AdvanceML_Discussion1

Ali Alghaithi

10/20/2020

Data

I can not use my poject data becasue it is not a time series. for this disscussuin we would need diifernt data. I choose to use data set from my Bunnies forecasting class call olist data set. the data is listed in Kaggle here: https://www.kaggle.com/olistbr/brazilian-ecommerce

After cleaning the data. I created a time series that's only has sales valume for each day.

- Reading the data

```
library(readr)
library(ggplot2)
library(kableExtra)

Sales_and_date_df <- read_csv("/Users/alialghaithi/Box/BF_Class/BF_Midterm/Sales_and_date_df.csv")
```

```
## Parsed with column specification:
## cols(
##   order_purchase_timestamp = col_date(format = ""),
##   sales_volume = col_double()
## )
```

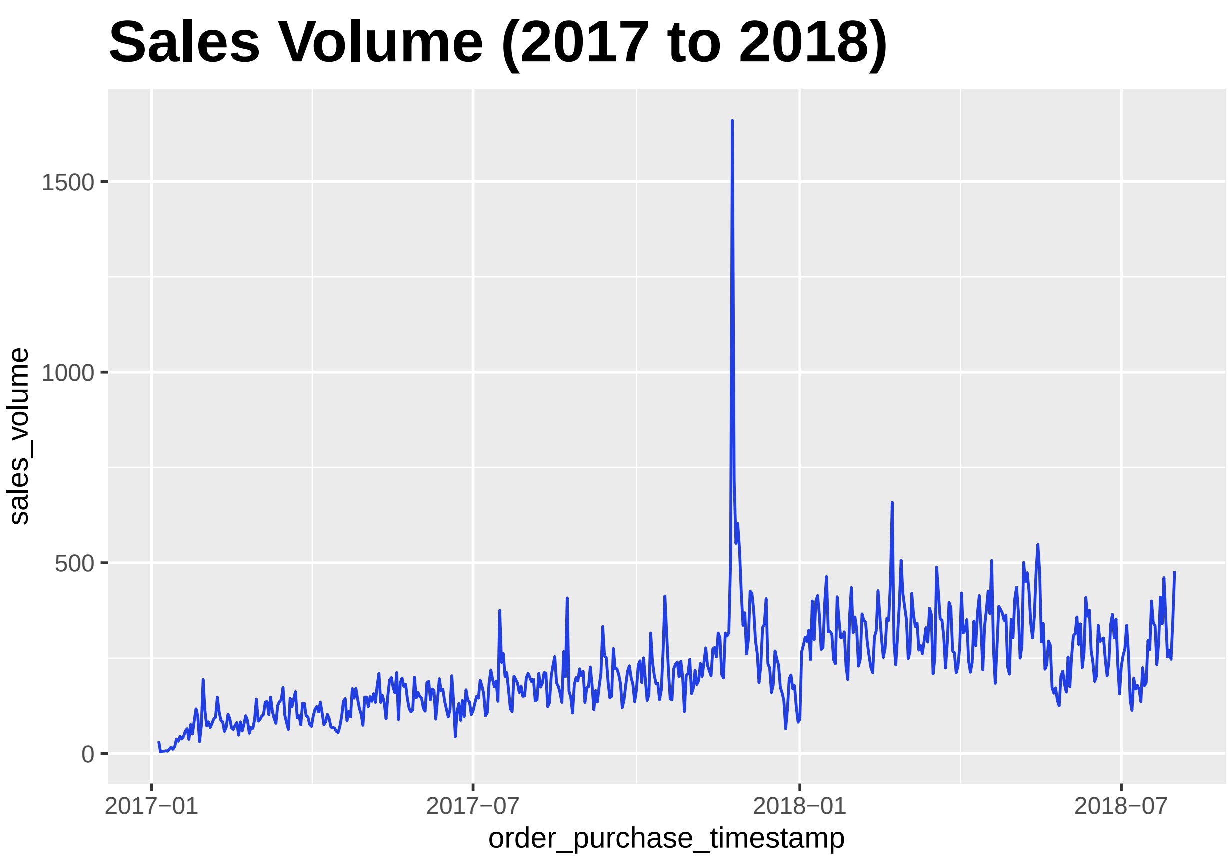
```
kable(head(Sales_and_date_df))
```

order_purchase_timestamp	sales_volume
2017-01-05	32
2017-01-06	4
2017-01-07	6
2017-01-08	6
2017-01-09	7
2017-01-10	6

```
Sales_and_date_df <- Sales_and_date_df[Sales_and_date_df$order_purchase_timestamp >= as.Date('2017-01-0
                                        & Sales_and_date_df$order_purchase_timestamp < as.Date('2018-08-0


# Looking at the overall time series trend
p <- ggplot(Sales_and_date_df, aes(x=order_purchase_timestamp, y=sales_volume)) +
  geom_line(color = "#213ee2") +
  ggtitle("Sales Volume (2017 to 2018)") +
  theme(plot.title = element_text(size = 22, face = "bold"))

p
```



data preperation

I prepared the data that so we can forecast 74 days.

```
# data preperation
ts_data <- ts(Sales_and_date_df$sales_volume)
split_point1 = 500
split_point2 = 450
trainset1 <- ts_data[1:split_point1-1]
testset1 <- ts_data[split_point1:573]
testset2 <- ts_data[split_point2:573]
```

Methods Impimintation

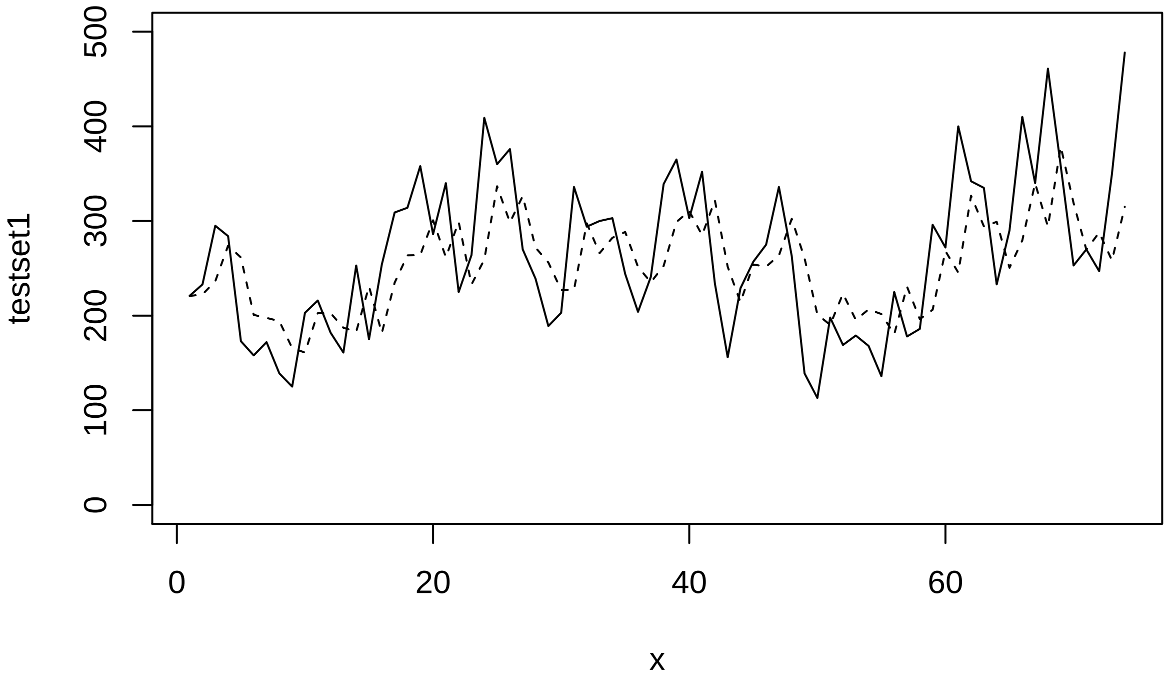
pure Arima

-After checing the time series, I decided to choose the order as listed in the model.

```
library(forecast)
```

```
## Registered S3 method overwritten by 'quantmod':
##   method            from
##   as.zoo.data.frame zoo
```

```r
dat.arima.model= arima(trainset1,order=c(4, 1, 2))
dat.test.Arima=Arima(testset1,model=dat.arima.model )
x=1:74
plot(x,testset1,ylim=c(0,500),type="l")+ lines(dat.test.Arima$fitted,lty=2)
```



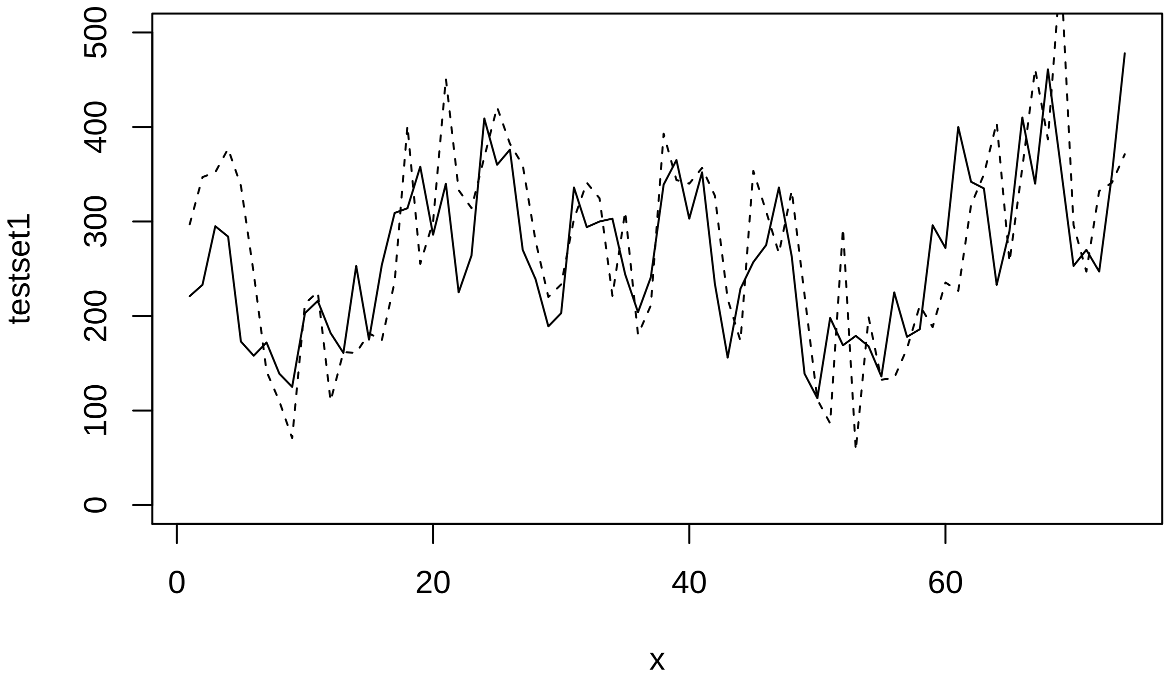

```
## integer(0)
```

```r
MSE_Arima=mean((testset1-dat.test.Arima$fitted)^2)
MAD_Arima = mad((testset1-dat.test.Arima$fitted))
```

pure ANN

For the parameter of ANN model, i Decide to go with the recommended parameters.

```r
dat.ANN.Model = nnetar(trainset1)
dat.ANN.Model.fore = nnetar(testset2,model= dat.ANN.Model)
one.step = subset(fitted(dat.ANN.Model.fore),start=51)
x=1:74
plot(x,testset1,ylim=c(0,500),type="l")+lines(x,one.step,lty=2)
```




```
## integer(0)
```

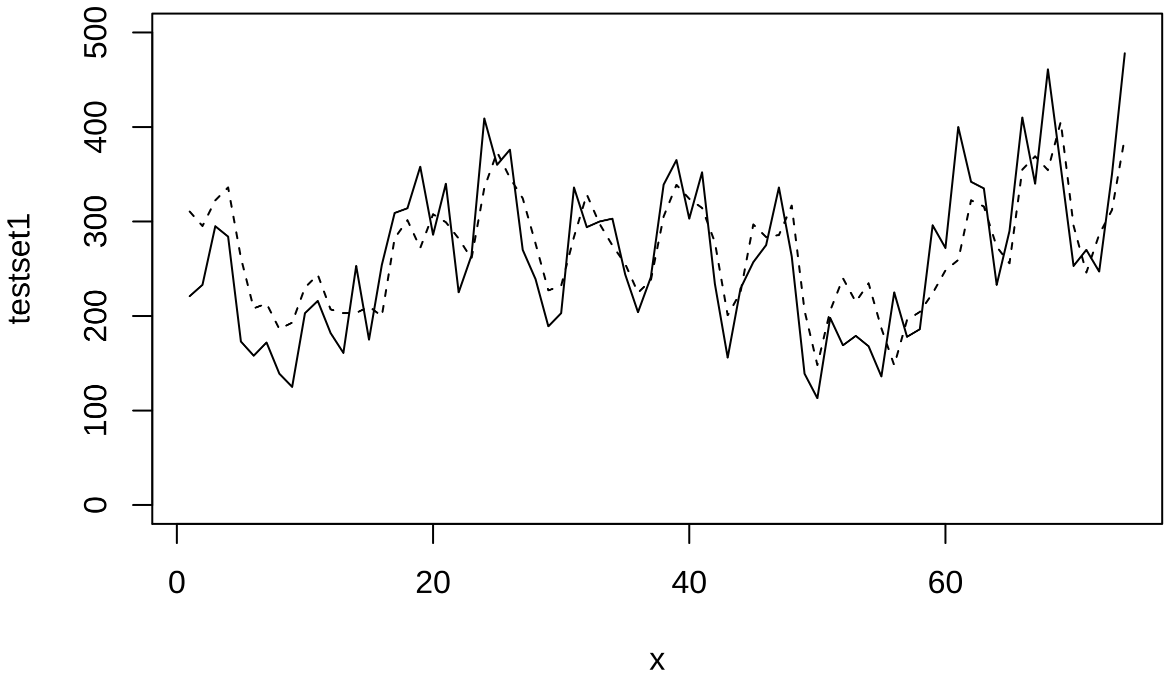
```
MSE_ANN <- mean((testset1-one.step)^2)
MAD_ANN<- mad((testset1-one.step))
```

Hybird method

```
dat.Hybird.step1= arima(ts_data,order=c(4, 1, 2))
rediduals = Arima(ts_data,model= dat.Hybird.step1)
res = dat.Hybird.step1$residuals

dat.Hybird.step2 = nnetar(res)
dat.Hybird.step2.fore = nnetar(res,model=dat.Hybird.step2)

hybird.one.step= subset(fitted(dat.Hybird.step2.fore),start=500)
x=1:74
plot(x,testset1,ylim=c(0,500),type="l")+ lines(x , rediduals$fitted+hybird.one.step,lty=2)
```




```
## integer(0)
```

```
MSE_Hybird <- mean((testset1-(rediduals$fitted+hybird.one.step))^2)
MAD_Hybird<- mad((testset1-(rediduals$fitted+hybird.one.step)))
```

MSE and MAD

```
MSE <- data.frame("Method"= c(" Arima","ANN","Hybird"),
                  "MSE" =c(MSE_Arima,MSE_ANN,MSE_Hybird),"MAD"=c(MAD_Arima,MAD_ANN,MAD_Hybird))
kable(MSE)
```

Method	MSE	MAD
Arima	4381.511	69.01759
ANN	5909.548	76.63182
Hybird	2395.235	45.51921

Conclusion

- From the findings, it shows that the Hybrid method does better than the other models based on the MSE and MAD, which agrees with the paper. This method works very well since we are only having the time and a value, but in many cases, we usually encounter time series with other variables. However, the ARIMA-ANN Hybrid Model does best at modeling the linear and nonlinear behaviors in the data set. The ANN-ARIMA hybrid model can overall achieve more accurate results. To have the ARIMA-ANN Hybrid Model more effective we will need more data points.Alos the model reduces the chance of overfitting witch is a great advantage of this model. I would also say that this model might not be best for catching trends in the time series and maybe applying TSLM would be better. for example, in the above data we can consider adding a new column called vacation where we determined the date of the vacations and help predict sales more efficiently in that case, but for ARIMA-ANN Hybrid molded it would not be able to do that. Fitting a linear model to each store time series (Sales) including trend, seasonality components and date of vacation might have better results.