DSO 522: APPLIED TIME SERIES ANALYSIS FOR FORECASTING

PROJECT REPORT FOR ELECTRICITY LOAD FORECASTING FOR DAY-AHEAD ENERGY MARKETS

GROUP 11

SHRINGAR SHARAN
ASTHA SRIVASTAVA
RAJAT GAUR
VARUN GUPTA

ABSTRACT

This report explores different techniques such as Naïve Forecasting, Holt-Winter's Exponential Smoothing and machine learning techniques for forecasting electricity load for Southern California Edison (SCE), which is a primary energy supply in Southern California. The Gradient Boosting Algorithm fetched the best results with lowest forecast error and puts forth a robust model. We then analyse the forecasts based on the results of the proposed methodology with the objective of minimizing the loss from the inaccurate forecasts of electricity demand. Lastly, we highlight some of the insights elicited by the analysis and elaborate on them further to provide actionable business recommendations in terms of bidding strategies for day-ahead energy markets that SCE participates in. These strategies help to mitigate risks associated with fluctuations in electricity demand.

INTRODUCTION

Electricity demand forecasting is pivotal not only for the energy companies but also for regulators and electricity markets. Energy companies utilize these forecasts to balance supply and demand on their grid. Load forecasts are crucial for ensuring that energy companies meet the demand of their customers and that customers are paying the lowest prices for electricity. The rise in competition in energy markets and the complex nature of load forecasting have bolstered the evolution of various price forecasting techniques and researches over the last two decades.

BUSINESS PROBLEM

Southern California Edison (SCE), the largest subsidiary of Edison International, is the primary electricity supply company for much of Southern California. It is one of the largest power utilities that serves 15 million Californians with electricity across a service territory of approximately 50,000 square miles. It participates in the California deregulated energy market (CAISO) and buys and sells everyday energy in the market in order to meet the demand of its customers. Having the right load and price forecast is imperative to ensure lowest prices for its customers. By buying too much unused energy, SCE must endure loss of revenue and face potential penalties by state regulators. On the other hand, by not buying enough, they risk having to buy energy in 'real-time' at a higher price. Most of the forecasted load is bought 16 hours before the start of the 'flow date' in the 'day-ahead' energy market. The flow date is defined as the day the energy is consumed by the customers. The rest of the load is bought in the real-time energy market in order to capture the real-time fluxes of demand.

OBJECTIVE

We aim to produce accurate forecasts for the day-ahead electricity demand so as to minimize loss faced by SCE due to over or under prediction. We also aim to estimate the loss company has to endure due to over prediction or under prediction of load. We also incorporated external factors such as weather forecasts, information on holidays, day of the week information and others in our model to check whether it improves the accuracy. The objective is to build upon a holistic forecasting method which is both accurate and would enable SCE to minimize losses by implementing actionable insights elicited by our analysis.

Forecast Horizon: In terms of importance, the accuracy of the 16-hour forecast is the most crucial because that is when SCE buys most of its energy. The 16-hour forecast is 16 hours before the start of the day they buy energy. SCE participates in the day-ahead market every day and must run this forecast every morning. Since the 16-hour forecast is most important, we have focused more of the analysis on this. Moreover, SCE wants us to predict in hourly intervals as it's the most accurate among all other forecast horizons considered.

Peak Hours: Peak hours are defined as the hours of the day with the highest load. Forecasts for peak hours are very important because it's crucial for SCE to meet the demand during these peak hours. Using domain knowledge, we define peak hours to be usually between 6PM to 9PM.

Through our analysis, we expect to understand the dynamics of electricity supply and demand and how it affects both prices and load. We also want to accurately forecast these series and find additional factors which would help us in our endeavour so as to be able to share our analysis with the company.

DATA DESCRIPTION

We have the following data available from SCE:

Historical Load Series: The historical load data is available in hourly intervals for each day starting from 2014 till 2019. The load is measured in MWhs.

Real-Time and Day-Ahead Energy Market Price Series: The historical prices for real-time markets and day-ahead energy markets are available for each day starting from 2017 to 2019 and is measured in USD. This data has been used to estimate the cost of under-prediction and over-prediction associated with Load forecasting.

Temperature and Other External Data: We have actual temperature data available from 2014 that SCE tracks for its 5 sub regions in Southern California, namely,

- CQT: LA Downtown/USC Area
- RIV: Riverside
- LAX: LA Airport
- TRM: Jacqueline Cochran Regional Airport
- WJF: General Wm J Fox Airfield Airport

There are both the high and low temperatures for these sub regions. We have also incorporated other external factors from publicly available sources such as those related to weather such as **Relative Humidity, Visibility and Dew Point Temperature** and those related to **Day of the week and Holiday information**. These datasets are available across the five sub-regions mentioned above.

OVERVIEW OF ANALYSIS PERFORMED

We started off by collecting the internal load, pricing and temperature data and external data such as weather and holiday data. Next, we moved on to explore the data and make visualizations to understand the nature of the time series data we are dealing with. The following summarizes the steps entailed to analyze the data:

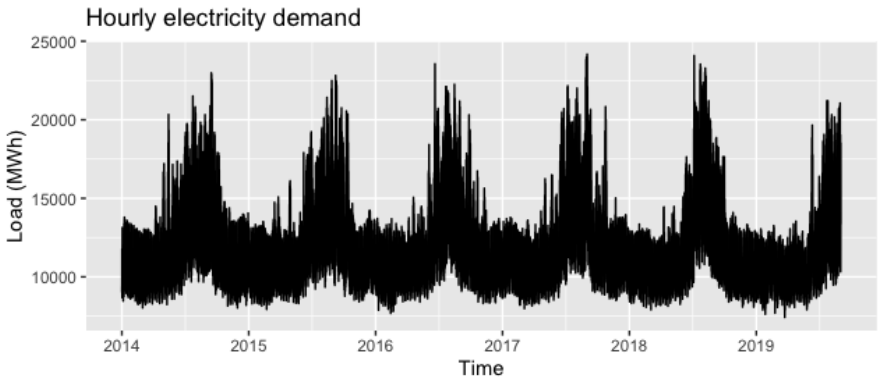
1. **Exploratory Data Analysis:** We explored the Load and Temperature data and made visualizations to aid in future analysis.
2. **Data Wrangling:** This step involved cleaning the data and imputing those records which have missing values.
3. **Constructing Indicator Variables:** Created new variables from existing data
4. **Forecasting Techniques:** We explored different forecasting methodologies to find the best one which fits our data well and helps us to build a forecasting model with high accuracy.
5. **Results and Highlights:** We highlight some of the important insights elicited form the analysis.
6. **Business Insights and Recommendations:** Lastly, we provide actionable business recommendations in terms of bidding strategies.

EXPLORATORY DATA ANALYSIS

We started off by analyzing and visualizing the available load and external variables data. We decomposed the time series data to identify trends, seasonality and other patterns present. Here are the findings from the exploratory data analysis that we performed:

Electricity Load Data

- From the plot of decomposed data, the trend seems to be slightly decreasing in nature (Appendix 1).
- The electricity load data exhibits three levels of seasonality, namely, **daily, weekly and annual** (Appendix 1).
- The distribution of load data is highly right skewed with some very high values mostly during the holidays (Appendix 2).



Departed Load in March 2019

We observe a sudden dip in electricity demand in March 2019. After enquiry, we find that this anomalous behavior is not attributed to a decrease in overall consumer demand but to customers who left SCE because of Community Choice Aggregation (CCA) created by Pasadena. CCAs act as a local community entity that serve load for certain cities. Many cities in Southern California have their own CCAs. However, the CCA Load departure data is an estimate based on estimates, which implies that SCE has an approximate idea when CCAs will depart, about how many Service Areas (SA) will

participate, how the billing cycles are staggered, and approximately how much Load is expected to decrease.

Participants have the option to 'Opt-in' or 'Opt-out' of the program at different times irrespective of the original transition. For instance, for a participant who opts-out in August 2019 after the original departure in May 2019, the cumulative net opt-in/opt-out Load would then be added to the CCA departure Load. When a participant opts-in or opts-out, their Load is added or removed depending on their billing cycle. About 1% of the 10% commercial customers that opted-out in March 2019 "opted back in" in June 2019. Estimating for the whole year, about 9% of the load departed due to CCAs in 2019.

Temperature Data

- Hourly mean temperature exhibits annual seasonality (Appendix 3).
- It is found to be highly correlated with Load data with a correlation coefficient of around 0.3818 for the year 2018 and 0.298 for the entire series from 2014 to 2019 (Appendix 4). We observe some non-linear relationship between load and temperature.

DATA PREPARATION

Data Cleaning

To make our data suitable for modeling, we performed data cleaning and prepared the data. The steps entailed are as follows:

- For load and temperature data, we imputed the missing values with local averages when local maximum and minimum temperatures were not expected and there were few (8 or less consecutive) missing data points.
- In temperature data, larger chunks of data were filled with temperature averages from the same hour when stratified by month.
- In load data, larger chunks of data were filled with load information from a similar day's hourly data from the same day of the week, with a similar temperature profile during the same time of year

Constructing Indicator Variables

We created other variables using existing data we had (Appendix 5). The steps taken are:

- Included Day of the Week information from Monday to Sundays.
- Divided hourly timestamps into day, month and year variables by encoding with integers.
- Created variables for hourly mean temperatures from high and low hourly temperature data. Also created hourly humidity and visibility variables from respective data for five different sub-regions in SCE territory in Southern California, namely, CQT, RIV, LAX, TRM, WJF.
- Created lag variables of 48 hours, 72 hours, 1 week, and 1 year.

FORECASTING METHODOLGIES

Based on the initial analysis of electricity demand, we observed that noise, level, trend, and seasonality were present in the electricity load data. Moreover, three levels of seasonality are present within the data, namely, daily, weekly and annual. Naïve Forecasting served as the baseline model for setting forecast performance targets. Some of the other methods we tried are Classical Time series Decomposition, Holt-Winter's Exponential Smoothing, Holt's Additive, Holt's Multiplicative models and Machine Learning algorithms such as Gradient Boosting. We tried Holt's Additive and Multiplicative models since they are known to handle long-term additive and multiplicative trend in the data respectively. Holt-Winter's model is known to handle data with additive or multiplicative trend or seasonality. We used MAPE as the evaluation metric and found that Gradient Boosting gives the best results. Because of the presence of multiple levels of seasonalities, we only used methods which would help us account for these multiple levels of seasonalities.

We were unable to use ARIMA as it does not address multiple levels of seasonalities present in the data. Moreover, linear regression models only explore linear relationship between the predictors and the target variable whereas there might be non-linear patterns in the data which could not be captured by this model. Hence, we opted for a model such as Gradient Boosting which would explore the linear as well as non-linear relationships within the lagged versions of the data itself and also of the other indicator variables.

The collected data was partitioned into training data and testing to evaluate the performance of these forecasting techniques. We tried different combinations of training and validation data for each of these methods. We present results for some of these combinations. Moreover, we have used **rolling forward** predictions for each of these methods using step size as 24 hours (1 day) since we need to predict electricity demand each day for the next day.

Forecast Horizon: The forecast horizon is 40 hours since SCE buys energy in the day-ahead electricity market 16 hours before the flow-date so we need to forecast demand for the entire next day (24 hours). Hence, we make hourly predictions till 40 hours ahead but calculate the error only for the last 24 hours which comprise of next day load predictions.

Training and Validation Period: We have taken three years' worth of training data and one year as the length of the validation period with rolling forward predictions for each of the 365 days. We try out different training and validation periods.

Tabular results of the predictions and MAPE for each forecasting method have been attached separately. Since we are predicting in hourly data points, the results files contained too many data points to be able to include the results here.

Naïve Forecasting

We implemented Naïve Forecasting method and evaluated it twice once for 365 days of the year 2017 and the other for the year 2018. For the former, we selected the period from 2014 to 2016 (3 years) as the training period, and 2017 as the validation period whereas for the latter, we selected 2015 to 2017 (3 years) as our training period and 2018 as the validation period. The forecast horizon in each case was 40 hours. The MAPE for the former was 15.32% whereas for the latter it was 14.85% for the validation sets. This served as baseline for the rest of the models. We discarded those models which

has errors greater than those of the Naïve Forecasting approach. Plot of observed vs predicted values for 1 week of naïve forecasts are in Appendix 6.

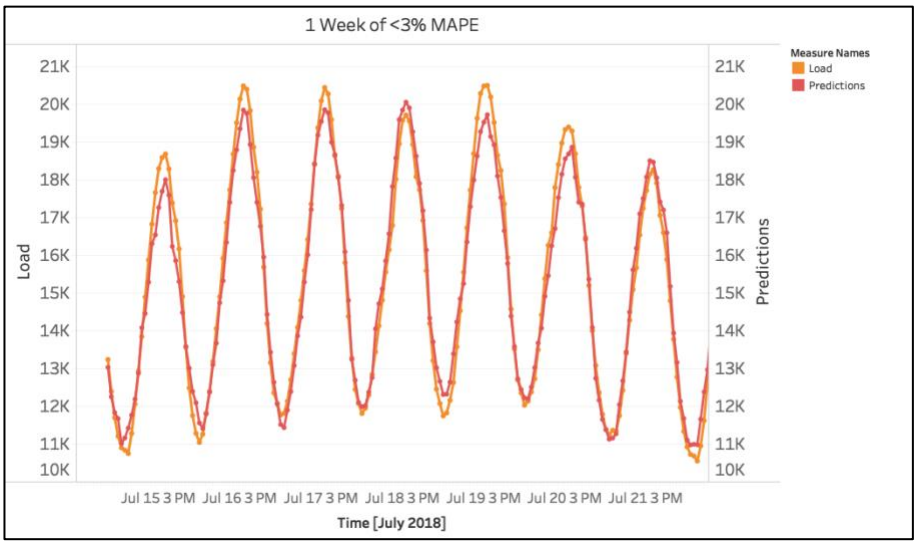
Holt-Winter's Exponential Smoothing

Before implementing this method, we checked whether the trend in the data over all years and for individual years is significant or not, we concluded that it was indeed significant. The training data was taken from 2015 to 2017 while the validation period was entire 2018. While implementing Holt-Winter's, we allowed R to generate the optimal type for noise, trend, and seasonality and found that multiplicative trend, additive noise and no seasonality. We can clearly see the problem with Holt-Winter's that it was unable to address the multiple level seasonality adequately and hence we discard this method for predictions.

Gradient Boosting Algorithm using XGBoost

All of the other models tried above do not account for multiple levels of seasonalities present in the load data. This algorithm gave the best results of all the forecasting methods we tried so far and is also robust. We used 13 independent variables and the dependent variable was Load. For the first iteration, we used 2 years of training data from 2016 to 2018 whereas the forecast horizon was 40 hours and the validation period was 2019 using rolling forward prediction. We repeated this prediction for total of 3 times by shifting the training period by one year and taking the validation period as 2017, 2018 and 2019 one by one. Out of all the variables, **'Lag load 48 hours', 'Lag load 1 week', 'Lag load 1 year' and 'Mean Temperature'** were found to be the most important as highlighted in the feature importance plot in Appendix 7.

This proves that Load series exhibits high autocorrelation with its 2 days, 1 week and 1 year lagged values. Moreover, this model captures the non-linear relationships between the predictor and response variables. The summary of Gradient Boosting model, its parameters and predictor variables can be found in Appendix 8.



The above plot depicts how closely the predicted values from Gradient Boosting follow the observed Load values. We also analyzed the residuals of the forecast (Appendix 8) and found that some

forecasts deviate from normal. ACF plot further shows that some trend or seasonality might be left which we could further incorporate in our model. This could one of the points for further improvement. Except for one or two places, residuals seem to have constant spread.

RESULTS AND ANALYSIS HIGHLIGHTS

We have used Mean Absolute Percentage Error (MAPE) as the evaluation metric for the forecasting techniques we tried. We calculate MAPE for each of the 365 days of the year and then take the average of those values to get a final MAPE for these forecasting methods. Naïve forecasting served as a baseline model to set performance targets. We have tried many different methods but are only discussing the ones which have MAPE lower than that from Naïve forecasts. The following is a summary of the some of the methods we implemented for forecasting future Load values and their corresponding MAPE:

Model Name	Average MAPE (2018)
Gradient Boosting (XGBoost)	**3.78%**
Classical Time Series Decomposition	10.06%
Naïve Forecasting	14.85%
Holt Winter's Exponential Smoothing	14.75%
Holt's (Additive)	45.22%
Holt's (Multiplicative)	45.22%

For interpretation of results and further analysis, we are only considering results from the Gradient Boosting model as the final prediction for each of the years 2017, 2018 and 2019. The average MAPE for each of these years is **3.89%, 3.78% and 4.49%** respectively (Appendix 9).

We utilize **day of the week** and **peak hours** to draw insights form the model results. Below are some of them:

- Forecast error (MAPE) varies greatly with the **day of the week**. For instance, we found that MAPE on average is highest on **Sundays and Mondays**. On inspection we found the reason behind this is that lagged loads for Mondays typically tend to be from weekends since fluctuations over weekends tend to be higher. We have included a heat map for MAPE over day of the week and average MAPE over all days of the week in the Appendix 10. Also, average MAPE for holidays (6.8%) tend to be much higher than that for non-holidays (3.7%).

- Highest errors are caused during **peak hours** as fluctuations are higher during this period. From the plot of observed vs predicted values in Appendix 11, it is evident that the troughs are captured well by the model but not the peaks.

- Autocorrelation is highest with **lagged load of 2 days (0.84), lagged load of 1 week (0.82) and lagged load of 1 year (0.81)**.

- We plotted the distribution of MAPE for hourly data over the years 2017, 2018 and 2019 and included the plot in the Appendix 12. The median MAPE for all the years is 3.35% which is quite close to the business target of 2-3% (Appendix 13). The threshold for MAPE was

disclosed by SCE. From the plot in Appendix 12, we can see that **MAPE is highest during summers for both the years 2018 and 2019**. This can be possibly attributed to the usual high fluctuations in the summers and also to the contribution of the customers who opted back in June 2019 after opting out in March 2019.

- We also note that 2019 has the highest MAPE (4.49%) due to two possible reasons:
  - CCA Departed load in March 2019 as explained above
  - Fewer data points in 2019. We have data only till September 1, 2019.

BUSINESS INSIGHTS AND RECOMMENDATIONS

Electricity demand forecasting is imperative for the producers' production arrangement and optimal scheduling of hydro energy production. Energy companies extensively utilize load forecasts to formalize bidding strategies and buy optimal load to ensure lowest prices for customers. If SCE over-predicts and buys more load, they can face loss of revenue and potential penalties imposed by the state regulators whereas if they under-predict and buy less energy than actual demand, they risk having to buy energy in 'real time' at a higher price.

Cost Analysis of Under/Over Forecasting of Load

We hereby present a cost analysis of the case when we are under/over predicting load values. To accomplish this, we first looked at the percentage of days we have overpredicted and underpredicted during the entire year for each of the years 2017, 2018 and 2019 by comparing the forecast with the actual data. The table below reveals this information. Moreover, we looked at the percentage of days wherein we achieved exact prediction during peak hours and the percentage of days wherein we predicted within between +1 and -1 standard deviations of Load values during peak hours. We were able to achieve this 88.52% of the time for 2019.

Under/Over Prediction

Year	% Days Under Predicted	% Days Over Predicted
2017	46.98%	53.02%
2018	45.76%	54.24%
2019	45.50%	54.50%

Peak Hour Prediction

Year	% Days Peak Hour Exact Prediction	% Days Peak Hour +/-1 Prediction
2017	50.27%	81.04%
2018	54.79%	90.13%
2019	52.86%	**88.52%**

We then moved onto calculate total base cumulative cost of electricity demand (ideal case considering exact prediction with no forecast error). For this, we looked at the average day-ahead energy prices for each of the years 2017, 2018 and 2019 and calculated the total cost by using the below formula:

Total Cumulative Cost for year **t** = Average Day-Ahead Price for **t** * Total demand for **t**

To estimate the **cost of penalty** for under-predicting load (due to higher prices in real-time energy market), we took the difference in average price per unit of real-time and day-ahead electricity and multiplied that with the amount of load underpredicted over the entire year. For over-prediction of load, we consulted SCE stakeholders and leveraged their domain knowledge in this regard. We hereby provide cost estimates considering over and under predictions for each of the three years of 2017, 2018 and 2019 (Appendix 14).

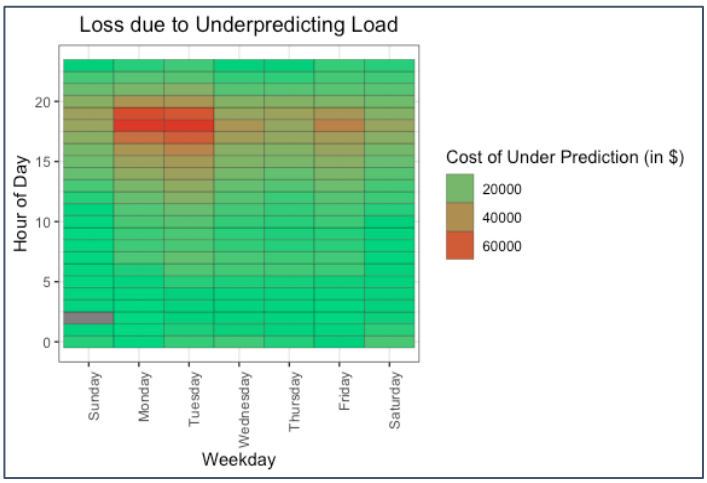
Cost Comparison

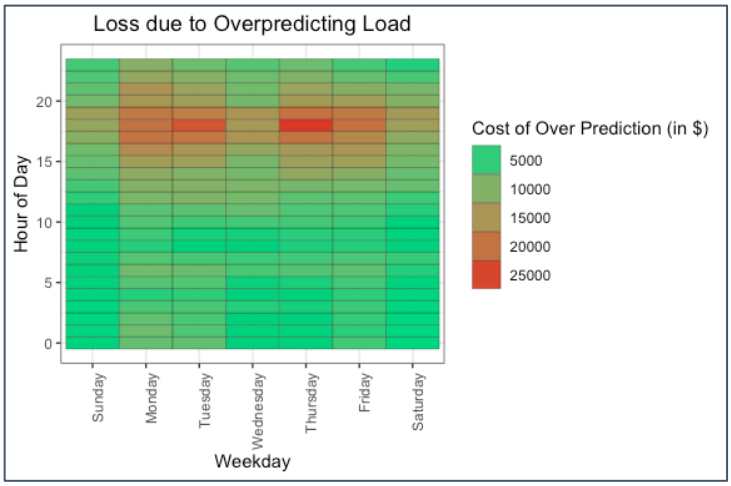
Year	Baseline Forecast (million $)	Selected model forecast (million $)	Baseline Cost/Day (million $)	Selected Model Cost/Day (million $)
2017	6268.72	5,718.45	17.17	15.67
2018	5348.59	4,948.22	14.65	13.56
2019	2681.03	2,406.73	10.99	9.86

Business Recommendations

We propose actionable business recommendations which are based on the cost estimation explained above. We also found that day of the week and peak hour predictions provided most insightful business implications and hence we have based our business recommendations on these two factors.

The below figure depicts a heat map of the estimate of the loss suffered by SCE due to over and under forecasting of load.





We hereby propose bidding strategies for SCE to enable them to minimize losses in the day-ahead electricity market specifically considering the forecasted load during the peak hours the prices in real-time energy market (Appendix 15) for each day of the week.

- **Sunday:** MAPE is usually observed to be high on Sundays due to under-prediction. We observe higher demand at night possibly due to the weekend effect. However, the price in real-time energy market (Appendix 15) is usually low. Hence, we don't observe much loss on this day. So, it's preferable to buy energy is the ballpark of the forecasted load value.

- **Monday:** MAPE is highest on Mondays and load is typically under-forecasted during peak hours of 5PM – 9PM as depicted in the figure above. Hence, the loss due to under-forecasting is quite high. Moreover, real-time prices are also observed to be high during this time. Hence, it is advisable to not underbuy as being slightly on the higher end won't possibly lead to significant loss.

- **Tuesday:** Peak hour duration is highest during this time (4PM – 8PM). We also observe highest real-time prices on Tuesdays and hence the cost of under-forecasting is quite high. It's preferable to not underbuy during this time as having some extra energy on hand won't possibly lead to much loss.

- **Wednesday:** Demand is significantly lower than other days and hence loss due to under or over prediction is not much. It's possibly safe to buy energy in the ballpark of the estimated value.

- **Thursday:** We observe significant loss due to over-prediction during peak hours as demand is typically lower on this day. It would be advisable to slightly underbuy energy for this day and adjust strategy over time.

- **Friday:** Fridays tend to be typically overpredicted possibly due to the weekend effect. It's preferable to buy at or lesser than the forecasted demand for this day.

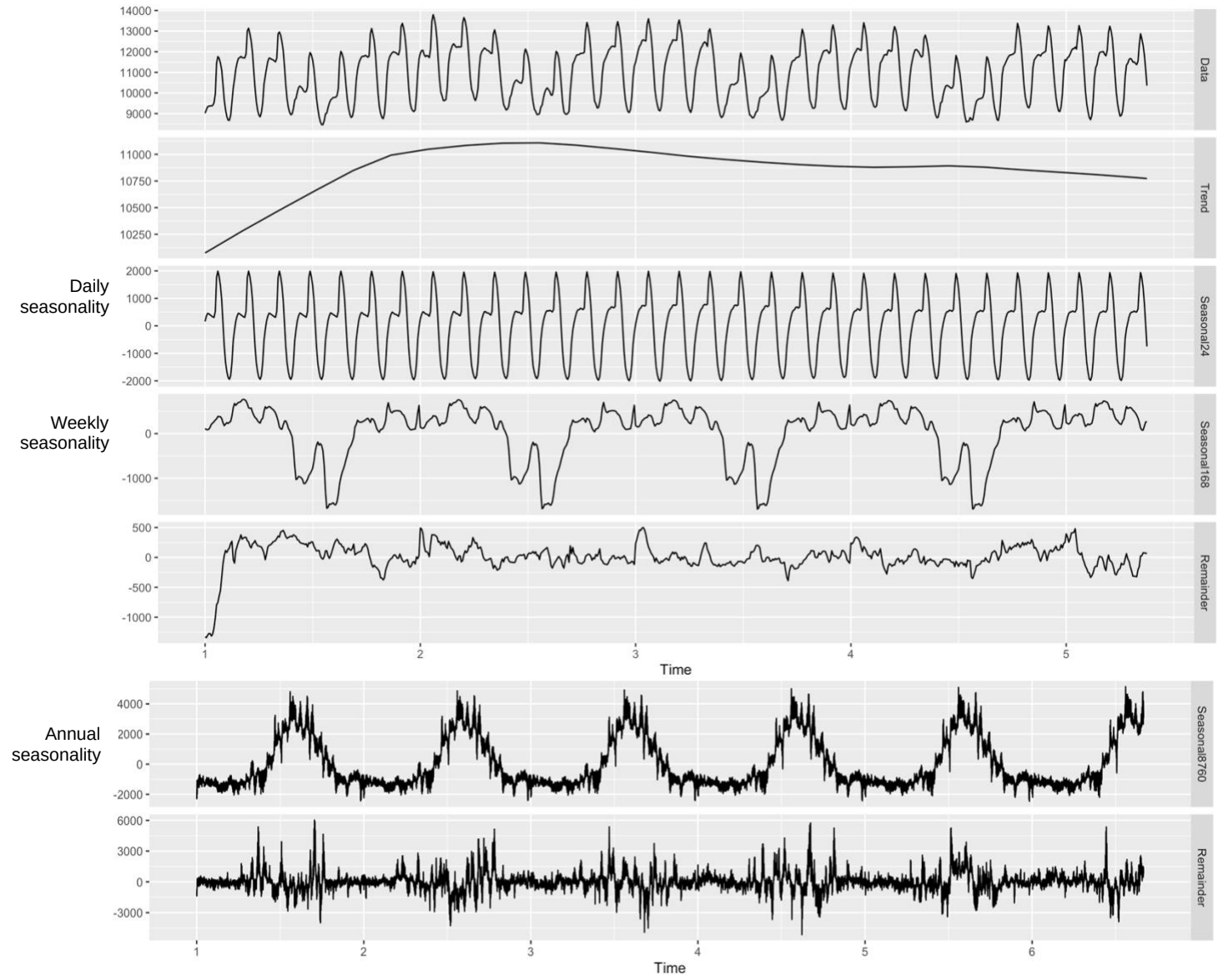
- **Saturday:** Saturdays do not exhibit unusual behavior and demand is typically lower on this day. Hence, it's advisable to buy around the estimate of the forecasted load.

CONCLUSION

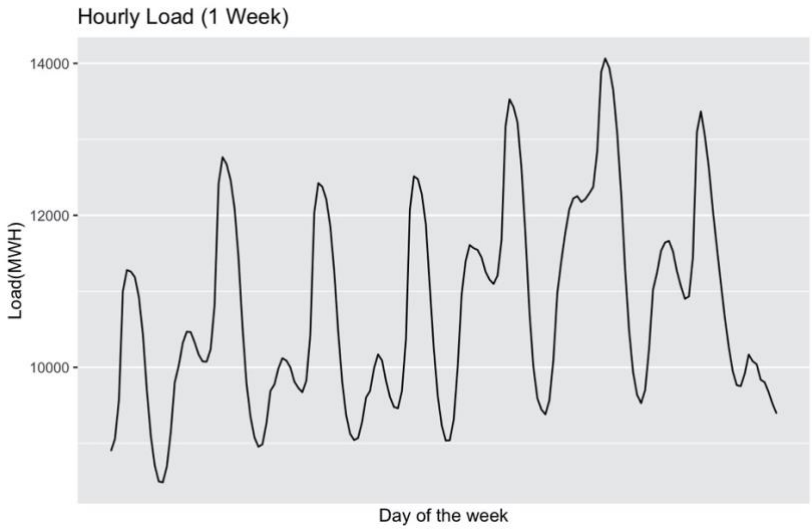
This report investigated the efficacy of some of the forecasting methodologies such as Naïve Forecasting, Holt-Winter's Exponential smoothing and Gradient Boosting Algorithm for forecasting electricity demand for Southern California Edison. Accurate forecasts are imperative for them to provide electricity to their customers at lowest prices and also to minimize losses from under or over predictions of load values. From our analysis, we conclude that Gradient Boosting algorithm performs the best over all periods of data. Our forecasts using this proposed method are only off by 3.78% from actual values for 2018 and 4.49% for 2019. Moreover, we identified that forecasts vary greatly by day of the week and peak hours, hence we need to take this into account. It is essential to address these points and come up with bidding strategies for this period so as to minimize loss. Sundays and Mondays usually tend to exhibit higher forecast error, hence we proposed buying strategies in terms of overbuying or underbuying from forecasted demand. It is also advisable to be careful around Holidays as demand fluctuates more during this period.

APPENDIX

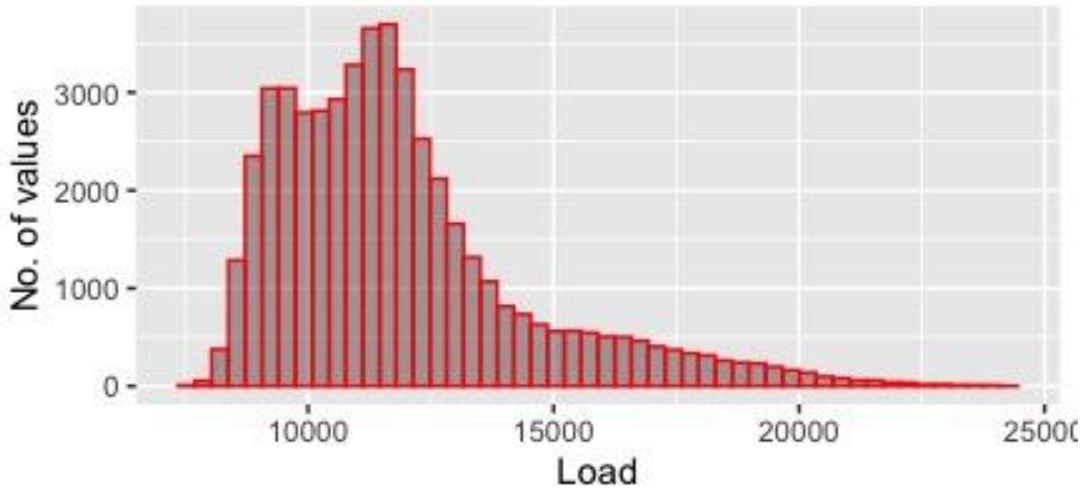
Appendix 1: Decomposition of Load data into trend, 3 levels of seasonalities and noise



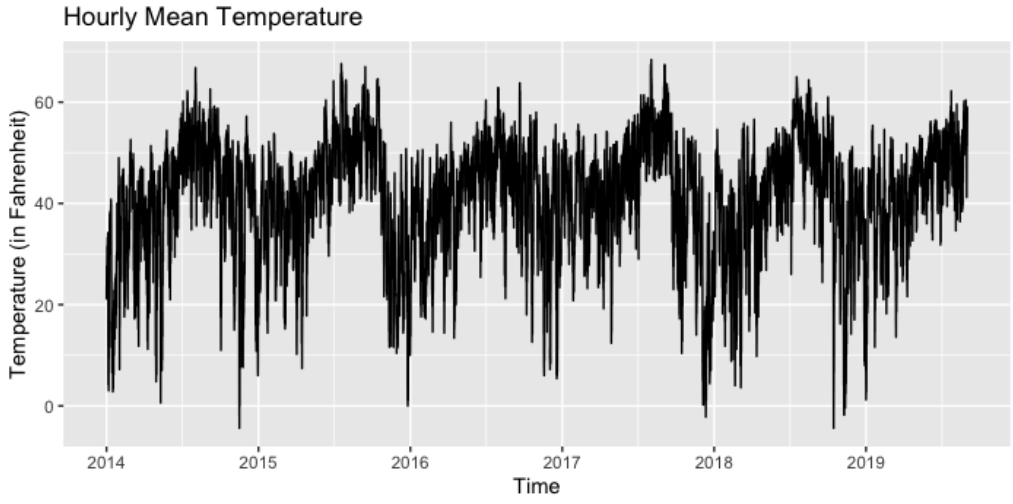
The above figure shows the decomposed Load data into trend and seasonal components. Besides the annual seasonal, the Load data also exhibits daily and annual seasonalities.



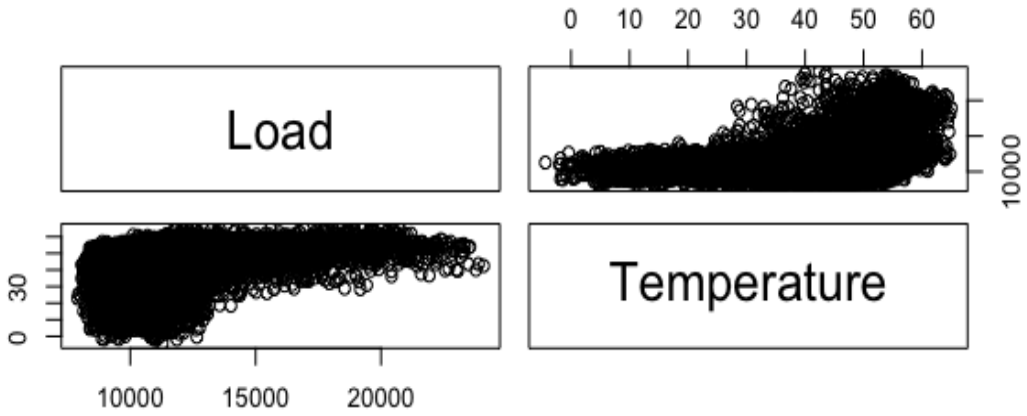
Appendix 2: Load distribution



Appendix 3: Time plot of Hourly Mean Temperature



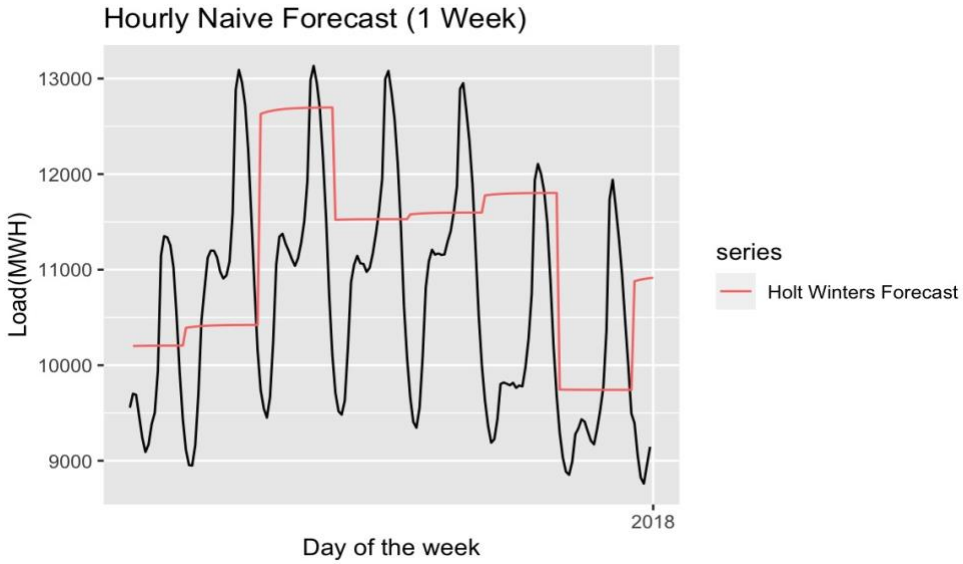
Appendix 4: Correlation between load and mean temperature for 2018



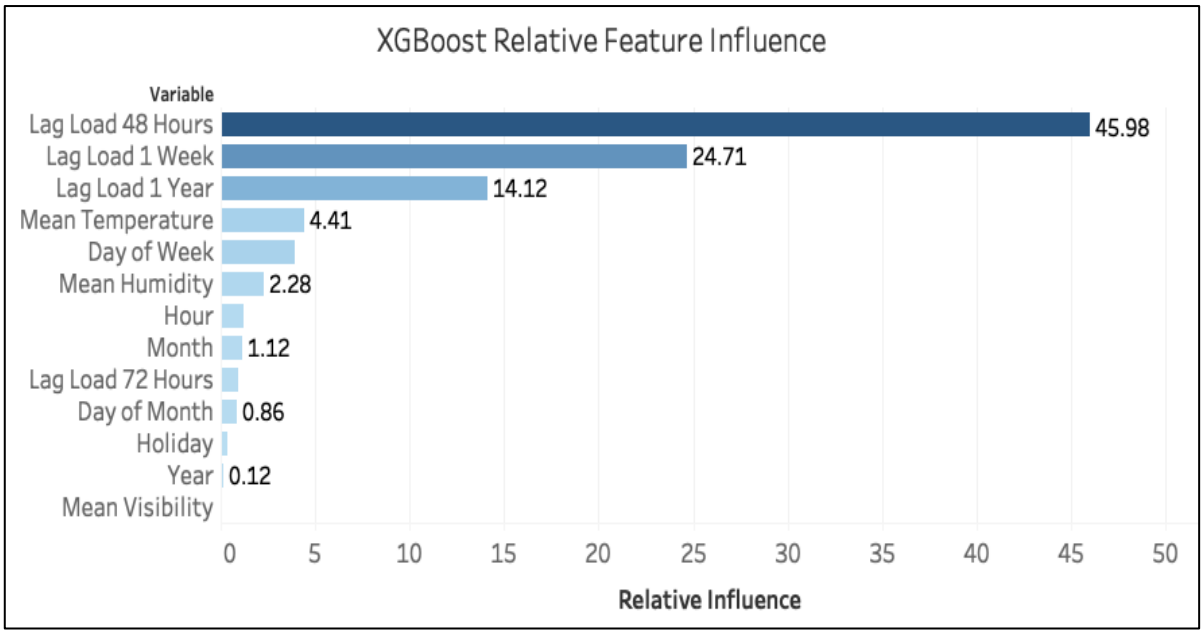
Appendix 5: Screenshot of final cleaned data with external variables used for modelling

Date	weekday	month	hour	year	day	Holiday	HourlyDewPc	HourlyRelati	HourlyVisibil	HourlyDewPc	HourlyRelati	HourlyVisibil	HourlyDewPc	HourlyRelati	HourlyVisibil	HourlyDewPc	HourlyRelati	HourlyVisibil	HourlyDewPc	HourlyRelat
01/01/14 8:00	2	1	8	2014	1	TRUE	46	90	46	10	47	10	25	66	25	14	32	14	40	83
01/01/14 9:00	2	1	9	2014	1	TRUE	42	66	42	13	39	13	26	53	26	23	26	23	42	71
01/01/14 10:00	2	1	10	2014	1	TRUE	42	58	42	13	29	13	28	43	28	22	19	22	40	51
01/01/14 11:00	2	1	11	2014	1	TRUE	37	40	37	13	22	13	29	39	29	24	16	24	38	40
01/01/14 12:00	2	1	12	2014	1	TRUE	43	50	43	11	16	11	27	36	27	25	15	25	34	30
01/01/14 13:00	2	1	13	2014	1	TRUE	51	72	51	9	13	9	19	30	19	28	17	28	38	30
01/01/14 14:00	2	1	14	2014	1	TRUE	51	72	51	10	12	10	25	25	25	29	16	29	46	49
01/01/14 15:00	2	1	15	2014	1	TRUE	49	62	49	13	13	13	20	25	20	30	17	30	48	51
01/01/14 16:00	2	1	16	2014	1	TRUE	49	62	49	14	14	14	22	31	22	30	16	30	49	53
01/01/14 17:00	2	1	17	2014	1	TRUE	50	72	50	16	21	16	22	30	22	29	19	29	50	65
01/01/14 18:00	2	1	18	2014	1	TRUE	53	84	53	15	27	15	24	45	24	31	22	31	43	58
01/01/14 19:00	2	1	19	2014	1	TRUE	53	84	53	15	32	15	26	52	26	24	28	24	48	72
01/01/14 20:00	2	1	20	2014	1	TRUE	54	83	54	15	33	15	23	60	23	21	26	21	49	80
01/01/14 21:00	2	1	21	2014	1	TRUE	53	87	53	16	39	16	24	69	24	23	31	23	49	83
01/01/14 22:00	2	1	22	2014	1	TRUE	55	90	55	15	42	15	24	75	24	21	28	21	49	86
01/01/14 23:00	2	1	23	2014	1	TRUE	53	87	53	15	46	15	24	77	24	20	25	20	48	86
02/01/14 0:00	3	1	0	2014	2	FALSE	51	87	51	15	46	15	25	80	25	21	29	21	44	79
02/01/14 1:00	3	1	1	2014	2	FALSE	51	75	51	15	47	15	25	83	25	24	31	24	34	44
02/01/14 2:00	3	1	2	2014	2	FALSE	40	53	40	15	50	15	26	75	26	23	35	23	28	32
02/01/14 3:00	3	1	3	2014	2	FALSE	32	39	32	16	54	16	25	77	25	23	33	23	27	31
02/01/14 4:00	3	1	4	2014	2	FALSE	31	37	31	15	54	15	25	79	25	22	35	22	26	29
02/01/14 5:00	3	1	5	2014	2	FALSE	31	42	31	14	60	14	25	79	25	24	36	24	26	29
02/01/14 6:00	3	1	6	2014	2	FALSE	35	54	35	14	58	14	25	80	25	23	44	23	26	32
02/01/14 7:00	3	1	7	2014	2	FALSE	33	45	33	13	60	13	26	82	26	27	36	27	26	32
02/01/14 8:00	3	1	8	2014	2	FALSE	36	47	36	16	56	16	27	77	27	25	47	25	27	31
02/01/14 9:00	3	1	9	2014	2	FALSE	29	25	29	18	41	18	29	62	29	28	28	28	28	26
02/01/14 10:00	3	1	10	2014	2	FALSE	27	18	27	19	34	19	32	50	32	29	22	29	29	22
02/01/14 11:00	3	1	11	2014	2	FALSE	25	14	25	18	26	18	23	48	23	30	19	30	29	18
02/01/14 12:00	3	1	12	2014	2	FALSE	43	38	43	20	22	20	23	53	23	32	18	32	26	15
02/01/14 13:00	3	1	13	2014	2	FALSE	48	49	48	21	21	21	24	49	24	32	19	32	25	13
02/01/14 14:00	3	1	14	2014	2	FALSE	44	41	44	21	20	21	22	41	22	33	19	33	41	26

Appendix 6: Observed vs predicted values for Naïve forecasts for 1 week at the end of 2017



Appendix 7: Feature Importance plot from Gradient Boosting Algorithm



Appendix 8: Gradient Boosting Model Summary

Predictors Used

- Date/Seasonality Dummies:
  - Day of Week
  - Year
  - Month
  - Day of the Month
- Holiday Indicator
- Lagged Variables
  - 48 Hour Lag
  - 72 Hour Lag
  - 1 Week Lag
  - 1 Year Lag
- Weather Data (Mean over 5 Zones)
  - Temperature
  - Relative Humidity
  - Visibility

Time Frame

Three years of historical data to predict next day

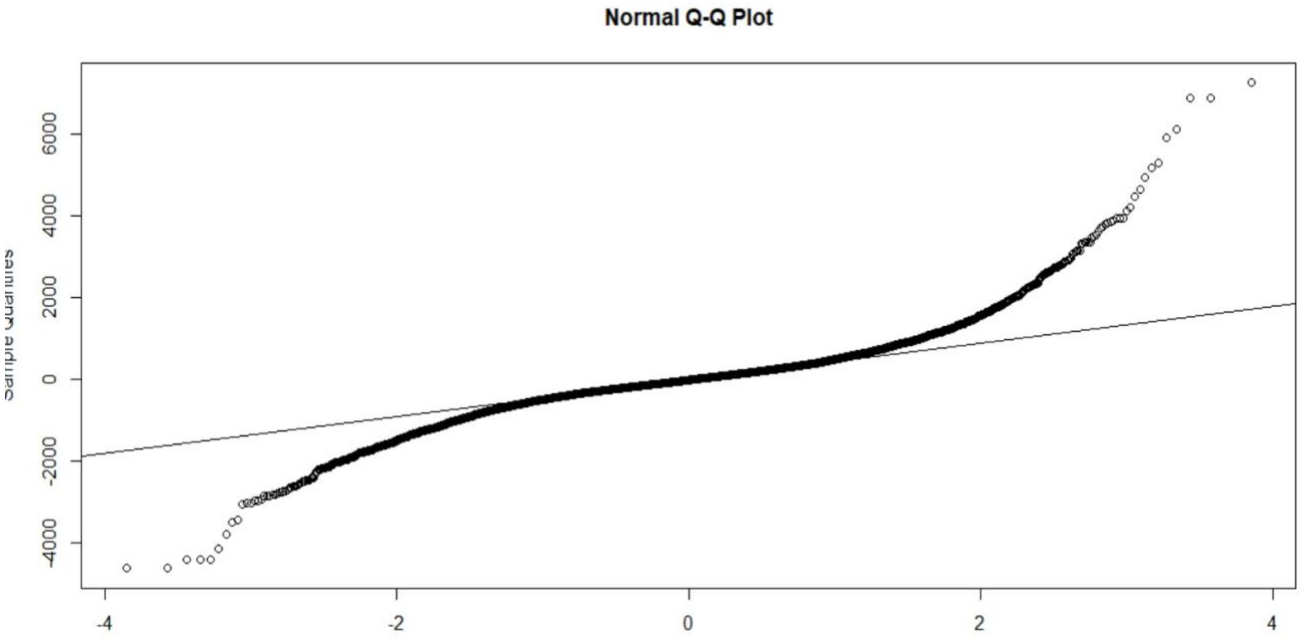
Model Structure

- Trees: 7500
- Interaction Depth: 4
- Shrinkage: 0.01

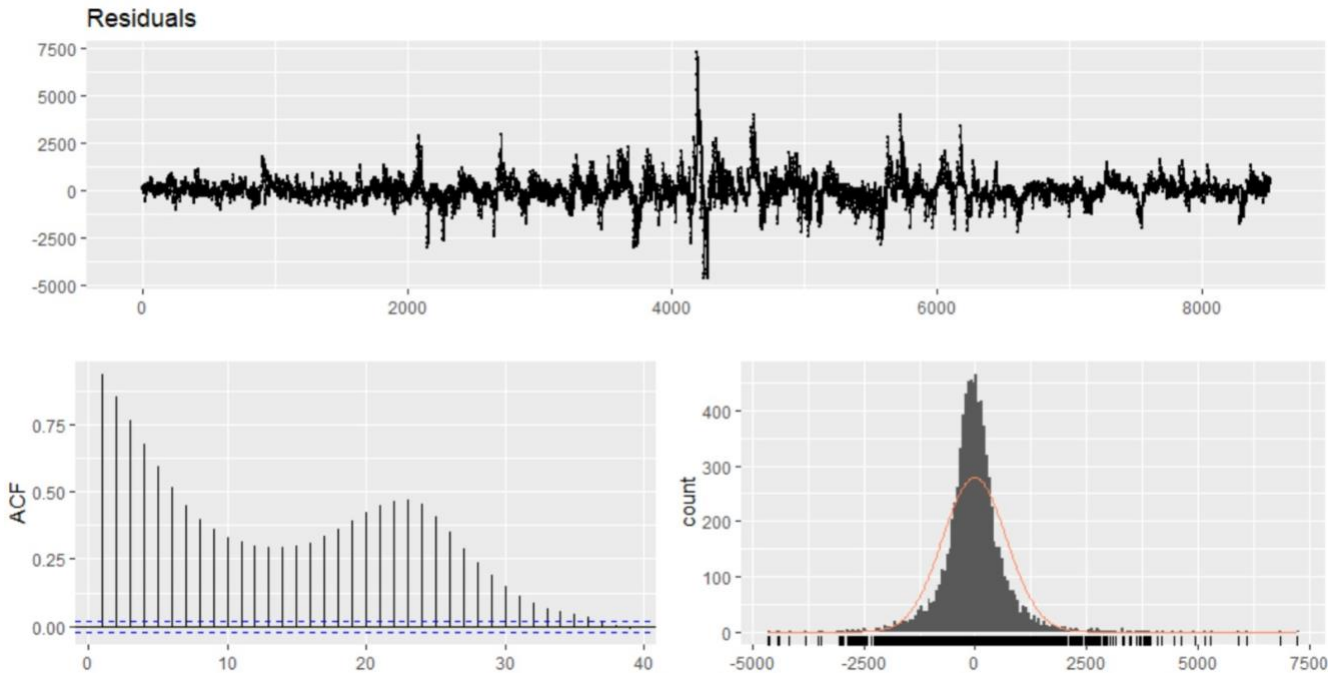
Considerations

- Cross Validation - Ideal Training Period and Parameters
- Computation Time
- Not Including 24 Hour Lag

Normality of Residuals



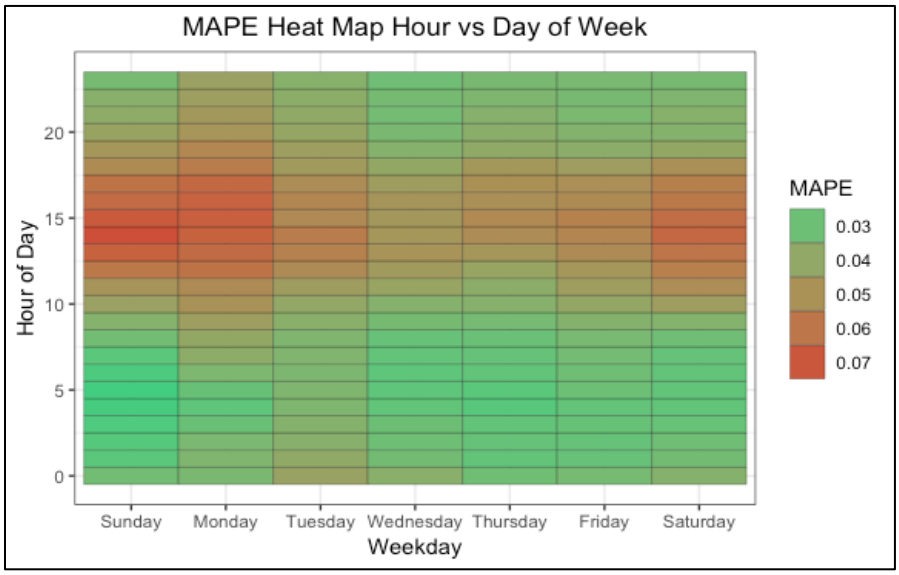
Constant Spread and ACF plot

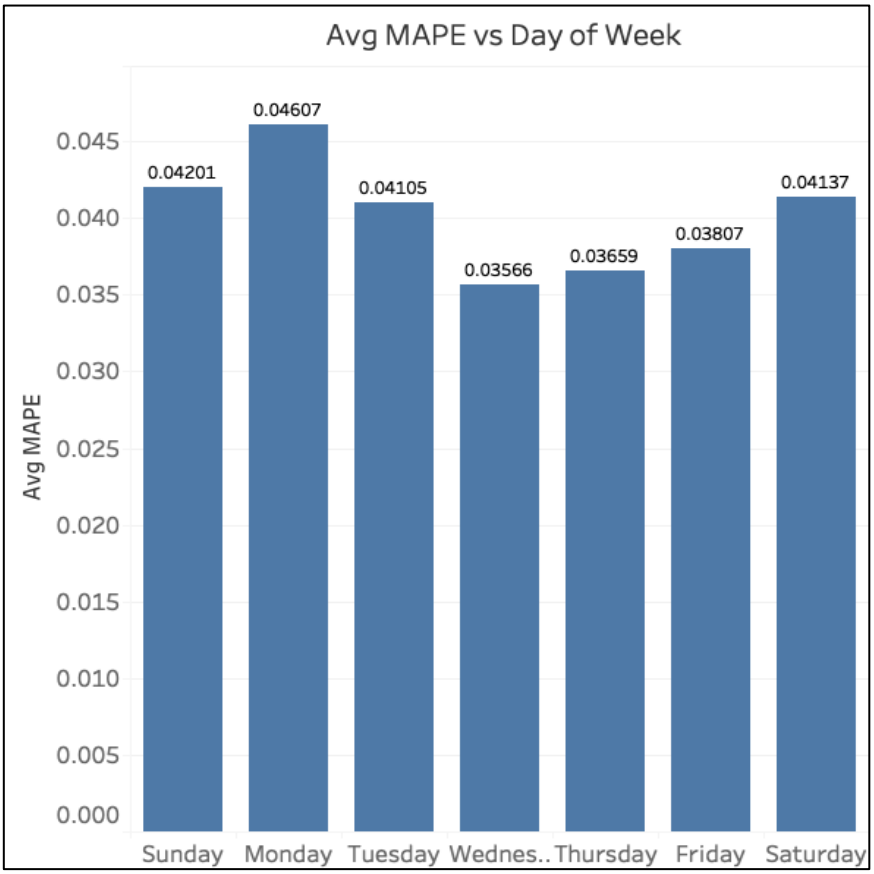


Appendix 9:

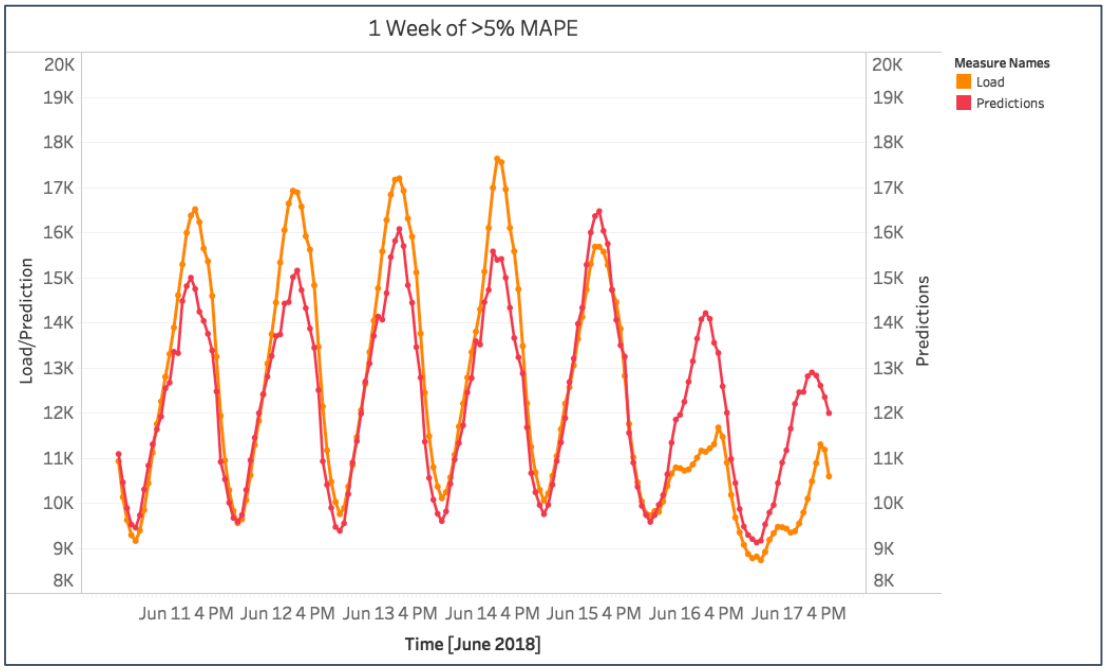
Year	Average MAPE for Gradient Boosting Model
2017	3.89%
2018	3.78%
2019	4.49%

Appendix 10:

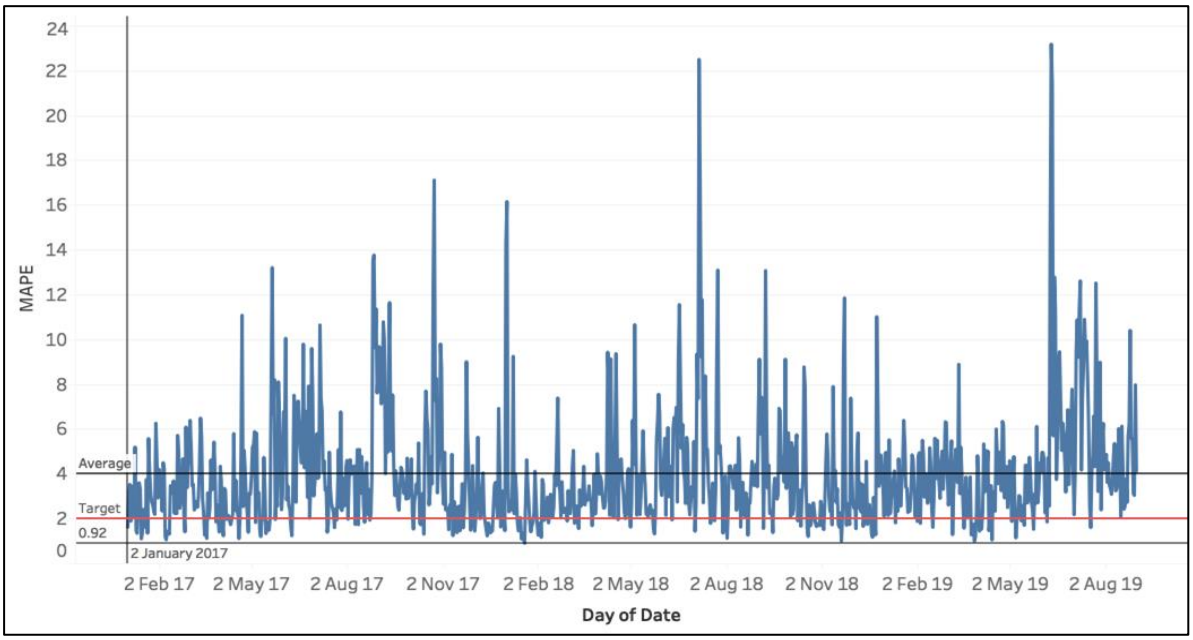




Appendix 11: Gradient Boosting – MAPE during peak hours



Appendix 12: MAPE over time for years 2017, 2018 and 2019 (Gradient Boosting)



Appendix 13: Summary statistics for MAPE over time

Summary Statistic	MAPE
Minimum	0.91%
1st Quartile	2.23%
Median	3.35%
Mean	4.01%
3rd Quartile	4.84%
Max	23.19%

Appendix 14: Cost analysis estimate for years 2017, 2018 and 2019.

Year	Cumulative Load Cost (in $)	Cumulative Load Cost + Penalty (in $)	Mean Cost/Day (in $)	Mean (Cost + Penalty)/Day (in $)
2017	5,514,921,802	5,718,448,285	15,150,844	15,710,023
2018	4,812,376,722	4,948,225,002	13,184,594	13,556,781
2019	2,336,787,245	2,406,732,709	9,576,997	9,863,659

Appendix 15: Real-time price data analysis

REFERENCES

1. Utility Data: http://www.energyonline.com/Data/GenericData.aspx?DataId=18&CAISO___Actual_Load

2. US Energy Information Administration: www.eia.gov

3. Southern California Edison Proprietary Load and Temperature Data