Anomaly Detection in Multi-Seasonal Time Series Data

Student: Ashton Williams

Student Email: williams.1525@wright.edu

Faculty: Dr. Soon Chung

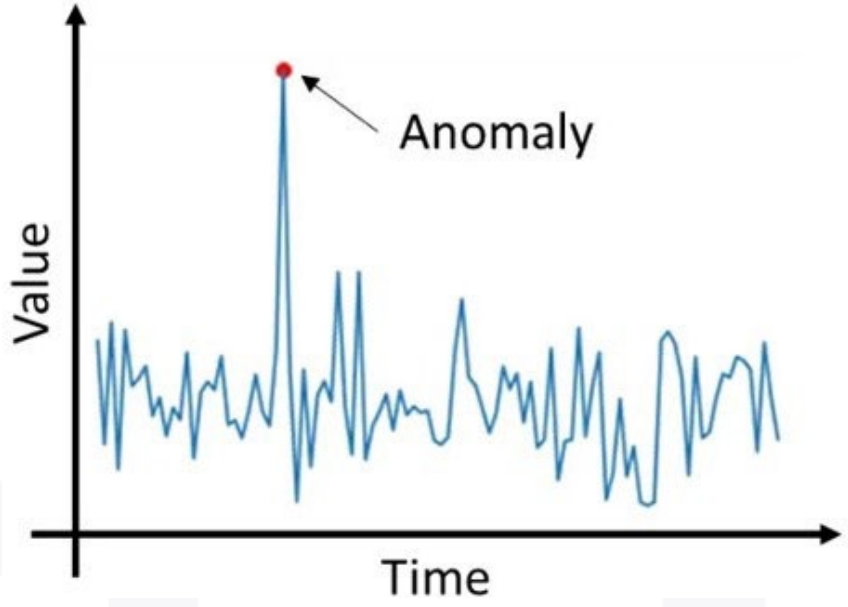
Faculty Email: soon.chung@wright.edu

AFRL Sponsor: Dr. Vincent Schmidt

AFRL Directorate: RH

PA #: AFRL-2023-5063

Introduction



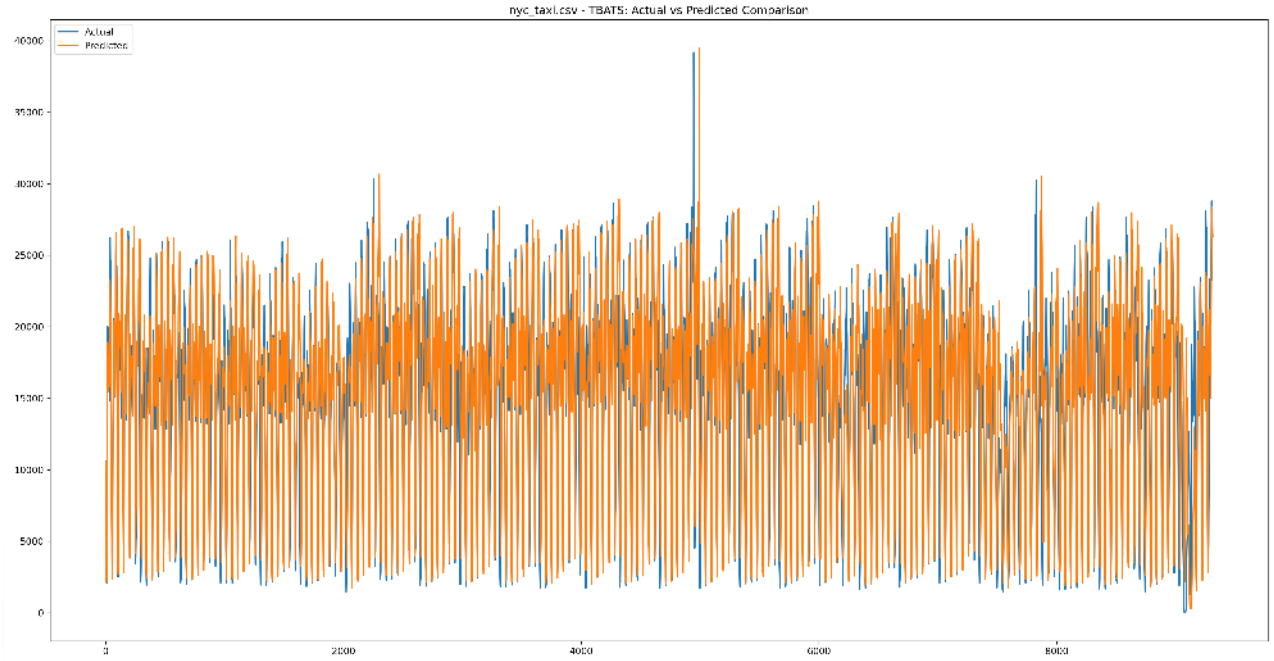
- A **time series** is a sequence of data points indexed in time order.
  - X-axis is time, Y-axis is the data value
  - Used to track change of a value over time
  - Usually at equally spaced points in time
- **Anomalies** (or outliers) are datapoints that significantly deviate from their expected value (or predicted value).
  - Anomalies contain useful information about the abnormal characteristics in a dataset
- **Seasonalities** are cycles that repeat regularly over a period
  - Hourly, Daily, weekly, monthly, yearly, etc.
- Most of today's big data are time series that contain both **anomalies** and **multiple seasonalities**
- **Early** and **accurate** detection of anomalies allow businesses to mitigate harmful effects
- Examples:
  - A bank can detect abnormal spending behavior and quickly lock your account
  - A hospital can detect abnormal results in medical data and notify professionals before it's too late.
- A model that can take advantage of **every seasonality** in a dataset can **improve** its anomaly detection accuracy
- Most common models today are only suited for **single seasonality**

Our Approach

- Developed a new multi-seasonal model for anomaly detection in time series data called the multi-SARIMA
- Extends the popular Seasonal Autoregressive Integrated Moving Average (SARIMA) model
- Utilizes a time series' multiple pre-determined seasonal trends
- Increases anomaly detection accuracy even more than the original SARIMA
- Requires more processing time
- Used our multi-seasonal model as the second step in the Two-Step approach
- The Two-Step algorithm consists of two steps:
  - Step 1: simpler model that labels data fast with less accuracy (MA, SIMA)
  - Step 2: complex model that labels data accurately but requires more time (SARIMA, TBATS, multi-SARIMA)
- Goal of the Two-Step approach is to reduce the false positive rate

How Anomaly Detection Works

- For real-time time series data, data forecasting models are used for anomaly detection
- Two parts to anomaly detection:
  - Data forecasts
  - Anomaly labeling



- Data forecasts are used to compare forecasted values to actual values (shown in image)
- An anomaly labeling metric called an “anomaly score” is calculated and used to determine if a given point is anomalous based on the comparison

The Models

- We implemented numerous popular models to compare the multi-SARIMA to including:
- Moving Average (MA)
  - Predicts future values as a **weighted sum** of **lagged residuals**
- Seasonal Integrated Moving Average (SIMA)
  - Extension of the MA model by considering **one seasonal component**
- Seasonal Autoregressive Integrated Moving Average (SARIMA)
  - Extension of the autoregressive integrated moving average (ARIMA) model
  - Incorporates **one** seasonal component into its forecasts
  - One of the **best** and **most common** time series forecasting models
- Hierarchical Temporal Memory (HTM)
  - a neural network-based machine learning algorithm derived from neuroscience
- Trigonometric seasonality, Box-Cox transformation, ARMA errors, Trend and Seasonal components (TBATS)
  - Currently one of the best and most common **multi-seasonal** time series forecasting models

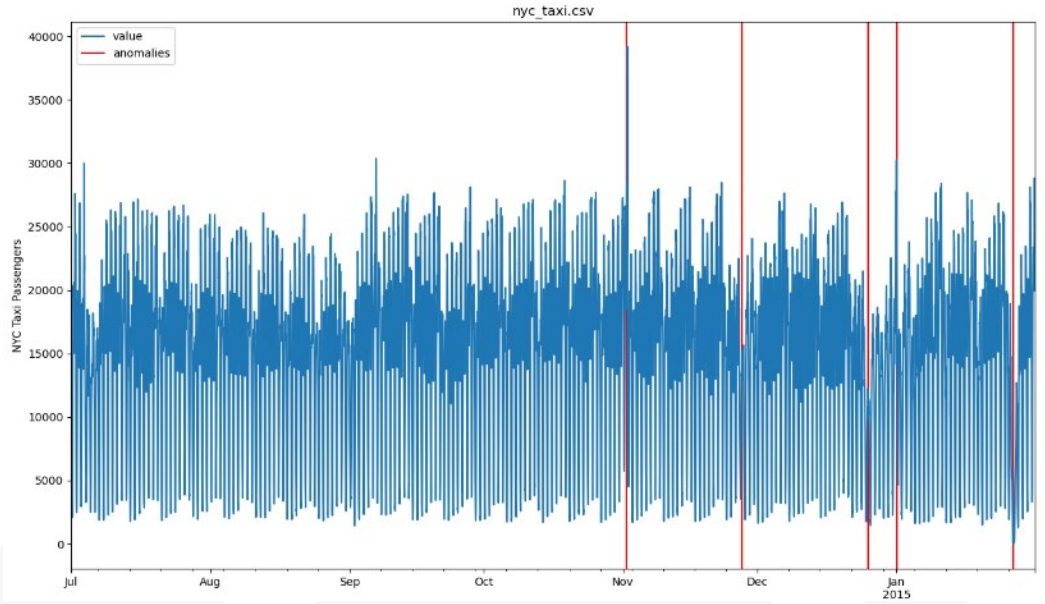
The Multi-SARIMA

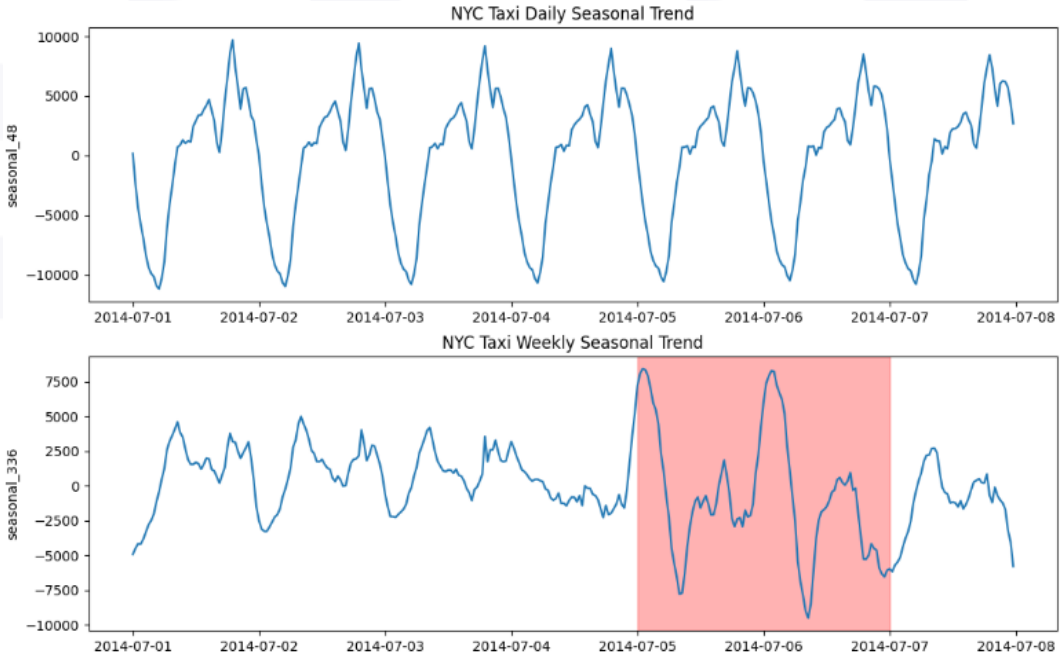
- Derived by extending the original SARIMA equation
- Extends the original SARIMA model by adding **another seasonality**
- Denoted by SARIMA$(p_1, d_1, q_1)_{m_1} * (p_2, d_2, q_2)_{m_2}$
- It predicts *X* by modeling **one** seasonally differenced series $\nabla_{m_2}^{d_2} X_t$ with **two** SARIMA$(p, d, q)_m$ models
  - The first SARIMA model is trained on three iterations of the **shorter** seasonal trend (m_1)
  - The second SARIMA model is trained on three iterations of the **longer** seasonal trend (m_2)
- Applies values from both models to the multi-SARIMA equation to get the prediction *X* at time *t*
- Expect multi-SARIMA to perform well when at least two strong seasonal components are present

Dataset Overview

- Evaluated all models on two Numenta Anomaly Benchmark (NAB) datasets and a synthetic one:
  - NYC Taxi
  - Synthetic Dataset 3
  - HotGym
- All 3 datasets are **univariate** time series datasets with **two meaningful seasonal trends** and **hand-labeled anomalies**
- Utilized MSTL to decompose our datasets into their daily and weekly seasonal components to confirm their seasonalities

Datasets	Interval	Seasonality #1	Seasonality #2	Anomalies	Total Values
NYC Taxi	30 min	Daily	Weekly	5	10,320
Synthetic Dataset 3	1 hour	Daily	Weekly	5	8,664
HotGym	1 hour	Daily	Weekly	5	3,887





Red area depicts the weekend

Single-Step Results

- Multi-SARIMA has more TP while maintaining fewer FP than SARIMA for every real dataset
- Multi-SARIMA had either the best or second-best results for every dataset
- Multi-SARIMA has the highest runtime, but:
  - Is the only model that combines results from two models
  - Trains over the two seasonal periods (weekly & daily)
- Multi-SARIMA is only algorithm that achieved the same TP rate (4/5) as HTM for the NYC Taxi dataset
- Multi-SARIMA is one of the only models to perform better than HTM for the HotGym Dataset
- Multi-seasonal algorithms performed the best for the synthetic dataset (TBATS/Multi-SARIMA)
- TBATS had either more or the same TP as SARIMA while maintaining less FP for every dataset

Top 2 detectors of each dataset highlighted in green

	NYC_Taxi Dataset				Synthetic Dataset 3				HotGym Dataset			
Detector	TP	FP	FN	Runtime (sec)	TP	FP	FN	Runtime (sec)	TP	FP	FN	Runtime (sec)
MA	2	654	3	2.002	2	3	3	2.084	4	195	1	0.722
SIMA	3	1587	2	3.444	5	105	0	2.349	2	591	3	0.965
SARIMA	2	1464	3	3.682	5	86	0	3.297	2	443	3	1.44
Multi-SARIMA	4	1425	1	1443.402	5	9	0	73.737	4	170	1	268.692
TBATS	3	1391	2	73.428	5	4	0	67.889	2	399	3	32.582
HTM	4	178	1	46.71	1	1	4	30.93	2	121	3	16.518

Two-Step Results

- All Two-Step models have less or the same FPs than their standalone first step from table 2
- Most Two-Step models have less FPs than their standalone second step results
- Multi-SARIMA as second step produced significantly less FPs than the original SARIMA as the second step for every dataset
- Only model that produced less FPs than the multi-SARIMA is TBATS for the Synthetic Dataset
- Second-step models had improved runtime
- TBATS does better as the second step than the original SARIMA for every dataset

Top 2 detectors of each dataset highlighted in green

	NYC_Taxi Dataset				Synthetic Dataset 3				HotGym Dataset			
Detector	TP	FP	FN	Runtime (sec)	TP	FP	FN	Runtime (sec)	TP	FP	FN	Runtime (sec)
MA + SARIMA	2	131	3	2.753	2	3	3	3.23	3	120	2	1.529
SIMA + SARIMA	3	1072	2	3.627	5	91	0	3.207	2	547	3	1.625
MA + Multi-SARIMA	2	93	3	787.783	2	1	3	69.95	3	53	2	94.329
SIMA + Multi-SARIMA	3	475	2	697.393	5	50	0	70.612	2	220	3	92.137
MA + TBATS	2	122	3	76.867	2	0	3	45.359	3	68	2	36.836
SIMA + TBATS	3	1156	2	79.023	5	4	0	47.08	2	306	3	35.939

Conclusion

- Multi-SARIMA **improves** SARIMA model by including multiple seasonal components
- Multi-SARIMA produced **better** anomaly detection results than the original SARIMA for every dataset we tested
- In most cases, multi-SARIMA **outperformed** every model including HTM and TBATS
- Showed multi-SARIMA is the **optimal model** for anomaly detection accuracy with the datasets and models we included
- Showed multi-SARIMA produces the **best** results when used as the second step in the Two-Step approach
- Showcased anomaly **detection potential** of multi-seasonal models like TBATS