DSCI 521 Project Proposal: Predicting Electricity Demand

Our Team

Our project team includes three members: Andy Pike, Arya Monani, and Valentina Ozornina. Each of us comes from a diverse background and skill set, which we intend to leverage throughout different phases of this project.

Andy has a background in business analytics and product management. His experience at a climate tech startup taught him how to use data strategically, working backwards from key decisions to determine how to gather information, analyze data, and create strategic insights. He also has domain-experience analyzing energy data, which is the focus of our project. He is interning with an energy company this summer that is providing the data for this project. Through this class, Andy seeks to gain a broader understanding of analysis in Python and to learn the basics of machine learning and time-series analysis.

Arya has a background in computer science and data analytics, with experience in project management from roles where he led e-commerce and Edutech initiatives, driving significant revenue growth through data-driven strategies. His expertise includes Python, SQL, Tableau, and Agile methodologies, complemented by his current roles as a Course Assistant and Teacher's Assistant at Drexel, where he mentors students and refines his leadership and communication skills. Through DSCI 521, Arya seeks to deepen his quantitative analysis skills, master advanced techniques like neural networks and probabilistic modeling in Python, and enhance his ability to deliver reproducible, impactful insights to prepare for advanced data science roles.

Valentina brings a strong foundation in mathematics, complemented by practical experience in product management and technical analysis. She has extensive experience working with relational databases, SQL Server, and VBA, and excels at translating market needs into actionable technical requirements. Through this course, Valentina aims to reinforce her Python skills while gaining hands-on experience with data analytics techniques, predictive modeling, and time-series analytics. She is also focused on strengthening her ability to communicate findings visually and clearly.

Our Topic: Predicting Electricity Demand for Commercial Customers

Background Context

The electric grid must always balance supply and demand in real-time. Imbalances risk widespread power outage, infrastructure damage, and substantial economic losses. In the United States, this critical job of balancing the grid is managed by regional, regulated, non-profit grid operators (see Figure 1).

Grid operators run a wholesale market to ensure there is sufficient energy to meet demand. Utilities forecast how much electricity their service areas need the next day and place bids to buy electricity in auctions. Suppliers, also known as generators, bid to sell an amount of electricity at a certain price to utilities. The grid operator selects the most competitive bids to meet utility demand. Winning suppliers receive dispatch instructions for the time periods in which they need to generate energy.

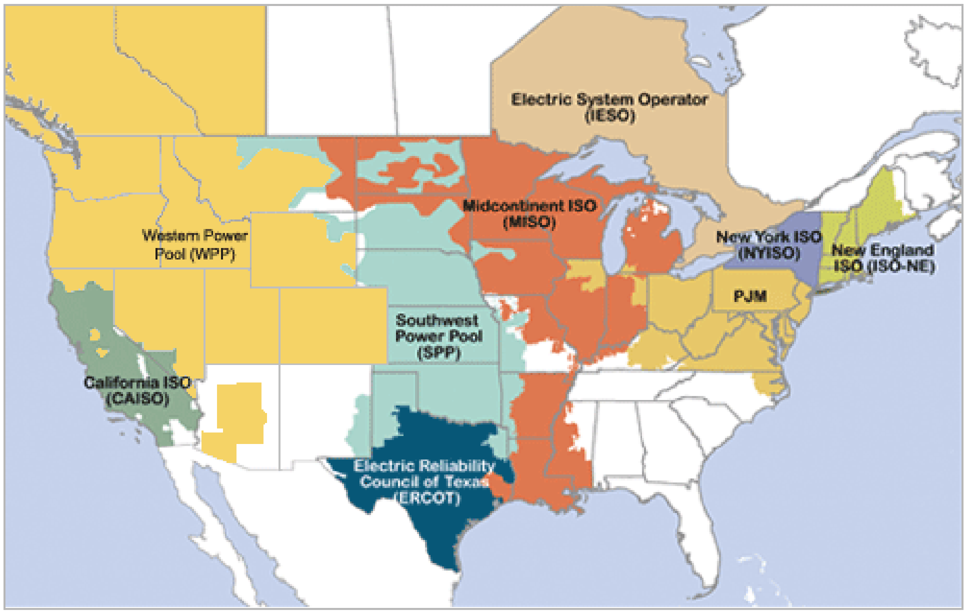


Figure 1: American Grid Operators

However, the actual supply and demand of electricity frequency differs from forecasts, so the grid operator needs to balance the grid in real-time. When demand spikes or supply is lower than expected, grid operators dispatch signals to fire up fossil fuel powered peaker plants or pay customers to reduce usage temporarily, which is known as demand response (see Figure 2).

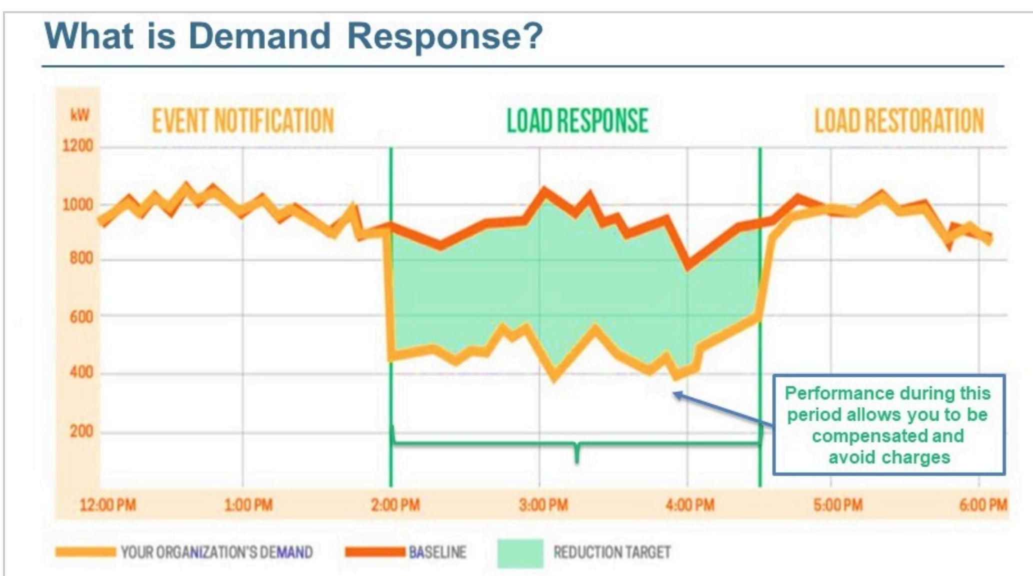


Figure 2: Demand response programs pay participants to reduce electricity during times of peak demand

Several key drivers are increasing the importance of demand response. First, renewable energy production is growing rapidly and its production is highly variable depending on weather conditions. Second, electricity demand is increasing due to industrial growth and the electrification of sectors like transportation. Third, weather is becoming more unpredictable due to climate change, which increases the cost and frequency of events destabilizing the grid.

Our client & analytical goal

Voltus is the leading platform that enables demand response in the United States and Canada by connecting business to energy markets (see Figure 3). Businesses like Walmart, crypto mines, and steel foundries consume a lot of energy but have flexibility to change their consumption. For a commission, Voltus bids their flexibility into the day-ahead and real time markets, tracks the reduction in demand using telematics, and ensures their customers get paid by the marketplace. To do this effectively, they need to accurately forecast how much energy their customers can curtail and bid that amount into the market. If they underestimate, they are not compensated for the extra energy reduction and have missed out on revenue. If they overestimate and reduce demand less than they promised, they are fined and risk losing the privilege of participating in the wholesale energy market.

Andy is interning with Voltus this summer, so they are willing to provide our team with a dataset for analysis. **Our goal is to predict customer electricity consumption and evaluate the efficacy of different predictive models. Improving the quality of predictions means Voltus can bid more accurately and maximize customer revenue.**

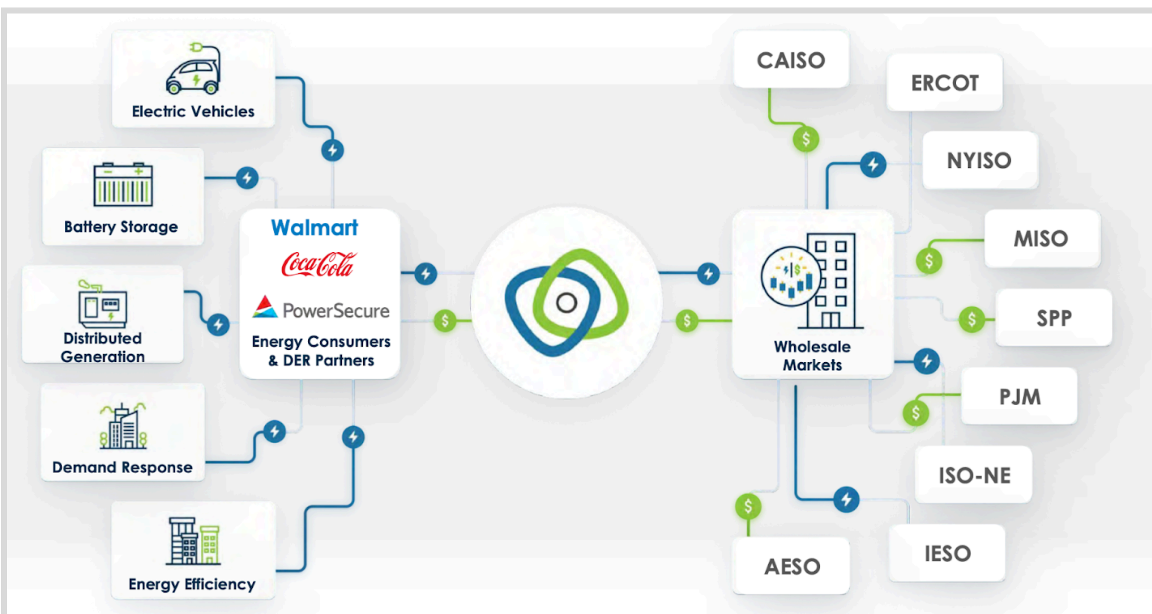


Figure 3: Voltus business model

Our Dataset

Voltus is providing our team with 2 years of historical energy data from 50 customer facilities. This data will be in 15 minute intervals, as shown below. This data will also include whether an interval was part of a demand response event. This is important so we can distinguish if customer consumption in an interval is business-as-usual versus intentionally reduced demand to earn compensation via Voltus.

Interval Beginning (EST)	kWh
5/2/22 0:00	1533.6
5/2/22 0:15	1490.4
5/2/22 0:30	1425.6
5/2/22 0:45	1414.8
5/2/22 1:00	1371.6
5/2/22 1:15	1414.8
5/2/22 1:30	1512
5/2/22 1:45	1512
5/2/22 2:00	1501.2

Voltus is also providing us customer information, such as zip code and the facility type. These features could help improve the quality of our predictions. For example, using facility zip codes, we can determine local weather conditions and analyze how conditions affect energy consumption. Voltus is still compiling our dataset and determining what information they can share without compromising customer anonymity, so we do not know exactly what information will be included.

Our Analytical Approach

Project Scope

Given our team's limited experience with machine learning and time series analysis, our approach will focus on learning foundational methods and applying them in a structured, incremental way. Our primary objective is to develop models that can reasonably forecast customer energy consumption using historical usage data, facility characteristics, and external factors like weather. We aim to compare a few well-established models to evaluate which performs best in this context, with an emphasis on interpretability and simplicity. To keep our scope focused and achievable, we will evaluate the performance of three core models:

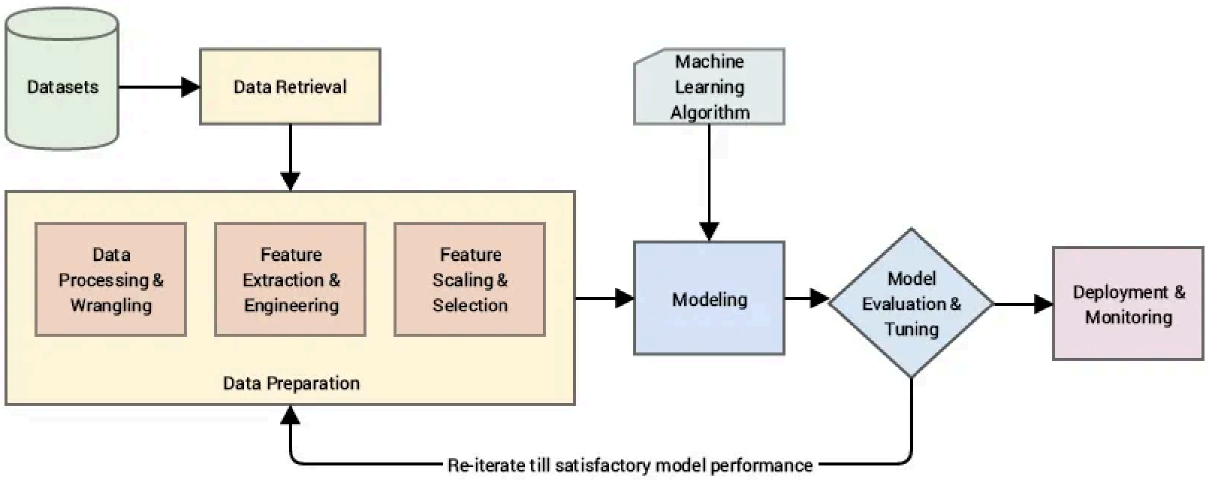
1. **Linear Regression**
   A simple and interpretable baseline model that can incorporate key features (e.g., time of day, day of week, temperature) to predict energy usage. This model will help us build intuition and establish a performance benchmark.

2. **Random Forest Regressor**
   A more flexible, non-linear model that can automatically handle interactions between features. It has strong performance for structured tabular data and can help us understand which features are most important for prediction.

3. **Exponential Smoothing (ETS)**
   A classical time series model that focuses on patterns in the consumption history of each facility (e.g., trend, seasonality). ETS will give us insight into the value of time-series-specific methods compared to feature-driven models.

Plan of Action

We plan to follow the data science workflow shown and detailed below:



1) **Data Exploration & Cleaning**
    a) Examine data structure, identify missing or anomalous values.
    b) Remove or flag periods influenced by demand response events
    c) Aggregate intervals into time-based features (hour, day, month, etc.)
    d) Merge weather data, using open-source APIs such as NOAA or OpenWeatherMap
    e) Perform unsupervised clustering before modeling to group facilities with similar consumption patterns. These clusters can inform separate model training per group and help Voltus set expectations for similar customers.

2) **Feature Engineering**
    a) Create features like lagged energy values, day-of-week indicators, and temperature effects.
    b) Tag business hours vs. off-hours to capture operational patterns

3) **Model Training & Validation**
    a) Split the data into training and testing periods, ensuring that testing data is chronologically later
    b) Train each model and evaluate performance using metrics like Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE)

4) **Model Comparison & Interpretation**
    a) Compare models based on accuracy, interpretability, and consistency
    b) Identify strengths and limitations of each approach

5) **Reflection & Documentation**
    a) Synthesize and visualize key insights into a presentation for the class.
    b) Upload our project code to GitHub (without compromising Voltus privacy requirements)

c) Include a README.md describing the data, how the project may be repeated, and a summary of key insights and areas for deeper analysis

Limitations of Our Analysis

Despite our best efforts to build robust predictive models, there are several limitations that affect both accuracy and generalizability:

1) **Weather Granularity**
   Due to customer anonymity and privacy constraints, we could only use weather data based on zip code, which may not capture localized conditions. This limits the precision of temperature-based features.

2) **Incomplete Feature Set**
   As we do not have access to several key variables that likely influence energy usage, such as occupancy levels, equipment operation schedules, building control systems and system upgrades etc. The absence of these inputs reduces our models' ability to capture rapid or irregular changes in demand.

3) **Operational Variability**
   Commercial electricity consumption is highly sensitive to internal operations, economic shifts, and business-specific behavior. Without access to internal business data, our models are unable to fully account for these variations.

4) **Assumption of Static Behavior**
   We assumed that historical consumption patterns would be predictive of future usage. However, businesses may evolve their operations significantly over time, which challenges model generalization without adaptive methods.

Opportunities for Future Work:

Future improvements could include integrating more granular external data such as high-resolution weather information, real-time electricity prices, and regional event calendars to better capture external drivers of demand. Access to anonymized customer-specific metadata—like equipment usage patterns, occupancy levels, or maintenance schedules—could enhance model accuracy. Additionally, incorporating more robust time-series segmentation techniques and adaptive modeling would help account for operational changes over time, improving the models' long-term reliability and responsiveness.