International Journal of Engineering & Technology, 7 (2.7) (2018) 1107-1110

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET

Research paper

Explore Big Data and Forecasting Future Values using Univariate Arima Model in R

S. Sagar Imambi[1]*, P.Vidyullatha[2], M.V.B.T.Santhi[3], P. Haran Babu[4]

[1,2,3]Department of Computer Science and Engineering,
Koneru Lakshmaiah Education Foundation , Vaddeswaram, Guntur, Andhra Pradesh
[4]Software engineer, HCL technologies, Hyderabad
**Corresponding author E-mail: simambi@kluniversity.in*

Abstract

Electronic equipment and sensors spontaneously create diagnostic data that needs to be stocked and processed in real time. It is not only difficult to keep up with huge amount of data but also reasonably more challenging to analyze it. Big Data is providing many opportunities for organizations to evolve their processes they try to move beyond regular BI activities like using data to populate reports. Predicting future values is one of the requirements for any business organization. The experimental results shows that time series model with ARIMA (3,0,1)(1,0,0) is best fitted for predicting future values of the sales.

Keywords: *ARIMA; Big data; Time-series; Univariate.*

1. Introduction

Data is generated continuously and at an ever-growing rate. Modern devices such as cell phones, electronic media and photo technologies to find out a medical diagnosis-will generate more and more new data, and that is to be stocked somewhere for some utility. Electronic equipment and sensors spontaneously create diagnostic data that needs to be stocked and processed in real time. It is not only difficult to keep up with huge amount of data but also reasonably more challenging to analyze it, particularly when it is not matching with notions of conventional data structure, to recognize useful patterns and draw useful information. With this big data, there is every chance to transform business, science, government and day to day life.[3].

Many industries have led the way in establishing their skill to collect and exploit data:

• Credit card companies watch every transaction their customers do and can observe wrong transactions with a high degree of precision using rules formed by processing billions of transactions.

• Mobile phone companies evaluate subscribers' calling patterns to find out, for example, whether a caller's regular contacts are on a different network. If that opponent network is giving an attractive offer that might affect the subscriber to defect, the mobile phone company can give the subscriber a gift to remain in her contract.

• For companies like Linked In and Facebook, information is their basic product. The assessment of these companies are heavily obtained from the data they collect and stock which includes more and more inherent value as the data increases.

Three attributes prominent in defining Big Data characteristics:

• Large volume of data: Rather than thousands or millions of rows, Big Data can be billions of rows and millions of columns.

• Complexity of data types and structures: Big Data emulates the mixture of new data origins, patterns and structures including digital traces being left on the web and other digital repositories for subsequent analysis.

• Pace of new data generation and progress: Big Data can describe high velocity information with fast data gulp and near real time analysis. Even though the volume of Big Data likely to draw the most attention, the variety and velocity of the data give a more apt definition of Big Data in general. (Big Data is sometimes expressed as having 3 Vs: volume, variety and velocity.)

Big Data may not be effectively described using only conventional databases or procedures because of its size or shape. Hence, Big Data problems need new tools and technologies to stock, control and visualize the business benefits. Thus, these new tools and technologies implement creation, manipulation and management of Big Data sets and the storage surroundings that house them.[1].

Another definition of Big Data comes from the McKinsey Global report from 2011:

Big Data is data whose scale, distribution, diversity, and/or timeliness require the use of new technical architectures and analytics to enable insights that unlock new sources of business value.

Big Data comes from multiple sources like social media, sensors, the Internet of Things, video surveillance and many sources of data that may not have been recognized data even a few years back. Since businesses combat to sustain with changing market requirements, some companies are exploring new ways to handle Big Data to their increasing business needs and more complex problems. As the Big Data is providing many opportunities for organizations to evolve their processes they try to move beyond regular Bl activities [5]. For example using data to populate reports and dashboards and move toward Data Science- driven projects, to identify opinions from customers and complex questions [2].

2. Methodology

2.1. Time Series Analysis

Time series analysis tries to model the hidden framework of raw data taken over a period. A time series, generally expressed as Y =a+ bX , is an ordered sequence of equally spaced values over time.

Main goals of time series analysis are

- To identify the distribution of data, and check the best fit model.
- To generate future values based on the time series.

Time series analysis can be applied in various domains like finance, biology, engineering, retail, and manufacturing and economics.

Usecase1-The customer data related to monthly sales of clothing retailer.

Big data tools are used to store the high volume of customer data. Time series can be applied here to forecasts needs of customer for the seasonal aspects, factors influencing the customer decisions in purchasing.

Usecase:2-To identify Spare Parts demand in a service organization.

Based on the services offered by company, and the types of services we are able to predict demand for the spare parts. Taking input variable as part number, part failure rate, service diagnostic effectiveness the complex ARIMA model was developed.

Use case 3- Stock trading:

To predict market opportunity, the relation between prices of two stocks is observed. The stock prices distribution shows how Company X and Company Y consistently changes. It may be in proportion or inverse proportion. Through time series analysis the right to invest can be easily predicted.

2.2. ARIMA MODEL

The statistical method ARIMA (Auto Regressive Integrated Moving Average) is very famous model to transfer function data in to stationary and to predict univariate time series data. . An ARIMA model decides a value in a stationary time series as a linear combination of its previous values, past errors. It can also find out the relation between and current and past values of two time series[1].

The ARIMA model was initially experimented by Box and Jenkins and ARIMA models, so they sometimes mentioned as Box-Jenkins models. [6][7].

According to Box and Jenkins(1976), the arima modelling fallows the three main stages they are

1. Identify
2. Estimate
3. Forecast
4. Identification stage

The first stage use the IDENTIFY statement to indicate the response of series and to find candidate ARIMA models for it. In this stage time series data is analysed and differentiated and calculate the autocorrelations, inverse autocorrelations (ACF) and partial autocorrelations(PACF). Stationary tests can be performed to determine if differencing is necessary. The output of this stage is usually suggests one or more best fit ARIMA model. It removes any trends or seasonality in the time series.

A time series is said to be stationary if it met the following features.

- The expected value (mean) of series should be constant.
- The variance of series should be finite
- The autocorrelation of series is constant over time.

A stationary time series is one whose statistical properties such as mean, variance, autocorrelation, etc. are all constant overtime.

In the second stage the ESTIMATE statement specify the ARIMA model to fit to the variable identified in the previous stage IDENTIFY statement. The ESTIMATE statement not only estimate the parameters of model but also produces diagnostic statistics. It helps to judge the competency of the model. By using Significance test for parameter estimation, the unnecessary terms are identified. Tests for white noise residuals indicate whether the residual series contains additional information that might be utilized by a more complex model. Repeat the estimation stage until positive result of diagnostic test with different models.

FORECAST statement is developed in third stage i.e., forecast stage. This generates the future values of time series and confidence intervals.

3. Experimental result

Our algorithm is implanted in R tool and we consider the time series data set of sales.

We identified the relation between Orders A target sales and Order B and target sales. The distributions of target sales are shown in fig.1. The log of sales and differentiated sales are visualized to make the data as stationary.

The fig. 4 shows ACF and PACF plots to evaluate the autocorrelations. We applied ARIMA model on the differentiated data set and the best-fit univariate model (1,0,0) is used to predict target sales values for 2018. .

Once best fit model has been determined, future values in the time series can be forecasted represented in fig 5.

CODE:

```
orders <-as.data.frame(read.csv("d:/ordera.csv"))
library(forecast)
data = ts (orders[,3],start = c(2001),frequency = 4)
plot(data, xlab ='year', ylab = 'Target sales')
plot(diff(data),ylab ='Differenced  Sales')
plot(log10(data),ylab ='Log ( Sales)')
par(mfrow = c(1,2))
acf(ts(diff(log10(data))),main='ACF  Sales')
pacf(ts(diff(log10(data))),main='PACF  Sales')
ARIMAfit = auto.arima (log10(data), approximation=FALSE,
trace=FALSE)
summary (ARIMAfit)
par(mfrow = c(1,1))
pred = predict(ARIMAfit, n.ahead = 36)
pred
plot(data,type="l",xlim=c(2001,2018),ylab ="Sales")
lines(10^(pred$pred),col="blue")
lines(10^(pred$pred+2*pred$se),col="orange")
lines(10^(pred$pred-2*pred$se),col="orange")
```

Predicted Arima Model-1 Summary:

```
summary(ARIMAfit)
```

```
Series: log10(data)
ARIMA(3,0,1)(1,0,0)[4] with non-zero mean

Coefficients:
   ar1          ar2       ar3     ma1     sar1    mean
  -0.1578     -0.5753  -0.6206  -0.543  -0.4219  0.5704
  s.e. 0.1245   0.0713   0.1175   0.145   0.1260  0.0017

sigma^2 estimated as 0.008769:
log likelihood=58.01

AIC=-102.02   AICc=-99.86   BIC=-87.36
```

Table 1: Error measures-Model

	ME	RMSE	MAE	MPE	MAPE	MASE	ACF1
Training set	0.00024	0.08	.068	-3.06	13.34	0.41	-0.028

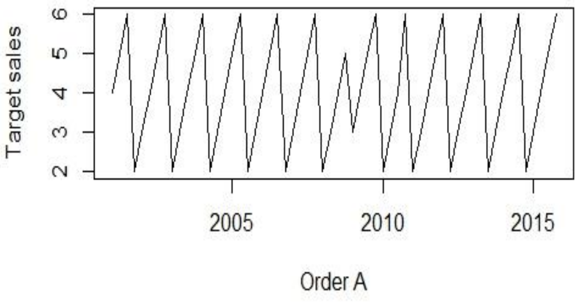

Fig. 1: Order A target sales distribution.

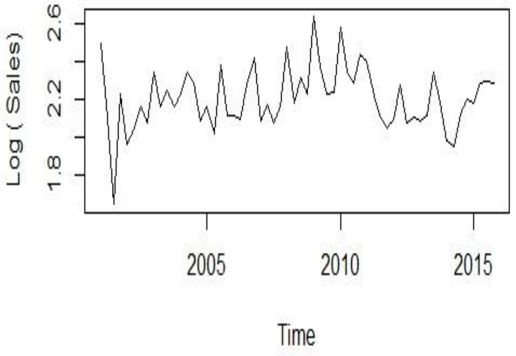

Fig. 2: Log values of target Sales.

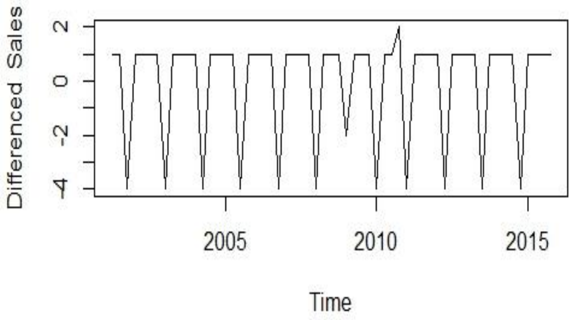

Fig. 3: Differentiated values of target Sales

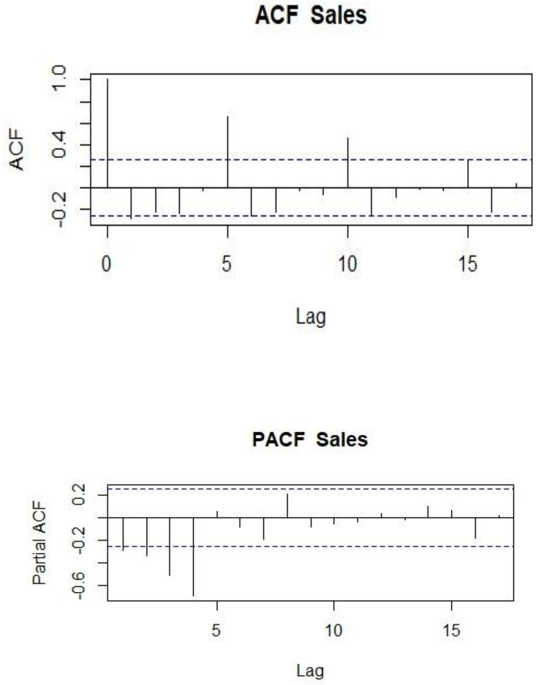

Fig. 4: Moving Average of Target sales

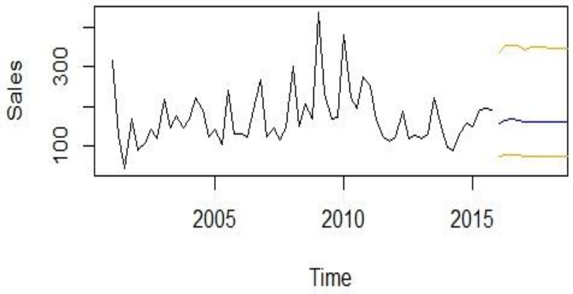

Fig. 5: predicted sales for 2018

Predicted Arima Model-2:

```
  summary(ARIMAfit)
  Series: log10(data)
  ARIMA(0,0,0)(1,0,0) with non-zero mean
```

Coefficients:

```
  sar1      mean
  0.2176    2.2056
  s.e. 0.1414   0.0259

sigma^2 estimated as 0.02643:
log likelihood=24.79
AIC=-43.57   AICc=-43.14   BIC=-37.29
```

Table. 2: Error measures-Model2

	ME	RMSE	MAE	MPE	MAPE	MASE	ACF1
Training set	0.0013	0.15	.122	-0.48	5.62	0.79	0.17

```
summary(ARIMAfit)
Series: log10(data)
ARIMA(0,0,0)(1,0,0)[4] with non-zero mean
```

Coefficients:

```
       sar1      mean
     0.2176    2.2056
  s.e. 0.1414   0.0259

sigma^2 estimated as 0.02643:
log likelihood=24.79
AIC=-43.57   AICc=-43.14   BIC=-37.29
```

Table 1 and Table 2 show the error measures of the two arima models for the training set. Fig 5 shows the future sales values in 2018 based on the training data.

4. Conclusion

Time series analysis is different from other statistical techniques in the sense that most statistical analyses assume the observations are independent of each other. Time series analysis implicitly addresses the case in which any particular observation is somewhat dependent on prior observations. Using differencing, ARIMA models allow non-stationary series to be transformed into stationary series to which seasonal and non-seasonal ARMA models can be applied. To identify the best fit model error measures MAE and other measures are considered. The best fit model is identified and future sales values are predicted using ARIMA (3,0,1)(1,0,0).

References

[1] Contreras, Javier, et al. "ARIMA models to predict next-day electricity prices." IEEE transactions on power systems, pp:1014-1020 (2003).

[2] S.Sagar Imambi et.al, "Analysing Customer Reviews using Opinion Mining,,International Journal of Advanced Research in Computer Science and Software Engineering, Vol 5,No.11,pp:12-15(2015).

[3] David Dietrich et al. "Data Science & Big Data Analytics: Discovering, Analyzing, Visualizing and Presenting Data", John Wiley & Sons, Inc.,pp:234-254 (2015).

[4] D. R. John Gantz, "The Digital Universe in 2020: Big Data, Bigger Digital Shadows, and Biggest Growth in the Far East," IDC (2013).

[5] Lakshmi Prasanna et al ,"Big data for Mobile applications in Retail market" ARPN Journal of Engineering and Applied Sciences ,Vol.10 No.18 (2015).

[6] Vidyullatha P et al, "Knowledge Based Information Mining on Data using Statistical Approaches", International Journal of Technology, Vol.8 , NO. 4(2016).

[7] P.Srikanth et al , "Comparative analysis of ANFIS, ARIMA and Polynomial Curve Fitting for Weather Forecasting " , International Journal of Science and Technology, Vol.9.No 15(2016).