ปฏิบัติการที่ 8
การใช้เครื่องมือวิเคราะห์ข้อมูลอนุกรมเวลา

วัตถุประสงค์

1. เพื่อให้สามารถใช้เครื่องมือวิเคราะห์ข้อมูลอนุกรมเวลาได้
2. เพื่อให้รู้วิธีการเพิ่มหรือพัฒนาเครื่องมือการวิเคราะห์ข้อมูลด้วยการพัฒนาคำสั่งขนาดเล็ก

1. ชุดข้อมูลปฏิบัติการ

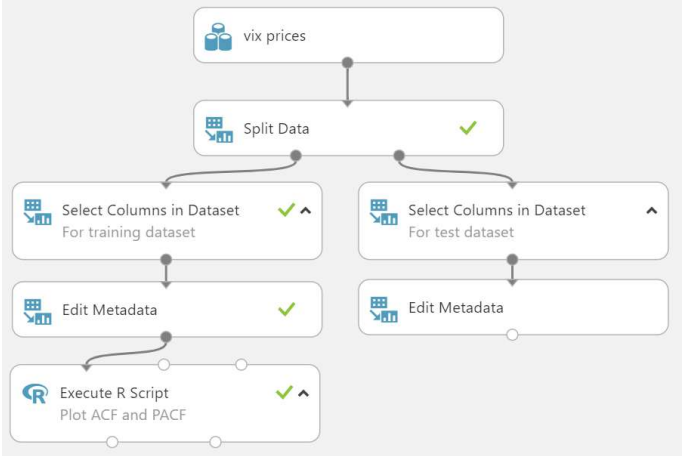
– ชุดข้อมูล VIX Prices (สำหรับการสาธิตและการฝึกปฏิบัติการ)

2. ขั้นตอนปฏิบัติการ

ขั้นตอนปฏิบัติการ มีดังนี้

1. นำเข้าชุดข้อมูล VIX Prices จากแฟ้มข้อมูล vix_prices.csv ตั้งชื่อชุดข้อมูลเป็น vix prices
2. ทำการสร้างการทดลอง โดยกำหนดชื่อการทดลองเป็น “Practice 8”
3. นำชุดข้อมูล vix prices เข้าสู่การทดลองโดยลากโมดูลชุดข้อมูลซึ่งที่อยู่ภายใต้ Saved Datasets → My Datasets ในหน้าต่างย่อย Modules มาวางบน Workspace
4. ตรวจสอบชนิดข้อมูลของตัวแปร date โดยจะต้องเป็นชนิดข้อมูล DateTime Feature หากไม่ใช่ให้ทำการแก้ไขชนิดข้อมูลให้ถูกต้องโดยใช้โมดูล Edit Metadata
5. ทำการแบ่งชุดข้อมูล vix prices ออกเป็น 2 ชุด คือ ชุดข้อมูลเรียนรู้ (Training Dataset) และชุดข้อมูลทดสอบ (Test Dataset) โดยใช้โมดูล Split Data (ภายใต้ Data Transformation → Sample and Split)
6. ให้ข้อมูลในชุดข้อมูลเรียนรู้เป็นข้อมูลหลังปี ค.ศ.2005 ส่วนข้อมูลในชุดข้อมูลทดสอบเป็นข้อมูลระหว่างปี ค.ศ.2005 ถึง 2019 ทำได้โดยเลือก Splitting mode เป็น Relative Expressions และกำหนด Relational expression เป็น \"date" < 1/1/2005
7. คลิก **RUN** เพื่อทำการประมวลผล
8. ในปฏิบัติการนี้จะสร้างโมเดล ARIMA ในการวิเคราะห์ข้อมูลอนุกรมเวลาเพื่อการทำนายค่าในอนาคต โดยทำการวิเคราะห์ข้อมูลราคาต่ำสุดของดัชนีตลาดซื้อขายอนุพันธ์ Chicago Board Options Exchange (CBOE)

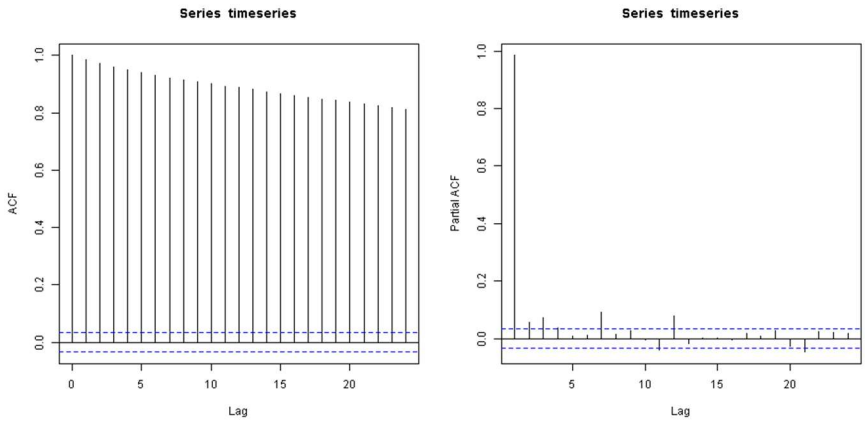
9. เลือกตัวแปรที่จะใช้ในการสร้างโมเดล ARIMA นั่นคือ ตัวแปร date และ vix_low โดยใช้โมดูล Select Columns in Dataset (ภายใต้ Data Transformation → Manipulation) ให้ทำการลากโมดูลดังกล่าวมาวางบน workspace
10. นำข้อมูลส่งออกจากโมดูล Split Data จากโหนดส่วนต่อประสานข้อมูลออกที่ 1 ซึ่งเป็นชุดข้อมูลเรียนรู้ เป็นข้อมูลนำเข้าของโมดูล Select Columns in Dataset
11. เพิ่มคำอธิบายให้ โมดูล Select Columns in Dataset นี้ว่า “For training dataset”
12. คลิกที่กล่องโมดูล Select Columns in Dataset ที่หน้าต่างย่อย Properties คลิก Launch column selector แล้วเลือกตัวแปรที่ต้องการนำมาใช้วิเคราะห์ จากนั้นคลิก ✓
13. ทำการเตรียมข้อมูลทดสอบสำหรับทดสอบโมเดล ARIMA โดยเลือกตัวแปรที่จะใช้ในการทดสอบ นั่นคือ ตัวแปร date และ vix_low โดยใช้โมดูล Select Columns in Dataset
14. นำข้อมูลส่งออกจากโมดูล Split Data จากโหนดส่วนต่อประสานข้อมูลออกที่ 2 ซึ่งเป็นชุดข้อมูลทดสอบ เป็นข้อมูลนำเข้าของโมดูล Select Columns in Dataset
15. เพิ่มคำอธิบายให้ โมดูล Select Columns in Dataset นี้ว่า “For test dataset”
16. คลิกที่กล่องโมดูล Select Columns in Dataset ที่หน้าต่างย่อย Properties คลิก Launch column selector แล้วเลือกตัวแปรที่ต้องการนำมาใช้วิเคราะห์ จากนั้นคลิก ✓
17. คลิก **RUN** เพื่อทำการประมวลผล
18. เปลี่ยนชื่อคอลัมน์ vix_low เป็น data ทั้งในชุดข้อมูลเรียนรู้และชุดข้อมูลทดสอบที่เป็นผลลัพธ์จากโมดูล Select Columns in Dataset โดยใช้โมดูล Edit Metadata
19. คลิกที่กล่องโมดูล Edit Metadata ที่หน้าต่างย่อย Properties คลิก Launch column selector แล้วเลือกตัวแปร vix_low จากนั้นคลิก ✓ เลือกตัวเลือก Label จากลิสต์ Fields และที่กล่องข้อความ New column names กรอกข้อความ “data” ทำเช่นนี้สำหรับทั้งชุดข้อมูลเรียนรู้และชุดข้อมูลทดสอบ
20. คลิก **RUN** เพื่อทำการประมวลผล
21. ต่อมาทำการคำนวณค่า autocorrelation function (ACF) และ partial autocorrelation function (PACF) ในที่นี้จะใช้ภาษาโปรแกรม R ในการพัฒนาชุดคำสั่งขนาดเล็ก โดยลากโมดูล Execute R Script (ภายใต้ R Language Modules)
22. เพิ่มคำอธิบายให้ โมดูล Execute R Script นี้ว่า “Plot ACF and PACF for low prices”
23. นำข้อมูลส่งออกจากโมดูล Edit Metadata สำหรับชุดข้อมูลเรียนรู้ จากโหนดส่วนต่อประสานข้อมูลออก เป็นข้อมูลนำเข้าของโมดูล Execute R Script (Plot ACF and PACF of low prices) แสดงตัวอย่างดังรูป



24. คลิกที่กล่องโมดูล Execute R Script (Plot ACF and PACF of low prices) ที่หน้าต่างย่อย Properties ในช่อง R Script ให้พิมพ์ชุดคำสั่งดังนี้

ชุดคำสั่ง	คำอธิบาย
dataset1 <- maml.mapInputPort(1)	- นำข้อมูลจากโหนดส่วนต่อประสานข้อมูลเข้าที่ 1 เก็บไว้ในตัวแปร dataset1
seasonality<-1 labels <- as.numeric(dataset1$data) timeseries <- ts(labels,frequency=seasonality)	- สร้างตัวแปร timeseries เพื่อเก็บข้อมูลจากคอลัมน์ data ให้อยู่ในรูปแบบอนุกรมเวลา
acf(timeseries, lag.max=24) pacf(timeseries, lag.max=24)	- คำนวณและแสดงกราฟของค่า ACF และ PACF โดยกำหนดให้จำนวนข้อมูลย้อยหลังที่พิจารณามากที่สุด 24 ข้อมูล ผลลัพธ์จะแสดงที่โหนดส่วนต่อประสานข้อมูลออกที่ 2

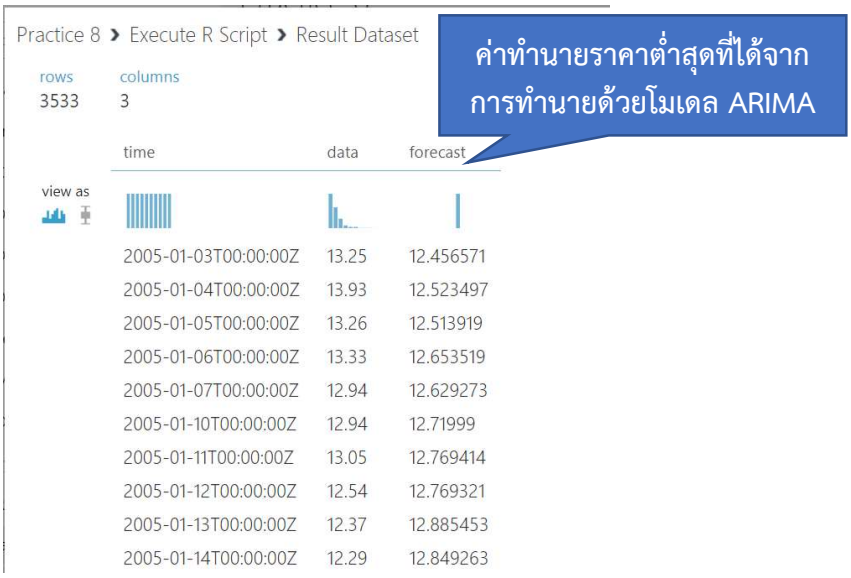
25. คลิก **RUN** เพื่อทำการประมวลผล แล้วดูผลลัพธ์จากข้อมูลออกที่ 2 ของโมดูล Execute R Script (Plot ACF and PACF of low prices) โดยคลิกที่โหนดส่วนต่อประสานข้อมูลออกที่ 2 แล้วเลือก Visualize จะปรากฏกราฟแสดงค่า ACF และ PACF ผลลัพธ์ดังรูป



26. ต่อมาทำการสร้างโมเดล ARIMA สำหรับการวิเคราะห์ข้อมูลอนุกรมเวลา ในที่นี้จะใช้ภาษาโปรแกรม R ในการพัฒนาชุดคำสั่งขนาดเล็ก โดยลากโมดูล Execute R Script
27. เพิ่มคำอธิบายให้ โมดูล Execute R Script นี้ว่า “ARIMA model for low prices”
28. นำข้อมูลส่งออกจากโมดูล Edit Metadata สำหรับชุดข้อมูลเรียนรู้ จากโหนดส่วนต่อประสานข้อมูลออก เป็นข้อมูลนำเข้าที่ 1 ของโมดูล Execute R Script (ARIMA model for low prices)
29. นำข้อมูลส่งออกจากโมดูล Edit Metadata สำหรับชุดข้อมูลทดสอบ จากโหนดส่วนต่อประสานข้อมูลออก เป็นข้อมูลนำเข้าที่ 2 ของโมดูล Execute R Script (ARIMA model for low prices)
30. คลิกที่กล่องโมดูล Execute R Script ที่หน้าต่างย่อย Properties ในช่อง R Script ให้พิมพ์ชุดคำสั่งดังนี้

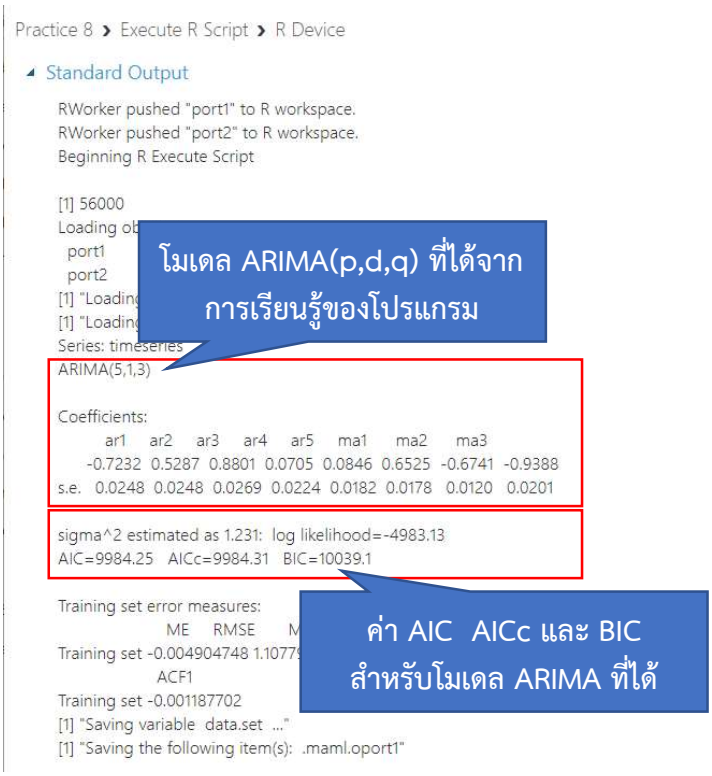
ชุดคำสั่ง	คำอธิบาย
library(forecast) library(ggplot2)	- นำเข้าไลบลารี forecast และ ggplot2
dataset1 <- maml.mapInputPort(1) dataset2 <- maml.mapInputPort(2)	- นำข้อมูลจากโหนดส่วนต่อประสานข้อมูลเข้าที่ 1 และ 2 เก็บไว้ในตัวแปร dataset1 และ dataset2 ตามลำดับ
seasonality<-1 labels <- as.numeric(dataset1$data) timeseries <- ts(labels,frequency=seasonality)	- สร้างตัวแปร timeseries เพื่อเก็บข้อมูลจากคอลัมน์ data ให้อยู่ในรูปแบบอนุกรมเวลา
model <- auto.arima(timeseries)	- สร้างโมเดล ARIMA สำหรับการวิเคราะห์ข้อมูลอนุกรมเวลา timeseries
numPeriodsToForecast <- dim(dataset2)[1]	- หาจำนวนข้อมูลในชุดข้อมูลทดสอบ เก็บไว้ในตัวแปร numPeriodsToForecast
fc <- forecast(model, h=numPeriodsToForecast) forecastedData <- as.numeric(fc$mean)	- ทำการทำนายค่าจำนวน numPeriodsToForecast โดยใช้โมเดล ARIMA ที่สร้างขึ้น
summary(model) autoplot(fc)	- แสดงรายละเอียดของโมเดล ARIMA ที่สร้างขึ้น และแสดงกราฟผลลัพธ์ของการทำนายค่าในอนาคต ผลลัพธ์จะแสดงที่โหนดส่วนต่อประสานข้อมูลออกที่ 2
output <- data.frame(time=dataset2$date, data=dataset2$data,forecast=forecastedData) data.set <- output attr(data.set$forecast, "feature.channel") <- "Regression Scores" attr(data.set$forecast, "score.type") <- "Assigned Labels" maml.mapOutputPort("data.set");	- นำผลลัพธ์การทำนายค่าและข้อมูลจริงในชุดข้อมูลทดสอบส่งออกไปยังส่วนต่อประสานข้อมูลออกที่ 1

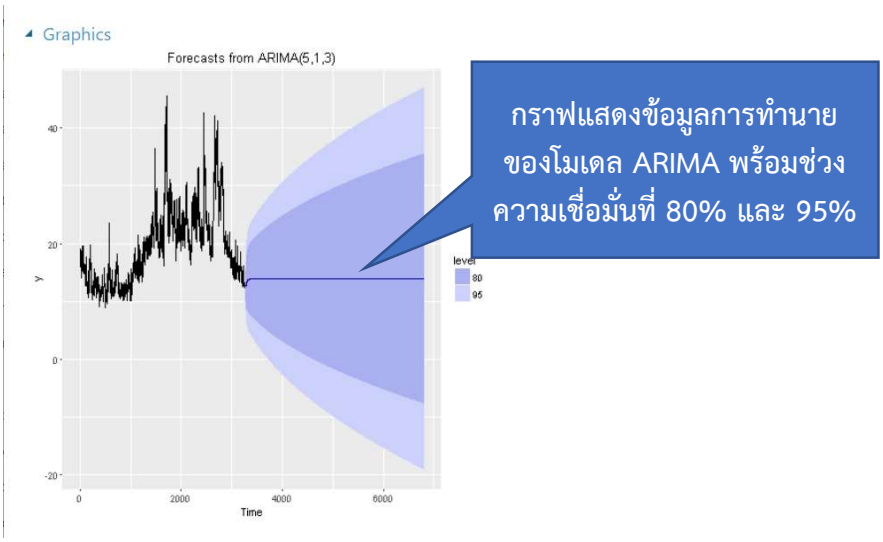
31. คลิก **RUN** เพื่อทำการประมวลผล
32. ดูผลลัพธ์จากข้อมูลออกที่ 1 ของโมดูล Execute R Script (ARIMA model for low prices) โดยคลิกที่โหนดส่วนต่อประสานข้อมูลออกที่ 1 แล้วเลือก Visualize จะปรากฏผลลัพธ์ดังรูป



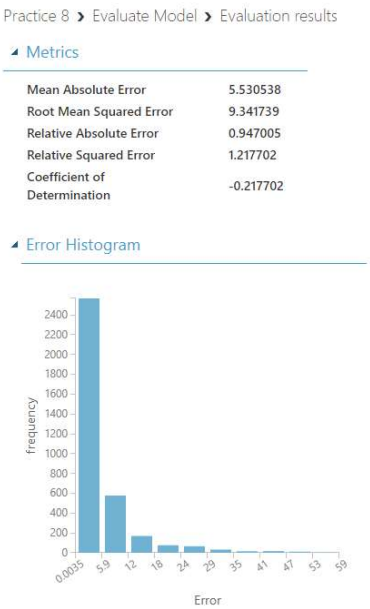
time	data	forecast
2005-01-03T00:00:00Z	13.25	12.456571
2005-01-04T00:00:00Z	13.93	12.523497
2005-01-05T00:00:00Z	13.26	12.513919
2005-01-06T00:00:00Z	13.33	12.653519
2005-01-07T00:00:00Z	12.94	12.629273
2005-01-10T00:00:00Z	12.94	12.71999
2005-01-11T00:00:00Z	13.05	12.769414
2005-01-12T00:00:00Z	12.54	12.769321
2005-01-13T00:00:00Z	12.37	12.885453
2005-01-14T00:00:00Z	12.29	12.849263

33. ดูผลลัพธ์จากข้อมูลออกที่ 2 ของโมดูล Execute R Script (ARIMA model for low prices) โดยคลิกที่โหนดส่วนต่อประสานข้อมูลออกที่ 2 แล้วเลือก Visualize จะปรากฏผลลัพธ์ดังรูป





34. ทำการทดสอบประสิทธิภาพของโมเดล ARIMA โดยใช้โมดูล Evaluate Mode (ภายใต้ Machine Learning → Evaluate) นำข้อมูลส่งออกที่ 1 จากโมดูล Execute R Script (ARIMA model for low prices) เป็นข้อมูลนำเข้า Scored dataset ของโมดูล Evaluate Mode
35. คลิก **RUN** เพื่อทำการประมวลผล แล้วดูผลลัพธ์จากข้อมูลออกของโมดูล Evaluate Mode โดยคลิกที่โหนดส่วนต่อประสานข้อมูลออก แล้วเลือก Visualize จะปรากฏผลลัพธ์ดังรูป



3. แบบฝึกปฏิบัติการ

ให้นักศึกษาทำแบบฝึกปฏิบัติการ ตามลำดับขั้นตอนต่อไปนี้

6. กำหนดให้นักศึกษาทำแบบฝึกปฏิบัติการนี้ ต่อจากการทดลองสาธิต โดยให้นักศึกษาใช้ชุดข้อมูล VIX Prices สร้างโมเดล ARIMA พร้อมทั้งคำนวณและแสดงค่า ACF และ PACF สำหรับการวิเคราะห์อนุกรมเวลา บนตัวแปร ราคาสูงสุดของ vix (vix_high) เพื่อทำนายค่าราคาสูงสุดของ VIX ตั้งแต่วันที่ 3 มกราคม ค.ศ.2005 ถึง วันที่ 15 มกราคม ค.ศ. 2019
7. สังเกตผลลัพธ์จากการค่า ACF และ PACF และบันทึกจำนวนค่าย้อนหลังที่ส่งผลต่อค่าราคาสูงสุดของ VIX อย่างมีนัยสำคัญ

   _____

   _____

8. สังเกตผลลัพธ์จากการสร้างโมเดล ARIMA บันทึกค่า hyperparameter ของโมเดล (p,d,q) ที่ได้จากโปรแกรม บันทึกสมการโมเดล ARIMA ที่ได้ พร้อมค่า AIC AICc และ BIC

   ARIMA(____,_____,_____) = _____

   _____

   AIC = _____

   AICc = _____

   BIC = _____

9. ทำการทดสอบประสิทธิภาพของโมเดล ARIMA (ใช้โมดูล Evaluate Mode) และบันทึกค่าวัดประสิทธิภาพ

   Mean Absolute Error = _____

   Root Mean Squared Error = _____

10. ศึกษาและอธิปรายผลการทดสอบประสิทธิภาพของโมเดลสำหรับการจำแนกข้อมูล

สิ่งที่ต้องส่งเป็นการบ้าน ภาพหน้าจอ Workspace ของนักศึกษาที่ใช้ทำแบบฝึกปฏิบัติการ โดยให้เห็นกล่องโมดูลทั้งหมดและชื่อ Workspace ซึ่งเป็นชื่อของนักศึกษา ตั้งชื่อไฟล์ในรูปแบบ Lab_08_id.jpg โดยแทน id ด้วยรหัสนักศึกษา ส่งผ่านเว็บไซต์ http://hw.cs.science.cmu.ac.th