Project in Time Series Analysis

Technion – Israel Institute of Technology

Raissa Chut Steinberg - 341241297
Kobi Amit - 206107344

Google Collab Notebook

Original data site

Introduction

The data chosen regards different daily measurements related to the weather of the city of Mumbai, India between January 1st 2016, and November 15th 2020 (1781 entries).

Our research question is how well we can predict future temperatures in Mumbai, we want to explore different forecasting models and determine the accuracy in which we can make predictions over time.

Each data entry includes various daily weather variables such as temperature, dew level, solar radiation, and humidity. For our analysis, we have chosen to focus on daily temperature and will later incorporate one of the other variables to enhance the accuracy of our predictions. The observed temperature series:

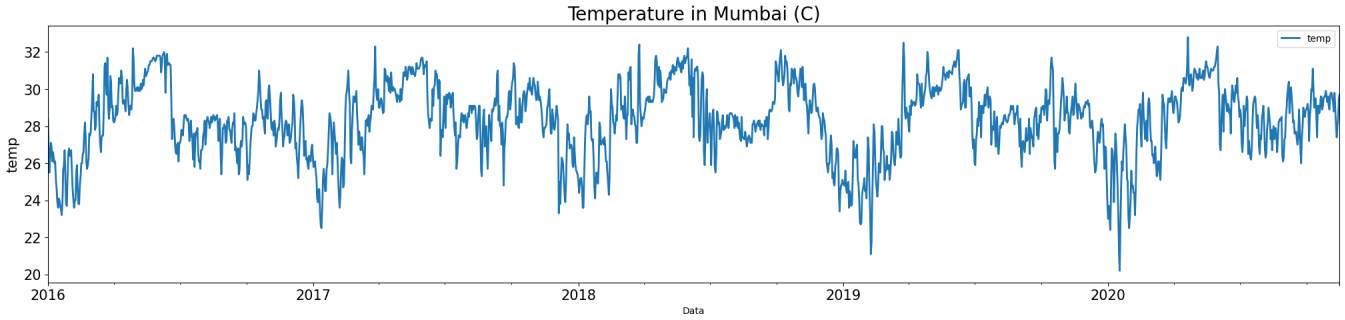


Figure 1: Observed daily temperature

We aggregated the data according to weeks to make the work with the used forecast models feasible, from now on the main data is an aggregated version of the temperature series seen above. Its distribution:

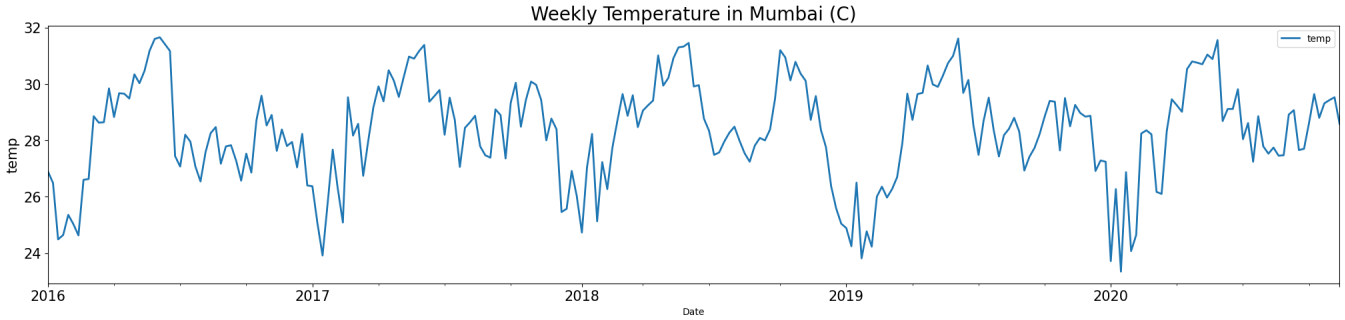


Figure 2: Observed weekly temperature

In order to get a better grasp of the data's underlying seasonality, trends and patterns we took an interest on a yearly decomposition of our weekly time series and on the distributions of the ACF and PACF of the differentiated series. The basic results we got were:

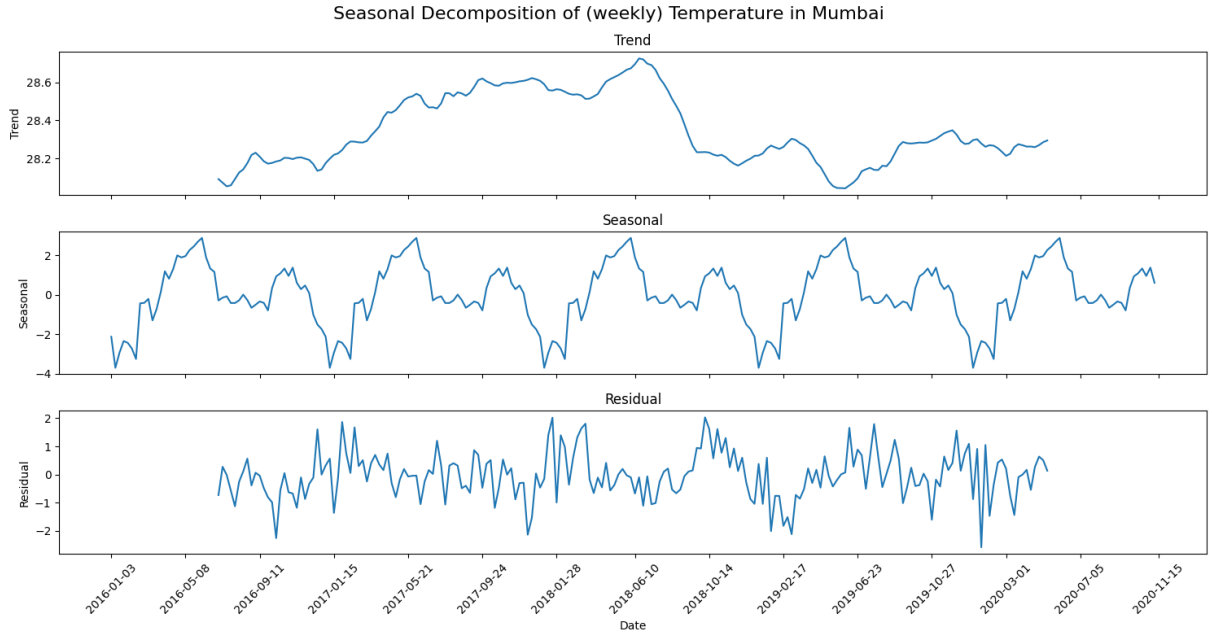


Figure 3: Seasonal decomposition of weekly temperature

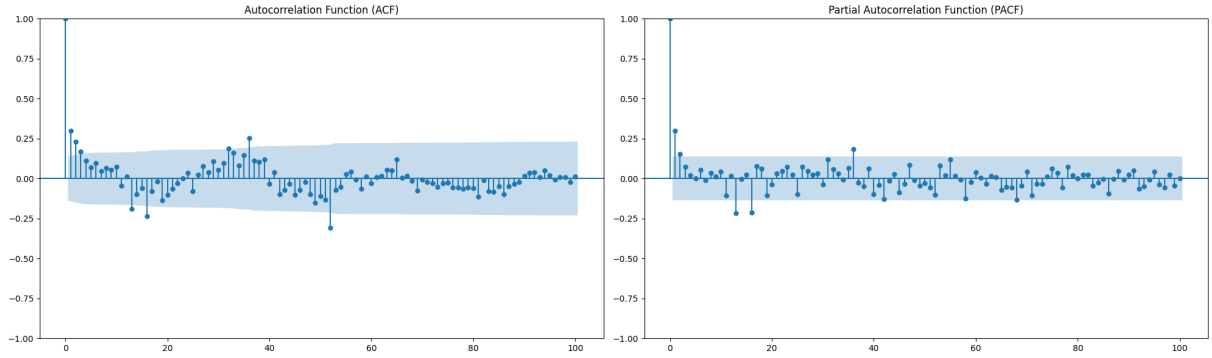


Figure 4: ACF and PACF of differentiated series

We can recognize from the trend distribution that there is a slight raise in temperature until mid-2018, after that the temperature drops, and more specifically Exponential Smoothing significantly low close do mid-2019. There is no constant trend other than these observations.

The yearly seasonality is easily noticed and the Residual graph gives us strong evidence that the differentiated series is stationary (the distribution is seemingly random and around 0), that is, the original series is yearly seasonal. The peaks on the ACF on lags 1 and 52 also support the seasonality explored.

Methodology

In this part, we will choose which statistical tools will be used throughout the project; explain the model fitting procedures and chosen parameters; the reasoning behind incorporating an exogenous variable and the process of checking for significant change points in the distribution of the time series.

The choice of working with weekly data instead of daily data as mentioned previously, came from encountering high complexity with the high-frequency (daily) data. This will allow the seasonal trends to be represented in a smoother way and lower short-term variation that might affect the performance of our models. The aggregation led us to get a dataset with 255 entries. Of which 178 were used as a training set, and 77 as a test set, to help us understand how well our models were defined.

Model Fitting

The time series models we chose to implement to explore and forecast the temperature in Mumbai are:

- **Seasonal Autoregressive Integrated Moving Average - SARIMA**

  The SARIMA model is an extension of the non-seasonal ARIMA model, and is designed to deal with seasonally patterned data. It models a given time series as a combination of three important components of time series, autoregressive (AR), integrated (I), and moving average (MA), along with a seasonal component (S). SARIMA, when correctly defined, can capture recurring patterns in the data at regular intervals, which makes it fit for temperature forecasting. We will explain the rationale behind choosing the best-fitting parameters shortly.

- **Prophet**

  The Prophet model is a forecasting tool designed by Facebook using an additive model to capture non-linear trends by fitting seasonal patterns yearly, weekly, and daily while also taking into account holiday effects. The design makes Prophet robust to missing data, capable of adjusting to shifts in trends, and resilient to outliers, ensuring reliable forecasts for real-world complex scenarios. It is especially effective for time series data with strong seasonal influences and past data, making it suitable for temperature forecast.

- **Exponential Smoothing**

  The Exponential Smoothing methods are known for their simplicity and accuracy in predictions based on historical data. It assumes that the forecast is a linear combination of past observations and weights different points in time depending on the purpose (importance of recent and far-away past data). We have chosen to work with the Holt-Winters smoothing method, given that our time series data has a significant seasonal component.

SARIMA

The SARIMA model, as mentioned above, takes into account 4 different components, it is defined with the help of 7 parameters.

$$SARIMA(p,d,q)(P,D,Q)_S$$

S is our seasonal component, given that we have a yearly seasonality and are working with weekly observations, $S = 52$.

Q and q are connected to the MA (Moving Average) component of the time series, when Q represents the seasonal Moving Average. Given the relevance of the first component of the ACF of the differentiated time series we will set $q = 1$ for an MA model. We can also note a sinusoid behaviour of the ACF, as well as peaks on weeks 16, 36, and 52. This leads us to believe that there is a seasonal MA component, so we will test $Q = 0, Q = 1$.

P and p are connected to the AR (Autoregrresive) component of the time series, when P represents the seasonal AR. Only the first component of the PACF appears to be relevant, this leads us to $p = 1$. Given the slight peaks on the PACF distribution, we will test $P = 0, P = 1$.

D represents the seasonal differentiation performed previously how we turned our series into a stationary one by differencing it at seasonal intervals (52 weeks/one year), so we set $D = 1$. d represents the non-seasonal differentiation, we see no relevant trend on the data, and this leads us to believe that $d = 0$, we tested $d = 1$ either way.

After testing all possible combinations of the above stated parameters, we observed which combination lead us to the best fitting model according to the BIC score from each model.

The values we got from each one:

Model	BIC	AIC
$SARIMA(1,0,1)(0,1,1)_{52}$	401.38	390.04
$SARIMA(1,0,1)(1,1,0)_{52}$	403.55	392.20
$SARIMA(1,1,1)(0,1,1)_{52}$	404.16	392.85
$SARIMA(1,0,1)(1,1,1)_{52}$	406.09	391.91
$SARIMA(1,1,1)(1,1,0)_{52}$	406.23	394.92
$SARIMA(1,1,1)(1,1,1)_{52}$	408.69	394.55
$SARIMA(1,1,1)(0,1,0)_{52}$	418.87	410.38
$SARIMA(1,0,1)(0,1,0)_{52}$	419.31	410.80

Table 1: BIC and AIC of the different possible models

Term	Coefficient
AR(1)	0.7585 (p=0.000)
MA(1)	-0.4133 (p=0.004)
SMA(1)	0.013 (p=0.013)

Table 2: Relevance of parameters chosen

The best-fitted model is $SARIMA(1,0,1)(0,1,1)_{52}$, achieving the lowest BIC score and carrying very promising coefficients on the parameters chosen. When SMA regards the Seasonal Moving Average term introduced to the model.

Exponential Smoothing

We have trained the model according to a 52 week seasonality, and chose the model of Holt-Winters with seasonality (as mentioned above), also known as "triple exponential smoothing", taking into account trends, level, and the seasonality of the data. The model automatically estimates the initial values of the components mentioned prior and utilizes them to get the best possible parameters in order to generate forecasts.

BIC	AIC
176.050	4.234

Table 3: BIC and AIC of the model

Smoothing Level	Smoothing Seasonal	Initial Level
0.3266	≈ 0	28.55

Table 4: Exponential Smoothing parameters

Prophet

The model was tailored to focus on yearly seasonality (52 weeks) given the temperature cycle throughout the year. Unlike the SARIMA model, it does not require manual tuning of the different components taken into account, given it estimates them directly from the given data. The model automatically estimates a linear trend that adapts over time and provides uncertainty intervals around the forecast, which enhances the readability of the predictions. Its design emphasizes ease of use, automatic adjustment to data patterns, and clear output, making it a robust choice for forecasting tasks that involve complex seasonal behavior.

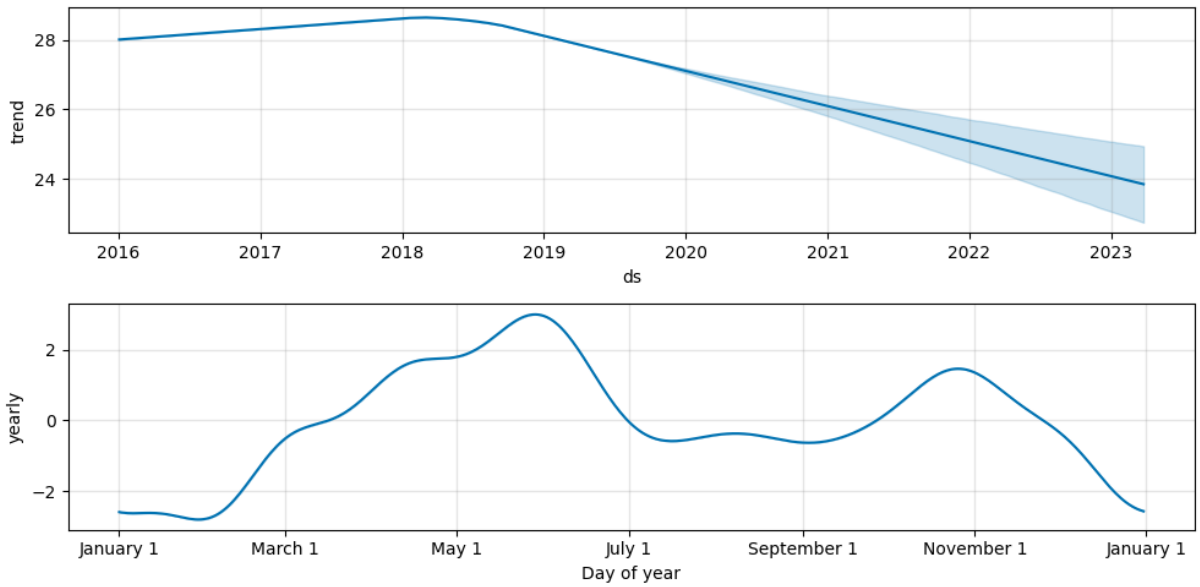


Figure 5: Decomposition of Prophet model

The model predicts that (the trend component) we see a gradual downward slope in the model's estimated average temperature from around 2016 to 2023, suggesting a "modest" cooling trend over time. The shaded interval widens toward the forecast horizon, indicating increased uncertainty further into the future. In the "yearly" component, the curves show a typical seasonal cycle: temperatures begin relatively higher in the early parts of the year, dip during the mid-year, then rise again towards late summer or early autumn before dropping of into winter. This pattern highlights the model view of how temperatures fluctuate over the year, with an amplitude of roughly a few degrees around the long-term mean.

Incorporation of an Exogenous Variable

After exploring the different time series models above, we will now try to incorporate an exogenous variable that may help us better forecast the data. The model of Exponential Smoothing does not support exogenous variables so we will focus on how it affects the models of SARIMA and Prophet.

The exogenous variable we chose to implement into our models was the dew level, that is given since correlation analysis indicated it had the strongest relationship to temperature among the available weather variables (e.g., humidity, solar radiation), making it a more promising regressor.

SARIMA

In addition to the standard SARIMA specification, we introduced "dew" as an exogenous variable in the model by including it in the exogenous regressor parameter when calling SARIMAX. After re-running the parameter grid search with this external regressor, the best performing configuration remains $SARIMA(1,0,1)(0,1,1)_{52}$, but now as a SARIMAX model with an external regressor. Our hope is that the incorporation of this exogenous variable will improve the forecasting performance of the model compared to the performance without it.

The values we got from each one:

Model	BIC	AIC
$SARIMA(1,0,1)(0,1,1)_{52}$	403.10	388.92
$SARIMA(1,1,1)(0,1,1)_{52}$	406.18	392.04
$SARIMA(1,0,1)(1,1,0)_{52}$	406.81	392.63
$SARIMA(1,0,1)(1,1,1)_{52}$	407.88	390.87
$SARIMA(1,1,1)(1,1,0)_{52}$	409.55	395.41
$SARIMA(1,1,1)(1,1,1)_{52}$	410.77	393.80
$SARIMA(1,1,1)(0,1,0)_{52}$	423.25	411.94
$SARIMA(1,0,1)(0,1,0)_{52}$	423.35	412.01

Table 5: BIC and AIC of the different possible models

Term	Coefficient
dew	0.1124 (p=0.042)
AR(1)	0.7090 (p=0.000)
MA(1)	-0.3744 (p=0.023)
SMA(1)	-0.9882 (p=0.919)

Table 6: Relevance of parameters chosen

Prophet

After training the Prophet model while taking into account the exogenous variable selected, we can observe the trend and yearly seasonality of the decomposition of the model aren't significantly different from the simpler model previously presented. In the bottom image we see the added value we get from adding the dew regressor to the model, its repeating wave pattern indicated a strong cyclical effect on the temperature variable and leads us to believe the effect of incorporating the regressor will result in better forecasting performance.

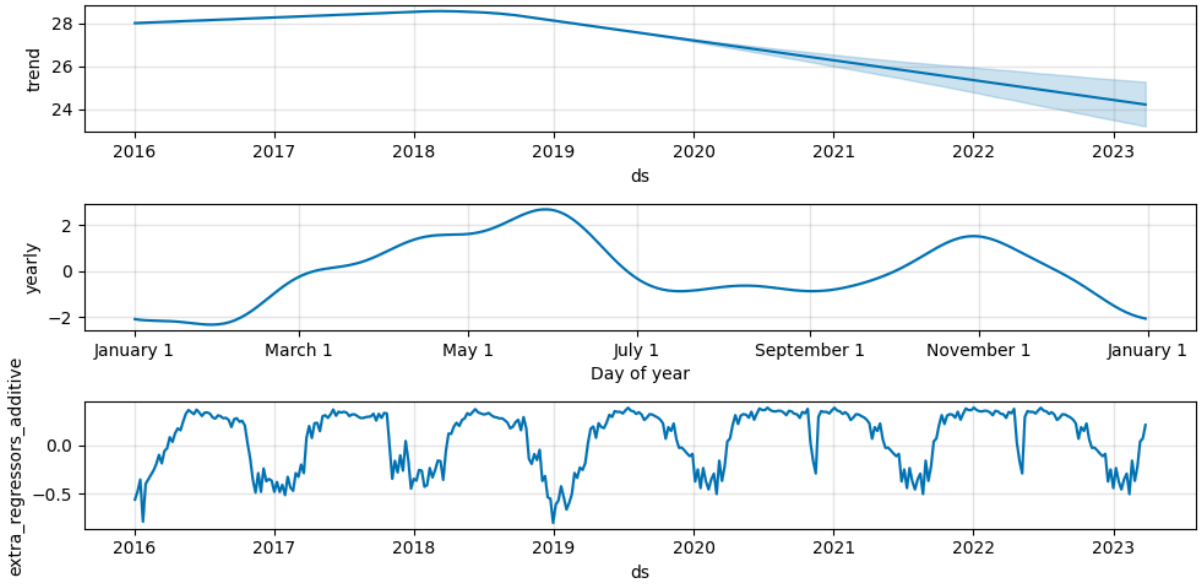


Figure 6: Decomposition of Prophet model with exogenous variable

Change-Point Detection

Many time series exhibit clear changes in distribution due to external factors, although in our series, there is no specific point where we see a significant change point, we are still interested in checking if there is a change in the distribution of the series over time. We have learned various tools throughout the course that make that possible like the Shewart Control Chart, SPRT, and CUSUM.

Firstly we will look for "obvious" change-points by using the Shewart Control Chart.

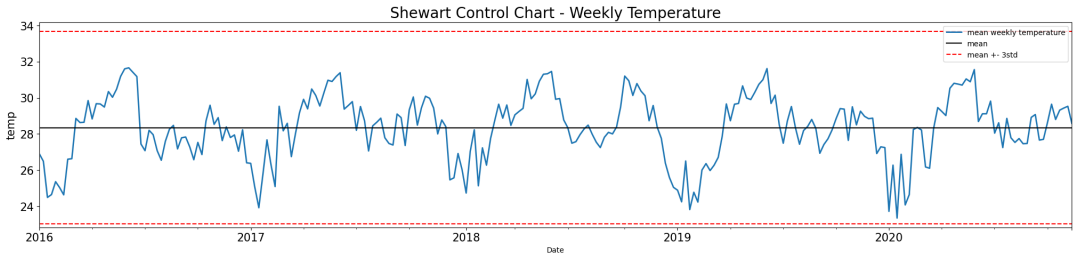


Figure 7: Shewart Control Chart

As seen above, all of the distribution stays within the bounds defined and in the beginning of 2017, 2019, and 2020 there is a significant drop in relation to the mean temperature.

To better understand distribution changes, we'll use the CUSUM method to detect abrupt deviations from the adjusted target (mean). This approach is more sensitive to small shifts than the Shewhart Control Chart, allowing for improved outlier detection.

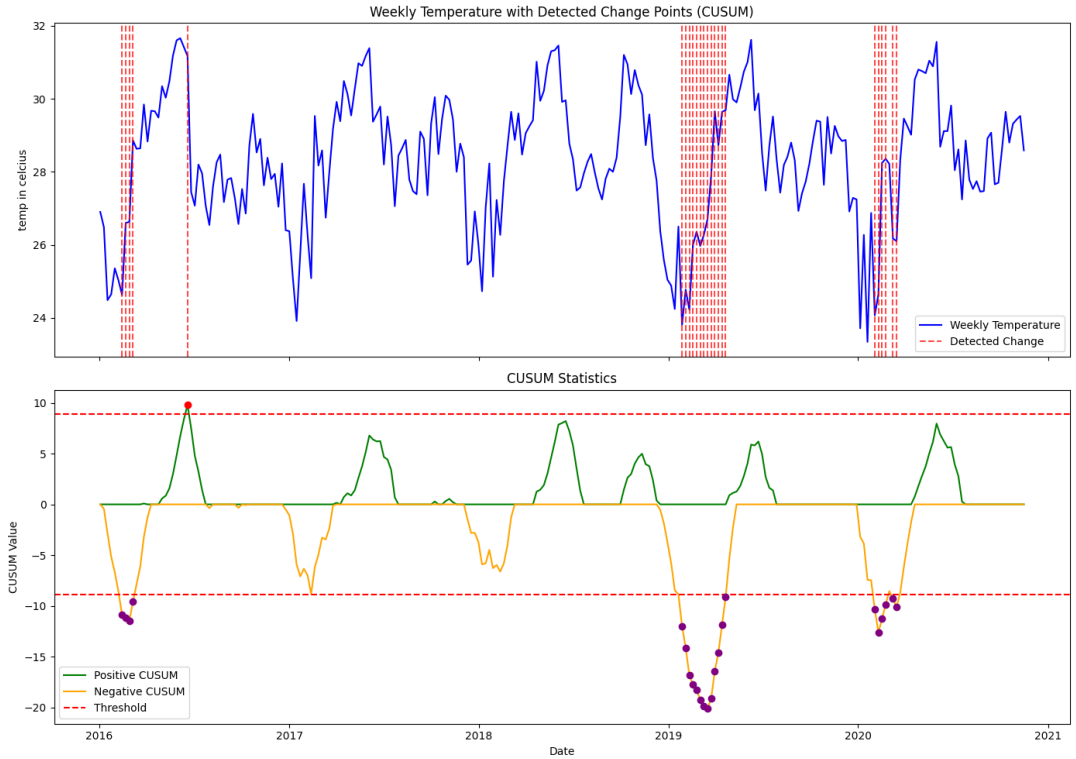


Figure 8: CUSUM ($k = 0.8 \cdot \sigma, h = 5 \cdot \sigma$)

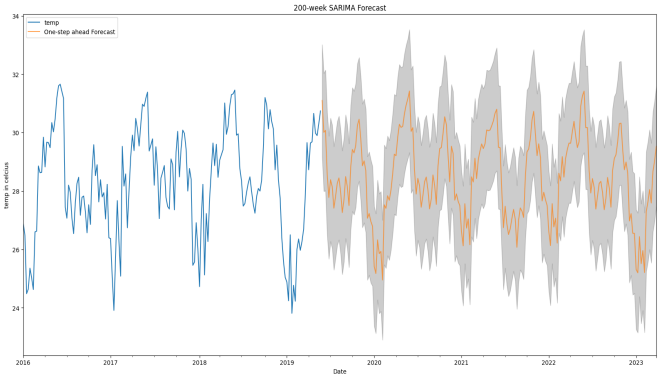
We were able to detect 24 change points, most of them being on the negative side of the variation. As expected by the statement previously, most of them take place around the beginning of the years 2019 and 2020.

Results and Discussion

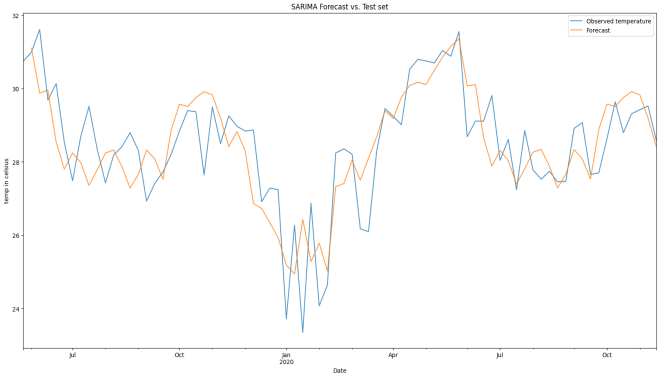
We will now discuss the findings from the analysis performed by the previously stated models learned

Results without exogenous variable

Firstly we'll observe the results of each of the models stated above without the exogenous variable incorporation.

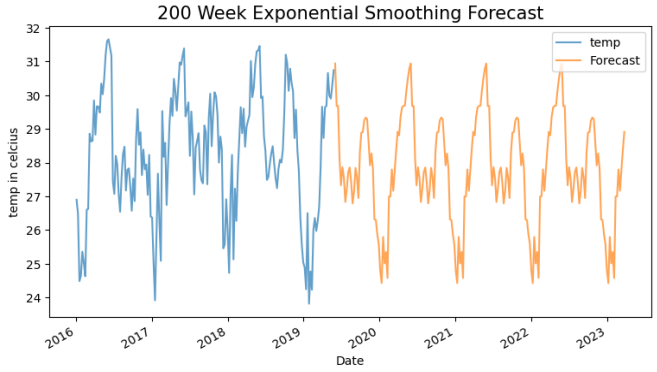


((a)) 200 week SARIMA Forecast

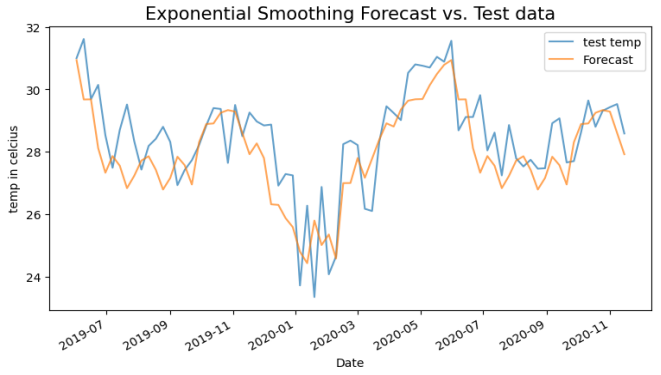


((b)) Comparison of SARIMA Forecast to test data

Above (left) we can observe a 200-week forecast using the SARIMA model and (right) a zoomed-in visualization of the forecast compared to the actual data (test set, not used on fitting). The **RMSE** of the SARIMA model was **0.994**.

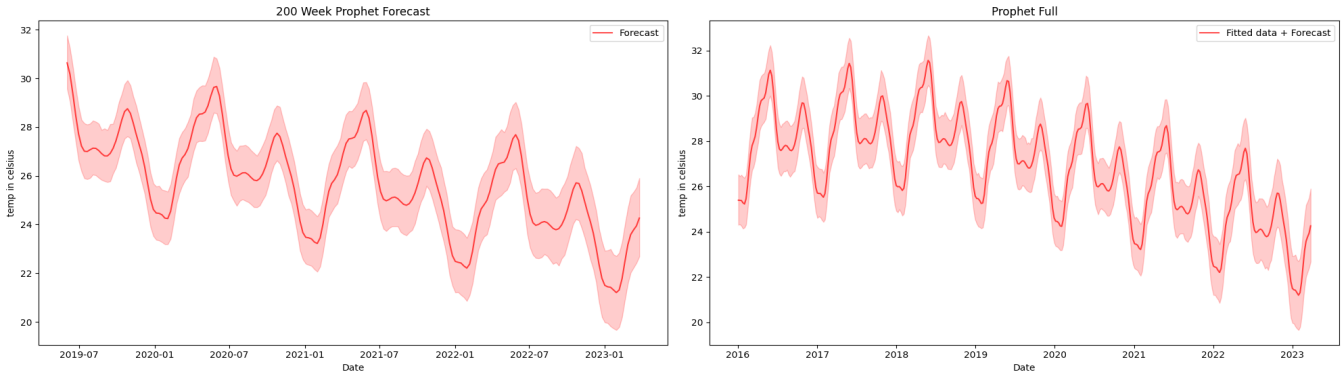


((a)) 200 week Exponential Smoothing Forecast

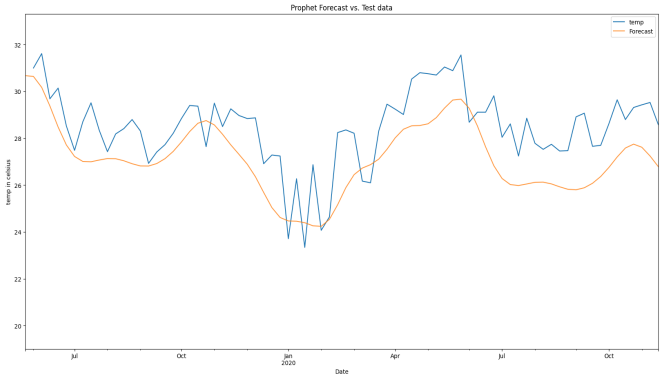


((b)) Comparison of Exponential Smoothing Forecast to test data

Above (left) we can observe a 200-week forecast using the Exponential Smoothing model, as well as a clear constant pattern on the forecast, which may stem from the time window used (200) in comparison with the training data size (178) or the parameters (α, β, γ) defined. And (right) a zoomed-in visualization of the forecast compared to the actual data (test set, not used on fitting). The **RMSE** of the Exponential Smoothing model was **1.099**.



((a)) 200 week Prophet Forecast

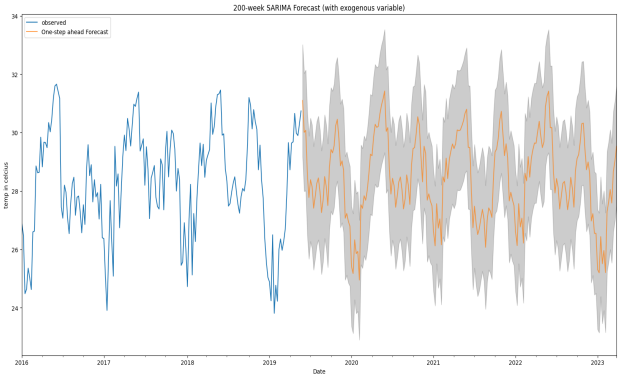


((b)) Comparison of Prophet Forecast to test data

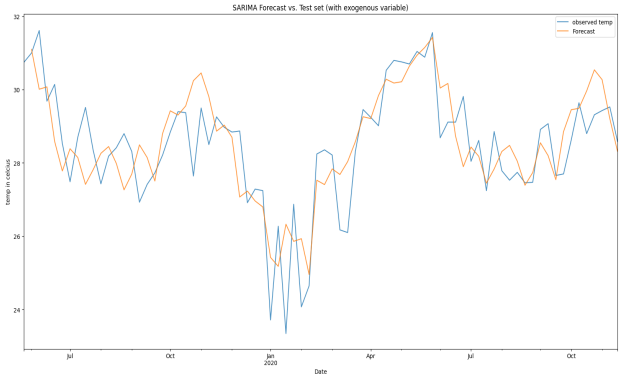
Above, we see a 200-week forecast using both the Prophet and Full Prophet models, along with their confidence intervals. The model predicts a downward trend, aligning with the earlier decomposition analysis. Below, a zoomed-in view compares the forecast to the actual test data, which was not used for fitting. The **RMSE** of the Prophet model was **1.678**.

Results with exogenous variable

After incorporating the best fit exogenous variable in our opinion, we will observe the results of the models of SARIMA and Prophet .

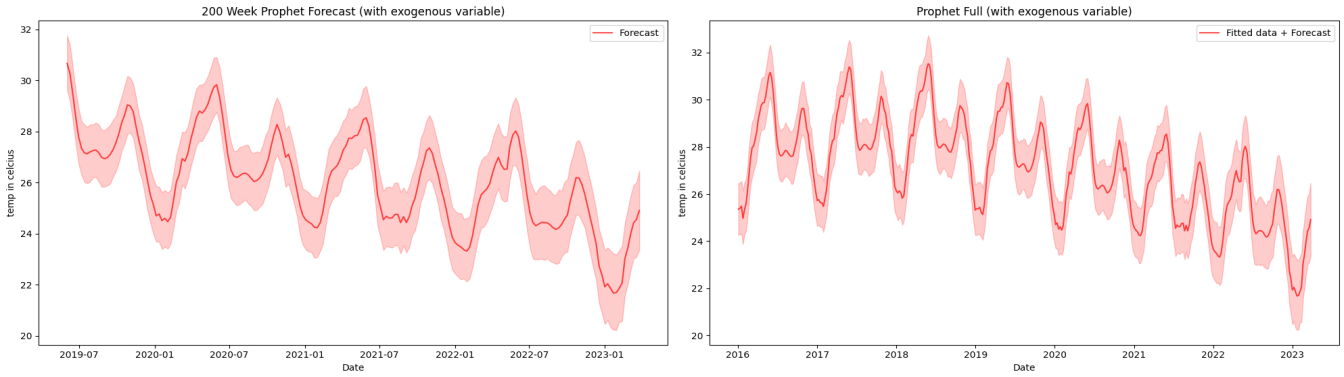


((a)) 200 week SARIMA Forecast - with exogenous variable

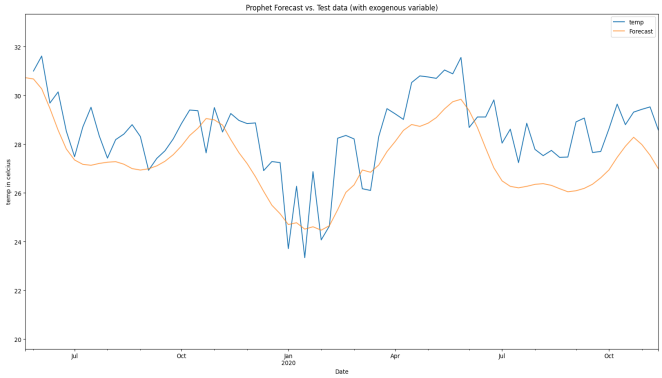


((b)) Comparison of SARIMA Forecast to test data - with exogenous variable

Above (left) we can observe a 200-week forecast using the SARIMA model with the incorporation of dew as the exogenous regressor and (right) a zoomed-in visualization of the forecast compared to the actual data (test set, not used on fitting). The **RMSE** of the SARIMA model was **0.989**.



((a)) 200 week Prophet Forecast - with exogenous variable



((b)) Comparison of Prophet Forecast to test data - with exogenous variable

Above, we see a 200-week forecast using both the Prophet and Full Prophet models with the incorporation of dew as the exogenous regressor, along with their confidence intervals. The model predicts a downward trend, aligning with the earlier decomposition analysis. Below, a zoomed-in view compares the forecast to the actual test data, which was not used for fitting. The **RMSE** of the Prophet model was **1.499**.

Model comparison and insights

We will compare how well each of the three chosen models SARIMA, Prophet, and Holt-Winters exponential smoothing, could forecast weekly temperature in Mumbai. We will also take interest in whether the incorporation of the exogenous variable of dew levels has lead us to better forecasting performance, that is, if we were able to increase the model's accuracies by taking this external factor into account.

Model	RMSE
SARIMA	0.994
SARIMA with exogenous variable	**0.989**
Prophet	1.678
Prophet with exogenous variable	1.499
Exponential Smoothing	1.099

Table 7: Relevance of parameters chosen

As observed, the best model within the models fitted is SARIMA, both with and without the incorporation of the exogenous variable.

It is also interesting to notice that the difference between the "bettering" of the RMSE of SARIMA and Prophet by adding the exogenous regressor, Prophet Smoothing' improves by 10.66% while SARIMAs' improves by 0.5%.

Conclusions

Finally, the main idea of this project was to show and explore how different time series models can capture and predict seasonal series like the temperature in Mumbai. By aggregating the data in a weekly manner, we were able to not take into account the short term noise that might have made the fitting of the models more complex, while maintaining the important seasonal trend we want to focus on.

Three different approaches/models were explored, SARIMA; Prophet and Exponential Smoothing. When SARIMA is the one who achieved the best results, with an RMSE value of **0.994**.

When regarding the dew level as well into our model fitting as an external regressor/variable we observe that the improvements' level in forecasting accuracy are significantly different between SARIMA and Prophet, indicating that the Prophet model may be able to better utilize the additional information the exogenous variable provides.

References

We have based our work mainly on the relevant lectures and tutorials.

Additional used sources were: