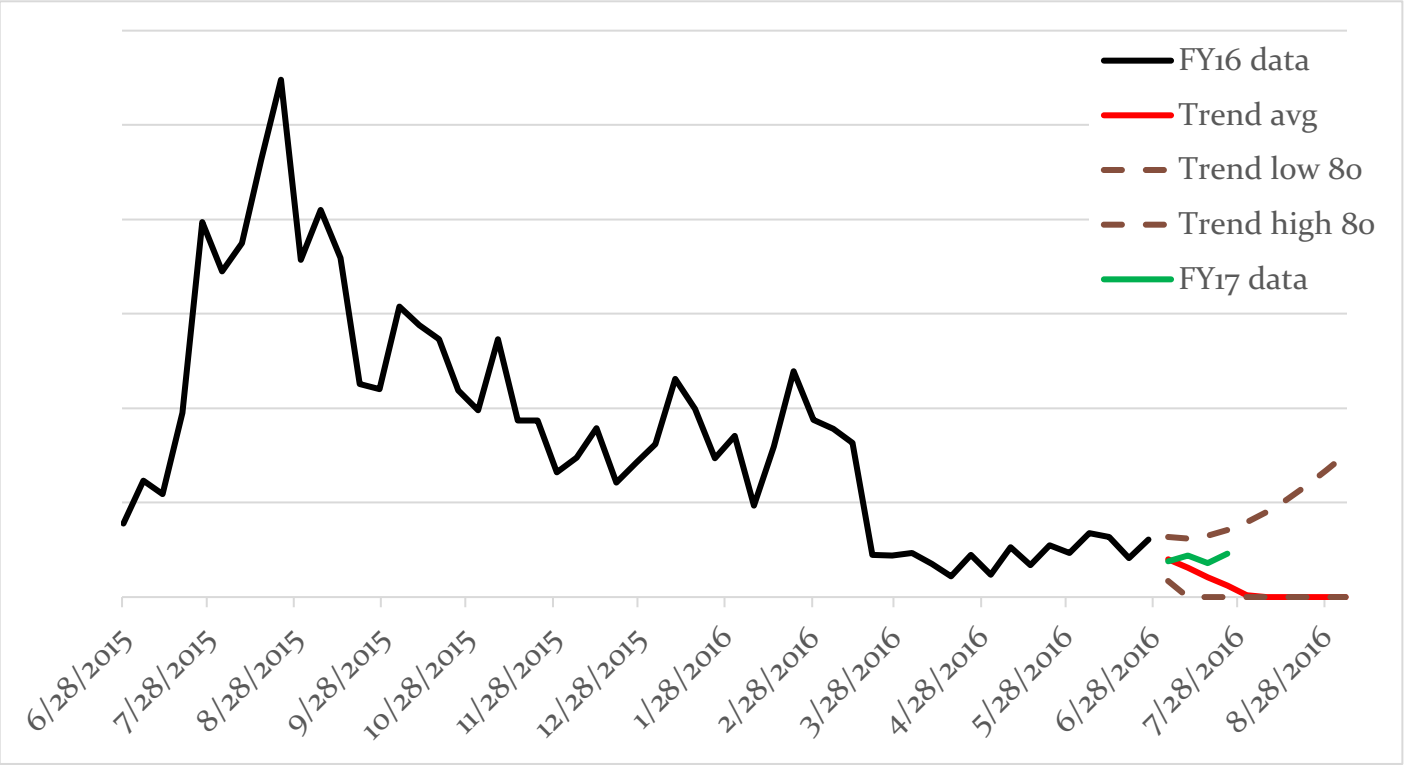


Forecasting Security Events

ADP Summer Internship 2016 & Capstone Project - December 15, 2016

TABLE OF CONTENTS

INTRODUCTION

Every company contains confidential data, also called sensitive data. If this data leaves the company network unencrypted, it can compromise the security of the company. This confidential data can be in any form – text, image, video etc. Here we deal with data in the form of emails. It is very common for employees to send emails within the company network and also outside it. But they should be particularly careful about the emails they send outside the company network as this could lead to potential "data leakage."

Companies are aware of the "data leakage" situation and employ various techniques to combat it. Access privileges are one such method where employees are prevented from accessing all the company information. They have access only to a part of that data, which pertains to their work. As soon as a new employee comes on board, he/she is made to go through tutorials which highlight the significance of keeping this data confidential since it contains trade secrets and other such information.

Another measure employed is to use a software tool called data loss prevention (DLP) that monitors and tracks all emails leaving the company network and inspecting them to find out how many and which emails contain unencrypted information. One such unencrypted email leaving the company network is called an event. This project deals with looking at the number of events that occurred over FY16 and forecasting them for FY17. This forecast* will help management determine where they stand in regards to "data leakage" and take appropriate action.

*Data along the y-axis (i.e. number of DLP events) has been removed to maintain confidentiality.

SCOPE

In–scope: To analyze and forecast FY16 data for security events. Also, this dataset only pertains to email data.

Out-of-scope: Since data for only one year is provided, it cannot be determined whether the data is seasonal and/or cyclical because they require data which encompasses more than 1 year. Additionally, any kind of data other than emails is not within the scope of this project.

ASSUMPTIONS

Before starting the process of forecasting, we make the following assumptions –

- Forecasting uses only information on the variable to be forecast. It makes no attempt to discover the factors which affect the variable's behavior.
- It is a wrong assumption that forecasts are not possible in a changing environment. The correct assumption would be: forecasting assumes that *the way in which the environment is changing,* will continue into the future.
- In general, the further into the future we forecast, the more uncertain we are.

DATA DESCRIPTION

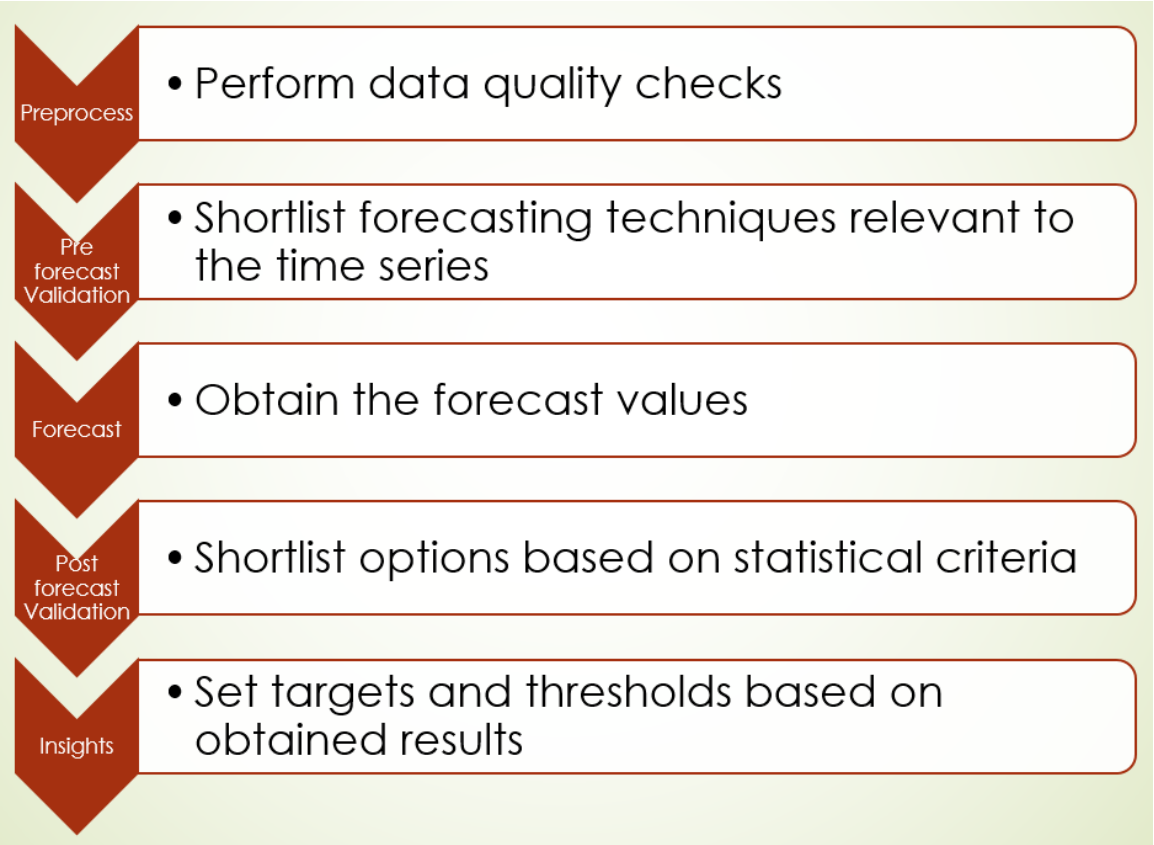
The project deals with forecasting wherein only 2 variables are required

- The variable to be forecast: Number of security events
- The timeline along which to forecast: FY17 using FY16's data

There are 250,000 records encompassing FY16 data

- On any given working day, numerous security events occur
- On weekends or holidays, there are no security events recorded

FLOW OF DATA ANALYSIS



DATA PREPROCESSING

Cleaning

- Some data points were blank. Missing value imputation was performed to fill in these values. This step is particularly important since R eliminates rows with 'nulls' and this would bias the forecast.
- Duplicate values were present. For example, an email sent out with the same timestamp from the same sender to the same recipient was recorded twice. This was later found to be an anomaly in the software but nonetheless eliminated from the dataset for the purpose of forecasting.
- Lastly, in the cleaning step, format checks were performed to ensure integrity of all columns and their corresponding datatype.

Summarization

- Another important step to note is the summarization step. When performing forecasting, it is extremely significant to have equi-distant data points because the number of data points have to be equal within any given interval of time.
- As noted earlier, we did not have records for the weekends and holidays so obtaining equi-distant data at the daily level was not feasible. Hence, we aggregated the data to a weekly level by taking an average over 7 days.

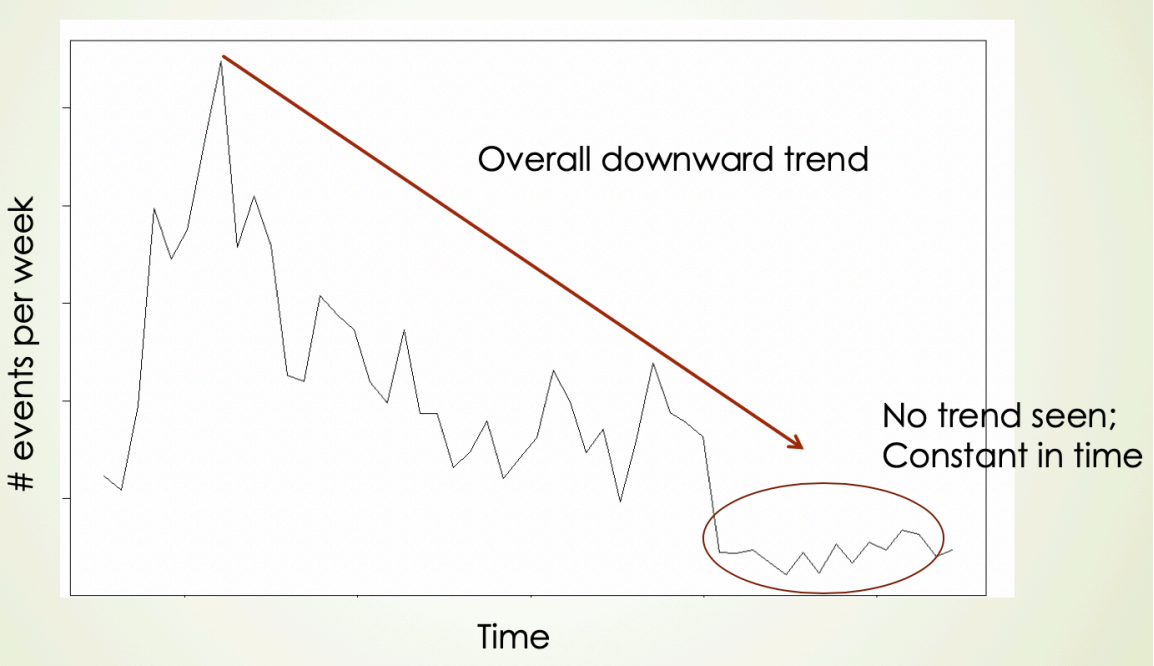
Visualization

- The last step in the pre-processing phase is to visualize the historical data in the form of a histogram to see the distribution of the data points.
- Using metrics like average, 10th quantile and 90th quantile, it helps us to find out how the events panned out it in FY16 and gain a better understanding of the same.

PRE-FORECAST VALIDATION

Plot Time Series

The first step under this phase is to visualize the data a little differently. Earlier, we plotted a histogram of the data but this time, we convert the data into a time series and plot it in order to identify its characteristics.



- In the above time series, we can see that the overall data has a downward trend and is definitely not stationary.
- But we also observe another interesting pattern. Towards the end of the time series plot, in the last 3-4 months, the events become stationary except for the random fluctuations and show no trend.

Shortlist Forecasting Techniques

The main purpose of this step is to **qualitatively** eliminate those techniques which will not work well with the data, thus putting our forecasting knowledge to use.

TECHNIQUES / ATTRIBUTES	Simple Exponential Smoothing	Holt's Exponential Smoothing	Holt Winter's Exponential Smoothing
Trend	NO	YES	YES
Seasonality	NO	NO	YES
Suitability to time series	✓	✓	X

Based on the above table, we can see that 2 forecasting techniques will work well with the characteristics identified in the previous step –

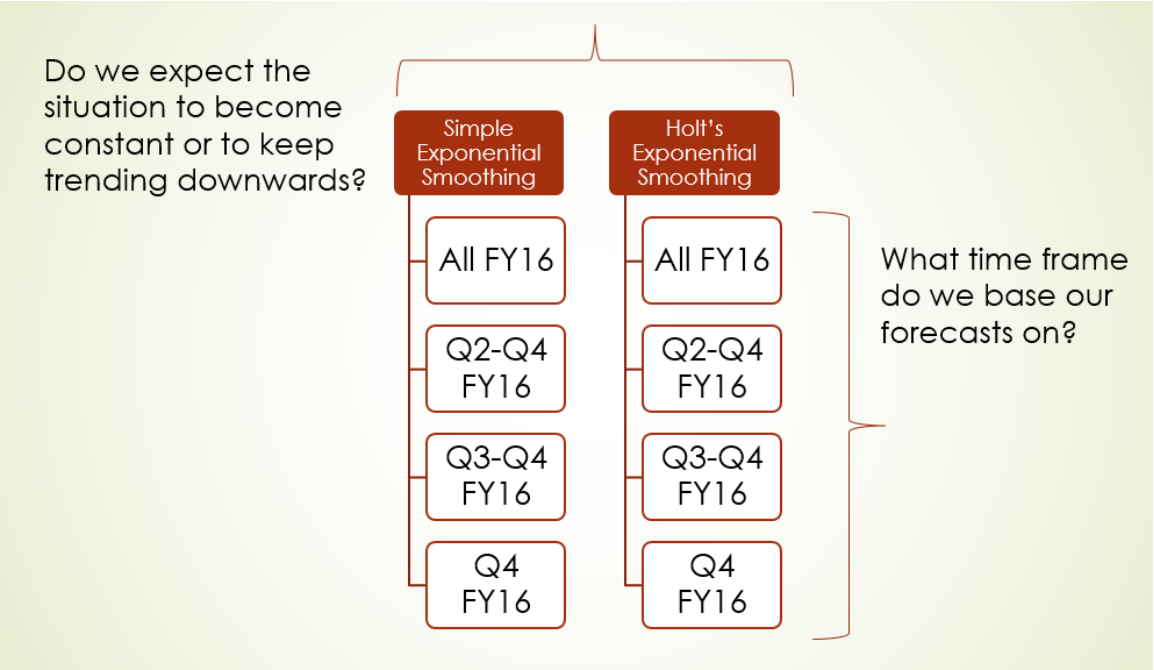
- Holt's Exponential Smoothing – since the historical data shows trend but no seasonality
- Simple Exponential Smoothing – since the last quarter of the data shows neither trend, nor seasonality

FORECASTING

Slice Data Input Based On Different Time Frames

In the previous step, we concluded that 2 forecasting techniques will be used. But there's one small step to perform before actually running the forecasting technique. That is to slice the data in different time frames and find out which one yields the best results. In other words, which time frame best represents the historical data.

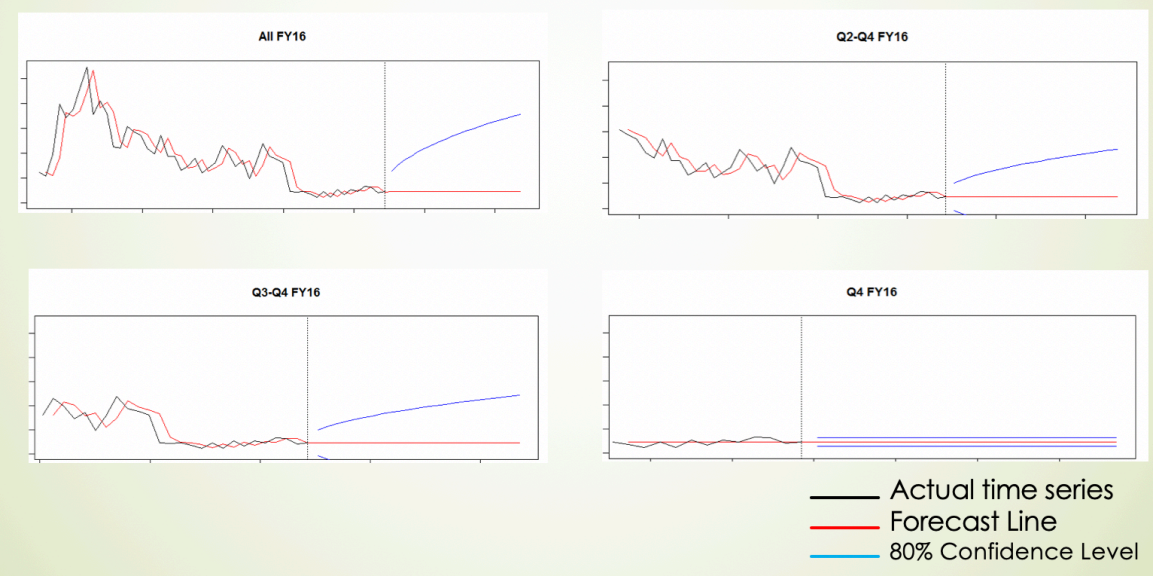
- Is it necessary to use the trending data to forecast when the last few months show a different picture altogether?
- Is the stationary data at the end misleading and just shows a slowdown of an otherwise rapidly progressing set of events?



We let the tool's statistical accuracy answer these questions for us and hence, we run 8 different forecasts in all.

Run Forecasting Techniques(refer to the Appendix for a Forecasting example)

In this step, we run the forecasting techniques against the data to obtain the 8 different forecasts, 4 of which are shown in the image below.



POST-FORECAST VALIDATION

In this phase, we **quantitatively** eliminate out of the 8 forecasts by looking at the forecast errors. Good forecasts have errors with the following properties –

- Centered around 0
- Follow a Bell Curve distribution
- Exhibit constant variance

The graphs seen above are for the four time frames using Simple Exponential Smoothing (assuming flat forecasts). As we can see from above, the following forecast is shortlisted –

- Flat forecast Quarter 4 FY16

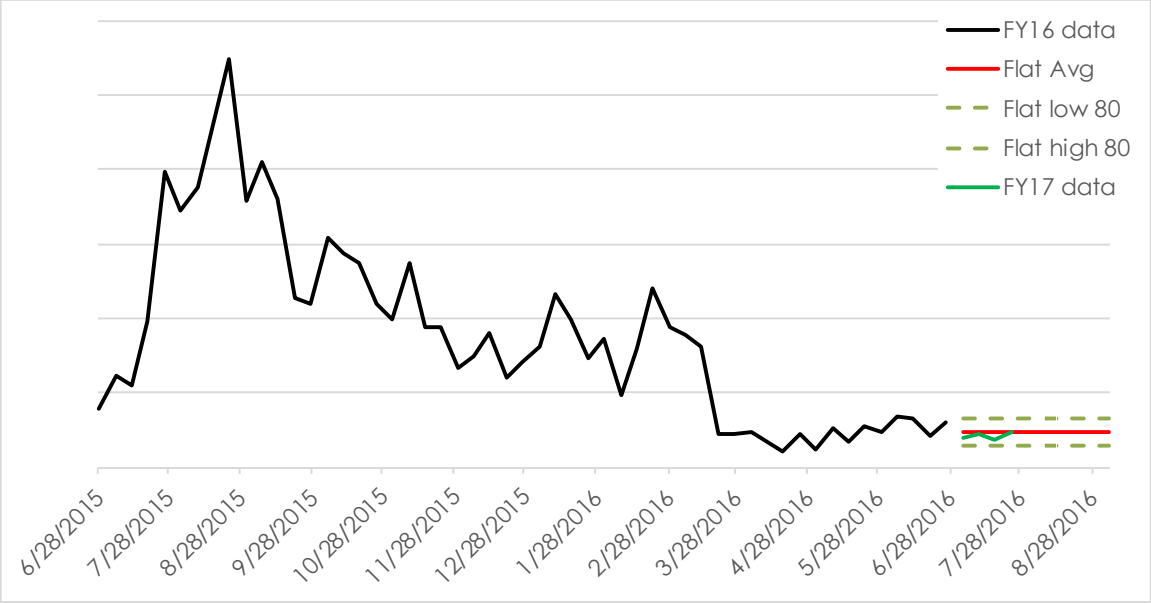
Additionally, we also look at the forecast errors for the four time frames using Holt's Exponential Smoothing (assuming trending forecasts). The graphs are not shown here. And the following forecast is shortlisted –

- Trending forecast Quarter 4 FY16

INSIGHTS

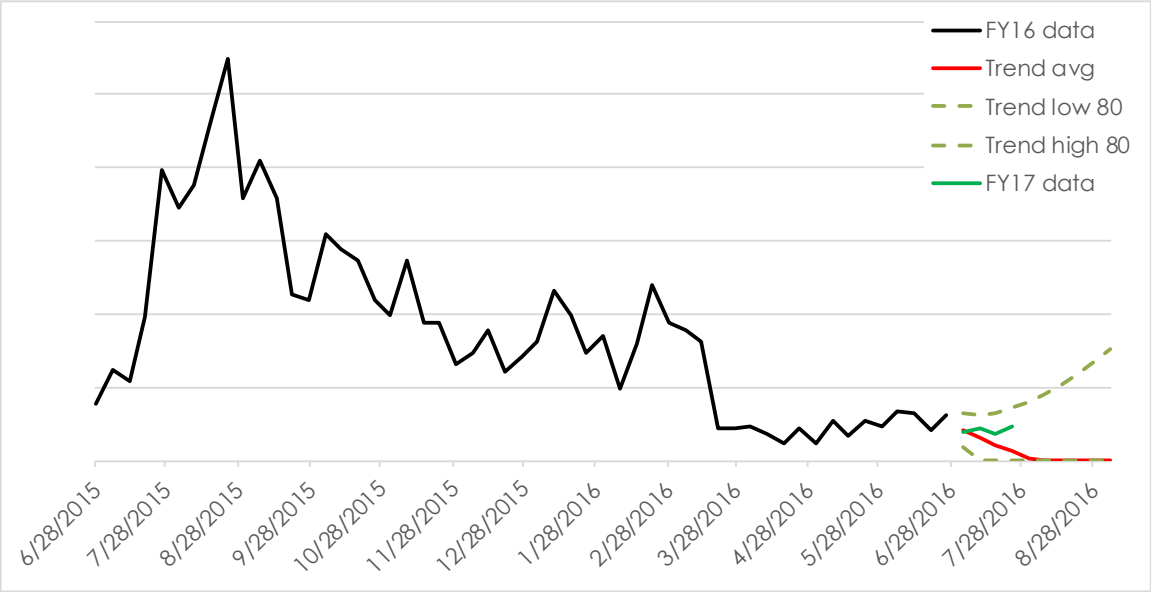
In the final phase of the analysis, we gain insights from the shortlisted forecasts and apply those insights to the business scenario. We assume 2 hypothetical situations that represent one of 2 ways that the management can base its decision upon.

Hypothetical Situation 1 – Management Decides To Remain Conservative



- If the management decides that they would like to stabilize in the future, we use the flat forecast to set the expectations. On validating the flat forecast against actual FY17 data, we see that FY17 data lies very close to the forecast line (in red).
- This means that the security events fall within the forecasted range and the company is doing well.
- The company does not need to implement any new security measures and can continue the way it is.

Hypothetical Situation 2 – Management Decides To Keep Progressing As Fast As They Were



- If the management decides that they would like to keep trending downwards just like in the past, we use the trending forecast to set expectations.
- In this case, we see that FY17 data is not falling as steeply as the forecast line, which means that progress is not being made fast enough and this should be a call for caution on the part of the company.
- In order to keep up with the forecast, the company will have to review its existing security mechanisms and find ways to improve it.

CONCLUSION AND FUTURE SCOPE

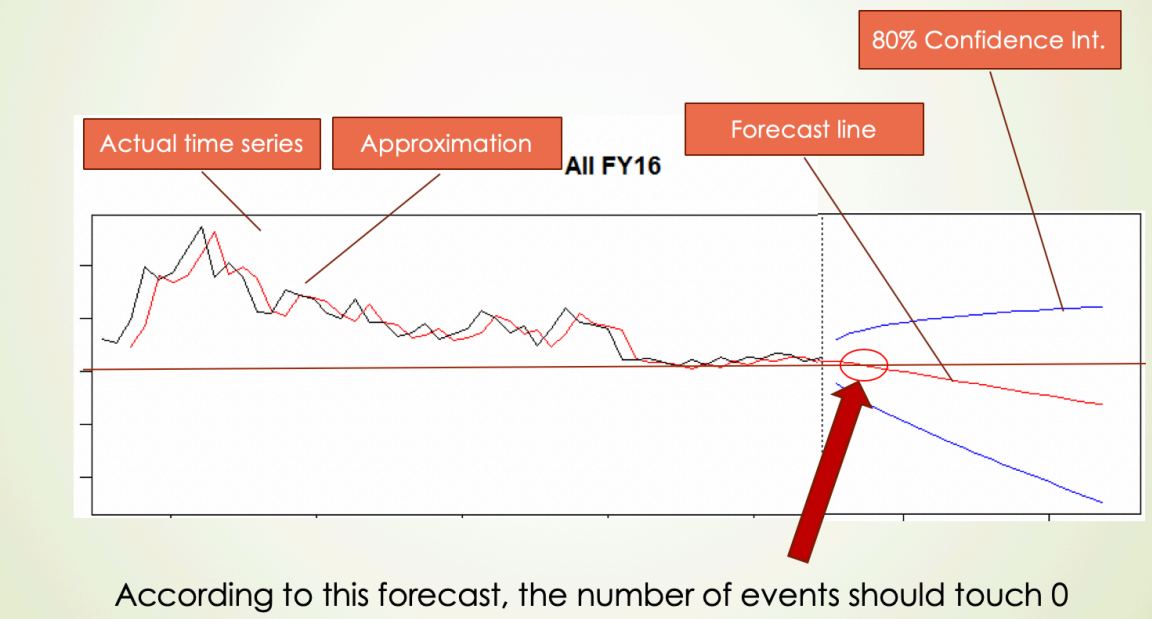
This project established a fact-based method to allow management to determine reasonable expectations of security events in the future. And therein lies the beauty of forecasting. It does not tell you what is going to happen in the future with a 100% probability. But what it does provide is a reasonable expectation so that you have a fair idea of what might happen and this added knowledge, in advance, can prove to be extremely advantageous from a business perspective.

Since these results are good for short term forecasting purposes, they can be used to forecast security events for the next month or two. The new real data for FY17 that will be generated can be fed back to the tool and new short-term forecasts can be obtained. Once data for FY17 has also been added to the current dataset, seasonality and the cyclic nature of the security events can be taken into consideration and more in-depth analysis can be performed.

Another interesting analysis that can be performed is to find out what factors affect the security events by performing regression analysis. The knowledge of how many events to expect in the near future (forecasting) coupled with the factors giving rise to these events (regression) will prove to be a powerful analytical combination and aid in reducing these security events in a very short span of time.

APPENDIX

Example of Forecasting



The above image explains the basics of forecasting –

- Actual Time Series – It is shown in black and are the actual data points from the dataset. It is a simple plot of the number of security events against time.
- Approximation – It is shown in red and is the approximation of the actual time series performed by the tool. The approximation follows the same pattern as the actual time series but with a small time lag.
- Forecast Line – The approximation, when projected into the future, becomes the forecast. In this example, the number of security events are forecasted to reach 0 in a few weeks' time. After 0, the forecast becomes negative. Since we cannot have negative number of security events, all forecasts after 0 are assumed to be 0 itself.
- 80% Confidence Interval – R also provides a confidence interval that accompanies each forecast. This parameter tells us that the forecast will lie within that interval with 80% confidence. We can change the confidence level depending on our requirements. A 95% confidence interval will mean a wider range. On the other hand, the lesser our confidence, the narrower the interval.