Analytics and Forecasting Energy Consumption Using Smart Meters

Team 014: Transatlantic Synergy

Andre Klein, Jenea Spinks, Tyler Nguyen, Austin Ikedi Okechukwu

MOTIVATION/INTRODUCTION

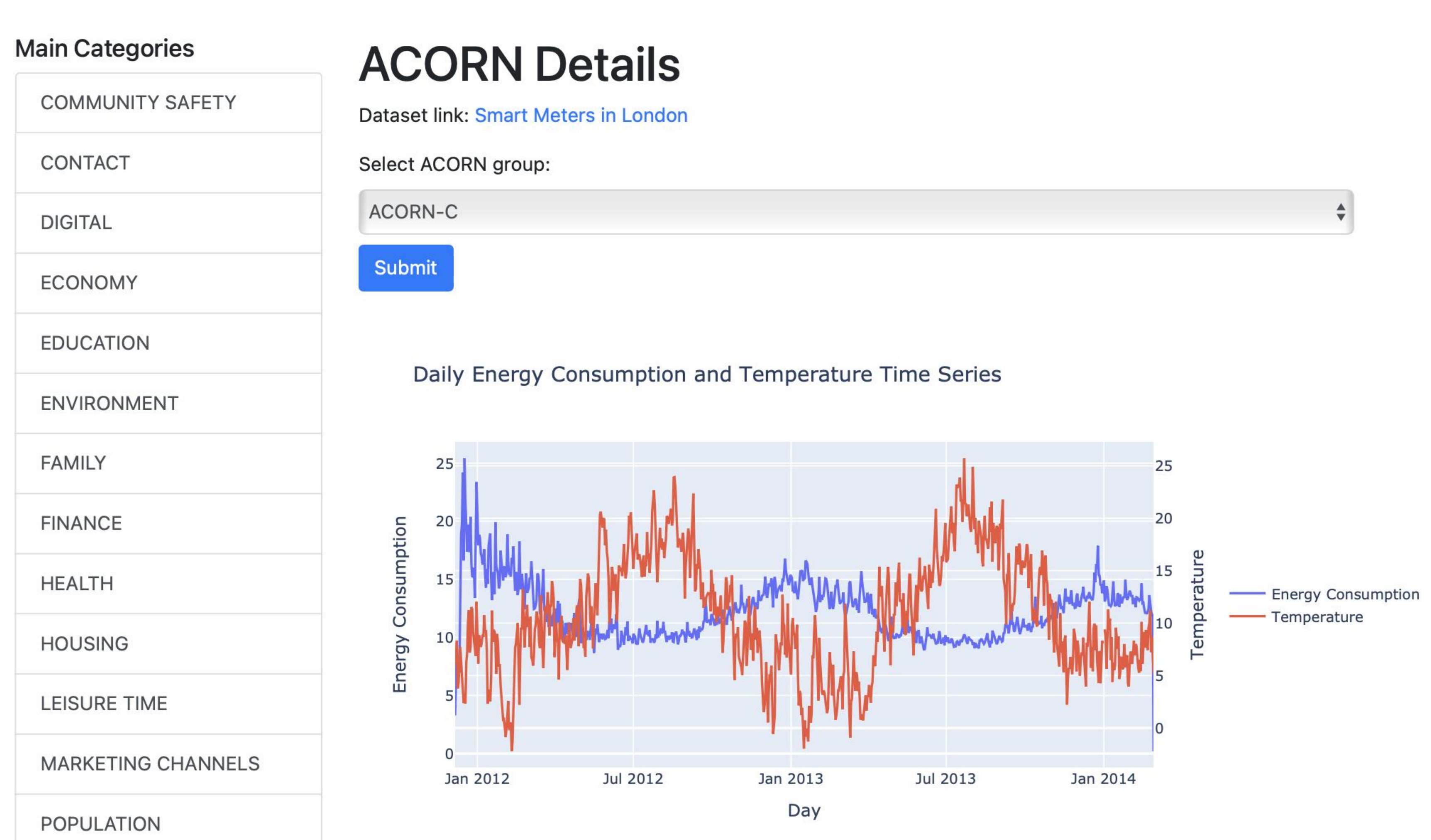
With the growing popularity of smart meters, it has become paramount to understand energy consumption patterns thereby giving energy providers and customers greater insights into energy usage. Our primary objective is to provide a solution that allows users to interact with historical energy consumption for different households via a website and allow the user to predict future energy consumption with different Time Series models (ARIMA, SARIMA, and STL).

DATA (THE LONDON SMART METER DATASET)

The energy consumption data (1 GB) is available on Kaggle to download and is a refactored version from the London Data Store. The dataset contains files with readings from 5567 residential households. It includes detailed profiles and attributes of acorn groups which segments the residents into different groups, household information including group classification, UK bank holidays, daily weather metrics from and daily measurements such as minimum, maximum, mean, median, sum, and standard deviation. Since the dataset involves energy meter and weather readings, it is largely temporal (time series).

APPROACH (ALGORITHM & VISUALIZATION)

In our analysis, we leveraged K-Means clustering to categorize smart meters into distinctive ACORN groups, revealing insights into demographic and consumption patterns which could inform tailored customer services and efficient energy distribution. With time-series analysis via ARIMA, we accounted for consumption trends and seasonal fluctuations, essential for forecasting and operational planning. The project's interactive website features multiple pages, each designed to display different aspects of our analysis, allowing users to explore ACORN group details, clustering outcomes, and predictive models with an interactive interface. By merging various data sources into a singular framework, we prioritized daily data resolution and discarded incomplete records, ensuring robustness in our analysis. Post data cleansing, we aggregated readings by block for a nuanced understanding, leading to the application of SARIMA and STL-ARIMA models upon noticing seasonal patterns in our data, thus improving our forecast's accuracy for energy usage.



CATEGORIES	REFERENCE	PEOPLE COUNT
Crime Survey for England	Alcohol	96
	Being attacked because of skin colour, ethnic origin or religion	51
	Being mugged and robbed	64
	Other economic activity	60
	Retired	154
	Self-employed	120
	Student	66
	Unemployed and seeking work	46
NS Socio-Economic Classifications	Full time students and SEC Unclassified	68
	Higher managerial, administrative and professional occupations	124
	Intermediate occupations	124
	Lower managerial, administrative and professional occupations	128
	Lower supervisory and technical occupations	96
	Never worked and long-term unemployed	35
	Routine occupations	68
	Semi-routine occupations	76
	Small employers and own account workers	125

Figure 1: Website page for K-Means showing how to create more clusters

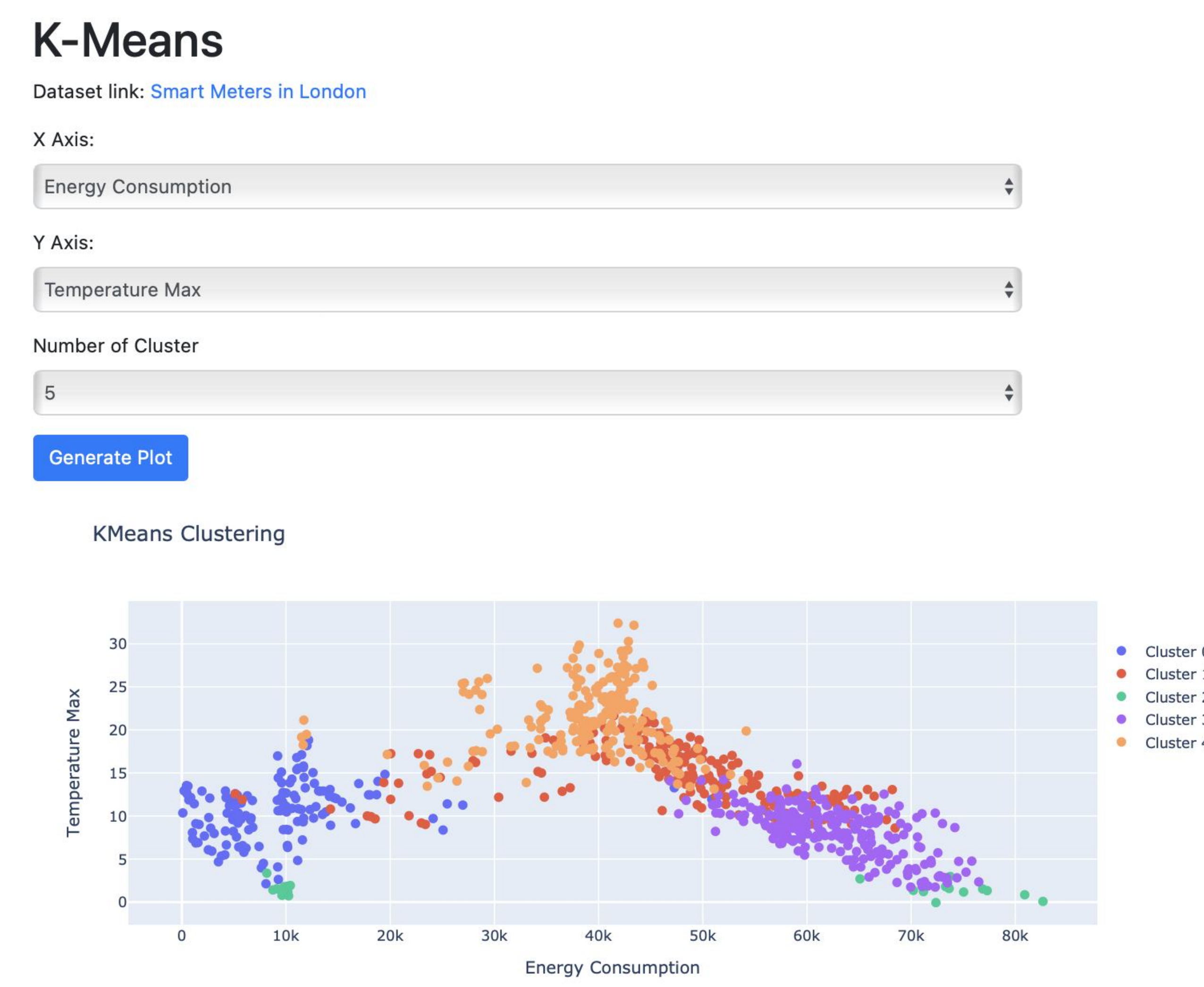


Figure 2: Website page showing the different characteristics of the dataset

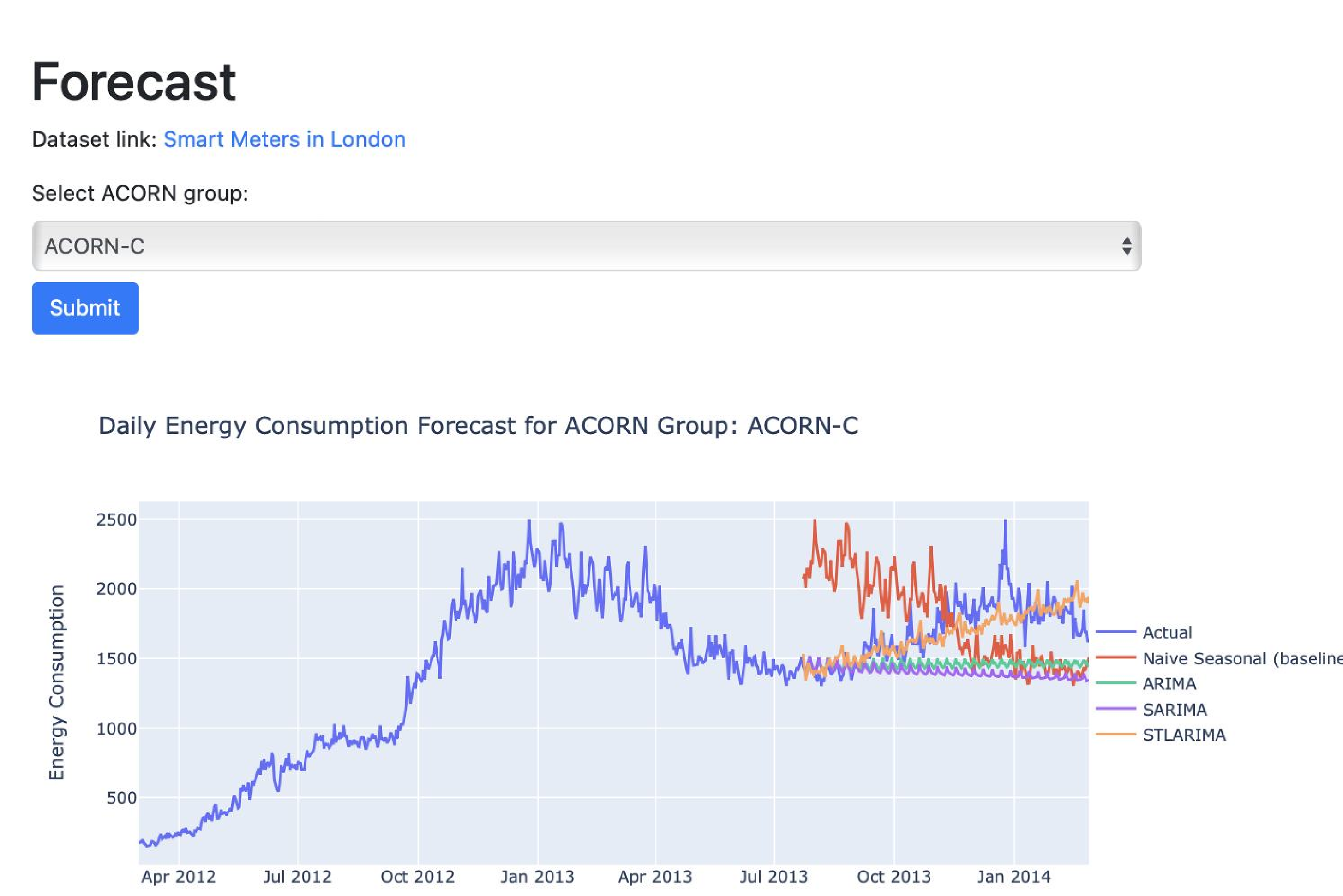
EXPERIMENTS AND RESULTS

After cleaning and exploring the data, we trained 3 different time series models and also created a baseline: ARIMA, SARIMA, STL-ARIMA and Naive seasonal respectively. They were all trained with data from March 1st, 2012 - July 22nd, 2013 and tested against the rest of the data. In the models, we tuned the hyper-parameters for trend and seasonal patterns on the daily aggregation and proceeded to forecasting on individual blocks. We evaluated all the models and decided to implement them for different blocks to allow the user to interact with the data on the website and see which models have lower Mean Absolute Error.

Despite ARIMA's robust performance, it failed to capture seasonal trends effectively, leading us initially to pursue SARIMA for better seasonality handling. After experimenting with various seasonal lengths, computational constraints limited us to a seasonal period of 7. Unfortunately, given this constraint, SARIMA's performance still was not satisfactory.

To address these challenges, we explored the STLForecast model, which combines seasonal decomposition with ARIMA to manage seasonal trends and residuals more effectively. Despite data limitations, STLForecast provided satisfactory performance, especially on smaller data segments, showing a notable decrease in mean absolute error compared to other models. Specifically, STLForecast's mean absolute error on individual blocks demonstrated significant improvements over SARIMA, ARIMA, and the naive-seasonal model. For example, in figure 2 ACORN-C STLForecast achieved a mean absolute error of 107.0 compared to 448.1 with the naive approach.

The KMeans clustering algorithm yielded reasonably good results in separating the data, but the inherent overlap of points in time-series data presented challenges. To better handle this, the K-Prototypes algorithm was introduced, as it effectively processes mixed data types without needing data encoding, achieving similar results to KMeans. Forecasting challenges occurred with ARIMA and SARIMA due to seasonality led us to adopt STLForecast, which is a more data-intensive model but provides better results for our purposes, even with a limited dataset that restricted our season length to one month. The team, relatively inexperienced with the complex dataset and variety of analytical methods, faced a steep learning curve. Nevertheless, we compensated for the initial lack of expertise by conducting thorough research, collaborating effectively, and leveraging available resources, maintaining a commitment to delivering insightful evaluations.



Mean absolute error, baseline	448.11
Mean absolute error, STLARIMA	106.98
Mean absolute error, ARIMA	260.28
Mean absolute error, SARIMA	315.47

Figure 3: Website page for forecasting showing each model and their error metrics