Energy Forecasting

ISDS 7075: Business Forecasting

Fall 2022

Benjamin Vinson, Mohammad Iqbal,

Sina Sedaghat Baghbani, and Morgan Doyle

Introduction

Electricity is a vital part of modern day life and holds great importance to the United States economy. The use of electricity holds a large range of applications in lighting, heating, cooling, refrigeration, computers, operating appliances, computers, electronics, machinery, and public transportation systems (1). Therefore, the demand for electricity has shown to increase greatly over time. This large demand for electricity requires electrical companies to meet the demand and supply equilibrium. As a solution, numerous companies attempt to estimate the future electrical demands to plan ahead effectively.

Energy forecasting plays an essential role in industry, as it provides a foundation to make decisions in a company's planning and operation. Projecting electrical demand is a complex task due to the involvement of various factors, such as weather conditions, calendar effect, economic activity, and electricity prices. Along with this, various methodologies may be used for predicting short-term, medium-term, or long-term electrical forecasting. These considerations make the use of standard forecasting methods insufficient for accurate forecasting. Hence, the need for accurate forecasting in electricity planning.

This work aims to predict hourly loads (in KW) for a utility company in fifteen different geographical regions with only temperature data and calendar information given as factors. The idea is to develop a short-term forecasting model that uses temperature and calendar information to predict electrical loads.

Data Understanding

Data Set

The dataset used in this prediction model contains hourly electrical measurements for 15 various sub-areas from January 2015 to June 2019 (see Figure 1 in Appendix). This results in around 250,000 observations. Unfortunately, below contains a list of the eight weeks within the data that are missing electrical load entries for the 15 locations:

1. March 6, 2016 – March 12, 2012
2. June 20, 2016 – June 26, 2016
3. September 10, 2016 – September 16, 2016

4. December 25, 2016 – December 31, 2016
5. February 13, 2017 – February 19, 2017
6. May 25, 2017 – May 31, 2017
7. August 2, 2017 – August 8, 2017
8. November 22, 2017 – November 28, 2017

Along with the missing weeks listed above, the electrical data from June 23, 2019 to June 29, 2019 are not shown in the data set and must be forecasted.

The calendar and weather data are also given. The calendar data contains specific holiday dates in each given year; the list of the specified holidays can be found in the Appendix in Figure 2. While on the other hand, the weather data includes hourly temperature readings from 9 different weather stations in the region (see Figure 3 in Appendix). The locations of both the 9 differing weather stations and the 15 different zones are unknown.

Pre-Modeling Analysis

In the beginning steps, the temperature and electrical load datasets were both graphed for an understanding of the trends. As shown in both Figure 4 and 5, the available data for the electrical loads are shown over the span of the given time span of 4 years. Figure 4 displays the electrical loads of the varying zones. Specifically, this figure highlights the trend of the data oscillating within a year span yet also increasing or decreasing over the total time span. Due to the overwhelming data, a specific hour and zone was graphed (see Appendix for Figure 5). This allowed the confirmation of the changing of data throughout the year and time span.

Along with plotting the electrical loads, the temperature values for each of the 9 stations were also plotted over the time span (see Figure 6 in Appendix). Overall, the temperature data shows an oscillation as well. Upon closer analysis, the temperature data was seen to have an opposite trend in the ending and beginning of the years than the electrical load. Specifically, Figure 6 shows the ending and beginning of the years to have the minimum temperatures of each year while that same time period shows to have the maximum electrical loads of each year. This led to the assumption of a correlation between temperature and electrical loads.

Data Cleaning

Prior to modeling, the data was reviewed for understanding and cleaning purposes. It was cleaned to ensure the ability to produce any forecast and/or regression for the modeling stage. The data contained a total electricity load of the 15 zones on every day in each month and year for each hour. In other words, after every 15 observations in the data set, a total was computed. This valuation was removed before modeling to ensure consistency amongst the data. Along with the removal of the totals, dummy variables (denoted by 0 or 1) were added for the incorporation of the dates, specified holidays, and the zone ids of each entry.

As stated above, the temperature data contained the temperature of 9 different stations for each hour. There is a lack of knowledge on the stations' locations in regards to the differing zones; therefore, the total for each hour of the zones were computed to use for the model. Along with this, the week of June 23, 2019 through June 29, 2019 contain no temperature entries. To alleviate this, each past temperature on that specific date and hour were averaged to place in those missing entries.

Modeling Methods

Boosted Tree Algorithm

In data modeling, there are a multitude of algorithms to use to input data. In this analysis, a decision tree type of modeling algorithm was used for the predictions. A decision tree model is a supervised form of machine learning; therefore, the data must be labeled and partitioned into at least two sets: training and testing (or validation). In simple terms, the concept of this algorithm imitates human thinking. This flowchart-like structure splits into multiple sub nodes, representing different decisions, factors, or potential outcomes (2). Specifically, the algorithm used in this analysis was a gradient boosted decision tree.

In terms of decision trees, boosting refers to the fact that each tree is dependent on prior trees. The algorithm uses the errors to improve the accuracy and efficiency of the model. These errors are the actual known values of the test set in comparison to the predicted values that the model formed based on how the training data was split into nodes and branches. Overall, this model is an iterative process with each tree depending on the previous tree.

Modeling Details

Through the use of the JMP statistical software, the data was modeled using a boosted tree algorithm. The data was randomly partitioned into 75% used for training and the other 25% was used for validation. In regards to the boosted tree parameters, the defaults that JMP provides were used: 200 layers on the tree, 19 splits per tree, and a learning rate of 0.137.

Multiple different models were run to not only analyze the importance of each variable, but also in attempts to better the model. Out of the multitude tested, there were some worth noting. The first model that was run contained all of the variables described in above sections: zone number, year, month, date, holidays, and temperature variables. Unfortunately, this model resulted in many negative output values. Logically thinking, the electrical load for a given hour should not be negative. From this, other models were tested.

Next, another model was tested in which the temperature was not used as a variable. The average RSquared for each hour prediction variable was averaged. This resulted in an average RSquared value of 0.969 for the training and 0.964 for the validation (see Figure 7 in Appendix for RSquared for each hour). This means that about 96% of the variance in the predicted energy loads can be explained by the variables used in this model. Along with the use of the average RSquared value, an average of the mean absolute percent error (MAPE) value was calculated to be 45.23%. This value measures the accuracy of a model by representing each entry in the dataset's average of absolute percentage errors. This shows how accurate the forecasted values are in comparison to the actual values that are known in the data. Due to this, the model showed a great improvement, but could still be improved further.

Analysis

After many iterations, a final model was produced with the use of all variables except days. This model resulted in an average RSquared of 0.989 for the training set and 0.986 for the validation set, along with a MAPE value of 28.84% (see Figure 8 in Appendix for RSquared for each hour). In comparison to the model previously described, the MAPE decreased 36% showing an increase of accuracy of the model.

Other than the three models mentioned, another model worth noting is the one in which the zone locations were excluded. The average RSquared of this model decreased to 0.044 for the training and 0.007 for the validation, meaning that now only about 4% of the variance in the predicted energy loads can be explained by just the temperature and dates. Through the removal of this variable, this large decrease highlights the importance of the zones in the model.

Results

The exact electrical load predictions can be found in the Excel file that corresponds to this report. Due to the unknown location of the 15 differing zones, the average of all the zones were determined. Figure 9 in the Appendix provides the average loads for each hour during the week of June 23, 2019 to June 29, 2019. The first three days of the week, June 23rd through June 25th, contain the maximum average electrical loads at hour 18. The following two days of the week, June 26th and 27th, show hour 19 with the maximum. Lastly, the days of the 28th and 29th show the maximum electrical loads being exerted at hours 16 and 17 respectively. Therefore, at the times and dates stated above, the electrical loads must be prepared for the maximum electrical use out of the day in each zone. On the other hand, the minimum electrical loads for that week shows at hour 5 for all the days except for June 27th. June 27th shows that the minimum electrical loads exerted is at hour 4. Overall, the graph highlights the hours that contain the most and least electricity use during the week of June 23, 2019 to June 29, 2019.

Conclusion

The analysis provided here used the boosted tree algorithm to forecast electrical loads of June 23, 2019 to June 29, 2019. Specifically, the following variables were implemented in the final model: zone number, years, months, temperatures, and holidays. The location of both the zones, where the electrical loads were recorded, and the stations, where the temperatures were recorded, were unknown. Due to this, averages were taken of the temperatures. The large increase of accuracy in the model in using both the temperatures and zone locations shows the importance of the two in this model. Therefore, the location of the zones and stations was a huge limitation in this forecasting model. Along with this, 8 weeks throughout the 4 year time span were missing. Some dates around the same timeframe as the dates that needed predicting. This analysis provides a way of forecasting electrical loads even given the limitations. Today, the forecasting for electrical loads for companies holds high importance. As shown in this analysis, many different factors can affect the state of electricity. As companies became exposed to the effects of COVID-19 on the economy, many companies found the importance of forecasting and prediction models. Due to this, accurate models must be implemented to predict these values to save both time and money for companies.

Appendix

	obsn	zone_id	year	month	day	h1	h2	h3	h4	h5	h6	h7	h8	h9	h10	h11	h12
2	1	1	2015	1	1	66025	64655	63898	64078	65734	63360	66977	68704	69949	66699	62759	58639
3	2	2	2015	1	1	16908	16505	16572	16928	17119	17782	18629	19410	19589	18666	17721	16429
4	3	3	2015	1	1	126314	123368	119247	117562	118398	121283	126786	132344	139996	142005	145164	144814
5	4	4	2015	1	1	136288	133110	128663	126846	127747	130860	136798	142796	151052	153220	156628	156250
6	5	5	2015	1	1	539	512	505	503	499	545	578	636	667	688	688	654
7	6	6	2015	1	1	6884	6651	6580	6709	7032	7385	7850	8358	9245	9571	9469	8970
8	7	7	2015	1	1	133143	129964	125772	124217	125375	128613	134581	140648	149186	151521	154578	153728
9	8	8	2015	1	1	136288	133110	128663	126846	127747	130860	136798	142796	151052	153220	156628	156250
10	9	9	2015	1	1	3179	3011	3008	2969	3276	3416	3561	3959	4209	4265	4280	4258
11	10	10	2015	1	1	75298	67423	64105	63916	75907	80044	80128	75907	69607	69943	66604	65680
12	11	11	2015	1	1	23394	22155	21431	21390	21619	22296	22992	23721	24165	24171	23773	22825
13	12	12	2015	1	1	90755	86754	84298	84340	86142	90265	94625	98273	102079	102530	101639	98541
14	13	13	2015	1	1	118433	112535	108490	107279	108925	112450	117889	123131	129442	131775	131468	128482
15	14	14	2015	1	1	20728	19721	19075	18896	19365	19470	21055	22290	23981	24819	25001	23830
16	15	15	2015	1	1	21846	21455	21053	21269	21885	21849	23231	24546	25307	24374	22816	21065
17	total		2015	1	1	976022	940929	911360	903748	926770	950478	992478	1027519	1069526	1077467	1079216	1060415

	h13	h14	h15	h16	h17	h18	h19	h20	h21	h22	h23	h24
2	54523	51960	50102	49364	51966	60480	64995	65275	65167	65055	65449	61156
3	15161	14510	13573	13193	14185	16864	18205	18290	17980	16959	16217	14805
4	141586	133343	127738	126884	135218	155491	157905	156895	153409	147031	137768	128078
5	152768	143873	137825	136904	145897	167770	170376	169285	165525	158642	148647	138193
6	614	601	581	562	615	674	748	718	692	660	605	577
7	8373	7487	6913	6690	7275	8713	9234	9160	8826	8277	7360	6391
8	149904	140775	134595	133520	142438	164148	167084	166000	162181	155252	145073	134415
9	152768	143873	137825	136904	145897	167770	170376	169285	165525	158642	148647	138193
10	4135	4005	3829	3808	3958	4403	4663	4587	4327	3959	3582	3243
11	65659	65932	65722	65680	64987	66184	65176	64966	64966	65407	73660	76537
12	21921	20826	20177	20128	21185	24880	26482	26516	26255	25433	23881	22081
13	93619	88559	84570	83113	88278	105115	111353	112782	111668	107977	101566	93259
14	123422	116908	111541	109659	117167	140692	148971	151149	150370	145129	136564	124318
15	22481	21067	20086	20079	21385	24125	24711	24350	23715	22445	21827	19740
16	19328	17673	16438	16386	17741	21611	23378	23614	23940	23321	23306	21507
17	1026262	971392	931515	922874	978192	1128920	1163657	1162872	1144546	1104189	1054152	982493

Figure 1. Layout of Electrical Load Data from Excel

	Holiday	2015	2016	2017	2018	2019
2	New Year's Day	Thursday, January 1	Friday, December 31, 2015	Monday, January 2	Monday, January 1	Tuesday, January 1
3	Birthday of Martin Luther King, Jr.	Monday, January 19	Monday, January 17	Monday, January 16	Monday, January 15	Monday, January 21
4	Washington's Birthday	Monday, February 16	Monday, February 21	Monday, February 20	Monday, February 19	Monday, February 18
5	Memorial Day	Monday, May 31	Monday, May 30	Monday, May 29	Monday, May 28	Monday, May 26
6	Independence Day	Monday, July 5	Monday, July 4	Tuesday, July 4	Wednesday, July 4	
7	Labor Day	Monday, September 6	Monday, September 5	Monday, September 4	Monday, September 3	
8	Columbus Day	Monday, October 11	Monday, October 10	Monday, October 9	Monday, October 8	
9	Veterans Day	Thursday, November 11	Friday, November 11	Friday, November 10	Monday, November 12	
10	Thanksgiving Day	Thursday, November 25	Thursday, November 24	Thursday, November 23	Thursday, November 22	
11	Christmas Day	Friday, December 24	Monday, December 26	Monday, December 25	Tuesday, December 25	

Figure 2. The Holiday List given for the Analysis

	obsn	station_id	year	month	day	h1	h2	h3	h4	h5	h6	h7	h8	h9	h10	h11	h12	h13	h14	h15	h16	h17	h18	h19	h20	h21	h22	h23	h24
2	1	1	2015	1	1	45	41	40	39	43	44	41	39	37	39	44	50	56	60	62	63	63	62	59	59	57	56	54	50
3	2	2	2015	1	1	55	55	54	50	48	44	42	42	45	56	62	63	64	64	64	64	61	55	49	49	48	47	49	50
4	3	3	2015	1	1	47	45	44	39	39	38	37	37	37	42	51	56	61	63	64	65	65	61	56	55	55	54	55	55
5	4	4	2015	1	1	53	51	49	45	43	41	39	39	39	41	53	58	60	61	61	61	60	57	54	44	43	39	39	39
6	5	5	2015	1	1	54	52	50	46	42	41	39	38	39	50	58	63	67	68	69	69	66	60	55	52	49	46	48	51
7	6	6	2015	1	1	51	51	49	48	49	50	47	42	45	52	56	59	60	60	61	60	59	56	48	48	43	42	40	42
8	7	7	2015	1	1	53	52	51	47	47	46	45	45	47	53	61	64	67	69	69	69	68	62	55	53	54	52	52	54
9	8	8	2015	1	1	54	53	50	49	44	44	42	39	40	48	53	60	64	65	68	69	65	60	55	51	49	47	50	50
10	9	9	2015	1	1	52	53	51	43	39	44	43	42	43	50	57	63	67	69	70	71	69	62	52	46	45	45	52	53

Figure 3. Layout of Temperature Data from Excel

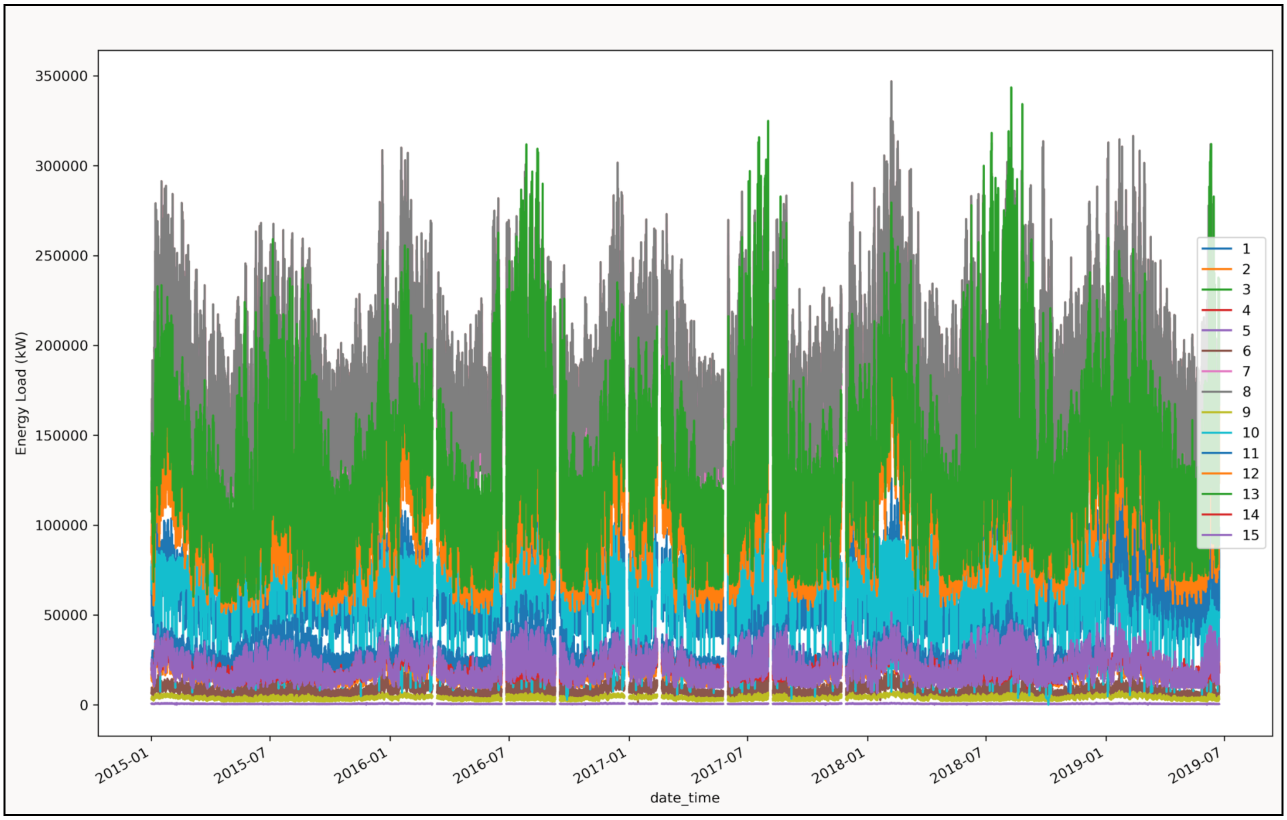

Figure 4. Data Exploration of the Different Zone Electrical Loads over the Time Frame of the Data

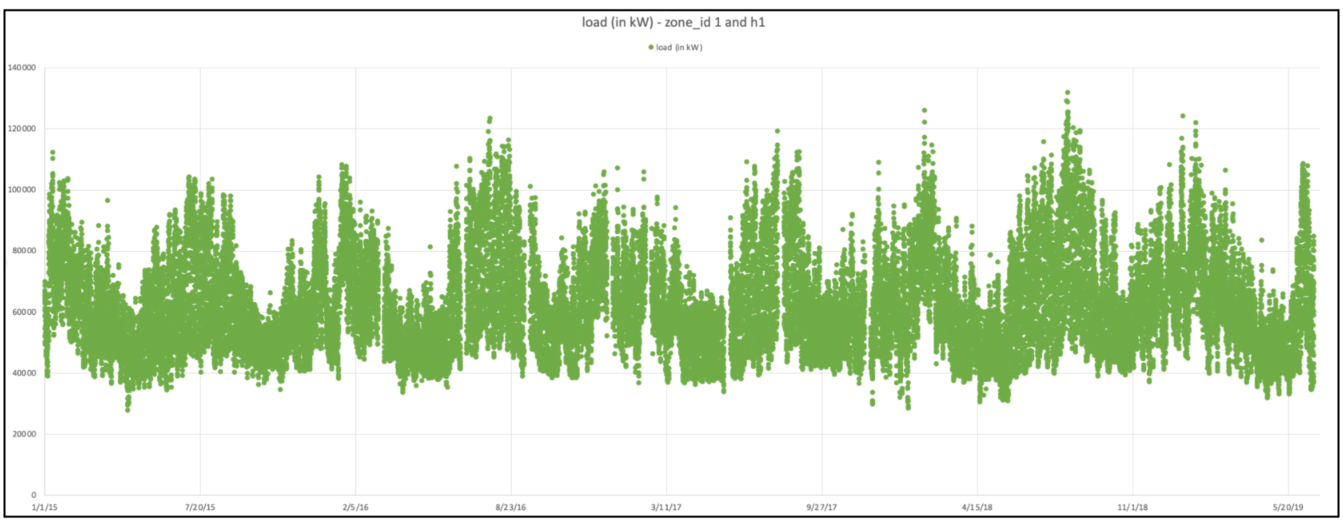

Figure 5. Data Exploration of the Electrical Load of Zone 1 at Hour 1 over the Time Frame of the Data

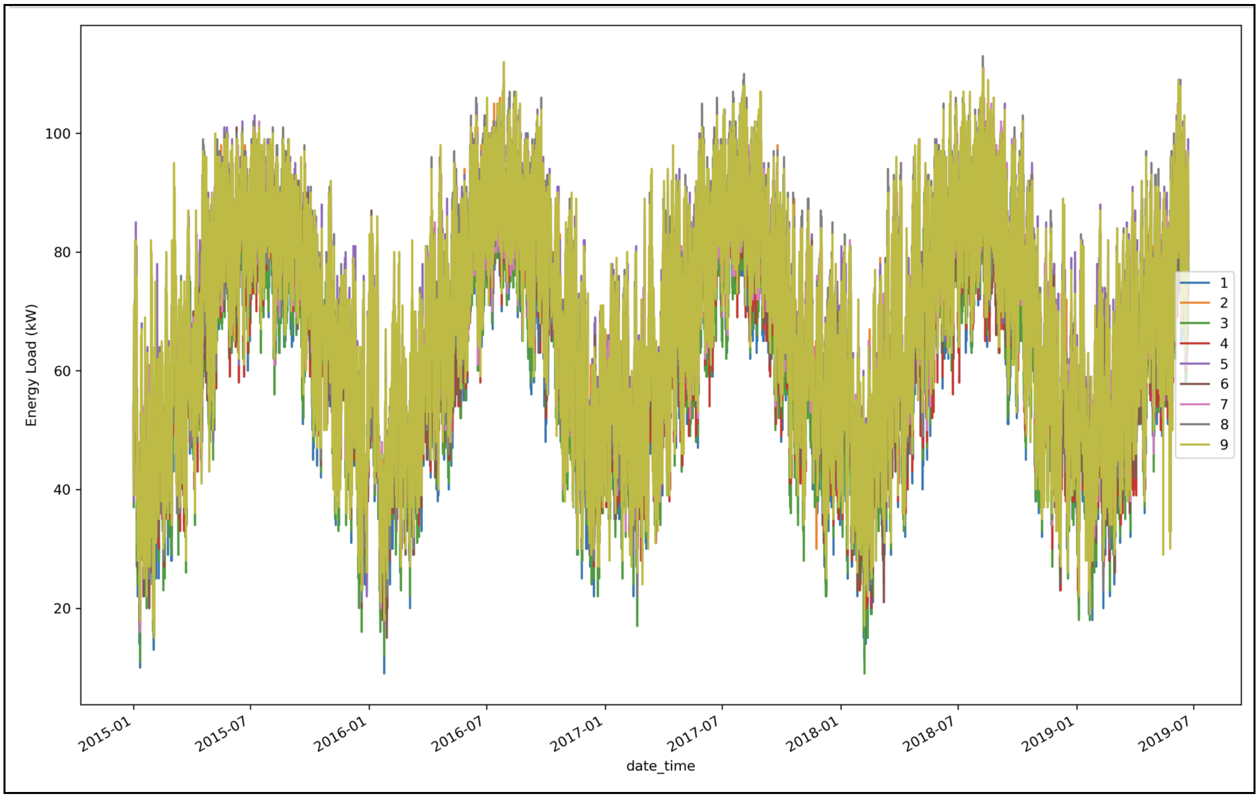

Figure 6. Data Exploration of the Temperature Data over the Given Time Frame

WITHOUT TEMP & WITH DAYS		
	Training (Rsquare)	Validation (Rsquare)
h1	0.971	0.967
h2	0.967	0.963
h3	0.966	0.961
h4	0.966	0.960
h5	0.964	0.958
h6	0.963	0.956
h7	0.958	0.951
h8	0.962	0.956
h9	0.971	0.967
h10	0.976	0.971
h11	0.975	0.971
h12	0.973	0.969
h13	0.971	0.967
h14	0.968	0.963
h15	0.966	0.960
h16	0.965	0.960
h17	0.966	0.961
h18	0.967	0.962
h19	0.970	0.965
h20	0.972	0.968
h21	0.975	0.971
h22	0.976	0.973
h23	0.976	0.972
h24	0.973	0.970
AVERAGE	0.969	0.964

Figure 7. RSquared Values for both the Training and Validation Model for Boosted Tree excluding temperature

WITH TEMP & WITHOUT DAYS		
	Training (Rsquare)	Validation (Rsquare)
h1	0.990	0.988
h2	0.990	0.987
h3	0.990	0.987
h4	0.990	0.987
h5	0.990	0.987
h6	0.987	0.984
h7	0.982	0.977
h8	0.984	0.979
h9	0.989	0.986
h10	0.990	0.988
h11	0.990	0.987
h12	0.989	0.987
h13	0.989	0.986
h14	0.989	0.985
h15	0.988	0.984
h16	0.988	0.984
h17	0.988	0.984
h18	0.988	0.984
h19	0.988	0.985
h20	0.989	0.986
h21	0.990	0.987
h22	0.991	0.988
h23	0.992	0.989
h24	0.991	0.989
AVERAGE	0.989	0.986

Figure 8. RSquared Values for both the Training and Validation Model for Boosted Tree excluding days

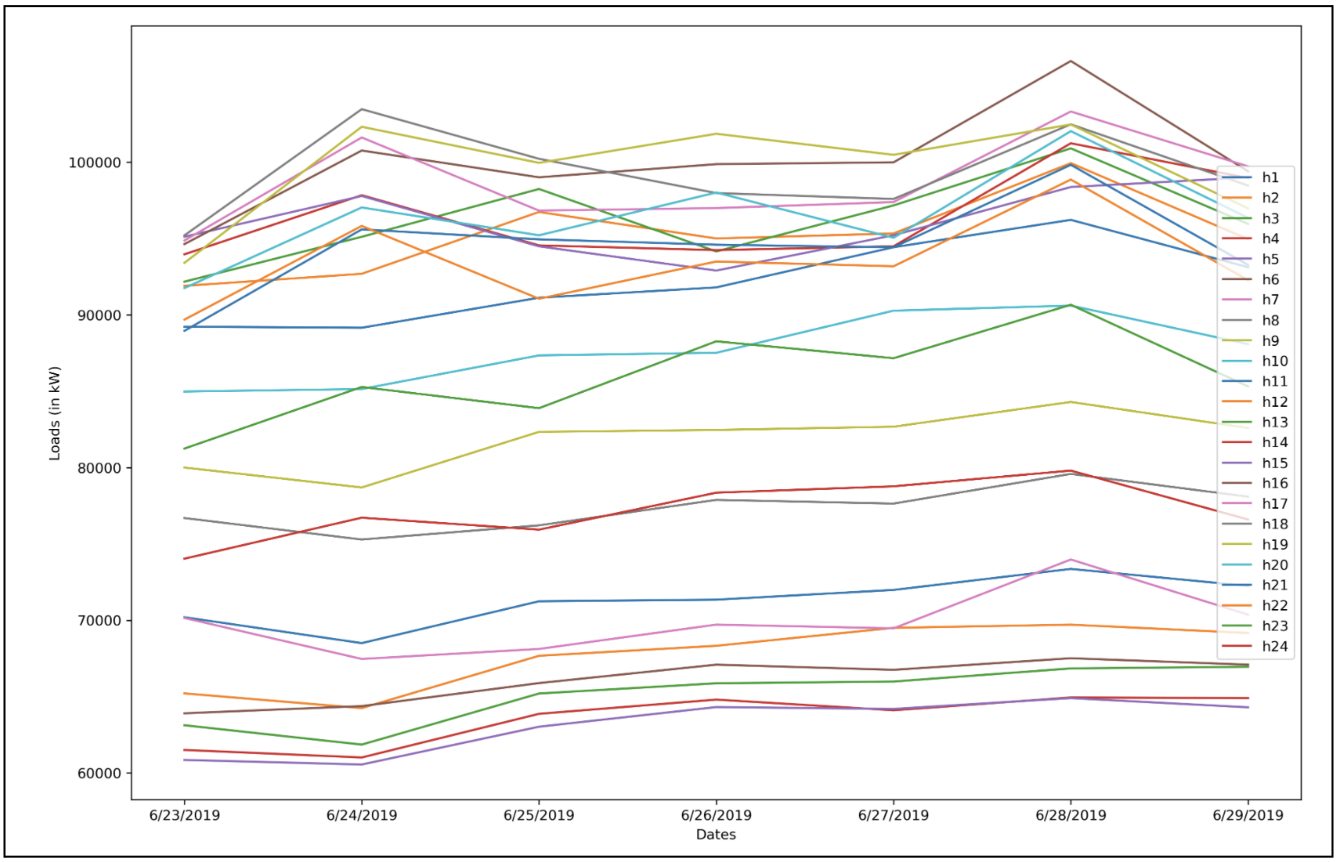

Figure 9. The Average Load of Each Hour during the week of June 23, 2019 to June 29, 2019

Works Cited

1. "U.S. Energy Information Administration - EIA - Independent Statistics and Analysis." *Use of Electricity* , U.S. Energy Information Administration, 3 May 2022, https://www.eia.gov/energyexplained/electricity/use-of-electricity.php.
2. "What is a Decision Tree." *Masters in Data Science,* 2U, Inc, 11 July 2022, https://www.mastersindatascience.org/learning/machine-learning-algorithms/decision-tree/