Forecasting Duke University Utility Usage

Data+

Team Members: Ethan Swartzentruber, Grace Llewellyn, Shota Takeshima
Project Manager: Billy Carson
Project Sponsors: John Haws, Gagandeep Kaur, Wendy Lesesne, Casey Collins

Introduction

Working with Duke's Facilities Department, we have developed a tool which allows the cleaning and forecasting of utility usage data, for use in accurate budgeting and planning of new buildings. Our data consists of utility usage measurements for the last three years from various buildings on Duke's campus, typically at the 15-minute level. The tool provides an implementation of four forecasting methods: a naive averaging of previous years, SARIMA, Holt-Winters exponential smoothing, and a 1-Dimensional Convolutional Neural Network.

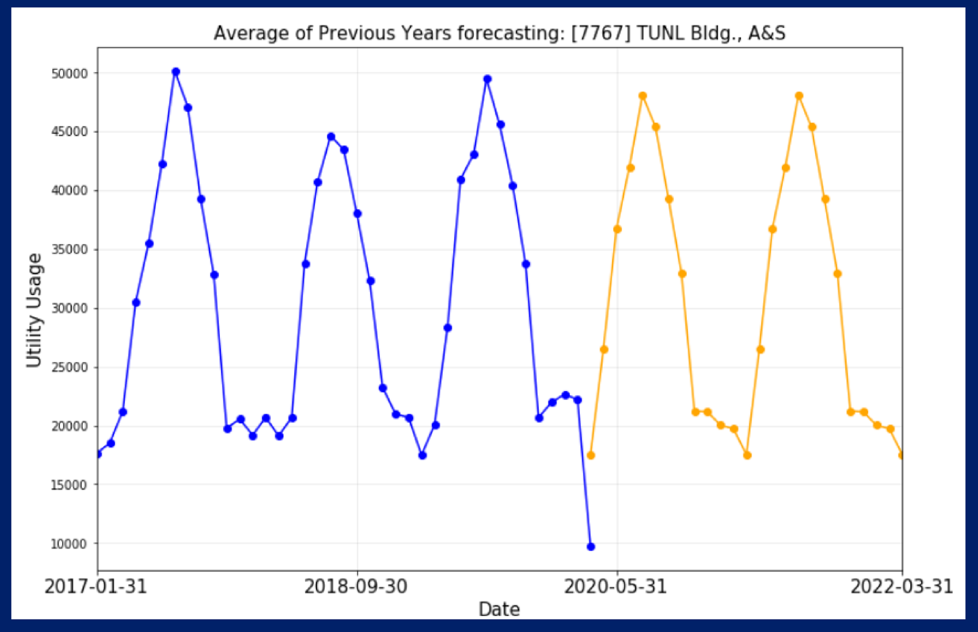
Forecasting

Training Data

Predictions

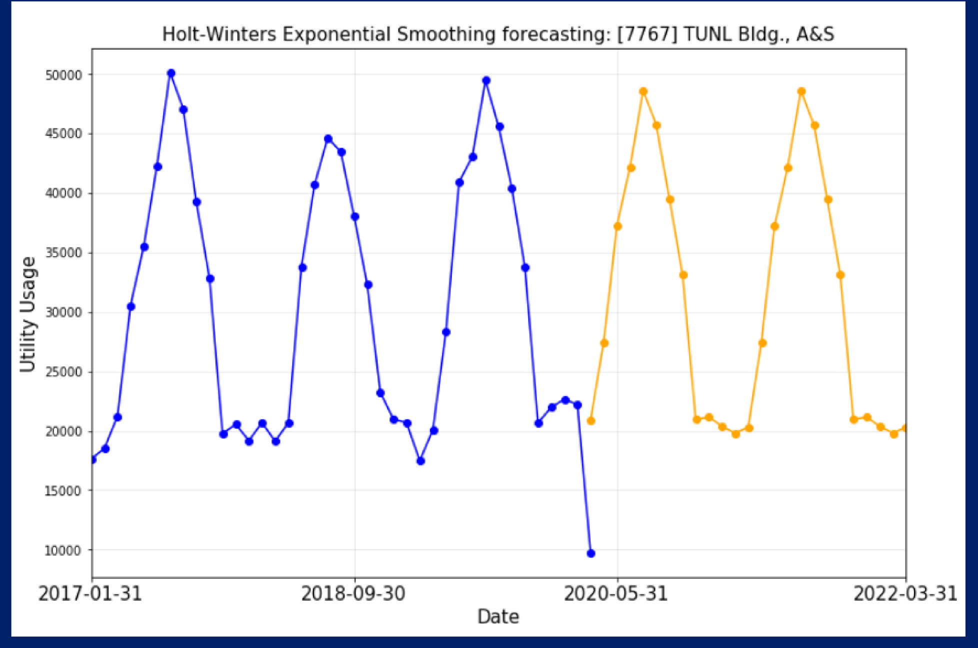
MAPE **M**ean **A**bsolute **P**ercentage **E**rror, the average absolute value of error percentage for each measurement

MASE **M**ean **A**bsolute **S**caled **E**rror, the ratio of the method's error to that of the naive model on the same data



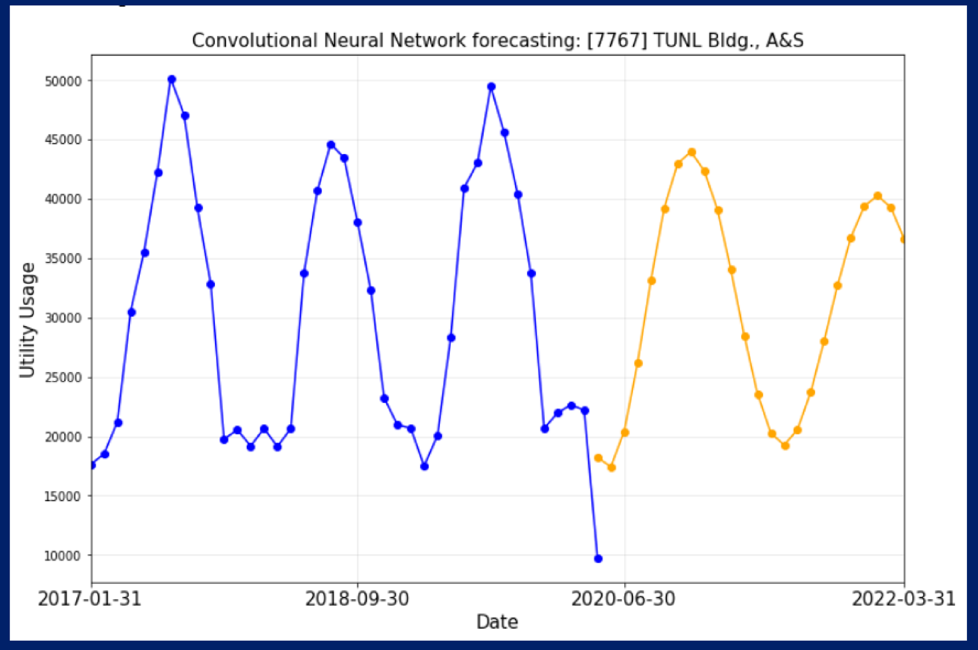
- Naive average of the same month in previous years of data.
- This is the "baseline error" used to calculate MASE. This is a common choice of naive model in time-series forecasting.

MAPE: 14.56% **MASE: 1.00**



- Holt-Winters method has exponential smoothing, meaning observations have exponentially decreasing effects on the prediction over time.
- This method is more sensitive to the seasonal nature of the data, and can account for the seasonality and "trend" (pattern between seasons) by decomposing the series.

MAPE: 15.28% **MASE: 1.92**

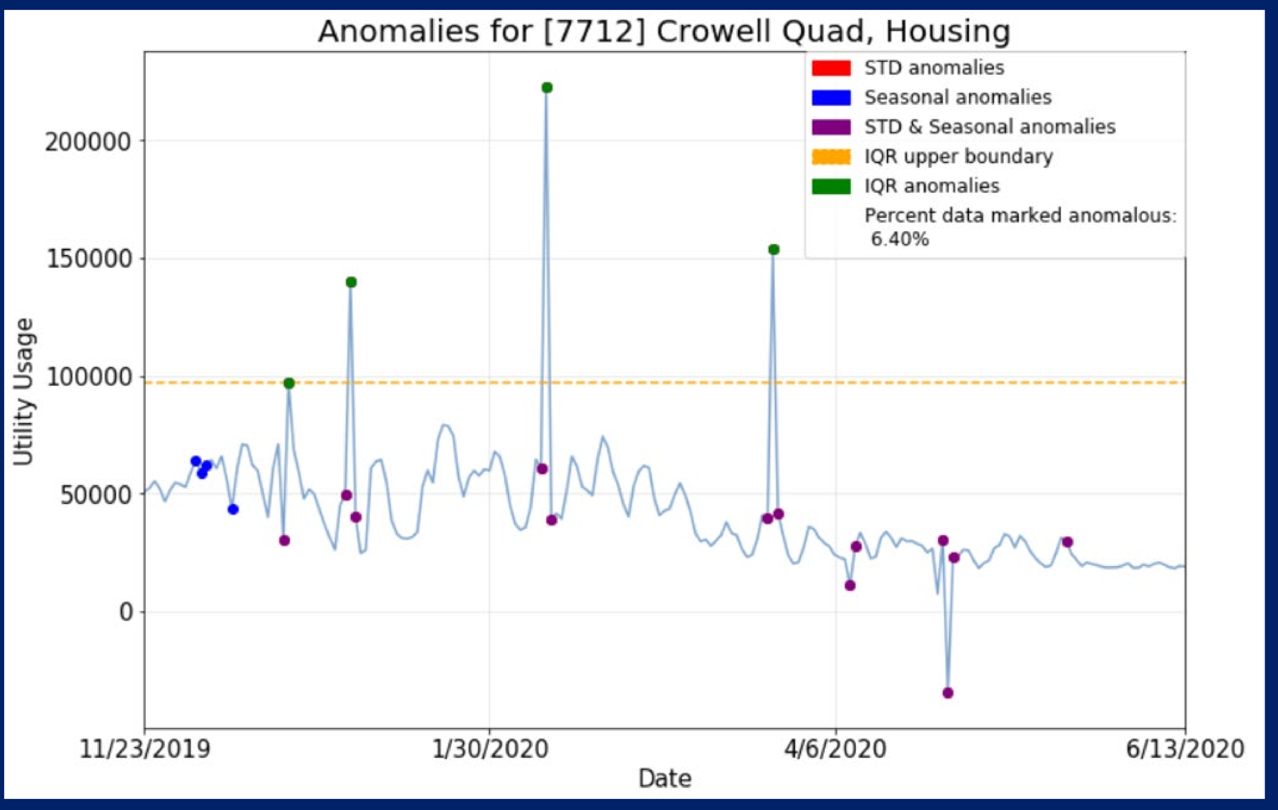


- NN(Neural Network) model predicts the target monthly usage using the last 12 months of usage measurements before it.
- In terms of predictive accuracy and computational load, Convolutional NN is the best performer compared to others we've tried, Long short-term memory (LSTM), and Gated Recurrent Unit(GRU).

MAPE: 19.20% **MASE: 2.17**

During the training process, we report error measurements by breaking data into test and training sets. Because NN-based methods use lagged measurements as predictors, the beginning of the data is not used, and they have very little data to train on - reported error values are typically high. This effect is less pronounced when training on the full data for actual predictions, but the flexibility of the method is still problematic for this smaller dataset. We expect that more complex models will begin to outperform naive ones with more data.

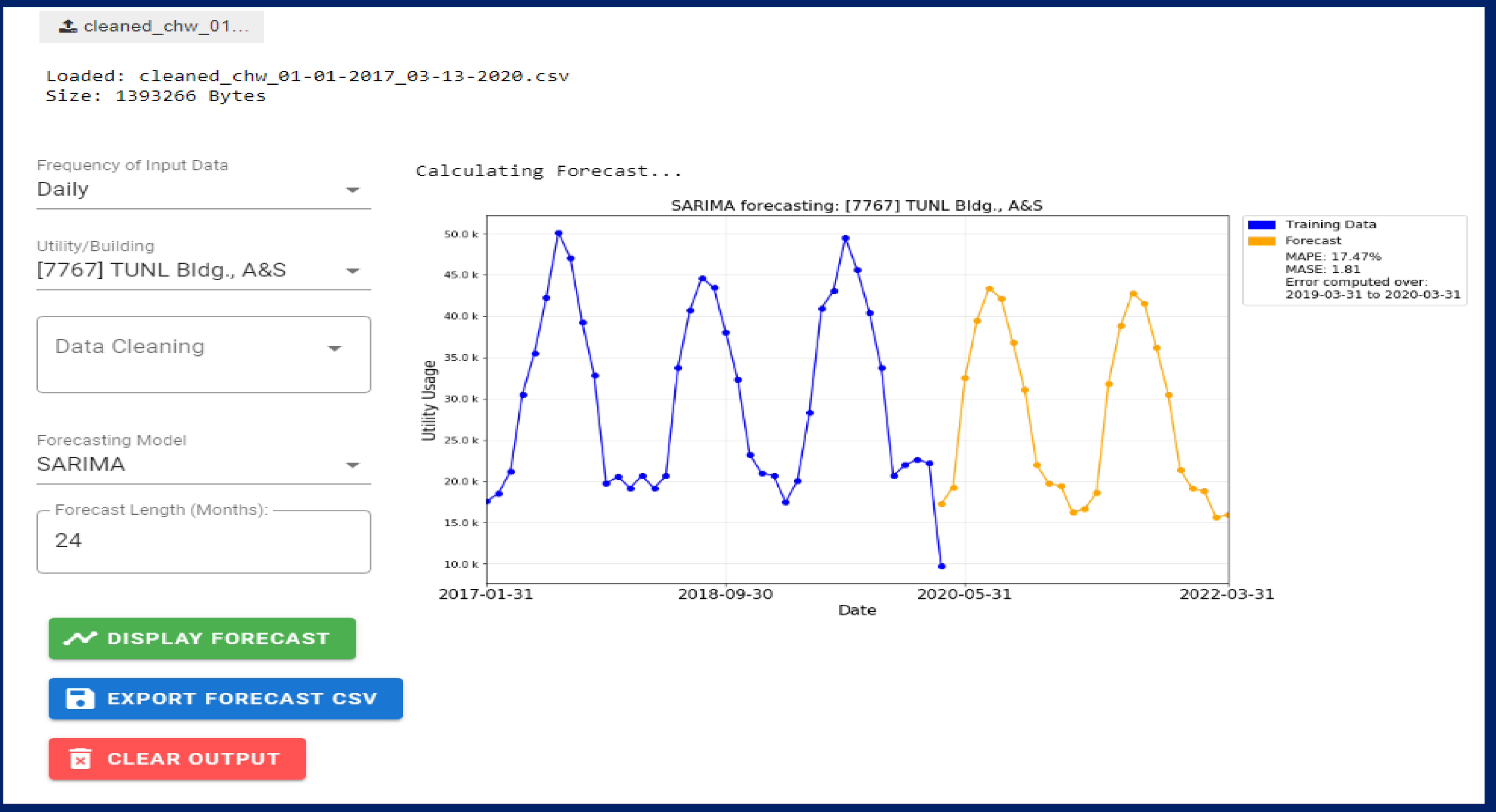
Data Cleaning



We applied three different anomaly detection methods to the data:

- The IQR (inter-quartile range) method produces a simple boundary which catches the tall, wide spikes in the data.
- The Seasonal ESD (extreme Studentized deviate test) method is conscious of the data's seasonal nature, and helps to remove measurements which are of a typical magnitude but are out of phase.
- The windowed STD (standard deviation) method helps catch smaller spikes, and allows for more consideration of locality than the other two methods.
- Data is imputed by averaging a data point's 10 nearest non-anomalous neighbors.

Tooling Approach



In order to ensure that Duke Facilities can run forecasting in a repeatable way as they collect more data in coming years, we developed a tool which allows the user to upload their own data, perform cleaning and forecasting operations, plot their results, and export the forecasted estimates to a CSV file.

Try it out! github.com/epswartz/utility-forecasting