CHAPTER - 4

SYSTEM IMPLEMENTATION

Forecasting is a technique that uses historical data as inputs to make informed estimates that are predictive in determining the direction of future trends. It is an important and common data science task in organizations today. Having prior knowledge of any event can help a company tremendously in the formulation of its goals, policies and planning. However, producing high-quality and reliable forecasts comes with challenges of its own. Forecasting is a complex phenomenon both for humans and for machines. It also requires very experienced time series analysts which as a matter of fact are quite rare.

4.1 INTRODUCTION TO FACEBOOK PROPHET

Prophet is a procedure for forecasting time series data based on an additive model where non-linear trends are fit with yearly, weekly, and daily seasonality, plus holiday effects. It works best with time series that have strong seasonal effects and several seasons of historical data. Prophet is robust to missing data and shifts in the trend, and typically handles outliers well.

It is a tool that has been built to address these issues and provides a practical approach to forecasting “at scale”. It intends to automate the common features of business time series by providing simple and tunable methods. Prophet enables the analysts with a variety of backgrounds to make more forecasts than they can do manually.

4.1.1 THE PROPHET FORESTING MODEL

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

$$y(t) = g(t) + s(t) + h(t) + \epsilon_t$$

- **g(t)**: piecewise linear or logistic growth curve for modelling non-periodic changes in time series
- **s(t)**: periodic changes (e.g. weekly/yearly seasonality)
- **h(t)**: effects of holidays (user provided) with irregular schedules
- ε_t: error term accounts for any unusual changes not accommodated by the model

Using time as a repressor, Prophet is trying to fit several linear and non linear functions of time as components

4.1.2 HOW PROPHET WORKS

Prophet is especially useful for datasets that:

- Contain an extended time period (months or years) of detailed historical observations (hourly, daily, or weekly)
- Have multiple strong seasonality's
- Include previously known important, but irregular, events
- Have missing data points or large outliers
- Have non-linear growth trends that are approaching a limit

4.2 APPROACH

Prophet follows an analyst-in-the-loop approach to business forecasting at scale. This approach begins by modeling a time series using the parameters specified by analysts, producing forecasts and then evaluating them. Whenever a performance issue or a need for human intervention crops up, these issues are flagged to human analysts so that they can then inspect the forecast and potentially adjust the model based on this feedback.

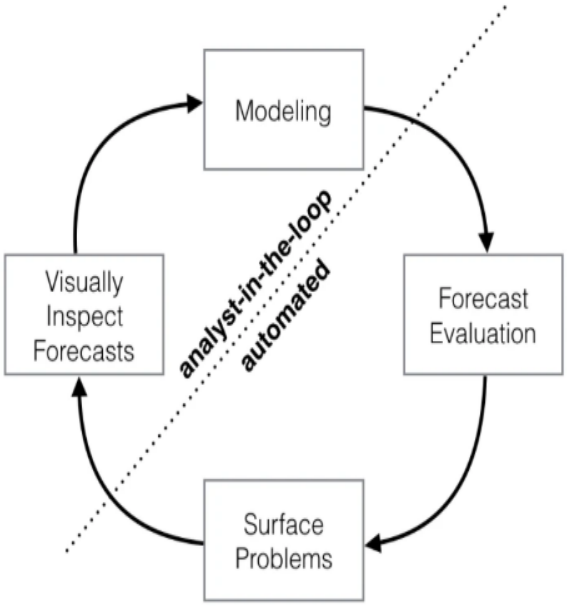

Figure 4.1 Schematic view of the analyst in the loop approach to forecasting at scale

COURTSEY:towardsdatascience.com

Figure 4.1 shows the schematic view of the analyst in the loop approach to forecasting at scale.

4.3 ADVANTAGE OVER OTHER TIME SERIES MODELS:

- The cool thing about Prophet is that it doesn't require much prior knowledge or experience of forecasting time series data since it automatically finds seasonal trends beneath the data and offers a set of 'easy to understand' parameters. Hence, it allows non-statisticians to start using it and get reasonably good results that are often equal or sometimes even better than the ones produced by the experts.
- Prophet modelling can be able to detect the Change Points in time series data.
- We can include the holidays (play-offs & super-bowls) in our data. Details has been added later.

- We can regularise the parameters by means of Bayesian optimisation with cross-validation.
- We can incorporate the multiplicative-seasonality and determine the uncertainty intervals in the data.

4.4 INSTALLATION ON PYTHON API

Prophet follows the model API. We create an instance of the Prophet class and then call its fit and predict methods. The input to Prophet is always a dataframe with two columns: ds and y. The ds (date stamp) column should be of a format expected by Pandas, ideally YYYY-MM-DD for a date or YYYY-MM-DD HH:MM:SS for a timestamp. The y column must be numeric, and represents the measurement we wish to forecast.

4.5 STEPS INVOLVED IN FORECASTING USING PROPHET

Using Prophet is extremely straightforward. Import it, load some data into a Pandas data frame, set the data up into the proper format and then start modelling or forecasting.

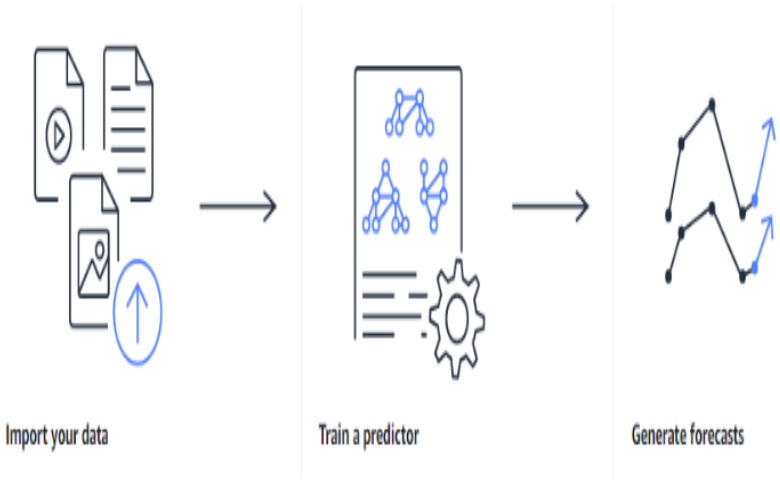

Figure 4.2 Steps involved in Forecasting

COURTSEY:d1.awsstatic.com

Figure 4.2 represents the steps involved in forecasting where the data is imported and predicator is used to train it followed by generating forecasts.

4.5.1 DATA COLLECTION

Datasets are collections of input data. The Crime dataset contains a summary of the reported crimes occurred from the year 2014 to 2019 that are collected from various sources like Kaggle dataset, Data.gov and data.worldin csv format.

Datasets contains the following columns:

- Id-unique identifier for the record
- Case number-record division number
- Date-date when the incident occurred
- Type-crime type
- Location description - description of the location where the incident occurred
- Arrest-indicates whether the arrest was made
- District-indicates the district where the incident occurred
- Ward-the ward where the incident occurred
- Community area-indicates the community area where the incident occurred
- X coordinate-x coordinate of the location where the incident occurred
- Y coordinate-y coordinate of the location where the incident occurred
- Year-year of the incidences occurred
- Updated on-date and time the where the record was last updated
- Latitude-the latitude of the location where the incident occurred
- Longitude-the longitude of the location where the incident occurred

- Location-the location where the incident occurred in a format that allows for creation of map and other geographic operations.

4.5.2 PREPARING THE DATA

Once the raw data is available, the complications like missing values has to be handled. A common occurrence in real-world forecasting problems is the presence of missing values in the raw data. A missing value in a time series means that the true corresponding value at every time point with the specified forecast frequency is not available for further processing. So the data should be cleaned and prepared in such a way that the data is suitable for the forecasting models.

4.5.3 MAKING PREDICTIONS

The input to Prophet is always a data frame with two columns: ds and y. The ds (date stamp) column should be of a format expected by Pandas, ideally YYYY-MM-DD for a date or YYYY-MM-DD HH:MM:SS for a timestamp. The y column must be numeric, and represents the measurement to forecast.

4.5.4 BACKTESTING

The concept of the Back testing is actually pretty simple. It is to keep the later part of the data as Test data so that the evaluation of the model is done against this Test data.

4.5.4.1 SPLIT THE DATA INTO TRANING DATA AND TESTING DATA

Before building the model, separate the data into two parts, one is Training data and another is Test data. Figure 4.3 shows the diagrammatic representation of how data is split into training and testing which is used to obtain the future forecasts.

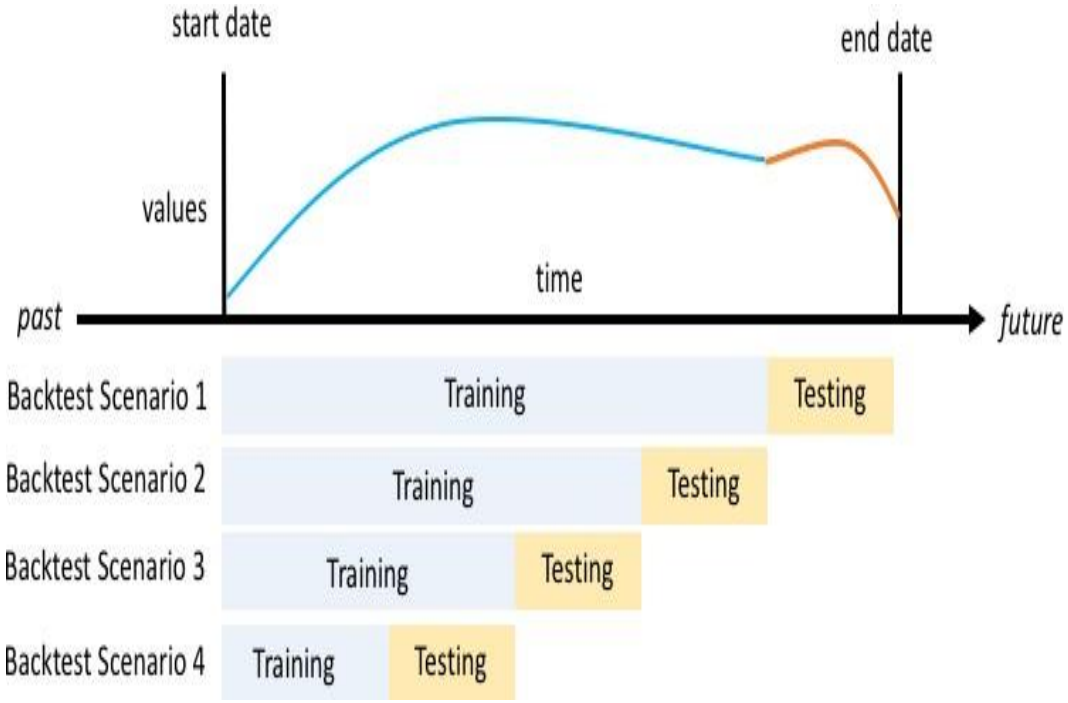

Figure 4.3Training data and Test data

COURTSEY:d1.awsstatic.com

4.5.4.2 BUILD A MODEL

Then, build the forecasting model based on the training data with the Prophet. Once the model is built, then we can use the model to forecast for the test data period. Prophet follows model API wherein an instance of the Prophet class is created and then the fit and predict methods are called. The model is instantiated by a new Prophet object and followed by calling its fit method and passing in the historical data frame.

4.5.4.3 FORECASTING

By default, Prophet uses a linear model for its forecast but a logistic model can also be used by passing it as an attribute. Predictions are then made on a data frame with a column ds containing the dates for which a prediction is to be made. The predict method will assign each row in future a predicted value which it names yhat.

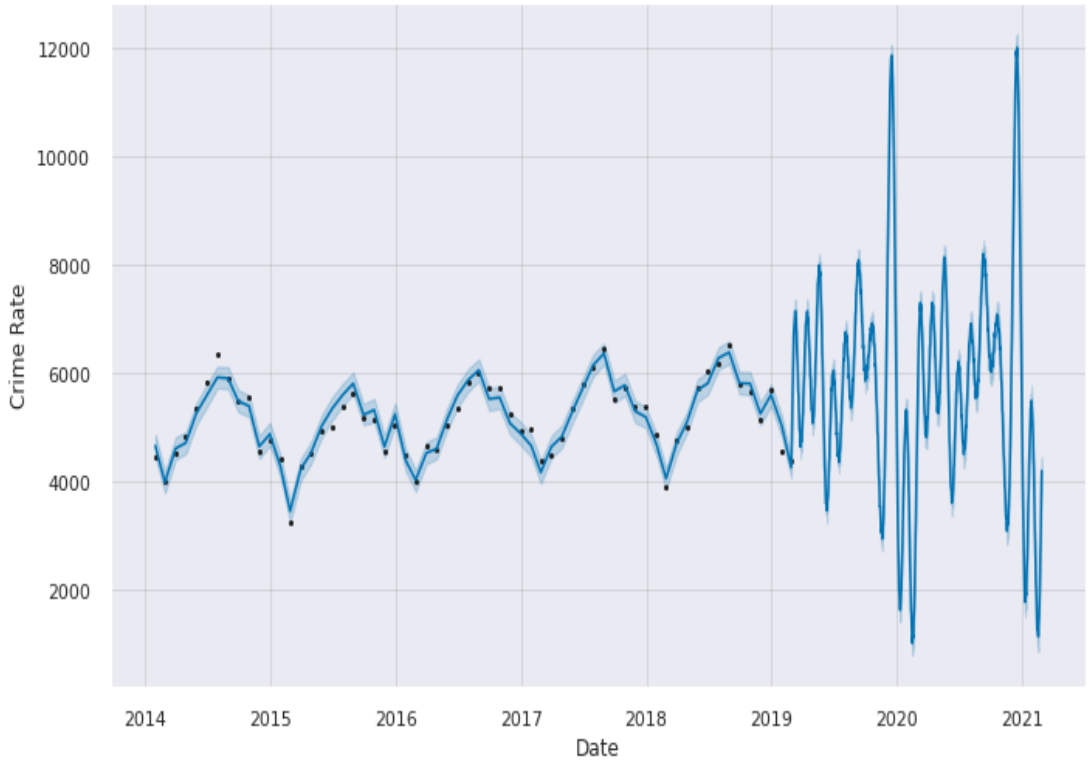

Figure 4.4 Forecasting

The black dotted lines represent the actual value and the blue line represents the predicted values in figure 4.4. The forecasted values in the Training period are very close to the actual data (black dotted lines). This is because the forecasting model was built based on this actual data. On the other hand, the forecasted values in the Test period seem to be farther from the actual values as this is the period that the forecasting model doesn't know the answers. The forecasted values tend to be further away from the actual values in the Test period. This is because the forecasting model assumes that whatever the trend and seasonality found in the Training data will be repeated in the future.

4.5.4.4 FORECASTING COMPONENTS

The forecasting component visualizations show that Prophet was able to accurately model the underlying trend in the data, while also accurately

modelling weekly and yearly seasonality. The below graph figure 4.5 shows the trend, monthly and weekly seasonality. There is a significant increase in the trend for the years 2014-2021.

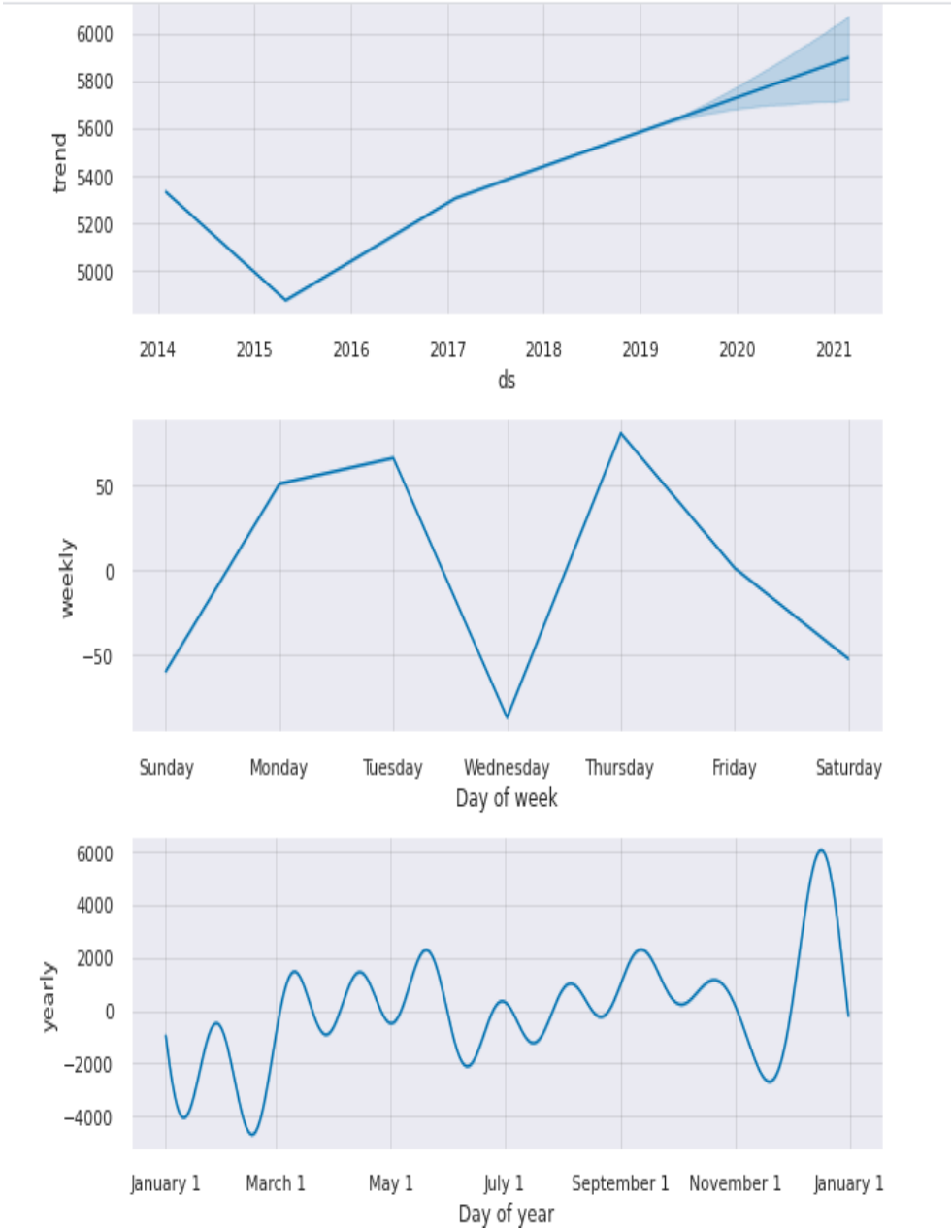

Figure 4.5 Forecasting Components

Therefore, it denotes that the crime count for Jewel theft murder is expected to increase for the upcoming years 2020 and 2021. The monthly seasonality shows the monthly trend of the crime. It shows that the November and January are the crucial months as most of the crimes are expected to happen around this time.

The weekly seasonality shows the weekly trend of the crime. It shows that the Wednesday is the day where least number of crimes are expected to happen. But Tuesday and Thursday are the crucial days as there is a hike in the crime rate for these two days.

4.5.4.5 TREND CHANGEPOINTS

Trend is modeled by fitting a piece wise linear curve over the trend or the non-periodic part of the time series. The linear fitting exercise ensures that it is least affected by spikes, missing data.

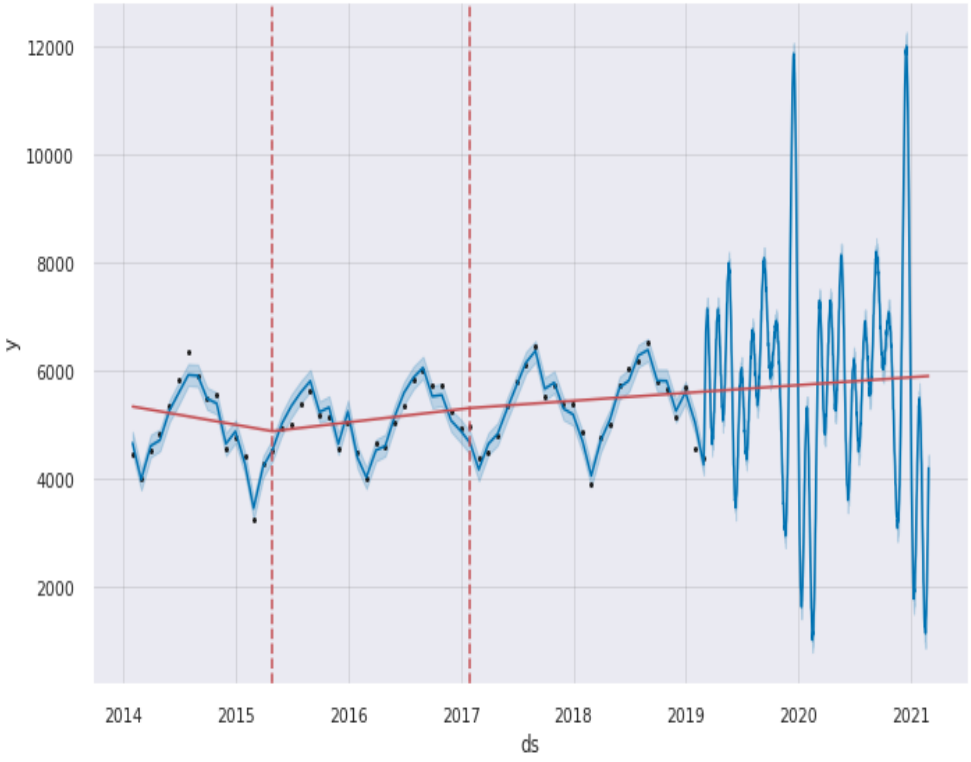

Figure 4.6 Trend change points

Change points are the points in our data where there sudden and abrupt changes in the trend. The change points parameter is used, when we supply the change points dates instead of having Prophet determine them. The red line figure 4.6 represents the trend change point at which the rate is allowed to change. The change point range usually does not have that much of an effect on the performance.

4.6 SUMMARY

Prophet is an extremely easy tool for analysts to produce reliable forecasts. From the above forecast, we can see that jewel theft murder crime rate is expected to rise in the start of 2020 and 2021 and there will be times where crime rate will be low as well. We can see that our predicted values quite matches with actual values, hence our model will prove efficient for foreseeing the future. Hence, Prophet make the entire forecasting process easy and intuitive and also gives a lot of options. The actual advantage of this model can only be assessed on large datasets but Prophet does enable forecasting a large number and a variety of time series problems which is truly forecasting at scale.