Homework 5

Abby Roberts

**Project Background**

The goal of this project is to predict the energy consumption of Chicago using different weather conditions, dates, and census information as inputs. This city was chosen due to having easily accessible power consumption data available online.

By predicting energy consumption, COMED, the power provider for Chicago, will be able to decide upon future energy decisions. For example, a day with a very high predicted energy consumption may require buying energy from another provider. Alternatively, a day with a very low predicted energy consumption may be able to sell their excess to other providers. Additionally, some energy generation units, like nuclear and coal, can take an extended amount of time to come online. Predicting future loads allows for the power provider to give the command to these slower ramping energy units to come online in time for peak loads. Finally, predicting generation may also impact their future planning, as more generation units may need to be built to keep up with consumption rates.

**Problem Formulation**

The goal of this project is to predict how much power a specific city (Chicago, Illinois) will use given weather, date, and census information.

**Data Strategy Plan**

For this project, three different data sets will be combined.

The first is the actual energy consumption of Chicago, available on Kaggle [1]. This dataset contains 66,497 observations of energy consumption pulled every hour, and it runs from January 1st, 2011, to August 3rd, 2018. It is provided by PJM Interconnection LLC, which is the regional transmission operator over COMED. The data contains the columns Date and COMED\_MW.

The second is the census data of Chicago, available on macrotrends.net [2]. This dataset contains 86 observations and predicted values for the GDP and per capita GDP of Chicago every year, and it runs from 1950 to 2035. It is provided by the United Nations World Population Prospects. The data contains the columns Date, GDP, Per Capita GDP, and Annual % Change.

The third is the actual weather conditions of Chicago, available through the NOAA Integrated Surface Dataset [3]. Data is provided with one file for each year, and the files for 2011-2017 were pulled to match the date ranges from COMED. Each file has approximately 15000 observations individually, with the combined dataset containing 103221 observations overall. It is provided by NOAA, the National Oceanic and Atmospheric Administration. The data contains 104 columns, of which Date, Report Type, Call Sign, WND (wind), TMP (temperature), and DEW (dewpoint) will be used.

**Summary of Data Cleansing and Exploratory Data Analysis**

The target variable of this analysis is power. The goal of this analysis is to predict the power needed for the Chicago area using additional information from census and weather data.

First, the individual year files of weather data needed to be combined into one larger file. After combining, the different wind, temperature, and dewpoint values needed to be reformatted. These are saved by NOAA as a “value, error code”, format, with the value being the actual number multiplied by 10, and the error code indicating if there was any errors with the value [4]. After formatting these numbers, rows with an error code indicating an issue were removed. Next, the relative humidity was calculated using the temperature and dewpoint. Finally, the values from NOAA are stored in metric format. These values were then changed to Imperial units. For example, temperatures in Celsius were converted to Fahrenheit.

The next step was to clean up the census data. First, the column Annual % Change was always null, so it was dropped from the dataset. Next, the population was calculated using GDP divided by GDP per capita. This information was rounded to the nearest thousands, as population data is very rarely listed down to the exact person. Finally, the Date column was converted to just a year, as the entries were yearly and not down to individual days.

The third step was to clean the power data. For this dataset, the timestamps are in Chicago local time, which may be either Central Standard Time or Central Daylight Time depending on daylight savings. These datetimes were converted to UTC to match the NOAA dataset and avoid future issues with duplicates.

After this, the power data, weather data, and census data could be combined. The power data and weather data were inner joined on date. The census data was then inner joined to the resultant dataset on year. The inner joins were chosen as the final dataset runs for many years with more than enough information, so it is preferable to drop dates without weather data or without power data rather than attempting to fill these minimal nulls.

Finally, some additional columns were added to the joined dataset. Date and time of year can be very influential on the power generation needed. To account for this, dummy variables were created for the different seasons based on the observation date. Using the seasons defined by NOAA, March, April, and May are spring; June, July, and August are summer; September, October, and November are fall; and December, January, and February are winter [5]. To make a numeric version of the actual date and time stamp of the observation, an epoch column was added to help the model to account for time and date better. Additional Boolean columns were added to account for holidays, day of the week, month, weekends, daylight savings, holiday weekends, time of day (day or night), dawn (within 1 hour of sunrise), and dusk (within 1 hour of sunset). Additional numeric columns were added to account for hour of day, day of the month, day of the year, and ratio of current day of the month to total days in the month.

In addition to date and time, temperature is very important to predicting power. Columns were added to account for the daily maximum and minimum temperatures in addition to the existing current hour temperature.

No outliers were found in this dataset. All parameters are derived from governmental datasets and are the official reported values. As such, there are also no missing values. Values with reported errors from the weather data were removed earlier. Since these datetimes with errors are removed, these datetimes are removed from the final dataset since the data is being inner joined. There are no missing values.

The final size of the dataset is 56,746 observations with 48 columns. These columns are date, temperature, wind direction, wind speed, relative humidity, power, year, GDP, population, season, time in epoch, and 4 dummy columns for season. The season column remains for now for future visualizations, but it will be removed before modeling begins. The target variable will be power, and the remaining columns are explanatory variables. The final dataset runs from January 1st, 2011, to January 1st, 2018.

**Data Visualizations, Ethics and Future Tools**

*Ethical Considerations*

The ethical implications with this tool lie in instances where the tool fails to predict generation correctly. If the tool underpredicts the power needed, the power provider will need to purchase power from the energy trading market. This could lead to added costs for both the power provider and the consumers in Chicago who rely on this for their power. If the tool overpredicts the power needed, the power provider will need to sell the excess on the market. Depending on the rates, this could mean selling the power generated for much less than the cost of creating it, which would again lead to additional costs for the provider and potentially the consumers.

This tool is not intended to decide on transmission, which decides what power goes where. Transmission of power can often contain historical biases against impoverished areas, making it more difficult for customers in those areas to access power at reasonable rates. This tool is intended to purely estimate the generation needed for the power grid as a whole.

This dataset does not contain any personal information. Weather data, including temperature, humidity, wind speed, etc., does not carry any inherent ethical or privacy concern. Power data consisting of a timestamp and total power used cannot be traced to any one consumer’s personal information and is a strict measurement of total power usage for the covered area. Census data could be discriminatory towards the homeless population, depending on the methodology it was collected on, if it did not adequately survey people with no home address. However, this dataset was derived from the US Census version, which intentionally adapted their methodology to account for this group [6]. Additionally, the data set used only looked at yearly totals, which cannot be traced to any one person or disturb their personal privacy.

A potential data bias within this dataset is that the data is somewhat old. The data runs from 2011 to 2017. If the power usage of Chicago has changed drastically post COVID, the model will not be able to adequately predict the power needed for the area. As such, before this model went into a production environment, additional data would be needed to account from 2018 to present day.

Within the United States, power generation is very highly regulated to maintain the integrity of the grid and prevent outages if at all possible [7]. Any tool used to predict power generation needed for an area cannot be the last line of defense, and any outputs will be reviewed and modified as needed by specially trained power operators. Additionally, all power companies are required to hold considerable amounts of power in reserve to use in case of an emergency.

*Data Visualizations*

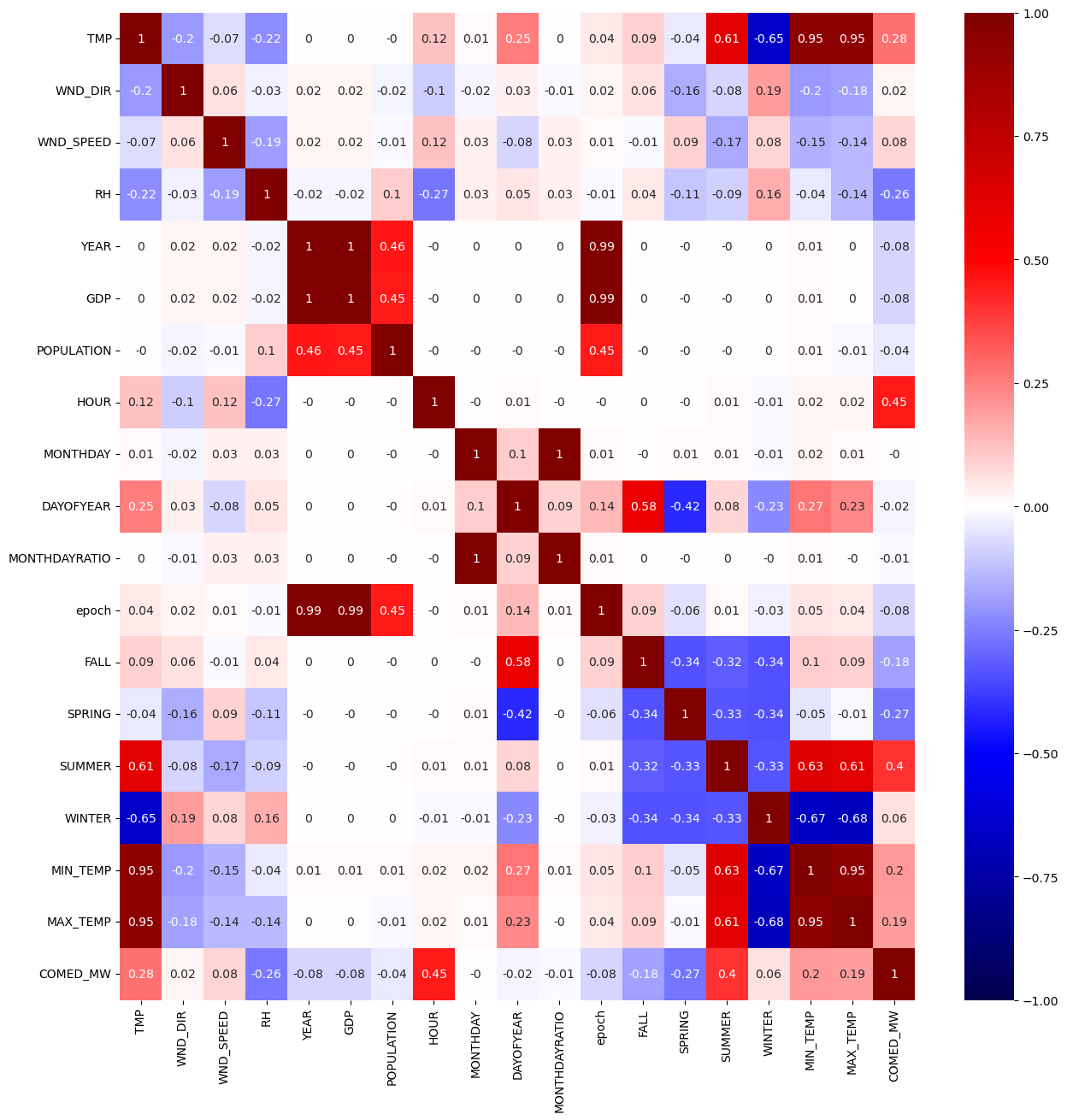


Figure : Correlation between the different variables in the scaled dataset

Figure 1 depicts the correlation map between all the variables in the scaled dataset. This map gives an overview of how different variables may be related. For example, temperature and winter have a strong negative correlation, while temperature and summer have a strong positive correlation. This makes sense, as temperatures tend to be lower in the winter and higher in the summer. GDP and population were only measured on a yearly basis, so it logically follows that both variables are strongly correlated to year. Looking closer at the target variable, COMED\_MW, the explanatory variables with the strongest correlation are temperature, relative humidity, hour, fall, spring, summer, minimum temperature, and maximum temperature. The seasons and temperatures make sense, as more generation is required in extreme cold and heat, while less generation is required in more moderate conditions. Hour makes sense, as depending on the time-of-day people need more or less power. For example, people generally use more power in the winter during off work hours, when they are home and may need the heat on. Relative humidity is odd, but it may relate to the Chicago area’s specific climate.

A blue graph with numbers

AI-generated content may be incorrect.

Figure : Power Usage by Date

Figure 2 depicts the actual power usage by date over time. Looking at the very cyclical pattern of energy consumption over time, it is apparent that time of year matters greatly for predicting the power needed. Summers have a much higher peak power usage than any other season, and winters have a smaller but also significant peak. This pattern repeats across every year used.

A graph showing a curve of temperature

AI-generated content may be incorrect.

Figure : Power Usage by Temperature

Figure 3 depicts the actual power usage by temperature. Looking at the pattern, it is apparent that power usage is much higher in hot temperatures, low in moderate temperatures, and slightly higher in cold temperatures. The relationship between temperature and power appears fairly strong, but it is obviously nonlinear. This may indicate that nonlinear methods may be better suited to predicting power needed.

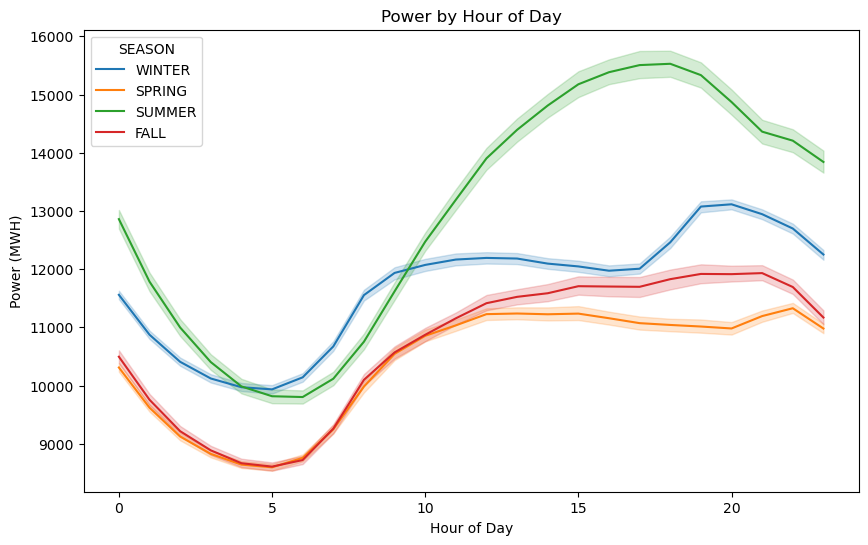


Figure : Power Usage by Hour of Day per Season

Figure 4 depicts the actual power usage by hour of day for each season. Each season has a unique trend for how power is used throughout the day. For example, summer sees one large peak in the afternoon, while winter sees one peak in the late morning and one peak at night. These trends indicate that the hour-by-hour power generation needs are different depending on time of year, and that the model needs to account for seasonality as well as time of day.

**Analysis and Business Insights**

Since the goal of this project is to predict the power needed for the Chicago area, different methods for regression were used. This allows a user to input the various different explanatory variables and output how much power it is predicted that the area will need.

*Model 1: Linear Regression*

The first method of analysis was linear regression. This gives a very general baseline for predicting the power needed. It does not depend on the random state of the machine and requires no additional tuning. It can also help to determine whether or not the relationship between the target variable and the explanatory variables is linear or non-linear.

The linear regression model resulted in a test set R2 of 0.577. This indicates that 57.7% of the variance in the test data is accounted for by this linear model. This is not a great fit, but it does indicate that there is some linear relationship present. To look more closely at the results, the residuals vs the predicted values were plotted in Figure 5.

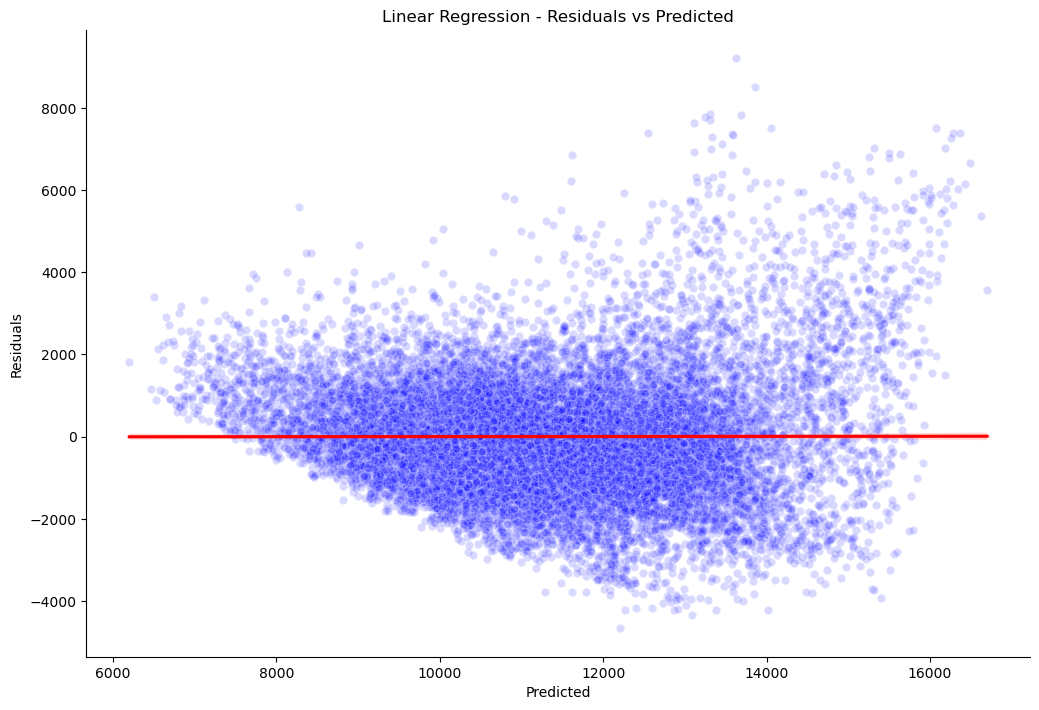


Figure : Linear Regression Residuals vs Predicted

Looking at the residuals plot, there is an obvious trend in the residuals. As the predicted value increases, the variability in the residuals increases. This indicates that the linear model has less variability in predicting times that require less power and more variability in predicting times that require higher power. Additionally, this indicates that a linear model is not appropriate for fitting this dataset, and that a non-linear model may be better able to account for the dataset.

*Model 2: Gradient Boosting Regression*

The next method of analysis was gradient boosting regression. Since the residuals plot from linear regression indicated that a non-linear model may fit better, this method was chosen due to its non-linear nature. Additionally, gradient boosting is a very common type of regression chosen for its accuracy in examples like Kaggle competitions. Maximizing accuracy is important in this scenario, as under or overpredicting the power needed results in additional costs to consumers.

The gradient boosting regression model results in a test set R2 of 0.985. This indicates that 98.5% of the variance in the test data is accounted for by this gradient boosting regression model. This model was chosen after some hyperparameter tuning to optimize for R2. The finalized hyperparameters after tuning can be seen in Table 1.

|  |  |
| --- | --- |
| Hyperparameter | Value |
| Learning Rate | 0.2 |
| Max Depth | 5 |
| Min Samples Leaf | 2 |
| Min Samples Split | 2 |
| N Estimators | 300 |
| Random State | 42 |

Table : Gradient Boosting Regression Model Final Hyperparameters

In addition to tuning the hyperparameters, the residuals for the gradient boosting regression model were reviewed. The residuals vs the predicted values for this model are plotted in Figure 6.

A blue and red dotted line

AI-generated content may be incorrect.

Figure : Gradient Boosting Regression Residuals vs Predicted

Looking at the residuals plot, the residuals are mostly randomized. There are in total more predictions at lower power levels than predictions at higher power levels, but the overall spread of residuals does not follow a trend relative to predicted value. This indicates that the model is likely to be appropriate for the data, and that it will be able to account for this dataset fairly well.

A graph of a number of blue squares

AI-generated content may be incorrect.

Figure : Top 10 most important features gradient boosting regression

Examining the model further, the most important features were reviewed in Figure 7. In order of importance from most to least, the top five most important features were temperature, hour, minimum temperature for the day, whether or not it is the weekend, and maximum temperature for the day. As expected, temperature is by far the most important feature used to predict power. Hour may help to account for what times people tend to be home and running electronics in the home, including heating and air. Whether or not it is the weekend may also help to account for this, as more people tend to be home on the weekend.

**Business Insights**

The goal of this project is to project the power needed for Chicago using weather and census data. Given the high accuracy of the gradient boosting regression model, this model could be used in actual production. As a business, this model could be sold as a product to the Chicago area power distribution company to help predict their future power requirements. The model may need additional fine tuning, as the data used only went up to early 2018, and the energy patterns in Chicago may have changed since then.

**Final Summary**

In summary, the goal of this project is to predict the power needed for Chicago using a variety of weather and census data. Predictions should be as accurate as possible, as over or underpredicting power can pass on additional costs to consumers or cause regulatory issues for the power company. Using gradient boosting regression, a final model was created with a R2 of 98.5%. This indicates that 98.5% of the variance in the data is accounted for by this gradient boosting regression model. On the whole, this model will be very accurate at predicting the power needs for Chicago given weather conditions and census data, but it could always be improved to incrementally increase R2 and get it even closer to 1.

# Bibliography

|  |  |
| --- | --- |
| [1] | R. Mulla, "Hourly Energy Consumption," [Online]. Available: https://www.kaggle.com/datasets/robikscube/hourly-energy-consumption?select=COMED\_hourly.csv. |
| [2] | Macrotrends, "Chicago Metro Area Population 1950-2025," [Online]. Available: https://www.macrotrends.net/global-metrics/cities/22956/chicago/population. |
| [3] | National Oceanic and Atmospheric Administration (NOAA), "Integrated Surface Dataset (Global) (Version Superseded)," [Online]. Available: https://www.ncei.noaa.gov/access/search/data-search/global-hourly?pageNum=1&startDate=2011-01-01T00:00:00&endDate=2017-12-31T23:59:59&bbox=42.010,-87.934,41.950,-87.874&pageSize=10&dataTypes=TMP&dataTypes=RH1&dataTypes=CIG&dataTypes=VIS&dataTypes=WND&data. |
| [4] | Visual Crossing, "How We Process the NOAA Integrated Surface Database Historical Weather Data," 10 February 2025. [Online]. Available: https://www.visualcrossing.com/resources/documentation/weather-data/how-we-process-integrated-surface-database-historical-weather-data/. |
| [5] | National Oceanic and Atmospheric Administration (NOAA), "Meteorological Versus Astronomical Seasons," 22 September 2016. [Online]. Available: https://www.ncei.noaa.gov/news/meteorological-versus-astronomical-seasons. |
| [6] | United States Census Bureau, "How the 2020 Census Counts People Experiencing Homelessness," 26 June 2020. [Online]. Available: https://www.census.gov/library/fact-sheets/2020/dec/2020-census-counts-homeless.html. |
| [7] | Federal Energy Regulatory Commission, "Major Orders & Regulations," 22 November 2024. [Online]. Available: https://www.ferc.gov/major-orders-regulations. |