Enhancing Sales Forecasting Accuracy for CX's Compact Crane Using Time Series Analysis and Machine Learning Techniques

Abhinav Venkatraman
av-430

ABSTRACT

This report addresses the critical challenge faced by CX, a leading manufacturer, in accurately forecasting sales of their flagship product, the compact crane. The sales team has been struggling with production inefficiencies due to repeated overproductions and underproductions caused by inaccurate sales forecasts. The current forecasting model, which utilizes a simple moving average of the last three months, fails to capture the variability and potential seasonality in the sales data, leading to unreliable forecasts.

To tackle this challenge, advanced time series analysis and machine learning techniques were employed using IBM SPSS. The analysis involved exploring seasonal patterns and employing sophisticated models to improve the accuracy of the sales forecasting model. The objective was to provide a more robust and reliable forecast that would enable the production team to schedule the appropriate number of compact cranes one month in advance.

The findings revealed compelling insights. Seasonal dependencies were identified, suggesting the presence of recurring patterns influencing sales. While various time series techniques, such as Holt's method and exponential smoothing were initially explored, they did not yield satisfactory results in capturing the underlying patterns and improving forecast accuracy. However, the adoption of an ARIMA model proved to be effective in addressing the challenges of sales forecasting for CX's compact crane. Leveraging the identified seasonal dependencies and incorporating other relevant factors, the ARIMA model demonstrated significant improvements in forecast accuracy and production planning.

In conclusion, the implications of these findings are significant for CX. By adopting the improved sales forecasting model, the company can optimize their production scheduling and improve their production planning. This will lead to improved operational efficiency, cost savings, risk mitigation and overall better business outcomes.

SCOPE AND LIMITATIONS

The analysis focuses on the sales data which was provided for the period between January 2005 and September 2017. The dataset solely includes historical monthly sales data without any additional information.

One notable aspect is the absence of missing data within the provided dataset, ensuring good data quality and reliability. This allowed us to work with a complete and consistent dataset, minimizing potential challenges associated with missing data imputation or data cleansing.

However, the limitation of the analysis is the absence of other variables beyond the sales data which restricted the scope to time series forecasting. As a result, alternative analytical approaches such as regression analysis, which rely on external factors, could not be employed. This limited the ability to explore the impact of market dynamics or other external influences on sales. Despite these limitations, the available data allowed for in-depth exploration and application of time series forecasting techniques to derive meaningful insights into the sales patterns and trends.

Table of Contents

1. INTRODUCTION

Time series forecasting has become an essential tool for businesses and organizations to predict future trends and make informed decisions. By analyzing historical data, time series forecasting methods enable us to identify patterns, detect seasonality, and capture underlying trends.
Time series forecasting has a rich history, with the earliest developments tracing back to the mid-20th century.

One widely employed technique is exponential smoothing, which involves calculating weighted averages of past observations to generate forecasts. **Holt (1957)** extended exponential smoothing by incorporating trend components, making it suitable for data with a linear trend. (**Holt & Winters, 1960)** method further enhances this approach by incorporating seasonal components, allowing it to handle data with both trend and seasonality.**Gardner (1985)** emphasized the flexibility and simplicity of exponential smoothing models, making them suitable for a wide range of forecasting scenarios and found that the Holt-Winters method produced accurate and reliable forecasts for seasonal time series data.

The Box-Jenkins method, another prominent approach in time series forecasting, emphasizes the utilization of autoregressive (AR), moving average (MA), and autoregressive integrated moving average (ARIMA) models. AR models capture the relationship between an observation and lagged values of the series, while MA models consider the relationship between an observation and the residual errors. ARIMA models combine both autoregressive and moving average components, with the integrated part addressing non-stationarity through differencing. The effectiveness of these models has been extensively explored in the literature, with seminal works such as **Chatfield (1975)** and **(Box et al,1994)** providing comprehensive guidance on their application in practical forecasting scenarios.

Several studies have validated the capabilities of these forecasting methods across diverse domains. **(Makridakis et al. 1998)** conducted a comprehensive evaluation of various forecasting techniques and highlighted the robustness of exponential smoothing methods. Similarly, **(Hyndman & Athanasopoulos,2018)** emphasized the usefulness of ARIMA models in time series forecasting.

To assess the performance and validity of these models, various testing methodologies have been employed. These methodologies provide valuable insights into the forecasting capabilities of the models.
Holdout testing, which involves splitting the data into training and testing sets, has been widely used to evaluate the accuracy of forecasts. **Chatfield (1996)** showed that models like Holt's method, Holt-Winters method, and ARIMA models perform well in terms of minimizing error metrics such as MAE, MSE, and RMSE**.** The use of cross-validation techniques, such as k-fold cross-validation, has also been found to provide robust evaluations of model performance by considering different training and testing subsets **Taylor (2003).** This paper also demonstrated the effectiveness of k-fold cross-validation in assessing the predictive accuracy of time series forecasting models.

Diagnostic tests are crucial for assessing the quality and appropriateness of time series models. The Box-Ljung test, proposed by **(Box and Ljung,1970)**, is commonly used to assess the presence of autocorrelation in residuals, ensuring that the model captures the serial dependence adequately. **(Makridakis et al. 1998)** also highlighted the importance of diagnostic tests like the Box-Ljung test in evaluating the adequacy of time series models.

Furthermore, the t-test is employed to evaluate the significance of model coefficients, ensuring that the estimated parameters are statistically significant and reliable. The autocorrelation function (ACF) and partial autocorrelation function (PACF) plots are useful diagnostic tools for identifying the appropriate lag structure of the model. These plots provide insights into the significance and strength of the lagged relationships in the time series data **(Brockwell & Davis, 2016).**

In the next section, we will apply these testing methodologies to evaluate and validate the effectiveness of the chosen models.

2. DATA ANALYSIS & PRESENTATION OF INFORMATION

In this section, we will present the information, evidence, and analysis conducted to address the challenges of sales forecasting for CX's compact crane product. All figures and tables are presented in the appendix.

2.1 Data Exploration

To begin the analysis, an exploratory data analysis was conducted to understand the characteristics of the sales data. The descriptive statistics are presented in *Table 1*, which shows that the mean monthly sales 20284.915 was with a standard deviation of 6370.759. There are 153 sales figures.

Figure 1 shows the time series plot of the sales data, which indicates an upward trend and seasonal pattern. There is also the presence of stationarity.

2.2 Initial Model Selection

During the initial phase of the analysis, a range of models were explored to tackle the sales forecasting challenge. These models encompassed various approaches such as simple exponential smoothing, Holt's Additive and Multiplicative methods, as well as Holt-Winter's method. While these models provided a foundational level of forecasting capabilities, they were deemed inadequate in capturing the intricate patterns and seasonal dependencies observed in the sales data.

The most accurate model, the simple seasonal Log exponential method, had an RMSE of 2193, Normalized BIC of 15.452, and a stationary R-squared of 0.505. Although it initially showed promise, diagnostic tests revealed unfavorable results. The Ljung-Box test indicated a lack of fit at a significance level of 0.272, and the delta (seasonal) coefficient exceeded the threshold of 0.05 suggesting instability in the model's ability to capture seasonal variations effectively.

Considering these diagnostic test outcomes, the simple seasonal exponential method was ultimately dismissed as a suitable forecasting model for the sales data at hand.

2.3 EXPLORING ARIMA MODELS

2.3.1 Making the Data Stationary

To utilize ARIMA models effectively, it was imperative to address the issue of non-stationarity in the sales data as these models rely on the assumption of a stationary time series.

The following steps were undertaken to transform the data into a stationary form:

1. The sales variable was subjected to a Log transformation to mitigate the effects of high variance. This resulted in a new variable. As illustrated in *Figure 2*, the variance was significantly reduced. Although, it was evident that a seasonal pattern and a persistent trend were still present.
2. To eliminate the remaining trend, a first-order differencing technique was applied. This procedure stabilized the mean and diminished the trend component. However, as depicted in *Figure 3*, the seasonal pattern persisted.
3. To address the remaining seasonality, a seasonal differencing operation was performed, creating yet another variable. The resulting differenced series exhibited no discernible patterns or trends, indicating the achievement of stationarity, as shown in *Figure 4*.

2.3.2 Reading the ACF and PACF correlograms

To determine the appropriate order of the (AR) and (MA) components in the ARIMA model, the ACF and PACF plots generated for the transformed data were examined as shown in *Figure 5*.

It can be observed that the autocorrelation gradually diminishes as the lag increases. On the other hand, the PACF plot exhibits a sharp cut-off after a certain 2 lag numbers. The gradual decay of the ACF and the cut-off in the PACF beyond the second lag suggest that a model with two autoregressive terms is sufficient to capture the relevant patterns in the data. Based on these observations, it is concluded that an AR model of order 2 would be appropriate.

2.3.3 Identifying the Optimal Model

To determine the most suitable model for our data, a comprehensive evaluation of various ARIMA models was conducted. While the initial analysis suggested that an AR(2) model would be the best fit, it was crucial to thoroughly explore alternative models to ensure optimal selection.

To assess the goodness of fit for each model, several key tests were employed-

1. Box-Ljung test significance level must be over 0.05.
2. T-test significance level for the coefficients should be lower than 0.05.
3. Normalized BIC should be as low as possible.
4. ACF and PACF spikes should be within the confidence limit.

it was found that three models displayed favorable results across all tests. The details of these models are presented in *Table 2*.

It is important to note that although the AR (2) model initially appeared to be the top contender, further analysis revealed that not all coefficients met the significance criterion of 0.05. This limitation also affected the ARIMA (101)(201) model.

Ultimately, the model that successfully passed all the tests and exhibited the most promising performance was identified as the ARIMA (100)(200) model. The ACF and PACF of this model is shown in *Figure 6.*

The subsequent section will delve into the results provided by this selected model.

2.3.4 Results of the ARIMA (100)(200) model

The forecast generated by the model, shown in *Figure 7*, provides insights into the future sales trends for a period of 7 months beyond the available data, extending until April 2018. The forecast considers the underlying patterns and dependencies captured by the model, offering a glimpse into the anticipated sales performance. This analysis enables us to assess the reliability and accuracy of the ARIMA (100)(200) model in forecasting future sales and guiding decision-making processes.

3. CONCLUSION

In conclusion, this report presented a comprehensive analysis of sales data and employed various data analysis techniques to forecast future sales trends. The initial exploration of the data revealed an upward trend along with a seasonal pattern, highlighting the need for robust forecasting models to capture the complex patterns and dependencies present in the sales data.

Through the exploration of different modeling approaches, we identified the strengths and limitations of each technique. While exponential smoothing methods provided basic forecasting capabilities, ARIMA models showcased their superiority in capturing both trend and seasonal patterns, making them a more suitable choice for our sales forecasting task.

To ensure the accuracy and reliability of our models, rigorous testing and evaluation were conducted. Diagnostic tests were performed to assess the goodness of fit for each model. Based on these tests, the ARIMA (100)(200) model emerged as the most suitable choice, passing all the diagnostic tests and demonstrating favorable forecasting performance.
The forecasted sales trends provided valuable insights for future planning and decision-making. Overall, this analysis highlights the significance of employing advanced time series techniques to predict the sales of CX's compact crane.

4. RECOMMENDATIONS

Based on the findings and analysis presented, the following recommendations are put forth:
Utilize the ARIMA (100)(200) model for sales forecasting: The model demonstrated superior performance in capturing the trend and seasonal patterns present in the sales data. Therefore, it is recommended to adopt this model as the primary forecasting tool for future sales predictions.

Explore additional variables and features: it is advisable to consider incorporating additional variables and features that may influence sales, such as marketing campaigns and economic indicators. By integrating these factors into the forecasting models, a more comprehensive understanding of the sales dynamics can be achieved, leading to more accurate predictions.

5. APPENDIX

5.1. Figures:

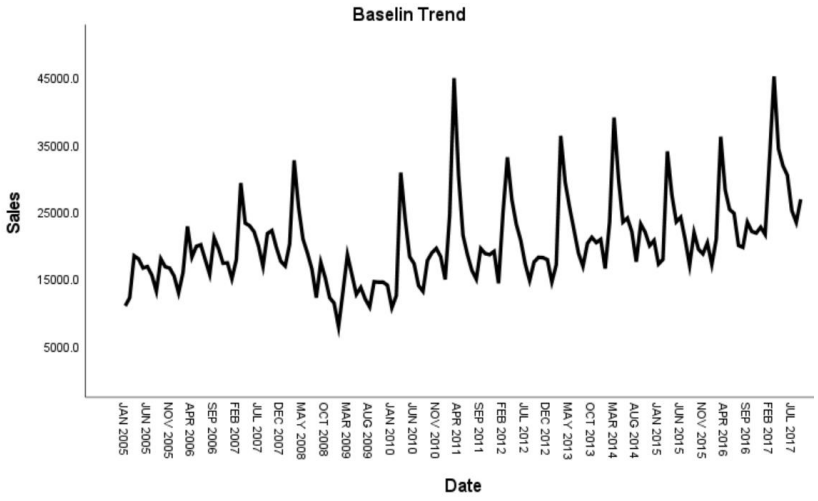

Figure 1: Sequence Chart - Baseline Trend

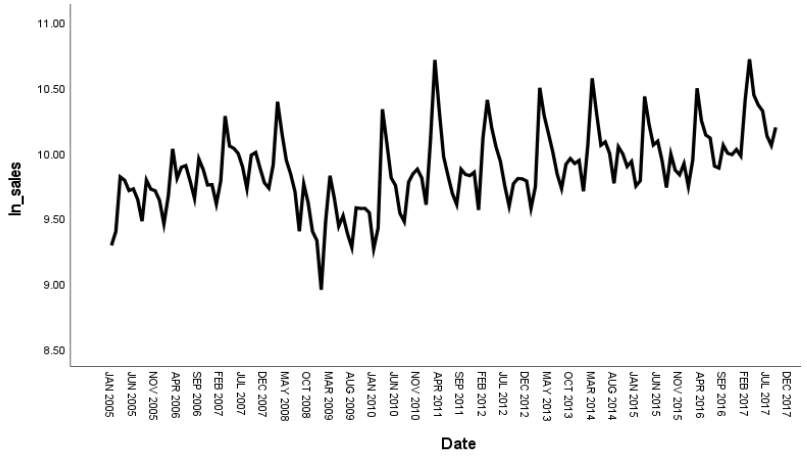

Figure 2: Sequence Chart - Trend after Log Transformation

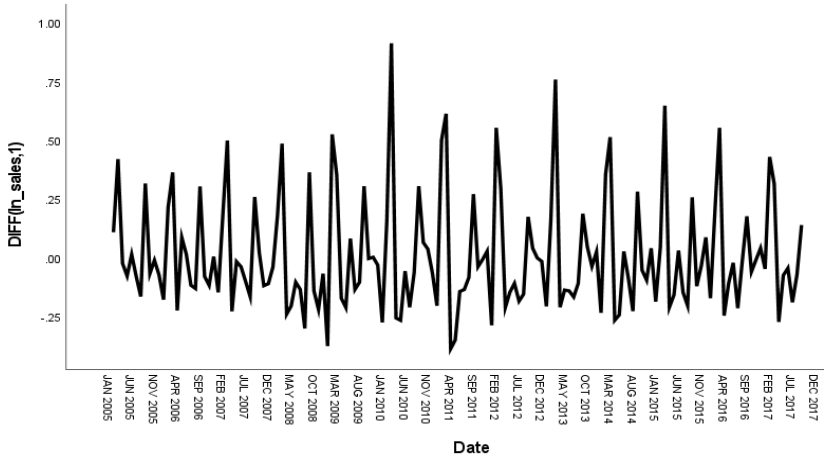

Figure 3: Sequence Chart - after first order differencing

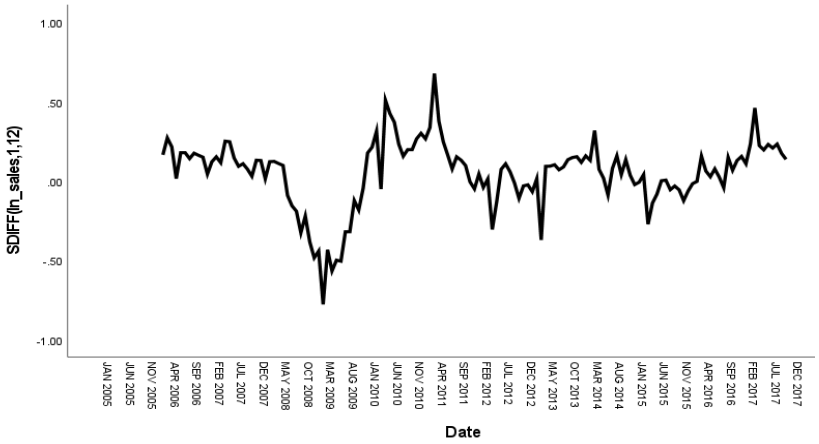

***Figure 4**: Sequence Chart - after seasonal differencing*

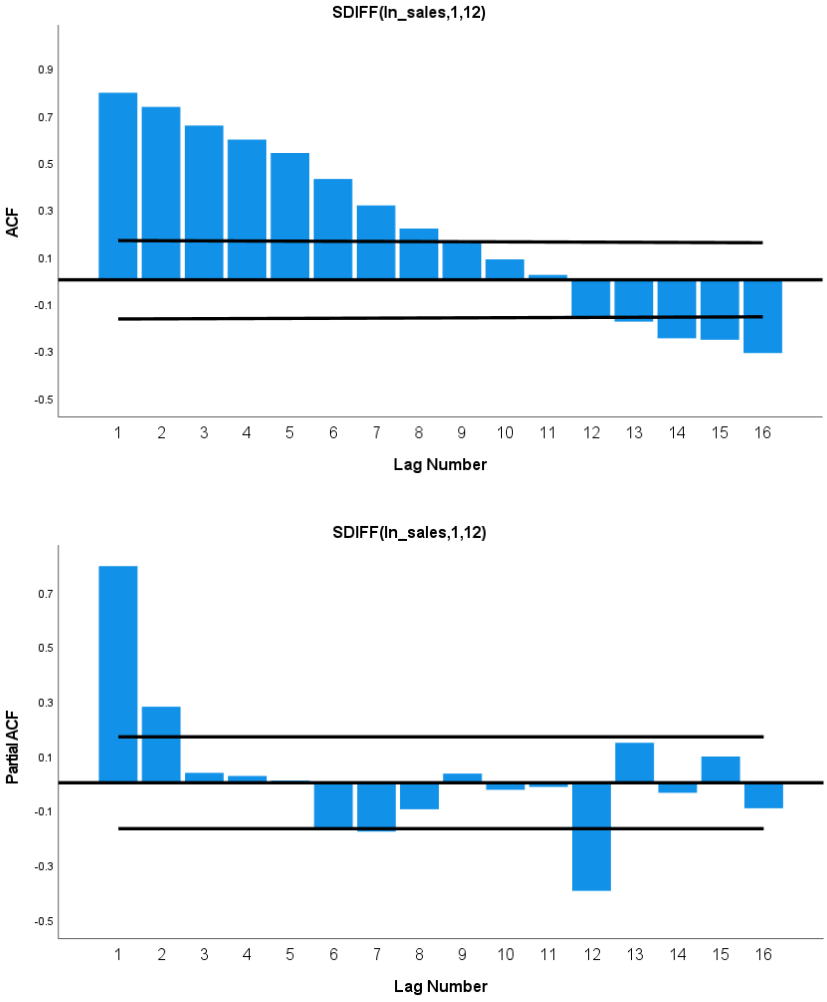

***Figure 5**: ACF PACF of ARIMA (100)(200) model*

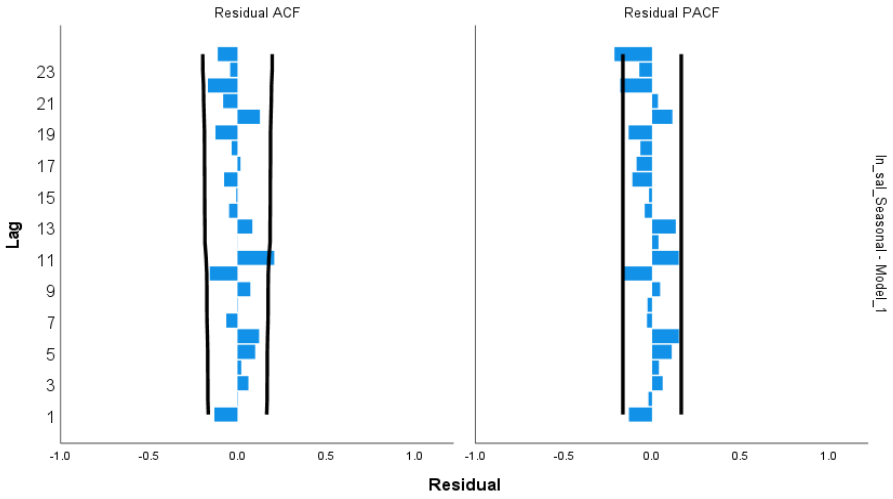

***Figure 6**: ACF & PACF of ARIMA(100)(200) model*

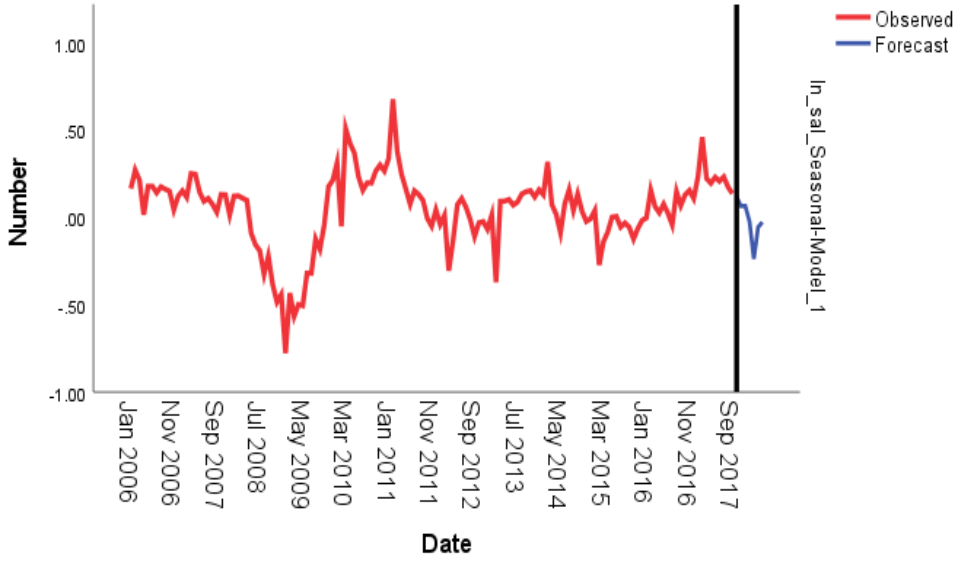

***Figure 7**: Forecast of ARIMA(100)(200) model*

5.2. Tables:

		Number	Minimum	Maximum	Mean	Std. Deviation
Sales		153	7709.0	45061.0	20284.915	6370.759

***Table1**: Descriptive Statistics*

Model	BIC	T-Test	Box-Ljung	ACF/PACF	R-Squared
ARIMA(200)(200)	-4.308	Constant (0.185) & AR Lag 2(0.031) over limit	0.354	Within Limit	0.759
ARIMA(101)(201)	-4.338	Constant (0.185)& MA Lag 1(0.028) over limit	0.332	Within Limit	0.776
ARIMA(100)(200)	-4.321	All coefficients under limit	0.126	Within Limit	0.752

Table 2: ARIMA Models: Diagnostic test results

6. REFERENCES

1. Gardner, E. S. (1985). Exponential smoothing: The state of the art. Journal of Forecasting, 4(1), 1-28.
2. Taylor, J. W. (2003). Short-term electricity demand forecasting using double seasonal exponential smoothing. Journal of the Operational Research Society, 54(8), 799-805.
3. Chatfield, C. (1975). The analysis of time series: Theory and practice. Chapman & Hall.
4. Chatfield, C. (1996). Time-series forecasting (3rd ed.). Chapman and Hall/CRC.
5. Box, G. E. P., Jenkins, G. M., Reinsel, G. C., & Ljung, G. M. (1994). Time series
6. Box, G. E., & Ljung, G. M. (1970). A note on the Box-Ljung test. Technometrics, 12(3), 743-744.
7. analysis: Forecasting and control (3rd ed.). Prentice Hall.
8. Makridakis, S., Wheelwright, S. C., & Hyndman, R. J. (1998). Forecasting: Methods and applications. John Wiley & Sons
9. Hyndman, R.J. and Athanasopoulos, G., 2018. Forecasting: Principles and Practice. [online] Available at: https://otexts.com/fpp2/
10. Brockwell, P. J., & Davis, R. A. (2016). Introduction to time series and forecasting (3rd ed.). Springer.
11. Holt, C. C., & Winters, P. R. (1960). Forecasting seasonal time series with exponential smoothing. Management Science, 6(3), 324-342.
12. Holt, C. C. (1957). Forecasting trends and seasonals by exponentially weighted moving averages. ONR Research Memorandum No. 52, Carnegie Institute of Technology