Predicting Electricity Usage: Deliverable 1

Data Extraction, Processing and Initial Modeling

Kevin Taylor | Kelly Du | Nathaniel Ho

Agenda

Introduction

Data

Variables

Pre-Modeling

Modeling

Results




Introduction

Problem

What will energy consumption look like after 2014?

Objective

Predict energy consumption after 2014 using data from 2011-2014

From:

Raw uncleaned dataset
'LD2011_2014.txt'

To:

Trained SARIMA model
with performance analysis

Value Creation

Evaluate the accuracy of a forecasting model on the data

Predict post-2014 energy consumption

Clean raw dataset so it is usable for other forecasting models

Data

Data Extraction

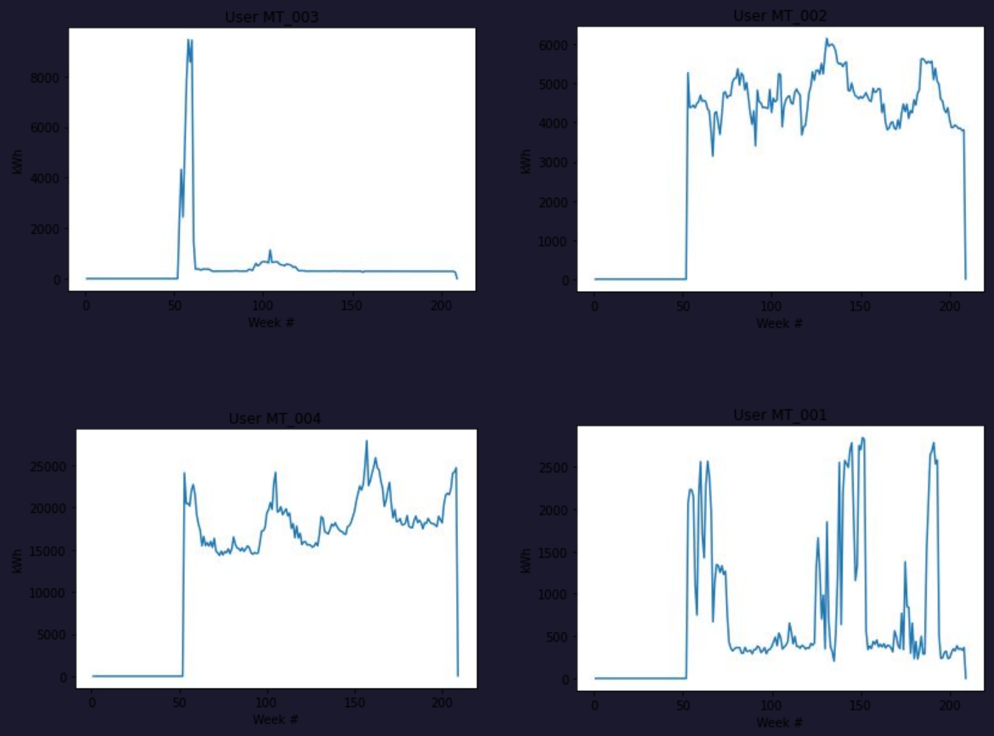
- Source: https://archive.ics.uci.edu/ml/datasets/ElectricityLoadDiagrams20112014#
  - Dataset is in a .zip file; extraction yields a .txt file
  - Text file delimited by ";"
  - Decimal values encoded as 0,00
  - Values in kW per 15 minutes - to convert to kWh, values must be divided by 4
  - Missing values ie) accounts created during the timeframe are encoded with zeros
  - Biannual time change results in either one hour of zeros or aggregation, depending on the season

- Shape of (140256, 370), or 140,256 rows and 371 columns (one for Datetime, 370 accounts)
- Head of raw data:

	Datetime	MT_001	MT_002	MT_003	MT_004	MT_005	MT_006	MT_007	MT_008	MT_009	...	MT_365	MT_366	MT_367	MT_368	MT_369	MT_370
0	2011-01-01 00:15:00	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0
1	2011-01-01 00:30:00	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0
2	2011-01-01 00:45:00	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0
3	2011-01-01 01:00:00	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0
4	2011-01-01 01:15:00	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0

Data Processing

- Data Extraction Process
  - Read entire dataset delimited by “;” into pandas DataFrame object
- Data Diagnostics
  - Dataset is clean (i.e. no duplicates, missing values, etc.)
  - Time change assumed not to affect overall trend (occurs every year)
- Modeling Preprocessing
  - Transform data into long pivot form
  - Divide all values by 4 to arrive at kWh, a unit of energy
  - Aggregate the dataset by account ID, year and week



Variables

Variables Overview

Target Variables	Predictive Variables

• Variable that is predicted in the forecasting model
• Weekly usage of electricity in kWh is our "y"
    • Denoted in DataFrame as *value*
	• Variables that predict target variable
	• Organized into direct or derived variables
	    • Direct variable: Directly from dataset
	    • Derived variable: Created by manipulating direct variables
	• All variables are direct

Predictive Variables Overview

- Initial model fit (SARIMA) does not use any exogenous variables
- SARIMA components (Autoregressive, Moving Average) use target variables at different lags, and weighted average forecast errors, as predictor variables, as well as a seasonal component.
- Future models ie) SARIMAX will use exogenous variables
  - ie) Holiday Week indicator, additional seasonality component ie) quarterly, monthly

Pre-Modeling

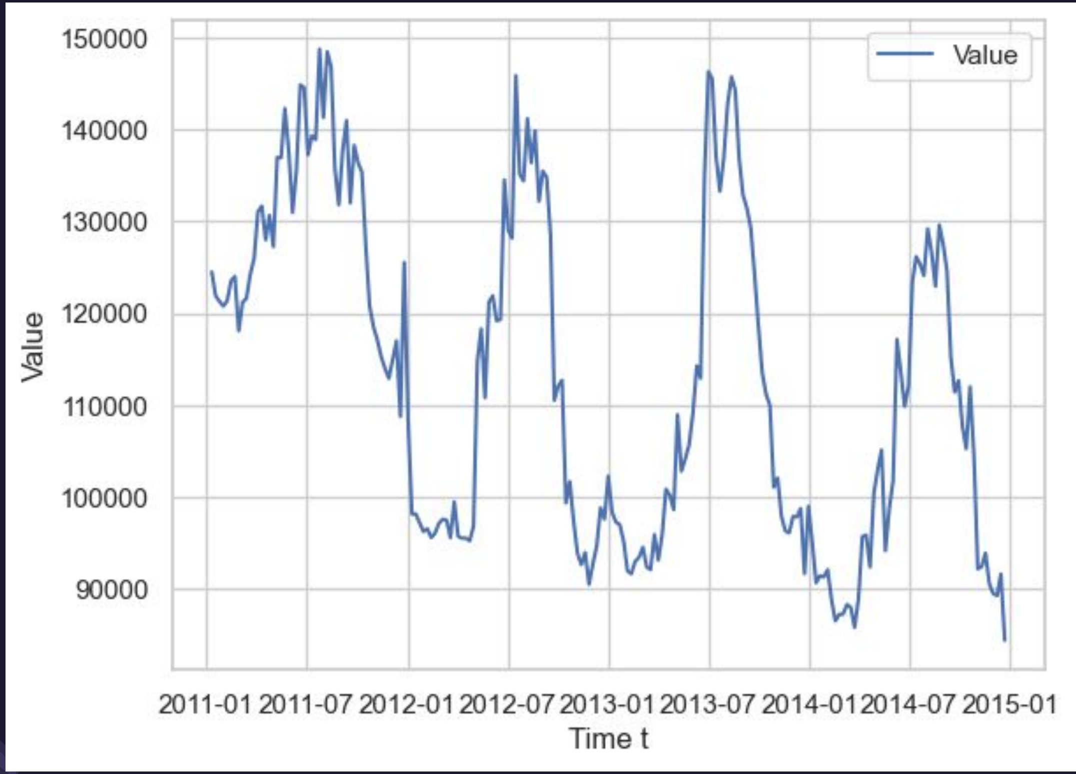
Pre-Modeling

- Exploratory Data Analysis
  - Identify trend and seasonality in the data for initial modeling
  - Perform this exercise on all accounts aggregated, then apply the chosen model to each individual account
- Pre-Modeling Utility Functions
  - Create utility functions for train-test split, MAPE, walk-forward validation, etc.

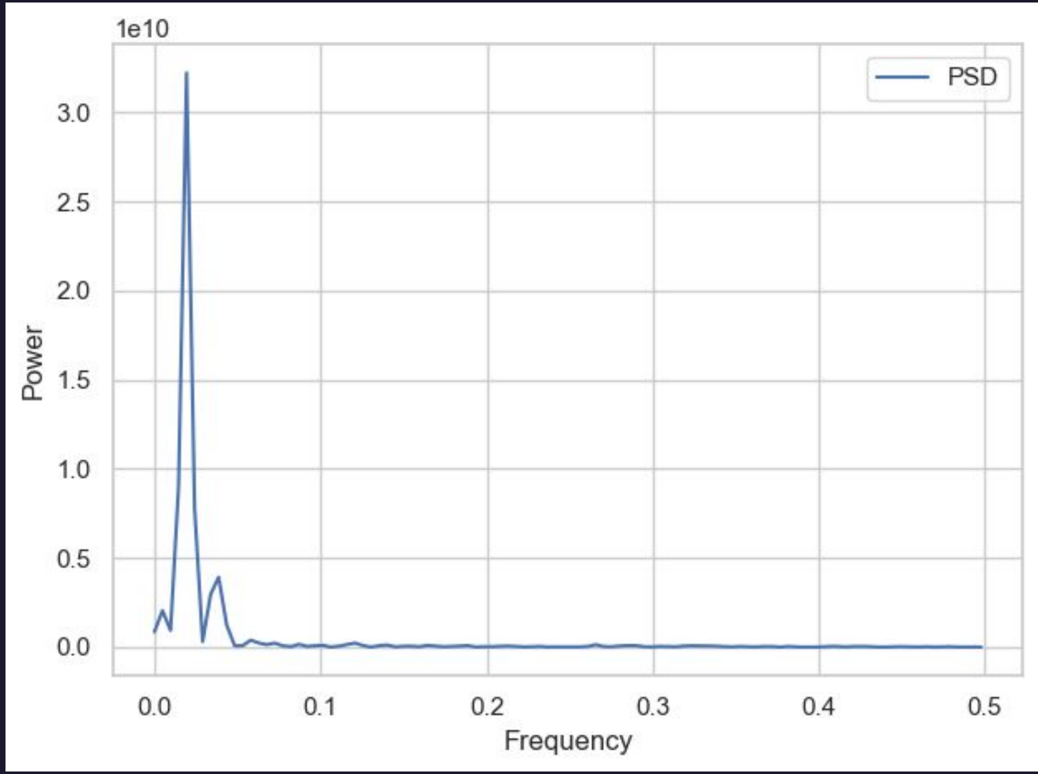
Pre-Modeling

Trend / Seasonality

Average Energy Use (kWh) - All Accounts

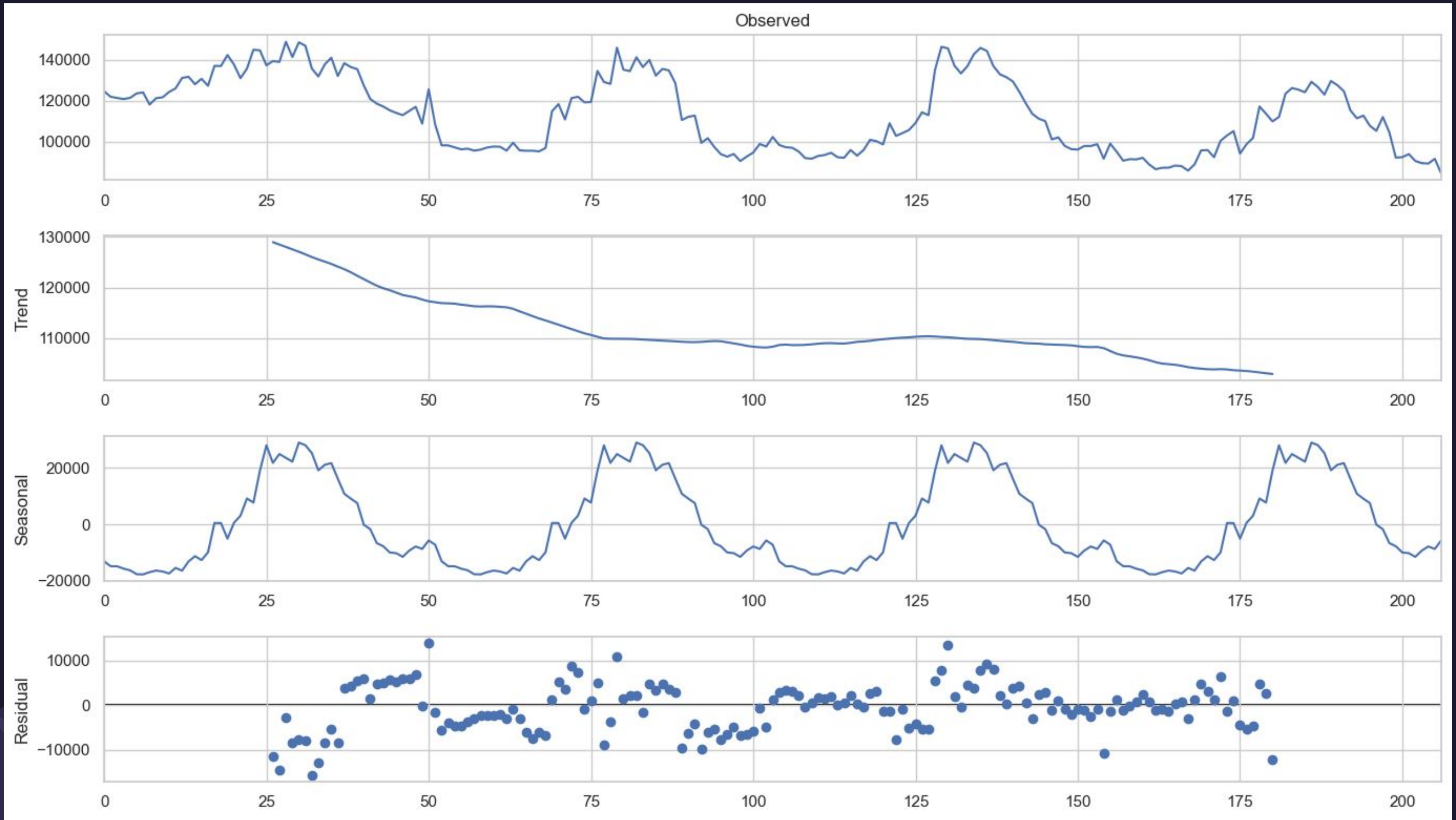


PSD of Average Energy Use



Time series plot and PSD indicate a first-order downward trend and yearly seasonality
Maximum power corresponds to a period of 51.75

Pre-Modeling
Seasonal Decomposition



- First order-trend indicates first-order differencing / integration component will be required
- Seasonal component has a yearly period
- Model parameters: SARIMA(0, 1, 1)(0, 1, 1)$_{52}$




Modeling

Modeling

- Algorithmic Solution Design
  - Fit an initial model with SARIMA
  - Evaluate MAPE on complete dataset and individual accounts
- Evaluation Metric
  - Mean Absolute Percentage Error (MAPE)
- Algorithmic Solution Finalization
  - Compare Predictions vs. Actual data to test the SARIMA model, MAPE calculations by account
- Future Modeling Work
  - Exogenous variables, development of SARIMAX model
  - Hyperparameters Selection of Algorithms
  - GridSearch to find the best parameters

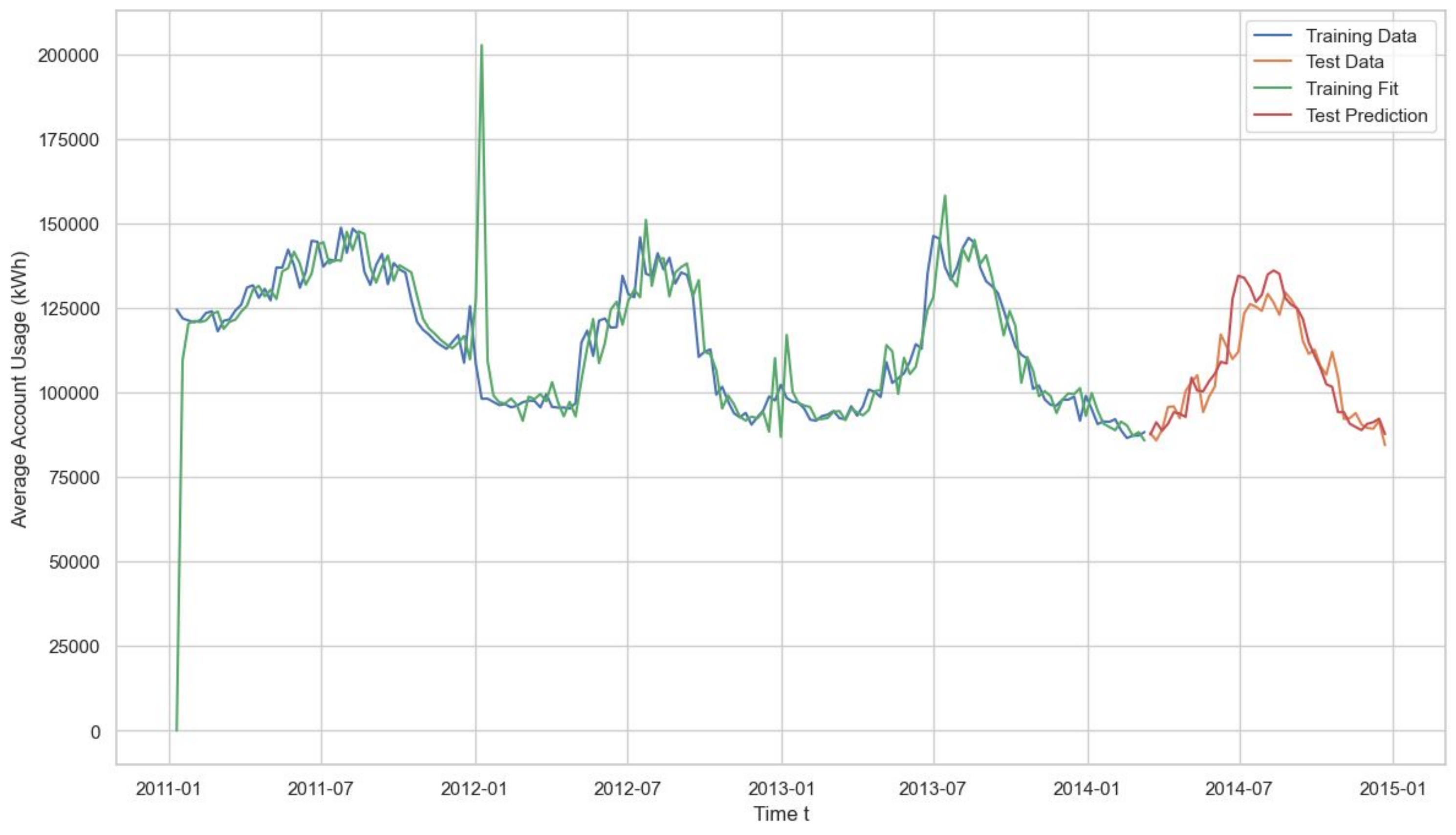



Results

Results: All Accounts Combined

First Pass at SARIMA(0, 1, 1)(0, 1, 1)$_{52}$

- Train-Test Split
  - Training Set: First 80% of Timeframe
  - Test Set: Last 20% of Timeframe
- Test Set MAPE = 4.5%

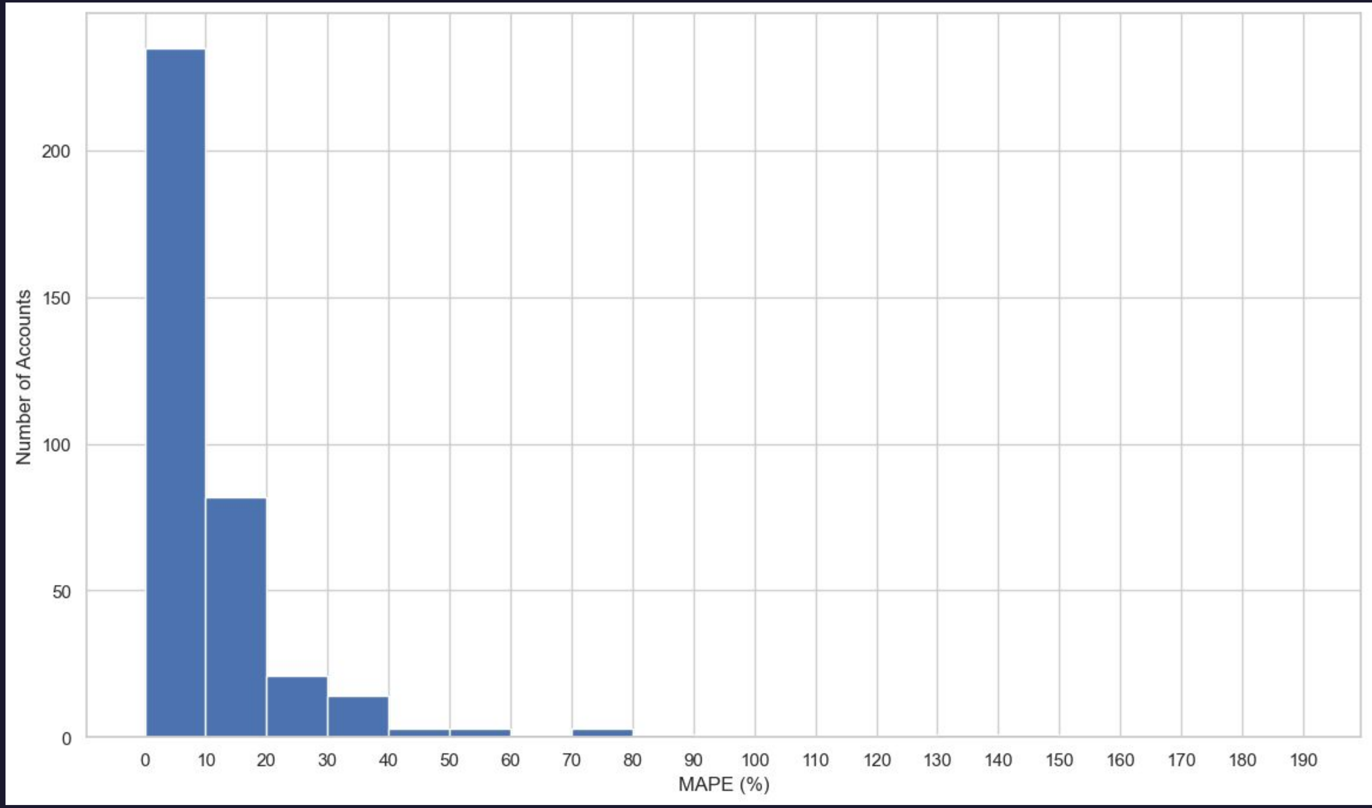


Results: Individual Accounts

First Pass at SARIMA(0, 1, 1)(0, 1, 1)$_{52}$

- Train-Test Split (Each Account)
  - Training Set: First 80% of Timeframe
  - Test Set: Last 20% of Timeframe
- Test Set MAPE varies greatly

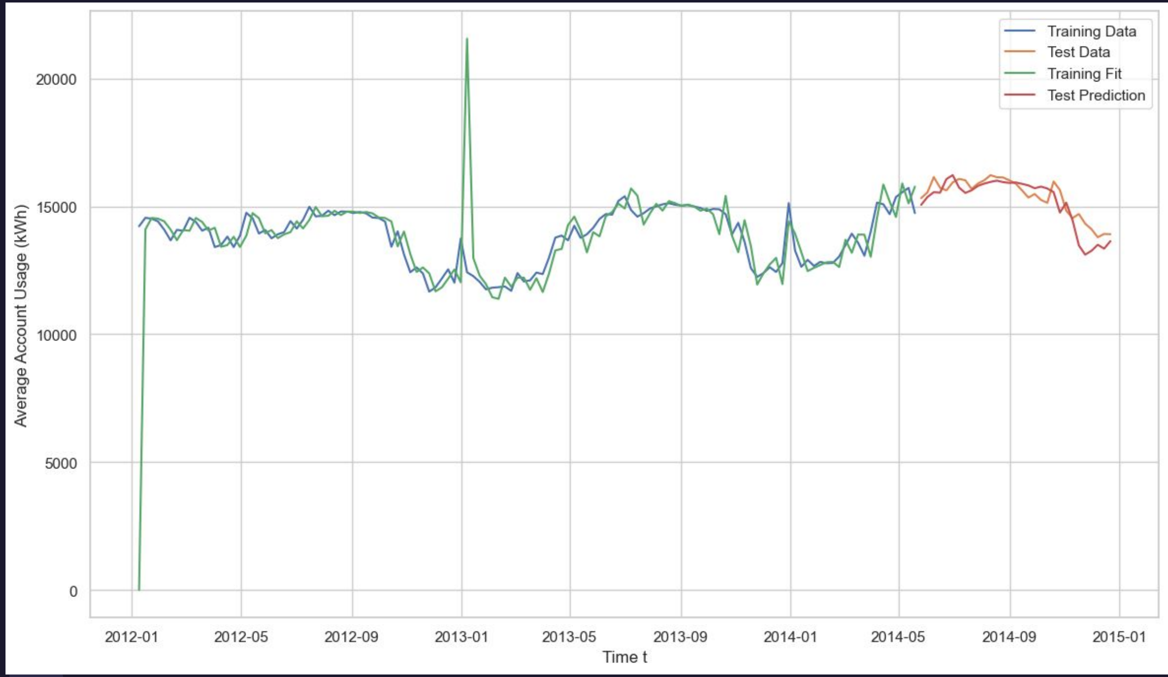
SARIMA Prediction MAPE: All Accounts



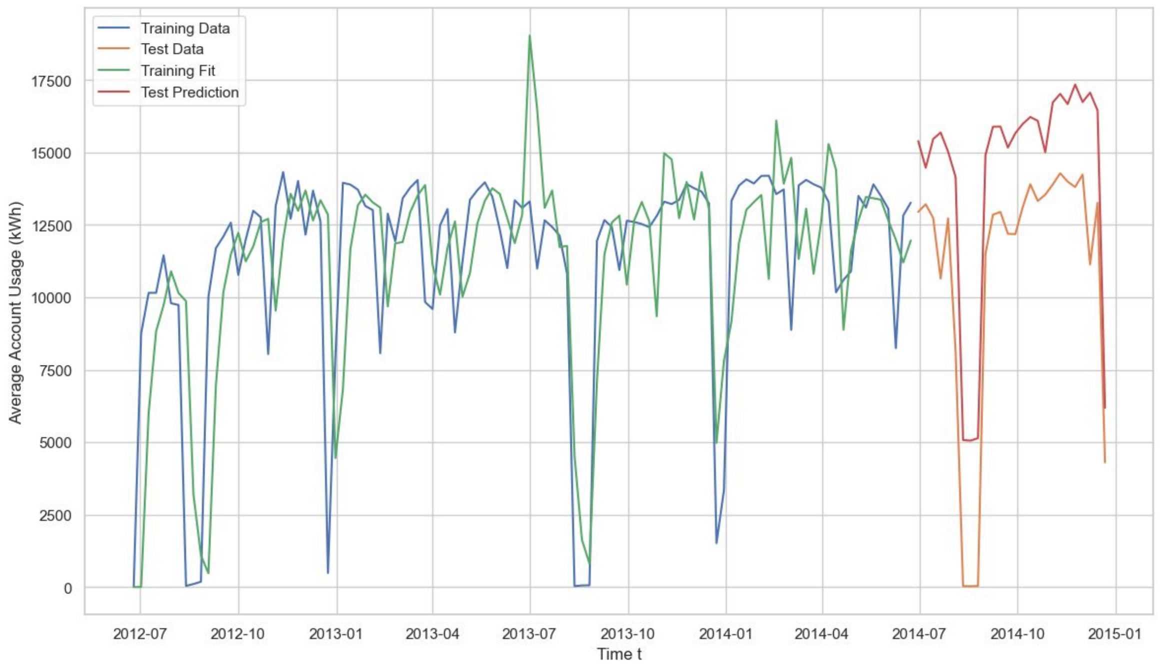
Results: Individual Accounts

Test Set MAPE varies greatly

Account MT_146 MAPE = 2.5%



Account MT_127 MAPE = 1700%*



*Skewed due to actual values close to zero, but not equal to zero

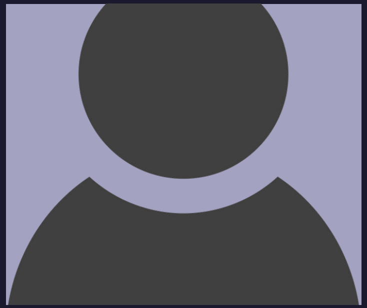
Results: Individual Accounts

Total MAPE: Assume *n* total accounts, each with *m* predictions

$$M = \frac{1}{n*m}\sum_{i=1}^{n}\sum_{j=1}^{m}\left|\frac{A_{ij} - F_{ij}}{A_{ij}}\right|$$

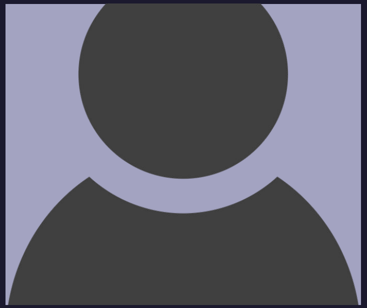
For this MAPE calculation, total test set MAPE = **15.6%**
Using the median instead of the mean, median MAPE = **6.5%**

Team



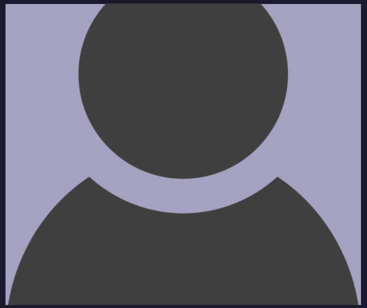
Kevin Taylor

MS Data Science Student



Nathaniel Ho

MS Data Science Student



Kelly Du

MS Data Science Student