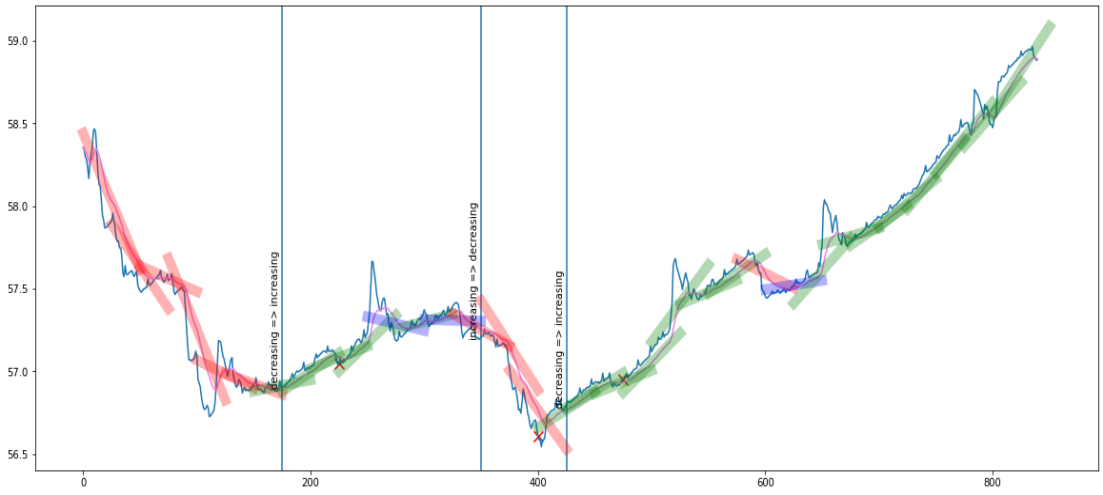


Trend, Seasonality and Change point Detection

A REPORT

ON

"Trend, Seasonality, and Change point Detection"

BY

Praneet Sai Madhu Surabhi ID: 2019A7PS0060H B.E Computer Science

AT

Greendeck Cliff.ai, Indore

A Practice School-I Station of

BIRLA INSTITUTE OF TECHNOLOGY AND SCIENCE, PILANI

June, 2021

A REPORT

ON

"Trend, Seasonality, and Change point Detection "

BY

Praneet Sai Madhu Surabhi ID: 2019A7PS0060H B.E Computer Science

Prepared in partial fulfilment of the Practice School-I Course

BITS F221

AT

Greendeck Cliff.ai, Indore

A Practice School-I Station of

BIRLA INSTITUTE OF TECHNOLOGY AND SCIENCE, PILANI

June, 2021

ACKNOWLEDGEMENT

I would like to extend my gratitude to Professor Tanmay Tulsidas Verlekar for his valuable guidance during the preparation of this report.

I would also like to thank my mentor from Greendeck Mr. Arpitanshu who spared time out of his busy schedule in guiding me through the project.

BIRLA INSTITUTE OF TECHNOLOGY AND SCIENCE, PILANI

(RAJASTHAN)

Practice School Division

Station: Greendeck Cliff.ai

Duration: 8 weeks **Date of Start**: 31st May 2021

Date of submission: 22nd July 2021

Title: Trend, Seasonality, and Change point Detection

ID Nos./Name(s)/Discipline(s) of the student(s):

Praneet Sai Madhu Surabhi ID: 2019A7PS0060H

Name of the Mentor: Arpitanshu, ML Engineer

Name of the PS Faculty: Prof. Tanmay Tulsidas Verlekar

Key Words: Metric, Machine Learning, Python, Time Series, Trend, Seasonality, Change points, Statistics, Mann-Kendall, ACF

Project Area: Machine Learning

Abstract: The purpose of this project is to deploy an online method for the detection of Trend, Seasonality and Change points in time series. It allows one to assess the factors which influence various metrics over time. The goal is to detect points of change in trend or seasonality and alert users the same. Finally, we need to create production ready code for all the methods.

Praneet

Signature of Student
Date: 22/07/2021

Signature of PS Faculty
Date:

Table Of Contents

1. Introduction

Cliff.ai is a business reliability platform. It monitors and tracks unexpected changes(anomalies) in metrics without creating dashboards or complex pipelines by running ML algorithms. With Cliff, one can instantly create data pipelines and monitor their metrics. On detection, an alert can be sent through various models such as Email, text, SMS, or Slack. Cliff also understands the normal behavior of a metric (instead of setting a threshold) and notifies instantly.

Trend detection and analysis helps firms understand strengths or weakness of a particular organization. It simply offers a method for businesses to project outcomes. It can also be used for failure analysis and as early warning indicator of problems. It can also be used to identify what caused the trend to increase or decrease or no trend.

Seasonality is a way to find periodicity in a time series. A seasonal pattern is affected by seasonal factors such as the time of the year or the day of the week. It helps us understand why there is such a pattern which repeats periodically.

This report is an attempt to implement trend detection methods as well as its improvement and application in businesses. It begins with explaining various terms of the project and we go further into the main algorithms. The report concludes by mentioning various concepts learned during this project.

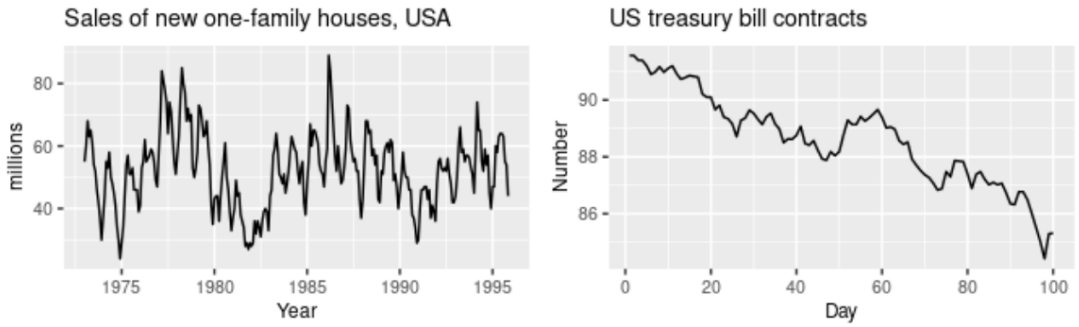
2. Useful Terms

Time Series: A time series is a series of data points indexed in time order. It is a sequence taken at successive equally spaced points in time. It is a discrete-time data. It allows one to see what factors influence certain variables from period to period. A time series can be taken on any variable that changes over time.

Trend: A *trend* exists in a time series where there is long term increase or decrease in data. It does not have to be a linear change.

Seasonality: A seasonal pattern occurs when a time series is affected by seasonal factors such as the time of the year or the day of the week. Seasonality has a fixed and known frequency.

Cyclic: A cycle occurs in a time series when the data exhibits rise and falls that are not of a fixed frequency. It occurs due to economic conditions. The duration of these fluctuations is usually at least 2 years.



1. Sales of new one-family houses, USA: The metric being measured here shows strong seasonality within each year, as well as cyclic behavior with a period of about 6-10 years. There is no trend in the given data.
2. US treasury bill contracts: It shows results from 100 consecutive trading days in 1981. Here there is no seasonality but an obvious downward trend.

3.Trend Detection

Mann-Kendall Test:

The Mann-Kendall test (MK test) is used to analyze data collected over time for consistently increasing or decreasing trends in Y values. It is a non-parametric test and hence works for all distributions but it should not have serial correlation. It can find trends with only few data points but the test would high probability of not finding a trend when one would be present if more points were provided.

1. The null hypothesis (H0) is that the data comes from a population with independent realizations and are identically distributed i.e., there is no monotonic trend in the time series.

2. The alternate hypothesis (HA) is that the data follows a monotonic trend (positive, negative or non-null).

The Mann-Kendall test statistic is calculated as:

$$S = \sum_{k=1}^{n-1} \sum_{j=k+1}^{n} \operatorname{sgn}(X_j - X_k)$$

with

$$\operatorname{sgn}(x) = \begin{cases} 1 & \text{if} \quad x > 0 \\ 0 & \text{if} \quad x = 0 \\ -1 & \text{if} \quad x < 0 \end{cases}$$

Working and in-depth analysis of Mann-Kendall test can be found here.

The python implementation of Mann-Kendall is a package called 'pymannkendall' which has multiple test functions. They similar input parameters. I tried to use the Regional MK Test

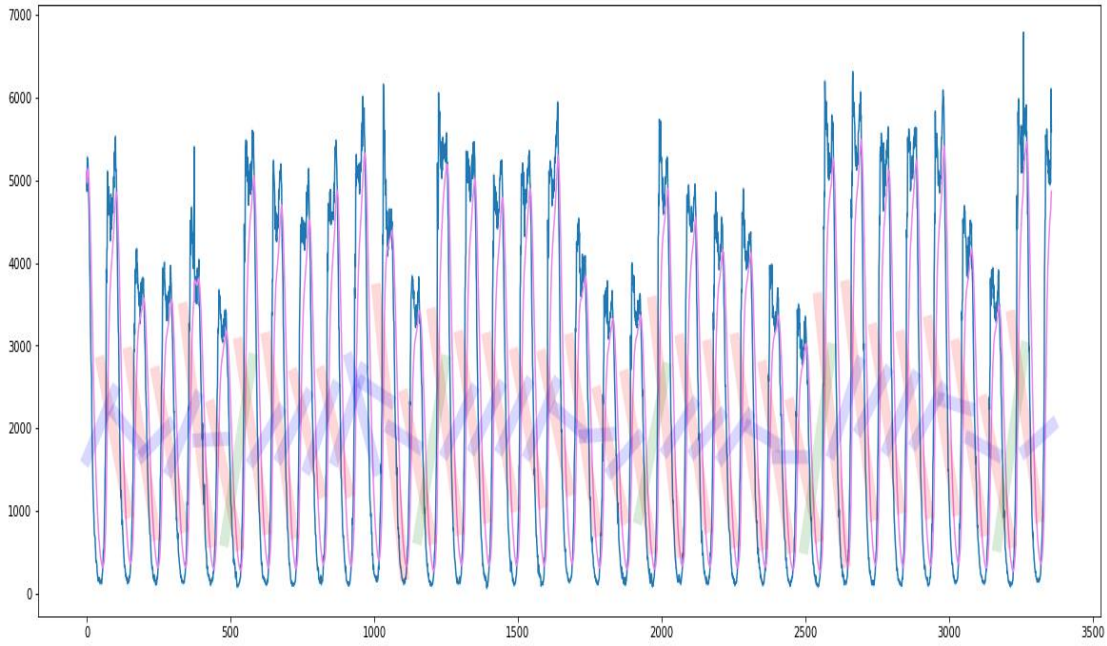
(regional_test) based on Hirsch (1982) proposed seasonal MK test, Helsel, D.R. and Frams, L.M, (2006) suggest a regional MK test to calculate overall trend in a regional scale.

The dataset which is a time series is first divided equally into multiple windows and the regional MK test is applied. Each window returns 9 values.

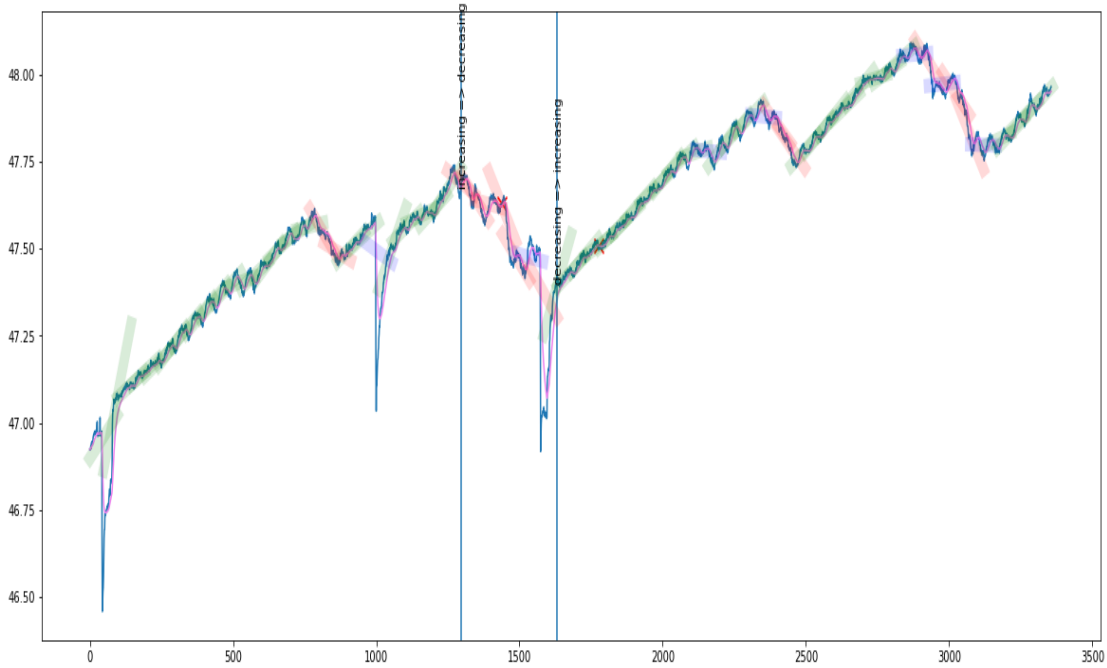
- trend: tells the trend (increasing, decreasing or no trend)
- h: True (if trend is present) or False (if the trend is absence)
- p: p-value of the significance test
- z: normalized test statistics
- Tau: Kendall Tau
- s: Mann-Kendal's score
- var_s: Variance S
- slope: Theil-Sen estimator/slope
- intercept: intercept of Kendall-Theil Robust Line, for seasonal test, full period cycle considers as unit time step.

Then we label each window on the graph as green for increasing trend, red for decreasing trend and blue for no trend.

The final goal was to detect the point where this trend is changing. Initially I fixed the number of continuous windows as 3. If there is a change in trend and it continues up to three windows then such a point needs to be notified.



This plot has no points of change in trend. In fact, it has no overall trend.



This graph has 2 points where the trend changes which is given by the 2 vertical blue lines.

UCB Algorithm (Upper Confidence Bound):

We needed a robust way to detect trend change points rather than fixing the number of windows. I tried to solve it using the UCB algorithm which is based on the Multi-Armed Gambit Problem.

Rather than performing exploration by simply selecting an arbitrary action, chosen with a probability that remains constant, the UCB algorithm changes its exploration-exploitation balance as it gathers more knowledge of the environment. It moves from being primarily focused on exploration, when actions that have been tried the least are preferred, to instead concentrate on exploitation, selecting the action with the highest estimated reward.

With UCB, 'A_t', the action chosen at time step 't', is given by:

$$A_t = \arg\max_a \left[Q_t(a) + c\sqrt{\frac{\log t}{N_t(a)}} \right]$$

where,

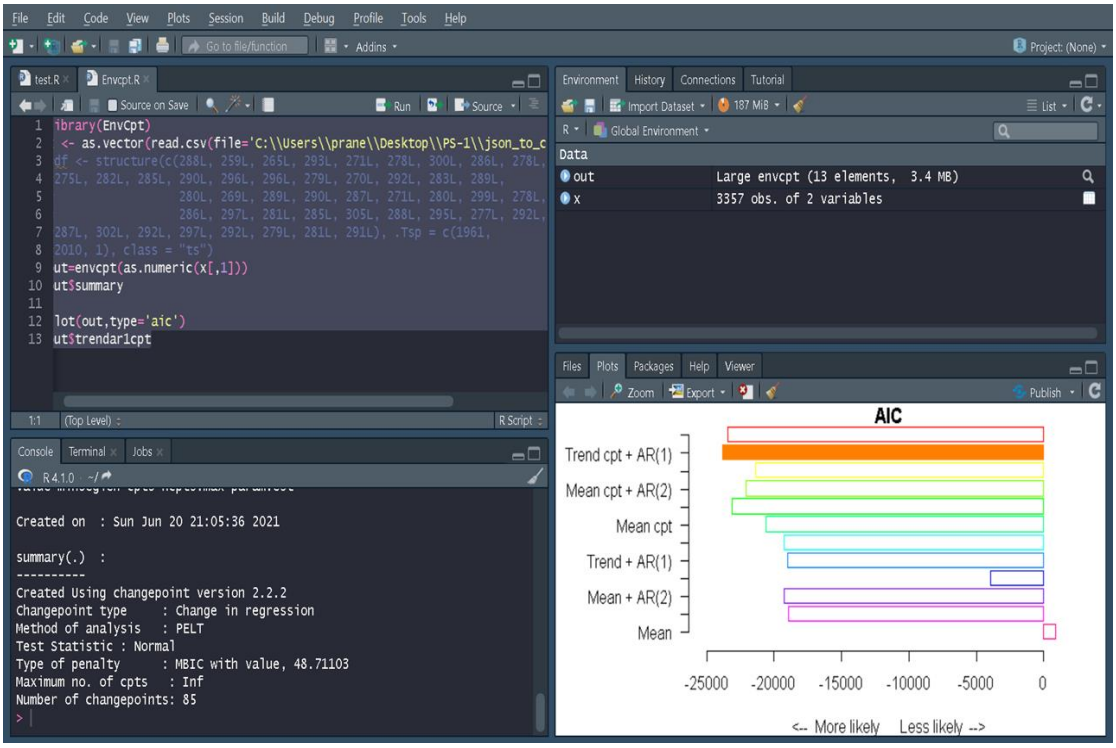
- $Q_t(a)$ is the estimated value of action 'a' at time step 't'.
- $N_t(a)$ is the number of times that action 'a' has been selected, prior to time 't'.
- 'c' is a confidence value that controls the level of exploration.

Upon implementing it does not seem to solve the problem in hand and hence I moved on to R language which had better and promising libraries.

R Language:

I tried implement a method from the package 'trendsegmentR' which is called trendsegment which detects linear trend changes and point anomalies for univariate time series. The function estimates the number and locations of change points in linear trend of noisy data. It may also contain some anomalies if any. It takes three steps- Tail-Greedy Unbalanced Wavelet (TGUW) transform, thresholding and inverse TGUW transform.

Then I implemented another package called EnvCpt which has a function called envcpt. It evaluates up to 12 different models and returns the model fits and summary of the likelihood for each model. This method gave me some promising results which can be improved for obtaining the ideal solution.



4. Learning Outcomes

- Python Basics
    - Pandas, NumPy, matplotlib
    - Hyperparameter Tuning, Training and testing in ML.
- Time Series
    - Trend
    - Seasonality
    - Stationarity - De-seasonalizing methods / De-trending methods
    - Normalizing / Standardizing
- Statistics related to Time Series
    - Normal Distributions Z-Score / 3-sigma clipping
    - T-Distributions
    - confidence / P-values
    - T-test
- Trend Detection Techniques
    - Mann Kendall
- Seasonality methods
    - ACF/PACF
- Change Point Methods
    - Explore methods - Bayesian analysis of change points (BCP), E-Agglomerative etc.

5.Linear Regression and Student's T test

Linear Regression:

Linear regression is a model that assumes a linear relationship between input variable x and an output variable y. Relationships are modelled using linear predictor functions whose unknown model parameters are estimated from the data. Such models are called linear models. Linear models are fitted using the least square approach.

Given a data set $\{y_i,\ x_{i1}, \ldots, x_{ip}\}_{i=1}^{n}$ of n statistical units, a linear regression model assumes that the relationship between the dependent variable y and the p-vector of regressors x is linear. The model takes the form:

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \varepsilon_i = \mathbf{x}_i^{\mathsf{T}} \boldsymbol{\beta} + \varepsilon_i,$$

The matrix notation is:

$$\mathbf{y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon},$$

where

$$\mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix},$$

$$X = \begin{pmatrix} \mathbf{x}_1^{\mathsf{T}} \\ \mathbf{x}_2^{\mathsf{T}} \\ \vdots \\ \mathbf{x}_n^{\mathsf{T}} \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1p} \\ 1 & x_{21} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{np} \end{pmatrix},$$

$$\boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{pmatrix}, \quad \boldsymbol{\varepsilon} = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}.$$

Student's t test:

The t-test is any statistical hypothesis test in which the test statistic follows a Student's t-distribution under the null hypothesis. A one-sample location test of whether the mean of a population has a value specified in a null hypothesis. It is used to determine if there is any significant difference between the means of two groups.

T test on slope of a regression line

Then t_{score} is given by:

$$t_{\text{score}} = \frac{\left(\widehat{\beta} - \beta_0\right)\sqrt{n-2}}{\sqrt{\frac{SSR}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}}.$$

Where, SSR = sum of squares of residuals

Beta0 = mean of initial window

n-2 degrees of freedom and $\widehat{\beta}$ is the least square estimator.

Linear regression with student's t test is said to be one of the best methods for detecting trend change points.

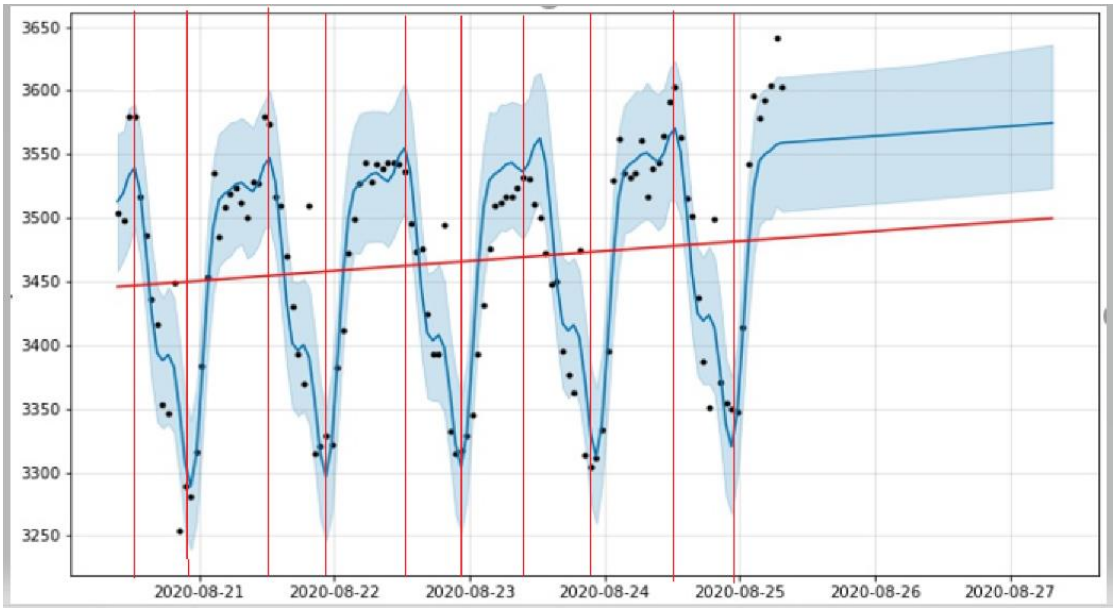
The algorithm:

- We start by taking an initial window of data and assume that trend does not change in this window.
- This window should be large enough to establish a good model and should not be small enough to not cover any trend change. We will fit a line to this window of data.
- Some points are picked after this window called test points.

- Using t-test we see if the trend of test points is statistically different from the previous trend.
- If so then we continue the test else we report the point as a trend change point.

Disadvantages of this method:

- Selecting the window size is the key to fitting the regression which was not simple for different kinds of data. It could not be made dynamic for each dataset.
- The threshold values could not be made Dynamic.
- This method could be applied for detecting local change points. It means that if the time series had any time series it would report false trend change points (one for every peak and dip). All vertical red lines are falsely reported as trend change points.



6.FB Prophet

GAM (General Additive Model):

The principle behind GAM is similar to that of the regression model. Unlike regression which uses an individual predictor for outcome, GAM uses sum of smooth functions to predict the outcome. The smooth functions here include functions describing trend component, seasonal component, holiday component and so on. **An example of GAM is fbprophet**. It is used for forecasting time series data where nonlinear data are fit with yearly, weekly, and daily seasonality, plus holiday effects.

Prophet uses a decomposable time series model with three main model components: trend, seasonality, and holidays. They are combined in the following equation:

$y(t) = g(t) + s(t) + h(t) + e(t)$

Here,

g(t) is a trend function which models the non-periodic changes. It can be either a linear function or a logistic function.

s(t) represents periodic changes i.e., weekly, monthly, yearly. A yearly seasonal component is modeled using Fourier series and weekly seasonal component using dummy variables.

h(t) is a function that represents the effect of holidays which occur on irregular schedules. ($n \geq 1$ days)

The term e(t) represents error changes that are not accommodated by the model.

- Growth Function g(t): Growth is typically modeled using the logistic growth model, which in its most basic form is

$$g(t) = \frac{C}{1 + \exp(-k(t-m))},$$

- Seasonality s(t): We rely on Fourier series to provide a flexible model of periodic effects. We can approximate arbitrary smooth seasonal effects with

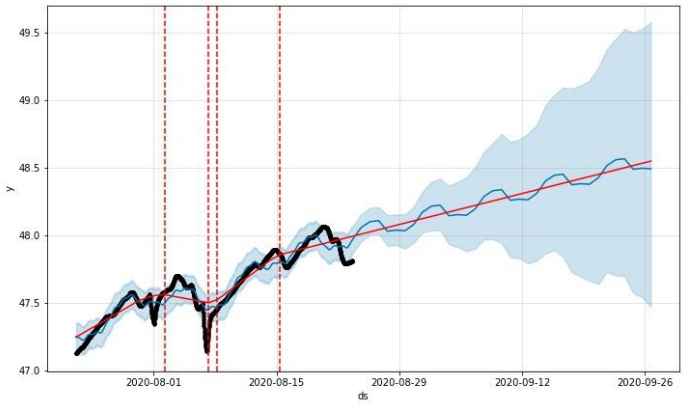
$$s(t) = \sum_{n=1}^{N} \left(a_n \cos\left(\frac{2\pi nt}{P}\right) + b_n \sin\left(\frac{2\pi nt}{P}\right) \right)$$

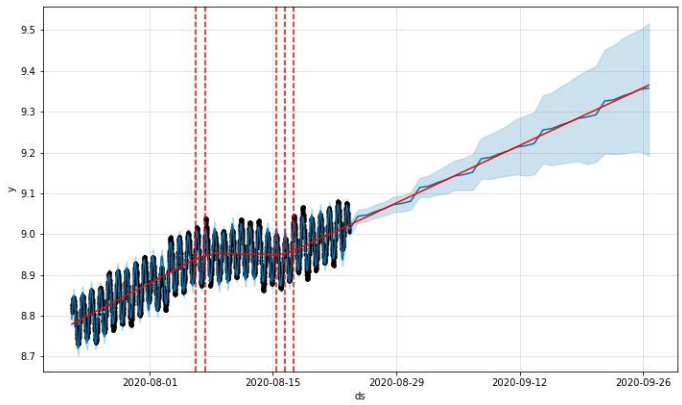
- Holidays h(t): Holidays and events provide a predictable and significant shocks to businesses around the globe. For example, Thanksgiving in the USA occurs on the fourth Thursday in November and an event like Super bowl can largely affect businesses which will be taken into account by fbprophet.

- Error e(t): It is simple the error component which could not be accommodated by the model.

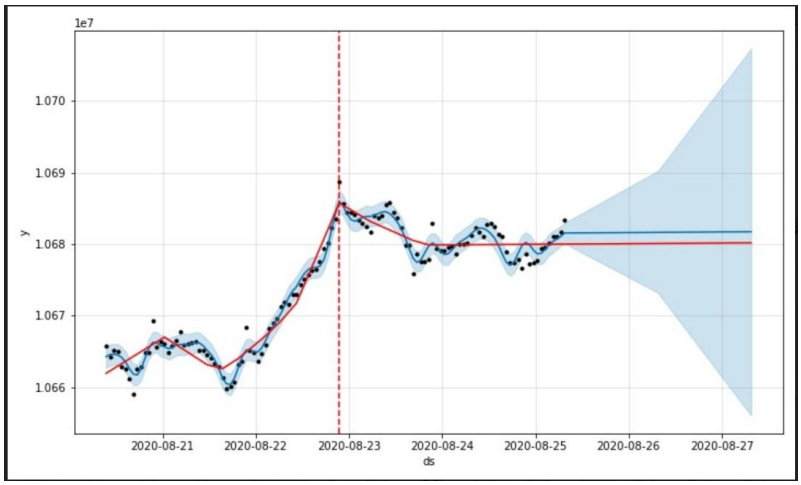
Results:

- Gave correct results for all the available datasets.
- Ran tests for over 150 datasets.
- Implemented it in an online manner by considering only 3 to 4 days of data and try to detect trend change points.
- The following graphs show the seasonality fit and the trend change points:





For the entire dataset



In an online scenario considering only the past 3 to 4 days of data.

CONCLUSION

In the past 2 months I have gained some work experience and learnt many Machine Learning algorithms which was completely new to me. I improved my coding skills and my communication skills. I came to understand what it means to be as team and work on a project.

LIST OF REFERENCES

1. https://www.cliff.ai/manifesto

2. https://otexts.com/fpp2/

3. https://cran.r-project.org/web/packages/trend/vignettes/trend.pdf

4. https://cran.r-project.org/web/packages/EnvCpt/EnvCpt.pdf

5. https://cran.r-project.org/web/packages/trend/vignettes/trend.pdf

6. https://cran.r-project.org/web/packages/trendsegmentR/trendsegmentR.pdf

7. https://towardsdatascience.com/the-upper-confidence-bound-ucb-bandit-algorithm-c05c2bf4c13f

8. https://peerj.com/preprints/3190.pdf#

9. https://en.wikipedia.org/wiki/Linear_regression

10. https://facebook.github.io/prophet/docs/quick_start.html#python-api

GLOSSARY

1. Time Series: A time series is a series of data points indexed in time order.

2. Trend: A *trend* exists in a time series where there is long term increase or decrease in data. It does not have to be a linear change.

3. Seasonality: A seasonal pattern occurs when a time series is affected by seasonal factors such as the time of the year or the day of the week.

4. Mann-Kendall test: The Mann-Kendall test (MK test) is used to analyze data collected over time for consistently increasing or decreasing trends in Y values.

5. Pandas: Pandas is a fast, powerful, flexible and easy to use open-source data analysis and manipulation tool, built on top of the Python programming language.

6. NumPy: NumPy is a library for the Python programming language, adding support for large, multi-dimensional arrays and matrices, along with a large collection of high-level mathematical functions to operate on these arrays.

7.R Language: R is a programming language and free software environment for statistical computing and graphics supported by the R Foundation for Statistical Computing.

8. Fbprophet: A python package by Facebook which is mainly used for time series forecasting and various other functions like trend changepoints and seasonality.