Anomaly Detection for Energy Data

Fei Wang, Crystal Ding, Gautam Moogimane

Data Science Masters Capstone 2019, University of Washington

UNIVERSITY of WASHINGTON

JLL

INTRODUCTION

Background:

Our sponsor Jones Lang LaSalle(JLL) helps their clients closely monitor energy consumption in order to reduce energy waste. JLL takes corresponding actions when anomalous energy consumptions are detected. Anomaly can be reported due to a number of reasons like malfunctioning equipment, faulty construction or even manual entry errors. We are tasked to architect a system that can detect actual and actionable anomalous points for our sponsor to act upon more efficiently.

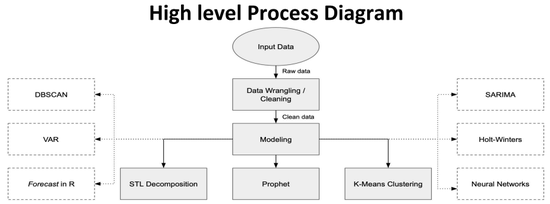
We used New York City Housing Authority (NYCHA) dataset to build models as proof of concept for our sponsor and delivered a data wrangling pipeline as well as three methods for detecting actual anomalous energy consumption data..

Objectives:

- Data Wrangling Pipeline: Provide an automated process to detect missing data at energy account level
- Outlier Detection Procedures: Architect automated, quick procedures to detect outlier values at energy account level that can be scaled out to thousands of accounts

APPROACH

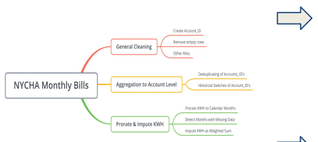
High level Process Diagram



- Raw input data from NYCHA was cleaned
- Multiple forecasting tools were tried, and some performed better than others
- STL decomposition, Prophet and K means clustering gave fast consistent results

DATA WRANGLING

- Summary of NYCHA public dataset:
  - Electricity Consumption Bill data from 2009 to 2018
  - 313,147 rows, 27 attributes
  - Attributes: Building_ID, Meter_Number, KWH_Consumption, Charges
- Result: 9 years of monthly time series data for ~2,000 accounts



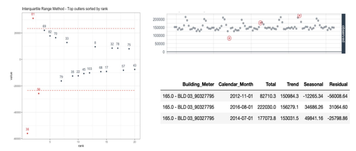
Building_Meter	Calendar_Month	Imputed_KWH
341.0 - BLD 04_7835072	2012-03-01	13496.774194
341.0 - BLD 04_7835072	2012-04-01	11652.413793
341.0 - BLD 04_7835072	2012-05-01	10779.586207
341.0 - BLD 04_7835072	2012-06-01	10240.000000
341.0 - BLD 04_7835072	2012-07-01	0.000000
341.0 - BLD 04_7835072	2012-08-01	12730.666667

MODELS

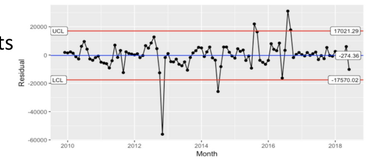
STL Time Series Decomposition

- Remove the trend and seasonal patterns of time series by decomposing it into 3 components
  - Trend, Seasonal, Residual
- Analyze the residual component via outlier detection methods
  - Individual Moving Range Chart (XmR Chart)
  - Interquartile Range (Tukey's box-and-whisker diagram) charts (3X and 6X)
  - Ensemble of 3 methods to determine & rank outliers

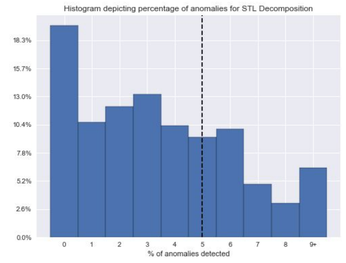
IQR Chart of Residuals



XmR Chart of Residuals



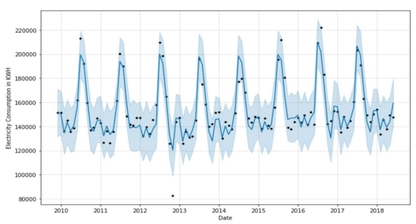
- ❖ The model was run for all the accounts with more than 50 data points
- ❖ A histogram was generated which highlights the % of anomalies detected for each account



Prophet

- Additive time series decomposition model that fits the signal as per the equation **y(t) <- g(t) + s(t) + h(t) + e**
- Trend, Seasonality, Holiday and Residual components are added to generate a forecast
- Holiday component is unique to Prophet, can be used to highlight days where spikes or drops are expected.

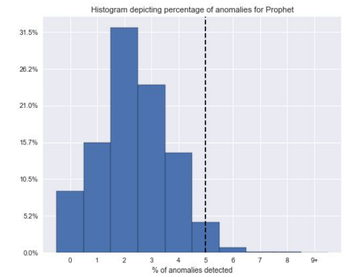
Sample output for one account with 3 outliers



Advantages of Prophet

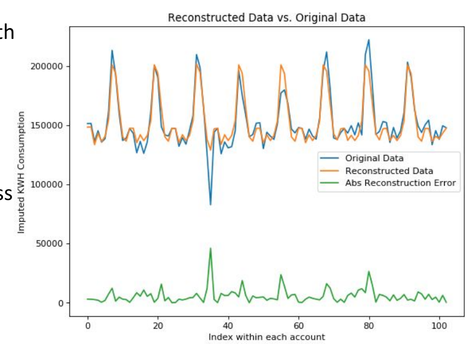
- No manual intervention means easy to automate
- Uncertainty interval generation is useful for outlier detection.
- Robust to missing data.

- ❖ Majority have < 5% total anomalies

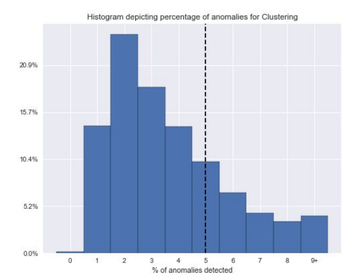


Time Series Clustering

- Segment time series trends into waveforms of 8 data points each with sliding window of 1 step
- Perform K-Means clustering (k=12) using squared euclidean distance to find nearest centroids
- Reconstruct time series trend by averaging nearest centroids that pass through each data points
- Detect anomalies by analysing reconstruction error greater than certain threshold (99th percentile)



- ❖ % of anomalies detected decreases exponentially in clustering model



CONCLUSION

Comparison of Models:

Methods	Runtime	Interpretability	Robustness to Missing Values	Implementation Ease
Decomposition	Low(~0.1s)	High	High	Low
Prophet	Low(~0.1s)	Low	High	High
Clustering	Medium(~0.2s)	High	Medium	Medium

Recommendations:

- If there are enough data points (50+) and % of missing values is low (<10%)
  - If prefer auto-tuning & high scalability, use Prophet
  - If prefer manual configuration, use STL decomposition
  - If prefer high interpretability, use STL decomposition or clustering
  - If data frequency is at monthly level or more granular, clustering technique works better
- If there is not enough data points (<50), apply IQR or XmR chart to detect outliers

CHALLENGES

- Biggest challenge that we faced was figuring out how to tune the models and evaluate its performance.
- Lack of any non anomalous training data added layer of complexity for outlier detection.
- No access to actual client data to implement multivariate models.

REFERENCES

- *Robert B. Cleveland, William S. Cleveland, Jean E. McRae, and Irma Terpenning. STL: a seasonal-trend decomposition procedure based on loess. Journal of Official Statistics, 6(1):3–73, 1990.*
- *F. Gullo, G. Ponti, A. Tagarelli, G. Tradigo, P. Veltri, A time series approach for clustering mass spectrometry data, J. Comput. Sci. 3 (5) (2011) 344–355. 2010.*
- *Taylor SJ, Letham B. 2017. Forecasting at scale.*

ACKNOWLEDGEMENTS

Our team would like to express our appreciation to Dr.Megan Hazen & Zhe Liu, for their guidelines throughout the last two quarters. We would also like to thank Linnea Paton from the project sponsor JLL, for her continued help and support to make the project a success.