**Introduction:**

As an Environmental Protection Agency (EPA), we are very concerned with residential energy usage. We hope to inform better policy decisions in the future regarding energy management, and if we are better able to understand and predict trends for a single house's energy usage, then we can scale up this data for larger populations. In this project, we will be using past time-series data to predict future power usage in the same house. We will be predicting using a simple naïve baseline model and a LSTM (Long Short-Term Memory) model on a training/validation set, evaluating results, and will be using the best-performing model to predict on a test set.

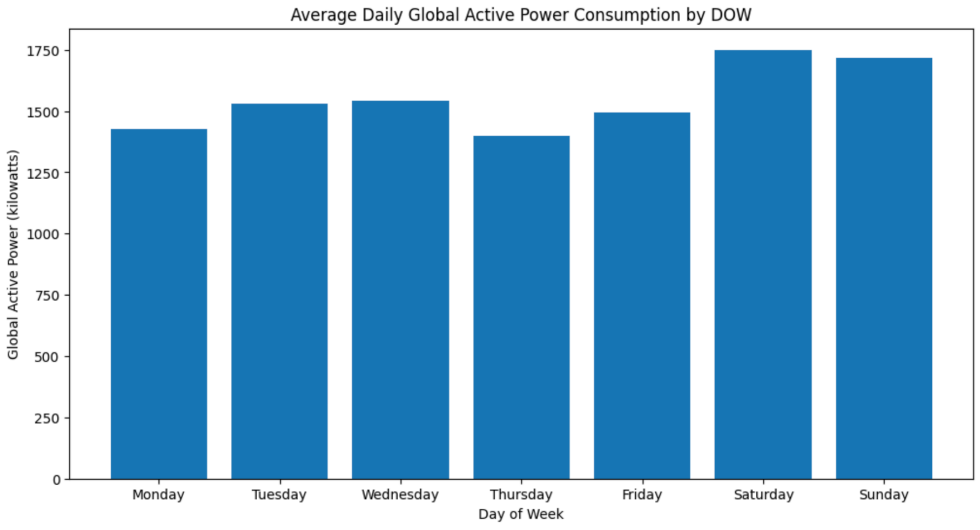
---

**Data:**

https://archive.ics.uci.edu/dataset/235/individual+household+electric+power+consumption

The data used in this analysis came from the UCI Individual Household Electric Power Consumption. Made available by the UC Irvine Machine Learning Repository, this dataset contains energy consumption data in one-minute increments gathered in a house located in Sceaux, France between December 2006 and November 2010. Data kept included the time and date of recording, global active and reactive power (kilowatt), voltage, global intensity (amp), and sub-metering measurements (watt-hour of active energy) found in different parts of the house. Some timestamps had reported null values, which were treated in this analysis as recordings of zero.

Exploratory data analysis found that while energy usage was on average fairly constant across the days of the week, Saturday and Sunday had the highest consumption. This trend can likely be explained by residents being at home much more often over the weekends, driving an increase in the amount of energy consumed.




---

**Methods:**

The data was aggregated from minutes to hours as our goal was to predict the next hour of power usage. The data was then split into a training set (2006-2008; 17911 observations), a validation set (2009; 8760 observations), and a test set (2010; 7918 observations).

The baseline model selected was a naïve last-observed model, which simply predicts each hour of power usage to be the exact same as the previous hour. This provides a clear and reasonable baseline as large spikes or drops are rare, meaning power usage does not vary greatly from one hour to another.

A LSTM (Long Short-Term Memory) model is a type of recurrent neural network that specializes in learning from and processing sequential data. It does this effectively by using a combination of gates and cell states to manage long-term dependencies. Upon selecting an LSTM model, scaling, and sequencing my data, I chose to use parameters of 50 hidden neurons, a dense, single layer output, and an adam optimizer as I felt these would work best with the time-series nature of my data. I ran this model with 20 epochs in mini-batch sizes of 64.

---

**Results and Analysis:**

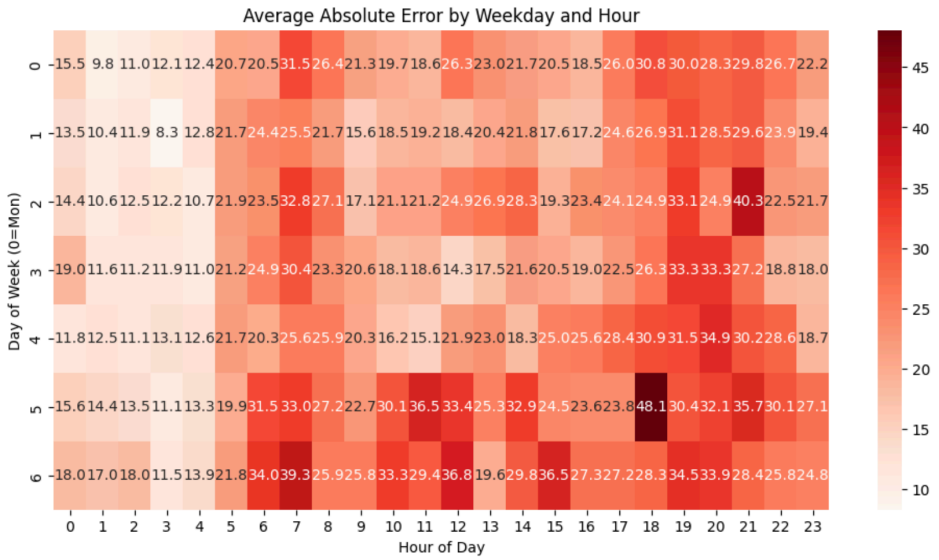
Upon running both my naïve baseline and LSTM model on my validation set, I received the following MAE and MSE results:

| Models | MAE | MSE |
|---|---|---|
| Naïve Baseline | 25.59 | 1537.74 |
| LSTM Model | 23.55 | 1064.38 |

The LSTM model performed the baseline model in both MAE and MSE. This means that on average, my model is 23.55 kW away from actual energy usage, which is an improvement of 2.04 kW compared to my baseline. This is a great result, leading me to choose to move forward with the LSTM model for my test set. My LSTM results on the test set are as follows:

| Model | MAE | MSE |
|---|---|---|
| LSTM Model | 22.85 | 980.29 |

Upon further research into where the model is erroring most commonly, I produced the following heatmap showing by day of the week and hour of the day. This shows that on average, the model is most inaccurate on weekends during the middle of the day to evening. This is consistent with earlier findings that energy usage was highest during these times.




---

**Conclusion:**

In conclusion, I recommend that the EPA move forward with adopting the LSTM model due to its improved performance over baseline models used. The tradeoff in selecting this model is that it is much more computationally taxing than the baseline model, as the naïve baseline model is very simple and quick to perform. However, I believe this is worth it for increased performance, especially if future iterations of the LSTM are better optimized. In order to do this, future models should be explored to try to handle high energy usage periods better, specifically weekend days.

Another future step could entail scaling up these models for a larger population. I found that there are 9,074 households in Sceaux, France in 2021. This is certainly more complicated than just scaling up for this number of houses, as there is obviously a great discrepancy in house size and energy usage among houses in any area. However, if the house used in this study were determined to be of average size, number of residents, and energy usage in the area, then it could provide a valuable baseline for the area.