because of the prior identification of temperature influencing load; the product of sun duration and potential solar irradiance difference and wind speed cubed difference because of sun and wind renewable energy power generation; humidity difference because it has been identified in other research to have a relationship with load [4].

The average MAPE of the model forecasting a year ahead: 4.005% - **see Appendix C** for results table.

As this was the strongest performing model of the linear regression models analysed, it will be incorporated into the system. The variables chosen for this model were arbitrary through prior analysis and the assumption that they provide a quantitative measure of an influencers to load demand. A less arbitrary method of choosing explanatory variables was investigated in the next chapter.

**2.8 LASSO Regression Model Analysis**

LASSO (least absolute shrinkage and selection operator) regression takes a large feature dataset and penalises the magnitude of the coefficients of the features, converging them to 0 and eliminating them from the model. Forward selection regression approach to variable selection has already been investigated in other research on the Northern Ireland load dataset [2]. Ridge regression was considered but it does not reduce model complexity. A ranked list of the most important variables to the load prediction can be produced using LASSO, enabling the creation of models using a specified number of variables that produces accurate load forecasts.

LASSO regression requires tuning model parameters. Alpha determines the weight given to the magnitude of coefficients when reducing the error by adding the residual sum of squares (linear regression error). To cross validate the alpha that produces a model with minimum MAPE, the models are trained on a random subset of the training year selected and tested on a random validation subset of training year. The alpha that produces forecasts with the smallest MAPE was then used as the alpha for feature selection.

All Explanatory Variables LASSO Regression Model

All explanatory variables in the dataset in **Appendix A** were included as candidate variables for LASSO regression feature selection.

The average MAPE of the model forecasting a year ahead: 9.349% - **see Appendix C** for results table.

SDLW Weather Correction Model LASSO Regression Model

SDLW Weather Corrected Linear Regression Model was the best performing linear regression model in prior analysis. Hence, building a LASSO model that chooses a subset of explanatory variables to correct the SDLW load forecasting prediction should produce more accurate forecasts. The variables used to construct the SDLW Weather Corrected Linear Regression Model were added to the dataset as candidate variable for LASSO feature selection.

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Test** | 2012 | 2013 | 2014 | 2015 | 2016 | 2017 | **Average** | **Variables** |  |
| **Training** |  |  |  |  |  |  |  |  |  |
| 2012 | 3.598 | 3.601 | 3.559 | 3.748 | 4.122 | 5.601 | 4.038 | 23 | 1.41 |
| 2013 | 3.712 | 3.325 | 3.225 | 3.584 | 3.948 | 5.486 | 3.880 | 35 | 0.11 |
| 2014 | 3.711 | 3.369 | 3.121 | 3.459 | 3.698 | 5.065 | 3.737 | 39 | 0.06 |
| 2015 | 3.939 | 3.688 | 3.440 | 3.435 | 3.881 | 5.456 | 3.973 | 27 | 0.71 |
| 2016 | 4.202 | 3.995 | 3.689 | 3.932 | 3.584 | 4.842 | 4.041 | 42 | 0.06 |
| 2017 | 4.819 | 4.905 | 4.471 | 4.916 | 4.217 | 4.213 | 4.590 | 46 | 0.01 |

The average MAPE of the model forecasting a next year’s load using the previous year to train the model was 3.805%.

***TABLE 2.4*** *The MAPE of the LASSO regression model with variables selected and fitted with a training year predicting the load of a test year. White cells: Next year load forecast. Blue cells (RIGHT): The average MAPE of the training year predicting test years, the number of explanatory variables in the model and the alpha used for feature selection.*

|  |  |
| --- | --- |
| **Variable** | **Occurrence (%)** |
| Temperature | 100 |
| Temperature-12hrs | 100 |
| Temperature-96hrs | 100 |
| Temperature over 24hrs | 33.33 |
| Temperature over 36hrs | 33.33 |
| Sun duration\*potential solar irradiance | 33.33 |

***Table 2.2*** *The top and bottom 3 occurring explanatory variables selected by LASSO regression built on 2012-2017 training years –* ***see Appendix D for full list***

Building a uniform linear regression model for the whole dataset based on the variables selected using LASSO regression for all models constructed, **see Table 2.2,** is not optimal for producing accurate load forecasts. It may exclude important variables that correlate to other variables which are eliminated early by LASSO feature selection Therefore, an approach was needed to determine the importance of the explanatory variables for producing accurate load forecasts for the whole dataset, not individual splits of the dataset.

Ranking Explanatory Variables using LASSO

A list of explanatory variables ranked by the order of elimination from LASSO regression by increasing the alpha incrementally was used to evaluate their ‘importance’ to creating a linear regression model that produces accurate load forecasts with n-features. The whole dataset was used to fit the model. Several iterations of the process were manually observed and the increments for increasing the alpha were decreased at lower alpha values (<1) to record atomic ranks for variables of least importance

|  |  |
| --- | --- |
| **Variable** | **Alpha** |
| Humidity Difference | 6000 |
| Cloud height | 500 |
| Wind direction | 200 |
| Potential solar irradiance | 0.05 |
| Temperature over 18hrs | 0.045 |
| Temperature over 96hrs | 0.03 |

**Figure 2.7 (a)** The top 5 most important (elimination alpha is highest) and least important (elimination alpha is smallest) explanatory variables - **See appendix D** for the full list **(b)** The MAPE performance of the number of explanatory variables selected for the lasso regression model

**(b)**

**(a)**

**(b)**

Visually in **Figure 2.7 (a)** the MAPE decreases unsteadily as the number of variables increases. The MAPE improvements plateau after 12 variables are selected. **In (b),** the variables eliminated first by LASSO feature selection are temperature related which is to be expected as they are highly correlated to other temperature variables. 2nd placed ranked variable ‘Cloud height’, unlike 1st ‘Humidity Difference’, was not recognised in prior analysis as being an important explanatory variable. This indicates that there are limitations to using LASSO to minimize MAPE errors, as LASSO minimizes the mean squared error (MSE).

SDLW Lasso Variable Selection Correction Model

A regression model was created with the 12 top ranked explanatory variables. The following model was created with the 12 most important variables: Humidity Difference , Cloud height , Wind speed^2 , Wind direction , Dayofyear ), Wind speed^3 , Temperature Difference , Day , Temp-48 Difference (, Holiday\_Alternate , Yearly cycle (sine wave) , Wind speed Difference :

The average MAPE of the model forecasting a year ahead: 3.859% - **see Appendix C** for results table. This is higher than the SDLW Weather Correction LASSO Regression Model by 0.054%. However, the SDLW Weather Correction Model uses a maximum of 46 input variables, while the LASSO model only uses 12 variables. The LASSO Variable Selection SDLW Correction Model provides a good balance between model complexity and good forecasting performance.

**2.9 System Decisions Made from Analysis**

The investigative processes undertaken in analysis is the functionality that the software solution should automate. The system should provide the flexibility to change the forecasting models’ parameters and the split of the dataset on which they are built and tested on. The visualisations of the model’s performance present in analysis should be generated on the fly by the system and not require an understanding of Python visualisation libraries. Models in the system should be able to be compared statistically with other models using the MAPE metric. The date characteristics of day entries should be provided to the user as an options to highlight them to notice load deviation trends types of day have compared to other days. The regression models correcting a displacement model’s predictions should be included in the solution as these models have the strongest forecasting performance in analysis. The analysis benefited both the understanding of the author in realising the challenges of thoroughly analysing the data and characteristics of the data.

**Citations**

[1] J. Foster, X. Liu and S. McLoone, "Load forecasting techniques for power systems with high levels of unmetered renewable generation: A comparative study", IFAC-PapersOnLine, vol. 51, no. 10, pp. 109-114, 2018.

[2] J. Foster, X. Liu, and S. Mcloone, “Adaptive sliding window load forecasting,” 2017 28th Irish Signals and Systems Conference (ISSC), 2017.

[3] E. Almeshaiei and H. Soltan, "A methodology for electric power load

forecasting," Alexandria Engineering Journal, vol. 50(2), pp. 137-144,

2011.

[4] Wemcouncil.org, 2019. [Online]. Available: <http://www.wemcouncil.org/wp/wp->content/uploads/2015/07/1230\_YingChen.pdf. [Accessed: 18- Apr- 2019].

**Appendix A**

*A list of the columns included in the SONI dataset. The variables added during a data cleaning process are in bold.*

|  |  |  |  |
| --- | --- | --- | --- |
| Date | Temperature over 18hrs | Humidity | **Week** |
| Load | Temperature over 24hrs | Sun duration\*potential solar irradiance | **Quarter** |
| Temperature | Temperature over 36hrs | Binary indicator sunny day | **Hour** |
| Temperature-6hrs | Temperature over 48hrs | Binary indicator windy day | **Minute** |
| Temperature-12hrs | Temperature over 72hrs | Potential solar irradiance | **Day** |
| Temperature-18hrs | Temperature over 96hrs | Weekday | **Dayofweek** |
| Temperature-24hrs | Wind speed | Yearly cycle (sine wave) | **Dayofyear** |
| Temperature-36hrs | Wind speed^2 | Yearly cycle (cosine wave) |  |
| Temperature-48hrs | Wind speed^3 | Daily cycle (sine wave) |  |
| Temperature-72hrs | Wind direction | Daily cycle (cosine wave) |  |
| Temperature-96hrs | Cloud height | Holiday\_Alternate |  |
| Temperature over 6hrs | Sun duration | **Year** |  |
| Temperature over 12hrs | Visibility | **Month** |  |

**Appendix B**

|  |  |
| --- | --- |
| **Holidays** | **SDLW Predicted Load - Actual Load Correlation** |
| Holiday New Year's Day | 0.739 |
| Derived Christmas Day | 0.748 |
| Holiday Easter Monday | 0.755 |
| Holiday Christmas Day | 0.82 |
| Holiday Boxing Day | 0.841 |
| Derived Easter Monday | 0.846 |
| Derived Battle of the Boyne | 0.901 |
| Derived Good Friday | 0.911 |
| Holiday Christmas Day (Observed) | 0.923 |
| Holiday May Day | 0.947 |
| Holiday Boxing Day (Observed) | 0.95 |
| Derived Late Summer Bank Holiday | 0.961 |
| Derived St. Patrick's Day | 0.963 |
| Holiday Good Friday | 0.966 |
| Derived Boxing Day (Observed) | 0.966 |
| Derived Spring Bank Holiday | 0.967 |
| Derived May Day | 0.969 |
| Holiday St. Patrick's Day | 0.971 |
| Holiday Late Summer Bank Holiday | 0.979 |
| Derived St. Patrick's Day (Observed) | 0.98 |
| Holiday Spring Bank Holiday | 0.981 |
| Holiday St. Patrick's Day (Observed) | 0.981 |
| Holiday Battle of the Boyne | 0.986 |
| Derived New Year's Day | 0.988 |
| Derived Boxing Day | 0.991 |
| Derived Christmas Day (Observed) | 0.997 |
| Holiday New Year's Day (Observed) | N/A |
| Derived New Year's Day (Observed) | 0.994 |

**Appendix C**

SDLW Linear Regression Model

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Test** | 2012 | 2013 | 2014 | 2015 | 2016 | 2017 | **Average** |
| **Training** |  |  |  |  |  |  |  |
| 2012 | 4.002 | 3.834 | 3.639 | 4.107 | 4.566 | 5.838 | 4.331 |
| 2013 | 3.998 | 3.822 | 3.628 | 4.099 | 4.559 | 5.835 | 4.323 |
| 2014 | 3.979 | 3.802 | 3.595 | 4.074 | 4.531 | 5.822 | 4.301 |
| 2015 | 4.003 | 3.839 | 3.642 | 4.109 | 4.568 | 5.838 | 4.333 |
| 2016 | 4.005 | 3.850 | 3.644 | 4.110 | 4.567 | 5.838 | 4.336 |
| 2017 | 4.018 | 3.868 | 3.666 | 4.127 | 4.585 | 5.847 | 4.352 |

SDLW Temperature Corrected Linear Regression Model Results

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Test** | 2012 | 2013 | 2014 | 2015 | 2016 | 2017 | **Average** |
| **Training** |  |  |  |  |  |  |  |
| 2012 | 3.950 | 3.702 | 3.487 | 3.940 | 4.351 | 5.627 | 4.176 |
| 2013 | 4.011 | 3.706 | 3.516 | 3.918 | 4.302 | 5.500 | 4.159 |
| 2014 | 4.013 | 3.709 | 3.518 | 3.919 | 4.301 | 5.498 | 4.160 |
| 2015 | 4.131 | 3.795 | 3.612 | 3.976 | 4.331 | 5.430 | 4.213 |
| 2016 | 4.088 | 3.765 | 3.578 | 3.953 | 4.312 | 5.448 | 4.191 |
| 2017 | 4.222 | 3.871 | 3.689 | 4.037 | 4.385 | 5.415 | 4.270 |

SDLW Weather Corrected Linear Regression Model Results

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Test** | 2012 | 2013 | 2014 | 2015 | 2016 | 2017 | **Average** |
| **Training** |  |  |  |  |  |  |  |
| 2012 | 3.815 | 3.579 | 3.393 | 3.814 | 4.215 | 5.771 | 4.098 |
| 2013 | 3.905 | 3.570 | 3.479 | 3.876 | 4.439 | 6.150 | 4.237 |
| 2014 | 3.858 | 3.603 | 3.343 | 3.763 | 4.050 | 5.617 | 4.039 |
| 2015 | 4.019 | 3.678 | 3.459 | 3.803 | 4.125 | 5.893 | 4.163 |
| 2016 | 4.411 | 4.305 | 3.915 | 4.243 | 3.832 | 5.080 | 4.298 |
| 2017 | 4.765 | 4.945 | 4.476 | 4.973 | 4.167 | 4.695 | 4.670 |

SDLW Weather Correction Model LASSO Regression Model

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Test** | 2012 | 2013 | 2014 | 2015 | 2016 | 2017 | **Average** | **Variables** |  |
| **Training** |  |  |  |  |  |  |  |  |  |
| 2012 | 8.385 | 9.093 | 9.732 | 9.962 | 11.022 | 14.341 | 10.422 | 42 | 0.02 |
| 2013 | 9.309 | 8.433 | 8.855 | 9.342 | 10.135 | 12.899 | 9.829 | 42 | 0.04 |
| 2014 | 9.545 | 8.551 | 8.481 | 8.905 | 9.433 | 11.984 | 9.483 | 43 | 0.02 |
| 2015 | 9.351 | 8.810 | 8.690 | 8.558 | 9.279 | 11.644 | 9.389 | 43 | 0.02 |
| 2016 | 10.022 | 9.207 | 8.714 | 8.667 | 8.620 | 10.615 | 9.308 | 41 | 0.12 |
| 2017 | 11.488 | 10.423 | 9.705 | 9.670 | 9.104 | 9.303 | 9.949 | 43 | 0.02 |

**Appendix D**

Table 2.5 Full

|  |  |
| --- | --- |
| **Variable** | **Occurrence (%)** |
| Temperature | 100 |
| Temperature-12hrs | 100 |
| Temperature-96hrs | 100 |
| Wind speed^3 | 100 |
| Wind direction | 100 |
| Cloud height | 100 |
| Visibility | 100 |
| Humidity | 100 |
| Yearly cycle (sine wave) | 100 |
| Holiday\_Alternate | 100 |
| Hour | 100 |
| Day | 100 |
| Dayofyear | 100 |
| Temperature Difference | 100 |
| Temp-48 Difference | 100 |
| Wind speed Difference | 100 |
| Humidity Difference | 100 |
| Temperature-24hrs | 83.33333 |
| Temperature-48hrs | 83.33333 |
| Temperature over 6hrs | 83.33333 |
| Wind speed^2 | 83.33333 |
| Weekday | 83.33333 |
| Week | 83.33333 |
| Temperature-18hrs | 83.33333 |
| Temperature-72hrs | 83.33333 |
| Wind speed | 83.33333 |
| Dayofweek | 83.33333 |
| Sun duration\*potential solar irradiance Difference | 83.33333 |
| Temperature-36hrs | 66.66667 |
| Sun duration | 66.66667 |
| Binary indicator windy day | 66.66667 |
| Yearly cycle (cosine wave) | 66.66667 |
| Daily cycle (sine wave) | 66.66667 |
| Daily cycle (cosine wave) | 66.66667 |
| Temperature-6hrs | 66.66667 |
| Quarter | 66.66667 |
| Temperature over 48hrs | 50 |
| Minute | 50 |
| Binary indicator sunny day | 50 |
| Temperature over 12hrs | 33.33333 |
| Temperature over 96hrs | 33.33333 |
| Temperature over 18hrs | 33.33333 |
| Temperature over 24hrs | 33.33333 |
| Temperature over 36hrs | 33.33333 |
| Sun duration\*potential solar irradiance | 33.33333 |
| Potential solar irradiance | 16.66667 |
| Month | 16.66667 |

Figure 2.7 (a) Full

|  |  |  |  |
| --- | --- | --- | --- |
| **Variable** | **Alpha** | **MAPE** | **MAPE Difference** |
| Humidity Difference | 6000 | 4.753937 | 0 |
| Cloud height | 500 | 4.748248 | -0.00569 |
| Wind direction | 200 | 4.750921 | 0.002674 |
| Wind speed^2 | 200 | 4.750921 | 0 |
| Dayofyear | 100 | 4.629771 | -0.12115 |
| Wind speed^3 | 100 | 4.629771 | 0 |
| Temperature Difference | 60 | 4.630165 | 0.000394 |
| Day | 50 | 4.520303 | -0.10986 |
| Temp-48 Difference | 40 | 4.536783 | 0.016481 |
| Holiday\_Alternate | 6.5 | 4.497308 | -0.03948 |
| Yearly cycle (sine wave) | 6.5 | 4.497308 | 0 |
| Wind speed Difference | 5 | 4.297932 | -0.19938 |
| Dayofweek | 4.5 | 4.294197 | -0.00374 |
| Temperature-24hrs | 4 | 4.298449 | 0.004252 |
| Temperature-96hrs | 4 | 4.298449 | 0 |
| Hour | 3.5 | 4.275841 | -0.02261 |
| Humidity | 2 | 4.27848 | 0.002639 |
| Temperature | 2 | 4.27848 | 0 |
| Weekday | 2 | 4.27848 | 0 |
| Year | 2 | 4.27848 | 0 |
| Temperature-36hrs | 1.5 | 4.274248 | -0.00423 |
| Temperature-18hrs | 1 | 4.274445 | 0.000197 |
| Temperature-48hrs | 0.9 | 4.274261 | -0.00018 |
| Temperature-72hrs | 0.85 | 4.274489 | 0.000229 |
| Sun duration\*potential solar irradiance Difference | 0.8 | 4.274345 | -0.00014 |
| Week | 0.7 | 4.228296 | -0.04605 |
| Visibility | 0.65 | 4.228266 | -3E-05 |
| Quarter | 0.6 | 4.228287 | 2.12E-05 |
| Binary indicator windy day | 0.45 | 4.225432 | -0.00285 |
| Temperature over 6hrs | 0.45 | 4.225432 | 0 |
| Binary indicator sunny day | 0.4 | 4.221397 | -0.00404 |
| Temperature-6hrs | 0.4 | 4.221397 | 0 |
| Wind speed | 0.35 | 4.221207 | -0.00019 |
| Daily cycle (sine wave) | 0.3 | 4.221079 | -0.00013 |
| Sun duration | 0.3 | 4.221079 | 0 |
| Temperature over 48hrs | 0.3 | 4.221079 | 0 |
| Daily cycle (cosine wave) | 0.2 | 4.21565 | -0.00543 |
| Yearly cycle (cosine wave) | 0.2 | 4.21565 | 0 |
| Month | 0.1 | 4.213028 | -0.00262 |
| Sun duration\*potential solar irradiance | 0.1 | 4.213028 | 0 |
| Temperature over 12hrs | 0.1 | 4.213028 | 0 |
| Temperature over 24hrs | 0.1 | 4.213028 | 0 |
| Temperature-12hrs | 0.1 | 4.213028 | 0 |
| Minute | 0.05 | 4.208086 | -0.00494 |
| Potential solar irradiance | 0.05 | 4.208086 | 0 |
| Temperature over 18hrs | 0.045 | 4.206988 | -0.0011 |
| Temperature over 96hrs | 0.03 | 4.207191 | 0.000204 |
| Temperature over 36hrs | 0.025 | 4.207225 | 3.39E-05 |
| Temperature over 72hrs | 0.025 | 4.207225 | 0 |