Electrical energy is an important component in our lives in today's world. We use it for things like powering our cars, supplying power for our lighting, and assisting in running our appliances at home. Furthermore, energy can be categorized into two subgroups renewable and non-renewable energy.

Renewable energy is when the energy source comes from a natural resource and is faster to replenish than non-renewable energy. However, they are less dense and harder to store than non-renewable energy. Some examples of renewable energy are wind, solar, and hydrogen. Non-renewable energy is when the energy comes from a source that cannot be replaced, they are typically more effective in terms of energy production. On the downside, if humans continue to utilize them at high rates this type of energy will no longer exist in the future. Some examples of non-renewable energy are coal, natural gas and nuclear.

Understanding the behavior of electrical energy over time can be a help maximize the economic benefits of electrical energy distribution, increase customer satisfaction, maximize energy production, and minimize electrical energy pollution. One strategy to achieve this is by using the historical data of three metrics, energy demand, generation and CO2 emissions, to forecast the behavior of energy. Energy demand is the amount of electrical energy consumed based on factors like people, government regulations and geography. Generation is the amount of electrical energy produced from one or more facilities. Lastly, CO2 emissions are the amount of carbon dioxide released from energy sources.

For this project, I will be using the historical data of those three metrics to forecast electricity generation and draw insights based on balancing authority. Balancing authority is the organization or institution that is responsible for managing the supply and demand for electricity based on geographic location or grid system. Specifically, the balancing authority that I will be using is ERCO (Electric Reliability Council of Texas), it’s one of the largest balancing authorities in the nation while only managing the power supply for Texas residents.

The two research questions that will be used to guide this project are: What variables, if any, have a linear influence on CO2 emissions? What forecasting methodologies (statistical, classical machine learning, or deep learning) will have the best performance when predicting the next 24 hours of net power generation?

The dataset that will be used for this project ERCOT hourly Grid Monitor dataset from the U.S. Energy Information Administration (EIA). According to the website, they are a “U.S. government agency that collects and analyzes, disseminates independent and impartial energy information to promote sound policymaking, efficient markets, and public understanding of energy and its interaction with the economy and the environment” (EIA,2024). Moreover, they are the statistical and analytics agency within the U.S. department of energy providing large amounts of data over different types of energy sources along with some key statistics and dashboards. EIA was chosen as the source for the data because it is government regulated which results in more reliable and accurate energy metrics.

Furthermore, EIA’s ERCOT dataset is a time series dataset that ranges from July 1st. 2015 from 1am to May 26th, 2024, at 12am. However, dates before December 10th, 2018, at 12am will not be utilized due to large amounts of missing data for portion of the column making hard to impute realistic data to maximize prediction accuracy and analysis. There’s 47,854 rows and 70 columns after excluding the date range mentioned previously. The dataset is updated on an hourly basis where each row represents an hour. The dataset includes multiple time-based columns like “Hour” which represents the hour number of the row, “Local Time”, which represents the date and time in central standard time, and “UTC time”, which represents date and time in UTC format. Additionally, other columns of the dataset are based on the three metrics of energy demand, generation, co2 emissions. For energy demand there are multiple groups like: “Demand” which measures the demand in Megawatts per hour (MWh), and demand of subregion which measures the demand based on the subregion of Texas. For generation, there are also multiple groups like: “Net Generation”, which measures the total amount of generation in MWh. Then there’s columns based on subgroups of energy generation like energy generated from coal, nuclear, oil, natural gas, solar, wind, and hydrogen, all in MWh. Lastly for co2 emissions: there’s “CO2 Emissions Generated” which represents the total amount of emissions generated per hour in metric tons, then there’s co2 emissions generated per subgroup like coal natural gas and oil.

A key consideration for this project is that no weather data will be appended to this dataset. Weather plays a significant role in both energy generation and energy demand, but there are no datasets available that align with the EIA’s ERCOT dataset.

References:

U.S. Energy Information Administration. (n.d.). Mission & overview. U.S. Energy Information Administration. Retrieved June 1, 2024, from https://www.eia.gov/about/mission\_overview.php