Michael Lee

CSCI E-63 Big Data Analytics Spring 2021

**Smart Meter Analysis Project**

**Problem Statement:** Smart Meters are a useful tool to measure energy consumption in households. This project aims to take the energy consumption data provided by Smart Meters and use Spark and Python to analyze and visualize factors which impact household energy consumption, such as socioeconomic factors, time, and weather conditions. In addition, a series of regression models is built to predict the energy consumption of a given household.

**Data Set:** Smart Meters in London dataset from Kaggle. This is a dataset which contains the daily and hourly energy consumption readings for 5,561 London households between November 2011 and February 2014. The demographic information and method of billing for the households is also provided, as well as the weather conditions during that time period in London.

**Technology/Feature:** This project uses Python and Spark to load and join data from different worksheets. Data visualization is performed using matplotlib. The Spark ML API is used to perform Linear Regression, Decision Tree Regression, Random Forest Regression, and Gradient Boosted Regression.

**Uses/Benefits:** The insights from the Smart Meter analysis will allow electricity utilities to understand what factors have an impact of electricity demand. A predictive model would allow utilities to better plan their power generation resources to be online during times of peak demand. This would allow the utilities to improve operational efficiency and reliability.

**Drawbacks/Challenges:** One of the challenges of this project was handling and processing the large amount of data, especially the hourly meter dataset which was over 10GB in size. Even using Spark, simple computations took an excessive amount of time to compute. Therefore, I decided to focus the main analysis with the daily energy consumption dataset and only use a subset of the hourly energy consumption dataset for analysis.

**Results:** Average energy consumption is higher during winter months, during weekends, and during evenings hours. More affluent socioeconomic groups have a higher average energy consumption. Regression analysis on the entire dataset did not provide meaningful results due to the large variance in household energy consumption. However, once I narrowed the scope of the regression analysis to focus on a single household’s energy consumption, I could build regression models to predict energy consumption.

**Project Files:**

Link to Dataset: <https://www.kaggle.com/jeanmidev/smart-meters-in-london>

Link to 2 minute Presentation Video:<http://bit.ly/CSCIE-63SmartMeterShortVideo>

Link to 15 minute Presentation Video: <http://bit.ly/CSCIE-63SmartMeterLongVideo>