Rishi Reddy

University at Albany | BFOR 516

Power Load Prediction

For the dallas forth worth metroplex

Table of Contents

[Executive Summary 2](#_Toc121344810)

[Business Problem 2](#_Toc121344811)

[Dataset 2](#_Toc121344812)

[Experiment 2](#_Toc121344813)

[Results 2](#_Toc121344814)

[Conclusion 3](#_Toc121344815)

[Business Problem 3](#_Toc121344816)

[Dataset Description 3](#_Toc121344817)

[Weather Dataset 4](#_Toc121344818)

[Power Load Dataset 5](#_Toc121344819)

[Population Dataset 6](#_Toc121344820)

[Per Capita Personal Income Dataset 6](#_Toc121344821)

[Dataset Limitations 7](#_Toc121344822)

[Dataset Correlations 7](#_Toc121344823)

[Machine Learning Model Selection 7](#_Toc121344824)

[Model Training Procedure 8](#_Toc121344825)

[Results 9](#_Toc121344826)

[Future Research 10](#_Toc121344827)

[Conclusion 10](#_Toc121344828)

[References 10](#_Toc121344829)

# Executive Summary

## Business Problem

The American power grid works around the clock to provide power to customers 24/7. The Electric Reliability Council of Texas (ERCOT) is responsible for monitoring Texas’ power grid to ensure that the state’s real time power load is met. This research project will utilize four machine learning models: Random Forest Regressor, Support Vector Machine Regressor, Multi-Layer Perceptron and K-Nearest Neighbors to accurately forecast the power load of the Dallas Fort Worth (DFW) Metroplex from Jan 2016-Dec 31st, 2020. Three independent variables: population, weather and per capita income will be used to train the models to predict the power load.

## Dataset

There were four datasets used for the prediction of the machine learning models as described here:

1. Power load of the North Central ERCOT region. This region contains DFW. This dataset provides the hourly power load from 2016-2021. It was aggregated to get the daily power load.
2. Average temperature of DFW. This dataset was pulled from NOAA and contains daily average temperatures during the observation period.
3. DFW per capita income and population. These datasets were obtained from FRED which provided income and population on an annual basis. They were manipulated to obtain a daily value which is vital for training the models.

## Experiment

The machine learning models were trained in 4 separate trails (control with 3 experimental ones) to determine which variables enable the models to make the best predictions of the power load. The trail setup is as follows:

Trial Evaluation Criterion

|  |  |
| --- | --- |
| **Trial** | **Variables Evaluated** |
| Control | Temperature |
| E1 | Population, temp |
| E2 | Per capita income, temp |
| E3 | Per capita income, pop, temp |

## Results

The accuracy of each machine learning model was evaluated using the Root Mean Square Error (RMSE) value. The normalized RMSE percentage error for each trial is shown below. The Random Forest performed the best with KNN performing the worst overall. KNN can accurately predict power load only in the control trial. MLP and SVM are okay at predicting power load but are not desirable.

Normalized model RMSE (%) for each Trial

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Trial** | **RF** | **KNN** | **SVM** | **MLP** |
| Control | 6.01 | 6.82 | 17.84 | 90.63 |
| E1 | 6.66 | 29.03 | 17.91 | 18.79 |
| E2 | 6.65 | 32.79 | 17.91 | 22.56 |
| E3 | 6.44 | 32.79 | 17.91 | 22.16 |

## Conclusion

The Random Forest Regressor model would be suitable for use in a production environment to predict the power load of any region in the US provided that the same datasets that were utilized in this experiment are used.

# Business Problem

Ever since Thomas Edison invented the light bulb, electricity has been vital to the growth of the American economy. Nicola Tesla developed the Alternating Current which enabled electricity to be transmitted across vast distances. This innovation spawned the construction of a nationwide power transmission and distribution system. His innovation also spawned the construction of the Hoover Dam and other large scale power generation facilities to meet the ever-rising power load (customer demand).

Unfortunately, the American power grid has a major problem. Power generating facilities must produce ONLY enough electricity to satisfy the power load because electricity cannot be stored. This is a challenging problem because power load fluctuates wildly throughout the day and year.

In the state of Texas, the Electric Reliability Council of Texas (ERCOT) is responsible for ensuring that Texas’ power load is met. ERCOT is a system operator, meaning that it does not own physical assets. Its mission is to monitor the state’s power load and transmission capacity. This is achieved by splitting the state into eight service regions such as the North Central and Far West regions.

This experiment will be focused on the prediction of the power load in ERCOT’s North Central service region from 2016-Dec 31st, 2020. This region was selected because it contains the Dallas Fort Worth (DFW) Metroplex which is one of America’s fastest growing metropolitan areas. The *weather*, *population* and *per capita income* will be used to predict the power load in this region. This experiment will also attempt to predict the power load using solely *population* and *per capita income*. The objective is to evaluate the impact of each mentioned variable on the accuracy of power load prediction. Certain variables will have a greater impact on power load prediction than others.

# Dataset Description

There are four datasets that contain the variables that will be used to predict the power load for the North Central region of Texas. These datasets include weather, power load, population and per capita income. The weather, population and per capita income datasets contain independent variables that will be used to train the machine learning model to accurately predict the power load during the survey period. In this experiment the power load serves as the dependent variable.

## Weather Dataset

The most important predictor of power load is the weather because power load is highly dependent upon the daily temperature. The weather dataset was pulled from the National Weather Service which provides the weather data averaged across the entire DFW Metropolitan area (1). Unfortunately, the website is outdated. On the website the researcher had to select “Daily Data for a month” to get the weather data for each month. Then they had to copy and paste this output into excel for every month from January 2016 to December 2021. The finalized spreadsheet has 2192 observations one for each day. For each day the following weather variables were obtained: Max Temp, Min Temp, Avg temp, Departure from Normal, HDD, CDD, Precipitation, New Snow and Snow Depth. For the summary statistics, DFW has mild winters with hot summers (please see temp graph below). The precipitation is about average for a major metro area. The mean temperature by year and month were averaged to generate the summary statistics. The monthly precipitation values were summed up by month and year. For the prediction of the power load only the precipitation and average temperature are of interest.

## Power Load Dataset

The power load data was gathered from the ERCOT website from 2016-2021. This site has power load data going back to 2002(2). They provide the power load for their eight regions which are: the South, West, Far West, North, North Central, Coast, South Central and East. These regions represent columns (variables) in the dataset. This experiment will only focus on the North Central region where DFW is located within. ERCOT provides power demand spreadsheets for each year. Thus, six separate spreadsheets were downloaded and appended using Python. Each spreadsheet has 8,760 observations with power consumption for each hour of the day. However, for the purposes of model training, the power load was aggregated by the day. This reduced the number of observations to 365 per year and 2192 observations during the survey period. In the chart below we can see that the average monthly power load peaks in summer. This was determined by taking the monthly average during the observation period. This is a vital dataset for which the power load predictions will be evaluated against.

## Population Dataset

This dataset was selected because DFW is a fast-growing metropolitan area. As a result, population should have a stronger impact on power load compared to most areas. More people should result in a greater base power load. This dataset was pulled from FRED. This dataset has the annual population of the DFW Metropolitan Area from Jan 2016-Jan 2021(3). However, this dataset is not compatible with the power load and weather data that will be used for model training because it has only 5 observations, one per year. Thus, the year-to-year population change had to be calculated and divided by 365 to obtain the population change by day. This provided the proper data format for the model training. Once this data manipulation was completed then the population of the DFW area was graphed as shown below. Note that the DFW population rose quickly during the observation period adding about 150,000 people each year.

## Per Capita Personal Income Dataset

The per capita income independent variable was chosen because higher income individuals on average consume more power. The researcher thought that this variable might help better predict increases in the power load. This dataset was also pulled from FRED (4). This dataset has the median per capita income of the DFW Metropolitan Area from Jan 2016-Jan 2021. Similar to the population dataset this dataset in its raw form is not compatible for use in model training. It was manipulated in the same manner as the population dataset to obtain the change in per capita income by day. This provided the proper data format for model training. Once this data manipulation was completed then the per capita income of the DFW area was graphed as shown below. Note that the per capita income rose sharply after the pandemic started (2020-2021). From 2016-2020 it rose from $52,000 to $61,000. From 2020-2021 it rose from $61,000 to $66,727.

## Dataset Limitations

The biggest limitation was that the power load data was for the North Central region of Texas which is far bigger than DFW. Ideally ERCOT would provide this data by metropolitan area instead of by service region. This experiment assumed that the power load for this service region consisted primarily of DFW’s and that the rural areas within this service region had a negligible impact on the power load.

## Dataset Correlations

There are five charts shown above across all of the datasets. It is of no surprise to the researcher that power load data is strongly influenced by the weather. The power load peaks in the summer months and is at its lowest during the fall and spring months. The power load is slightly higher in the winter months because heating systems are in use. The population and per capita income graphs are both trending upward at a rapid clip. However, their correlation to the power load is difficult to determine.

# Machine Learning Model Selection

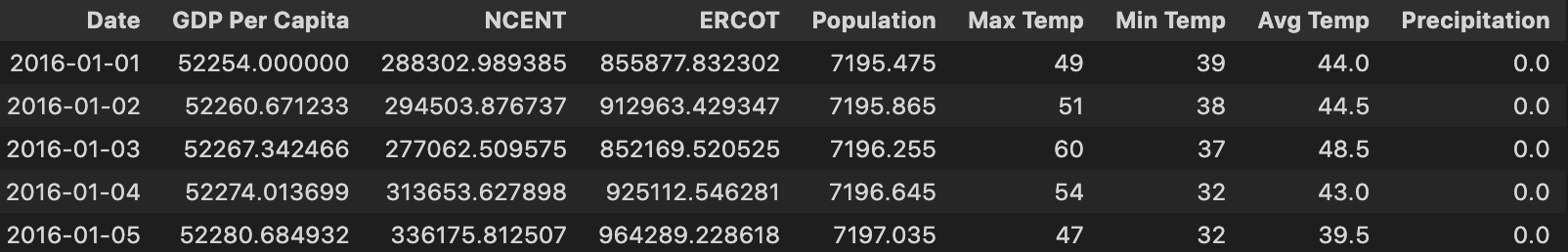
For the prediction of the power load for the DFW metroplex, four machine learning models were utilized. These models are the random forest regressor, support vector machine regressor (SVM), multi-layer perceptron (MLP) and K-nearest neighbors (KNN). Regressor algorithms were used because power load prediction is a continuous outcome. It is not a classification problem. The **KNN** uses the vote of the nearest neighbors to be able to predict an output. Providing this model with say a fall temperature, it will look at previously predicted power load values at the similar temperature to the one provided. **Random Forest** is a popular algorithm used to make predictions on complex datasets. Power load is influenced by numerous factors that cannot be accounted for in this experiment. This model performs well when faced with this type of challenge due to its use of numerous prediction trees. The **MLP** is a basic neural network that will map all the combinations of input variables in “hidden” layers before reaching an outcome. This is crucial because there are numerous combinations of predictors that can influence power load. The **SVM** makes predictions based upon the linear separability of data. As shown earlier, power load values can be linearly separated based upon the time of year. For ex: one can draw a line demarcating spring and summer power load into distinct groups.

The dependent variable in this experiment was the North Central power load with the population, per capita income and weather serving as predictor variables. The model training was split into four distinct trials to evaluate the effectiveness of each predictor variable in predicting the power load. The control trial trained the models on solely the average temperature since it has a strong impact on power load. The first experimental trial then evaluated average temperature and per capita income. The second experimental trial looked at average temperature and population. The final experimental trial evaluated the model on all of the predictor variables. The trials were performed in this manner because the researcher wanted to evaluate which independent variables most improved the prediction capabilities of each model.

# Model Training Procedure

The procedure for model training is identical to that used to train a continuous prediction dataset as demonstrated in Lab 10. The models were trained using Python within the Visual Studio Code environment. There were slight changes made to the code for model training.

The four datasets in the previous section were combined into a single spreadsheet for model training. During the combination of the datasets, observation dates in each dataset were aligned to ensure data integrity. Once combined the final spreadsheet that was imported into Python had the format shown below:



The predictor variables in each trial used to train the four models were as follows:

|  |  |
| --- | --- |
| Trial # | Predictor Variables List for each trial |
| Control | pred\_vars = ['Avg Temp'] |
| E1 | pred\_vars = ['Avg Temp', 'GDP Per Capita'] |
| E2 | pred\_vars = ['Avg Temp', 'Population'] |
| E3 | pred\_vars = ['Avg Temp', 'GDP Per Capita', 'Population', 'Precipitation'] |

For the model training the researcher decided to utilize the standard 75/25 train test split. In this experiment the first 1369 rows were used as training data with the remaining 456 rows as test data. The observations from Jan 1, 2016, to Sept 29, 2019, were split as training data while the observations from Sept 30, 2019, to Dec 31, 2021, were split as test data.

from sklearn.model\_selection import train\_test\_split

# use index-based sampling since we have time series data

train, test = train\_test\_split(power, test\_size=0.25, shuffle=False)

Finally, all four models were trained with the code provided in Lab 10. The training of each model went quickly since the combined dataset has 1828 rows. The Random Forest took the longest to train at just 1.5 seconds.

# Results

Since this research was focused on solving a continuous prediction problem, the Root Mean Square Error (RMSE) metric was used to evaluate the accuracy of each model for each trial. RMSE is an important metric to evaluate the error in each model’s prediction against the actual power load. The RMSE values in their raw form provide little context as to whether a model accurately predicts the power load. Thus, the normalized RMSE was calculated with the use of this equation: Normalized RMSE = RMSE/ (maximum power load – minimum power load) (5). The normalized RMSE for each model is on a scale of 0-1 with 0 indicating a perfect prediction. This scale was converted to a percentage for better data understanding.

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| |  |  |  |  |  | | --- | --- | --- | --- | --- | | **Normalized RMSE (% error)** | | | | | | **ML Algorithm** | **Control–Temp** | **population** | **per capita income** | **ALL Predictor Vars** | | KNeighborsRegressor | 6.82 | 29.03 | 32.79 | 32.79 | | Random Forest Regressor | 6.01 | 6.66 | 6.65 | 6.44 | | SVR | 17.84 | 17.91 | 17.91 | 17.91 | | MLP Regressor | 90.63 | 18.79 | 22.56 | 22.16 |   Figure 1  **RMSE Value**   |  |  |  |  |  | | --- | --- | --- | --- | --- | | **ML Algorithm** | **Control - TEMP** | **population** | **per capita income** | **ALL Predictor Vars** | | KNeighborsRegressor | 110599.5 | 97903.2 | 110599.5 | 23013.4 | | Random Forest Regressor | 21706.5 | 22452.5 | 22440.0 | 20262.1 | | SVR | 60397.1 | 60397.7 | 60396.8 | 60160.1 | | MLP Regressor | 305662.1 | 63358.0 | 76075.3 | 74724.4 | |

Figure 2

In Figure 1, the Random Forest Regressor accurately predicted the power load during the observation period with just a 6% error rate across all trials. During the literature review the researcher discovered that several authors were successful at using neural networks to predict power load. In the control trial the MLP failed to predict the power load. This is no surprise as an MLP needs multiple input variables. The performance of the MLP improved significantly in the three experimental trials at about 20% error. The SVM exhibited consistent performance across all groups at 17.01% error. The KNN’s performance was on par with Random Forest in the Control trial. However, performance degraded significantly in the three experimental trials. This was not surprising considering that power load data clumps strongly around temp.

Though a 6% error is a very low error rate, provided the magnitude of the power load, this is actually quite significant. The highest load day had a load around 500,000 MWH. A 6% error on this value is 30,000 MWH which provides a range between 470,000-530,000 MWH.

# Future Research

There are numerous factors that impact power load that were not included in this experiment due to limitations on data availability, time constraints and technical expertise. The researcher could have used the current datasets to predict peak load days by implementing classification models instead of regression models.

The DFW metroplex is a fast-growing metropolitan area. As population grows so does the power load that the grid has to support. But customers are purchasing solar panels and electric vehicles. How will these buying habits impact the power load? Will the power load increase at a slower rate or could it potentially decrease?

Another angle of research is to look at the power supply side. Over 1/5 of Texas’ electricity comes from wind with most power supply coming from natural gas and coal. Typically coal and natural gas plants supply base load with combined cycle plants used for peak loads. Can the same models used in this experiment accurately predict the power supply provided by each of these sources during the observation period?

# Conclusion

In this experiment the Random Forest model was successful at accurately predicting the power load of ERCOT’s North Central region provided temperature, population and per capita income. KNN also predicted the power load accurately when provided just avg temp. The researcher had a hypothesis that population and per capita income would improve the prediction of power load in addition to average temp. However, this hypothesis was proven false. RMSE values did not improve with these additional variables added.

The Random Forest and KNN models could be used to predict regional power load anywhere in the United States contingent that the average temperature and power load data are provided. However, Random Forest is viable if additional variables are desired to be tested. In this experiment data processing of weather and power load data took several weeks. It is recommended to establish a data processing procedure for power load prediction in a production environment if this data is not on hand.

# References

1. “Fort Worth/Dallas, TX Weather Forecast Office.” *NOWData - NOAA Online Weather Data*, National Weather Service, <https://www.weather.gov/wrh/Climate?wfo=fwd>.
2. *Hourly Load Data Archives*, Electric Reliability Council of Texas, https://www.ercot.com/gridinfo/load/load\_hist.
3. U.S. Census Bureau, Resident Population in Dallas-Fort Worth-Arlington, TX (MSA) [DFWPOP], retrieved from FRED, Federal Reserve Bank of St. Louis; https://fred.stlouisfed.org/series/DFWPOP, December 2, 2022.
4. U.S. Bureau of Economic Analysis, Per Capita Personal Income in Dallas-Fort Worth-Arlington, TX (MSA) [DALL148PCPI], retrieved from FRED, Federal Reserve Bank of St. Louis; https://fred.stlouisfed.org/series/DALL148PCPI, December 2, 2022.
5. Zach. (2021, May 10). *What is considered a good RMSE value?* Statology. Retrieved December 4, 2022, from https://www.statology.org/what-is-a-good-rmse/