## **Forecasting Energy Usage Per Source And Per Sector (Residential, Commercial, Industrial) For California**

## **Nima Hendi, Frank Dong, Yang Weng, Sean Lee**

### **1. Project Summary**

In this project, we want to develop a system to forecast energy usage per sector (residential, commercial, industrial) and per source for California. The approach we have decided to take is as follows:

1. Scrap and wrangle the data from U.S. Energy Information Administration and California Energy Commission website.
2. Preprocess the data and store it in a database
3. Build statistical and machine learning models to forecast energy usage per sector and per source
4. Validate and communicate the results using appropriate visualizations

### **2. Proposed Technical Approach**

We begin by gathering, scraping, and wrangling data. Once all appropriate data is scraped, we need to store it in a database. In this project, we are planning to use PostgreSQL to store the data and create a pipeline.

Since our data is mostly numerical and some categorization is needed, we plan to use logistic regression, MLP neural networks, and clustering techniques by scipy and PyTorch. The model will be trained using historical data and logistics related to California energy usage. Finally, we will be using validating techniques (cross-validation) to prove the accuracy of our model.

### **3. Data Sets**

### In this project, we will be using data from U.S. Energy Information Administration ([https://www.eia.gov/totalenergy/data](https://www.eia.gov/totalenergy/data/annual/)) and California Energy Commission website (<https://www.energy.ca.gov/data-reports>). We will be working on datasets that provide monthly energy usage, effects of temperature, and different needs of resources for different sectors. The data sets will be in CSV and EXCEL formats. Some of the data sets' names are U.S energy summary, U.S energy consumption summary, Industry energy consumption by category, Energy trade by category, Energy prices by category, and Industry emissions by source.

### **4. Experiments and Evaluation**

The performance of the machine learning model will be evaluated using metrics such as mean squared error and r-squared. We will be using cross-validation techniques such as k-fold as well. Furthermore, we will be using splitting techniques in our project. The model will be trained on historical data and tested on a held-out set of data. The model will also be compared to existing forecasting methods such as persistence forecasting. The results will be visualized using plots and tables and compared to existing knowledge about energy usage patterns in California.

### **5. Software**

In this project, we are expecting to use publicly available software such as VS Code, JetBrains, PostgreSQL desktop, GitHub, Anaconda Navigator, and Jupyter. We plan to use PostgreSQL to store, query, and wrangle our data. Additionally, we are expecting to use Python and R programming languages extensively for different purposes including training data, validating data, and visualization. Among the publicly available python libraries we plan to use are Pandas, NumPy, Matplotlib, PyTorch, and SciKit-Learn. Also, In R we plan to use Tidyverse, Tidypredict, and dplyr. Furthermore, we will need to write our own code to create additional components needed for prediction, wrangling, API, and web applications if needed.

### **6. Milestones**

During the last 2 weeks, we have been gathering data and communicating with Accenture to understand the basics of the problem. We have collected initial data sets and analyzed them. Based on our latest communication with Accenture, our project milestones are as follows:

● Winter

○ Weeks 8-10: Initial Data Sets Collection and Analysis

● Spring

Weeks 1-2 (3/29-4/9), Data Wrangling and Management

Weeks 3-4 (4/9-4/23), Data Analysis I and Data Analysis II

Weeks 5-6 (4/23-5/7), Prediction and Validation

Weeks 7-8 (5/7-5/21), Data Visualization

Weeks 9-10 (5/21-6/4), Presentation, report, LinkedIn article, and scientific publication in an academic journal