Sales Forecasting for Telecom Vertical Using ARIMA in R

Mangesh Patil[1st]
Dept. of Electronics Engineering
Terna Engineering College
Navi Mumbai, India

Renuka Chimankare[2nd]
Dept. of Electronics Engineering
Terna Engineering College
Navi Mumbai, India

Abstract— **The aim of this paper is to incorporate Time Series in order to predict sales growth and measure the demand and supply of optic fibre connections in a area. Thus, improving the conversion rates, Customer experience leading to good retention rates and the business model works on subscriptions. This prediction exhibits seasonal patterns, by analysing historical data in Time series to understand trends of seasonal as well as non-seasonal subscription numbers to forecast product Revenue & Supply of optic Fibre for next business period.. The data source used is the CRM data.**

Keywords— ARIMA, Time Series forecast, dickey fuller test, revenue prediction, sales forecast, forecast, stationary data, univariant time series forecast, R, autoarima.

I. INTRODUCTION

In Today's Growing Market, a smart decision is to understand the demand of customer and be prepared for the supply. It drives the Fundamentals of every business planning, procurement and production activities. We have analyzed the same to improve the user experience thus increase retention of users opting for Broadband Subscription. It is important to understand the demand because the service is been provided by natural resource like optic fiber and copper wire, which not only increase the cost of acquisition of customer but to also use these resources carefully. In this traditional business model, the use of data has been a key indicator to manage most things. With the help of data, we collected we analyzed the seasonal patterns understand the User behavior for a particular region and thus predicted the users for subscription for the next business quarter. this not only helped in planning but also to measure the future profitability. This was carried on past CRM data using ARIMA model in R.

II. RELATED WORK

Lot of research is happening in the field of Data Analytics. In our project we have used ARIMA model which is a forecasting-based algorithm that forecasts future values based on past data and its core components meaning the trends, seasonality and remainder associated with the dataset. This information can be used to understand the time series and predict future values.

III. AUTOREGRESSIVE INTEGRATED MOVING AVERAGE

We used only past values to predict future supply. In other words, we used univariate time series.

ARIMA stands for Auto-Regressive Integrated Moving Average. This algorithm proposes that only information from past values can predict future values. This model functions to predict x_t with values.

p (x_t | x_{t-1}, … ,x_1)
This Algorithm forecasts future values based on past values, its own deviation (latency) and predicted deviation (latency) so that this equation is used to predict future values.
To Incorporate ARIMA we need to breakdown its parameters p, q and d. where p stands for order of AR term, q stands for order of MA term and d is the number of differencing.
Auto Regressive is a linear regression model. The AR model is stated below:

$$AR(p) : x_t = \alpha + \sum_{i=1}^{p} \beta_i x_{t-i} + \epsilon$$

P value in the equation determines the number of past values. P will be taken into account for prediction. The higher the value of p the more past values will be taken into consideration. For e.g., if p value is 1 then the generalized equation will be $x_t = \alpha + \beta_1 x_{t-1} + \epsilon$. Therefore, the AR model can also be a linear combination of p past values.

MA model .

The moving average MA model is opposite of AR model. It depends on past forecast errors to make

$$x_t = \mu + \sum_{i=1}^{q} \Phi_i \epsilon_{t-1}$$

predictions

Thus, MA model can be defined as linear combination of q past forecast errors.

IV. STATIONARY & NON STATATIONARY DATA

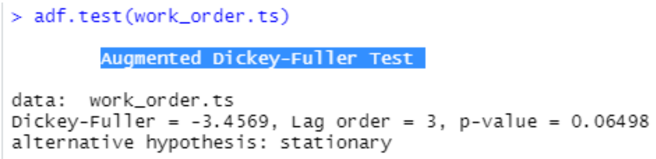
Stationary data means this data is independent of time. It does not depend on time variable.

If we want to predict basis only past values, then we require stationary data only. Not having stationary data will give you predictions unsatisfactory as your model would take the time variable into account.

To make a series stationary we need to at least difference it. i.e. subtracting the previous value from the current values. Thus, the value of d determines the minimum number to make a series stationary.

For stationary data the value of d is 0. To determine if the data is stationary or not, we can perform a Augmented Dicky fuller test.

The results of this test are stated below[1]



```
> adf.test(work_order.ts)

        Augmented Dickey-Fuller Test

data:  work_order.ts
Dickey-Fuller = -3.4569, Lag order = 3, p-value = 0.06498
alternative hypothesis: stationary
```

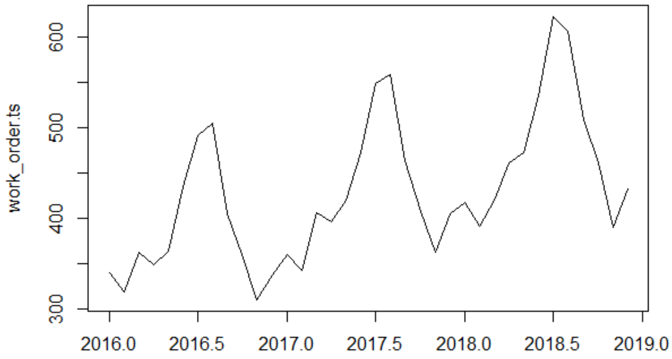
Fig [1] Stationarity check test

The hypothesis is if the p values is significantly less. The benchmark is 0.05. Thus we need to make this series as a stationary series.

In the most simpliest words, the non stationary data can be made stationary by taking the difference between the successive rows.

V. TRANSFORMING THE SERIES AS STATIONARY

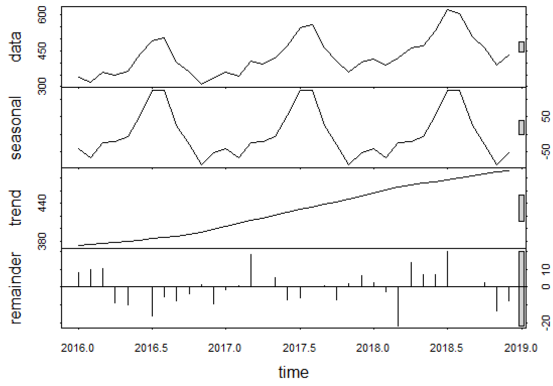
The Telecommunication sales connection data that we used is seen below [1]. This dataset used is non-stationary data.



Fig[2] Time series of the dataset.

The first step is to make this stationary data [2] in non stationary data and then decompose the series to analyse the seasonality & trend. On performing the dickey fuller test we found the lag order, value of d required to make the necessary transformation.

VI. DECOMPOSITION OF TIME SERIES



Fig[3] Decomposition of series

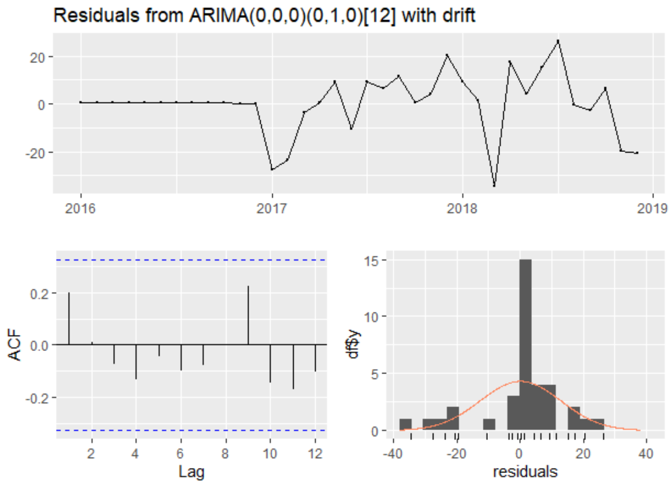
The Series is further broken down in core components. Seasonality i.e the cyclic behaviour of the series, Trends i.e overall rise or fall in the mean values, along with remainder which the noise or random varaiation associated with the data[3]. Remainder means whatever remains after removing trends and seasonality from the series. The Time series Model has assumptions that the data is stationary and only, this residual component satisfies the condition for stationary. The seasonality can be additive and multiple depending upon the data.

$$y = noise(t) = \epsilon_t$$

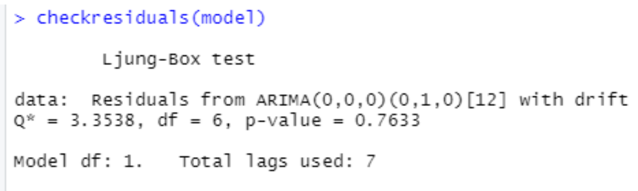
VII. EVALUATION OF THE MODEL

Step 1 is to forecast using a given model.
Step 2 is to then fit the model and get the fitted values.
Step 3 Based on these fitted values we can calculate residuals as the difference between the actual values and the fitted values.

Post which we can perform residual test. There are several tests that we can carry to check residual such as
a) Time plot: to know if the mean value is 0, if the series have any outliers, or check if we have constant variance.
b) ACF: To test whether residuals are correlated or not c) the Lung-box: if a series does not show any correlation then its white noise. To figure this out we can simply perform Lung-box test instead of looking at the ACF. If the series is white noise that means we have used all the information in our forecasting. If the Series is not white noise that means we haven't used all parameters and the residuals are correlated, thus we need to review & perform our forecasting method
d) histogram: to check whether these residuals are normally distributed or not.



Fig[4] plot of residuals, ACF and histogram

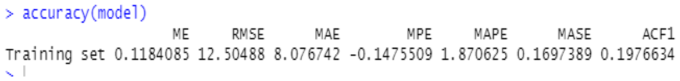


```
> checkresiduals(model)

        Ljung-Box test

data:  Residuals from ARIMA(0,0,0)(0,1,0)[12] with drift
Q* = 3.3538, df = 6, p-value = 0.7633

Model df: 1.   Total lags used: 7
```

Fig[5] residuals results.

From the figure, we can see that the ACF values are in the border range, the time plot is also hovering around mean 0 and the histogram is normally distributed.

The accuracy of this model, the MAPE value is 1.8 seen in the figure[6]



```
> accuracy(model)
                    ME     RMSE      MAE        MPE     MAPE      MASE      ACF1
Training set 0.1184085 12.50488 8.076742 -0.1475509 1.870625 0.1697389 0.1976634
```

Fig[6] accuracy of test data set.

Forecast error means difference between the actual & the observed values. Error in prediction is not mistake but the unpredictable part of the observation.

eT+h=yT+h−^yT+h|T

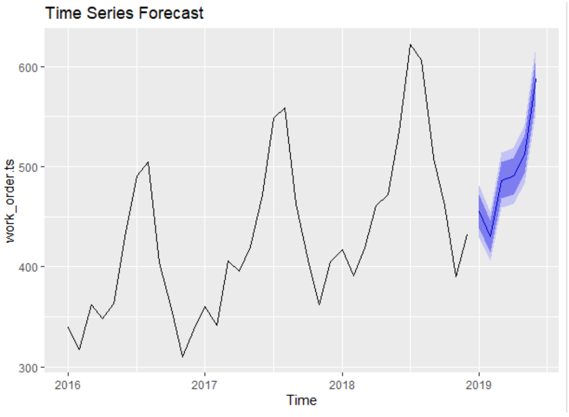
where the training data is given by {y1,…,yT}{y1,…,yT} and the test data is given by {yT+1,yT+2,…}{yT+1,yT+2,…}.
Error and residuals are not same. Residuals are calculated on training data set , while forecast are calculated on test data set. Residuals are based on one step forecast whereas errors are based on multi step forecast.

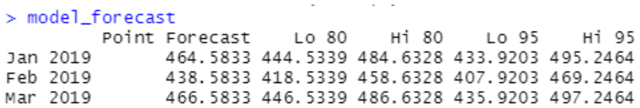
These errors are classified as

a) Scale dependent: MAE (Mean absolute error= mean(|et|),) minimizing this error will lead to forecast the median.
Root mean squared error RMSE = √m(e2t). Minimizing RMSE leads to mean forecasting.
b) Percentage error: Mean absolute error percentage MAPE=mean(|pt|). This error has advantage of being unit free and is therefore used majorly to evaluate the accuracy of forecast. For e.g., it does not consider the unit of the data like Celsius or Fahrenheit while forecasting temperature since temperature have arbitrary zero point.
c) Scaled error: it is an alternative forecasting percentage error calculator in cases with series having different units. MASE: mean(|qj|).
The training dataset used in this project had MAPE value 1.8 and thus this concludes that the model accuracy is best and good to go.

** The data used is not the company data as its confidential, but a similar data set is used for demonstration purpose of this paper.*

VIII. Results





```
> model_forecast
         Point Forecast    Lo 80    Hi 80    Lo 95    Hi 95
Jan 2019       464.5833 444.5339 484.6328 433.9203 495.2464
Feb 2019       438.5833 418.5339 458.6328 407.9203 469.2464
Mar 2019       466.5833 446.5339 486.6328 435.9203 497.2464
```

Fig 7: Forecasted series

A. *Equations*

IX. CONCLUSION

To conclude we were able to predict future sales basis the ARIMA model. To predict this, we must first plot the time series, parsing this series in its principal component. i.e. to determine if the data is stationary data or non-stationary data, observe the seasonality and trend associated with it and the remainder which is the white noise.

We can perform the Dicky Fuller test to analyse the stationarity characteristic of the series. The data set is made stationary by differencing method or logarithmic method post which we can then again perform the dickey fuller test after each difference to confirm the stationarity

Next steps is to forecast, we used the ARIMA model to build the forecast and fit the model to get fitted values. On checking the residues will tell us if we have used all the data pointers.

Thus, we could forecast the number of sales for future months with a accuracy of MAPE = 1.8

ACKNOWLEDGMENT

The Author shows appreciation to his research supervisor to guide him with this research paper. The Author also gives special thanks to the internship for teaching the valuable knowledge of process of the data set and its application, as getting through this research required more than academic support.

REFERENCES

[1] J.-H. B¨ose, et al., Probabilistic demand forecasting at scale, Proc. VLDB Endow. 10 (12) (2017) 1694–1705.

[2] Zimmermann, S.; Herrmann, P.; Kundisch, D.; Nault, B. Decomposing the Variance of Consumer Ratings and the Impact on Price and Demand. Inf. Syst. Res. 2018.

[3] Jadhav, V.; Chinnappa Reddy, B.; Gaddi, G. Application of ARIMA model for forecasting agricultural prices. J. Agric. Sci. Technol. 2017, 19, 981–992.

[4] Rangel-González, J.A.; Frausto-Solis, J.; Javier González-Barbosa, J.; Pazos-Rangel, R.A.; Fraire-Huacuja, H.J. Comparative Study of ARIMA Methods for Forecasting Time Series of the Mexican Stock Exchange. In Fuzzy Logic Augmentation of Neural and Optimization Algorithms: Theoretical Aspects and Real Applications; Castillo, O., Melin, P., Kacprzyk, J., Eds.; Springer: Cham, Germany, 2018; pp. 475–485.

[5] Ozturk, S.; Ozturk, F. Forecasting Energy Consumption of Turkey by Arima Model. J. Asian Sci. Res. 2018, 8, 52–60

[6] Meyler, A.; Kenny, G.; Quinn, T. Forecasting Irish Inflation Using ARIMA Models; Central Bank of Ireland: Dublin, Ireland, 1998.

[7] Geetha, A.; Nasira, G. Time-series modelling and forecasting: Modelling of rainfall prediction using ARIMA model. Int. J. Soc. Syst. Sci. 2016, 8, 361–372.

[8] Hyndman, R.J.; Athanasopoulos, G. Forecasting: Principles and Practice; OTexts: Melbourne, Australia, 2018

[9] Yang, W.; Wang, J.; Niu, T.; Du, P. A hybrid forecasting system based on a dual decomposition strategy and multi-objective optimization for electricity price forecasting. Appl. Energy 2019, 235, 1205–1225.

[10] Dickey, D.A.; Fuller, W.A. Distribution of the Estimators for Autoregressive Time Series with a Unit Root. J. Am. Stat. Assoc. 1979, 74, 427–431.

[11] Dickey, D.A.; Fuller, W.A. Likelihood Ratio Statistics for Autoregressive Time Series with a Unit Root. Econometrica 1981, 49, 1057–1072

[12] Wang, Y.; Wang, C.; Shi, C.; Xiao, B. Short-term cloud coverage prediction using the ARIMA time series model. Remote Sens. Lett. 2018, 9, 274–283.