

ระบบวิเคราะห์ข้อมูลอนุกรมเวลาด้วยเทคนิคทางสถิติและการเรียนรู้ของเครื่อง

THE TIME SERIES ANALYSIS SYSTEM THROUGH STATISTIC AND MACHINE LEARNING TECHNIQUES

พรทิวา วิศิษฐ์สรอรรถ¹

เอกสิทธิ์ พัทธวงษ์ศักดิ์²

บทคัดย่อ

การวิจัยนี้เป็นการศึกษามีวัตถุประสงค์เพื่อพัฒนาระบบวิเคราะห์ข้อมูลอนุกรมเวลาด้วยเทคนิคทางสถิติและการเรียนรู้ของเครื่องโดยเปรียบเทียบวิธีพยากรณ์ออกเป็น 5 เทคนิคประกอบไปด้วย เทคนิค AutoRegressive Integrated Moving Average (ARIMA) Seasonal AutoRegressive Integrated Moving Average (SARIMA) วิเคราะห์การถดถอยพหุคูณ (Multiple Linear Regression) ต้นไม้ตัดสินใจ (Decision Tree) และ ป่าสุ่ม (Random Forest) ซึ่งพัฒนาขึ้นมาเป็นรูปแบบของเว็บแอปพลิเคชันที่ช่วยพยากรณ์ข้อมูลในอนาคตได้หลากหลายรูปแบบ สำหรับงานวิจัยนี้ได้นำข้อมูลจากกรณีศึกษาต่าง ๆ มาประยุกต์ใช้สร้างแบบจำลองดังนี้ 1. ข้อมูลยอดขายรายวันของร้านกาแฟ 2. ข้อมูลจำนวนผู้ใช้บริการโครงการรถไฟฟ้าฟ้ามหานคร สายฉลองรัชธรรม 3. ข้อมูลจำนวนผู้ป่วยและผู้เสียชีวิตรายใหม่จากสถานการณ์ COVID-19 ในประเทศไทย ผลวิจัยพบว่าหากข้อมูลที่นำมาสร้างแบบจำลองมีอิทธิพลต่อฤดูกาลเทคนิค SARIMA ให้ผลการพยากรณ์ที่ดีกว่าเทคนิค ARIMA เช่น กรณีศึกษาการพยากรณ์จำนวนผู้ใช้บริการรถไฟฟ้าฟ้ามหานครสายฉลองรัชธรรมรายเดือนให้ค่า MAPE สำหรับเทคนิค SARIMA, ARIMA เท่ากับ 25.90% , 35.15% ตามลำดับ ส่วนการเปรียบเทียบเทคนิคระหว่าง Decision Tree, Random Forest และ Multiple Linear Regression สำหรับการพยากรณ์ยอดขายรายวันแยกตามรายสินค้าและจำนวนผู้เสียชีวิตรายใหม่จากสถานการณ์ COVID-19 ในประเทศไทยพบว่าเทคนิคให้ค่า MAPE เฉลี่ยต่ำที่สุดคือเทคนิค Random Forest 18.74% - 54.15% เมื่อเทียบกับเทคนิค Decision Tree 22.45% - 57.34% และ Multiple Linear Regression 28.08% - 156.86% ตามลำดับ

¹ นักศึกษาหลักสูตรวิศวกรรมศาสตรมหาบัณฑิต สาขาวิศวกรรมข้อมูลขนาดใหญ่

² ที่ปรึกษาสารนิพนธ์หลัก

คำสำคัญ: การพยากรณ์ อนุกรมเวลา บอซ-เจนส์กินส์ การวิเคราะห์เชิงถดถอย

ABSTRACT

This research was conducted to develop a web-based system for time series analysis using statistic and machine learning techniques. The five compared models of time series analysis included AutoRegressive Integrated Moving Average (ARIMA) and Seasonal AutoRegressive Integrated Moving Average (SARIMA), Multiple Linear Regression, Decision Tree, and Random Forest. The system was developed as a web application which could forecast future information in various datasets. In this study, three datasets were utilized for model development: 1. data on daily sales of a coffee shop, 2. data on passengers of the M.R.T Chalong Ratchadham Line, and 3. data on new cases and deaths of COVID-19 in Thailand. The results showed that when the data used for model development influenced SARIMA, it yielded better forecast in comparison to ARIMA. For example, a study forecasting the number of passengers of the M.R.T Chalong Ratchadham Line each month had MAPE of SARIMA, ARIMA representing 25.90%, 35.15% respectively. Moreover, the comparison between Decision Tree, Random Forest, and Multiple Linear Regression for forecasting the daily sales by product and new cases and deaths of COVID-19 in Thailand indicated that Random Forest had the lowest average of MAPE at between 18.74% to 54.15% compared to Decision Tree at between 22.45% to 57.34% and Multiple Linear Regression at between 28.08% to 156.86%, respectively.

Keyword: Forecasting, Time-series, Box-Jenkins, Regression Analysis

บทนำ

แนวคิดอนุกรมเวลาปรากฏเป็นพื้นฐานในแทบทุกกิจกรรมมีการประยุกต์ใช้การวิเคราะห์อนุกรมเวลาในทางปฏิบัติหลายอย่าง เช่น การคาดการณ์ทางเศรษฐกิจ การวิเคราะห์ตลาดหุ้น การพยากรณ์แผ่นดินไหว หรือการคาดการณ์ผลตอบแทน และอื่นๆ งานวิจัยนี้จึงมีวัตถุประสงค์เพื่อพัฒนาระบบวิเคราะห์ข้อมูลอนุกรมเวลาด้วยเทคนิคทางสถิติและการเรียนรู้ของเครื่องโดยแบ่งเป็น 1. Autoregressive Integrated Moving Average (ARIMA) 2. Seasonal Autoregressive Integrated Moving

Average (SARIMA) ประเภทแบบจำลองอนุกรมเวลา เพื่อจัดการกับข้อมูลที่มีลักษณะเกี่ยวข้องกับทางด้านเวลานั้นคือข้อมูลในอดีตสามารถส่งผลต่อข้อมูลในอนาคตได้ 3. เทคนิคการวิเคราะห์การถดถอยพหุคูณ (Multiple Linear Regression) เป็นการวิเคราะห์ความสัมพันธ์ในรูปแบบเชิงเส้นตรง (Linearity) ระหว่างตัวแปรอิสระ (Independent Variable) ที่มีมากกว่า 1 ตัวและตัวแปรตาม (Dependent Variable) 4. เทคนิคต้นไม้ตัดสินใจ (Decision Tree) และ 5. เทคนิคป่าสุ่ม (Random Forest) มาพยากรณ์ข้อมูล

ผู้ใช้งานสามารถเลือกใช้แบบจำลองในการพยากรณ์ข้อมูลในอนาคตที่มีความแม่นยำ งานวิจัยนี้แบ่งข้อมูลเพื่อสร้างแบบจำลองที่ต่างกันคือ 1) ชุดข้อมูลอนุกรมเวลาที่มีตัวแปรสร้างแบบจำลองคือ เวลาและค่าที่ต้องการพยากรณ์เพื่อใช้กับเทคนิค ARIMA, SARIMA 2) ชุดข้อมูลอนุกรมเวลาที่มีตัวแปรสร้างแบบจำลองคือ เวลา ตัวแปรอิสระร่วม และค่าที่ต้องการพยากรณ์ เพื่อใช้กับเทคนิคการวิเคราะห์การถดถอยพหุคูณ (Multiple Linear Regression) ต้นไม้ตัดสินใจ (Decision Tree) และป่าสุ่ม (Random Forest)

ระบบทำการประมวลผลสร้างแบบจำลองด้วยเทคนิคต่างๆข้างต้นจะแสดงผลเปรียบเทียบค่าความคลาดเคลื่อนใช้ตัวชี้วัดการประเมินผลด้วยค่าเฉลี่ยของค่าสัมบูรณ์เปอร์เซ็นต์ความคลาดเคลื่อน (Mean Absolute Percentage Error, MAPE)

วัตถุประสงค์ของงานวิจัย

พัฒนาเว็บแอปพลิเคชันเพื่อนำเสนอรูปแบบเปรียบเทียบการพยากรณ์ข้อมูลด้วยเทคนิค Autoregressive Integrated Moving Average (ARIMA) Seasonal Autoregressive Integrated Moving Average (SARIMA) เทคนิคการวิเคราะห์การถดถอยพหุคูณ (Multiple Linear Regression) เทคนิคต้นไม้ตัดสินใจ (Decision Tree) และวิธีป่าสุ่ม (Random Forest) โดยมุ่งเน้นให้แบบจำลองจากเทคนิคต่างๆ ข้างต้นสามารถวิเคราะห์คุณลักษณะสำคัญจากอนุกรมข้อมูลที่นำมาใช้และพิจารณาความสัมพันธ์ของแต่ละคุณลักษณะของข้อมูลในอดีตมาเพิ่มประสิทธิภาพการพยากรณ์ให้มีความแม่นยำและความคลาดเคลื่อนที่ต่ำที่สุดเพื่อนำไปใช้ในการพยากรณ์ในอนาคตที่มีความน่าเชื่อถือ

ขอบเขตงานวิจัย

1. ข้อมูลกรณีศึกษาที่นำมาใช้ในงานวิจัยคือ
 - 1.1 ข้อมูลยอดขายจริงรายวันของร้านกาแฟตั้งแต่เดือน มีนาคม ถึง สิงหาคม ปี พ.ศ. 2564 รวมไปถึงข้อมูลเพิ่มเติมดังนี้
 - 1.1.1 ข้อมูลอุณหภูมิตั้งแต่เดือน มีนาคม ถึง สิงหาคม ปี พ.ศ. 2564
 - 1.1.2 ข้อมูลวันหยุดตามประเพณีของสถาบันการเงิน ปี พ.ศ. 2564
 - 1.2 ข้อมูลจำนวนผู้ใช้บริการ โครงการรถไฟฟ้าฟ้ามหานครสายฉลองรัชธรรมตั้งแต่เดือน สิงหาคม ปี 2559 ถึง สิงหาคม ปี พ.ศ. 2564
 - 1.3 ข้อมูลจำนวนผู้ป่วย COVID-19 ระลอก 3 ตั้งแต่เดือน มิถุนายน ถึง ตุลาคม ปี พ.ศ. 2564
2. ทำการเตรียมข้อมูลให้สมบูรณ์ให้พร้อมสำหรับการนำไปใช้กับแบบจำลองต่าง ๆ
3. พัฒนาเว็บแอปพลิเคชัน
 - 3.1 สร้างแบบจำลองพยากรณ์ข้อมูลด้วยการเปรียบเทียบกับเทคนิคทางสถิติและการเรียนรู้ของเครื่องในรูปแบบต่างๆ
 - 3.2 ทำการวัดประสิทธิภาพของแต่ละแบบจำลองโดยพิจารณาจากค่าความคลาดเคลื่อนด้วยค่าเฉลี่ยของค่าสัมบูรณ์เปอร์เซ็นต์ความคลาดเคลื่อน (Mean Absolute Percentage Error, MAPE)
 - 3.3 สร้างตัวเลือกการพยากรณ์ข้อมูลจากการเลือกแบบจำลองที่พิจารณาประสิทธิภาพเปรียบเทียบก่อนหน้าสำหรับข้อมูลใหม่

ประโยชน์ที่คาดว่าจะได้รับ

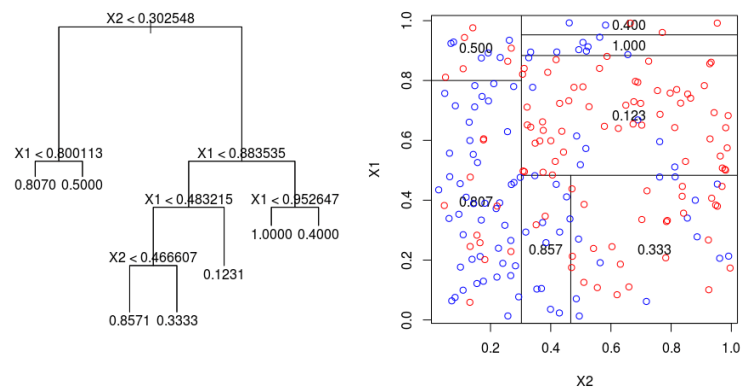
1. สามารถนำระบบพยากรณ์ข้อมูลที่มีประสิทธิภาพแม่นยำน่าเชื่อถือไปใช้ให้เกิดประสิทธิภาพในการวางแผนด้านการตลาด การจัดสรรทรัพยากร การคาดการณ์เศรษฐกิจ หรืออื่นๆ ตามจุดประสงค์ของข้อมูลในการใช้งานได้
2. ระบบพยากรณ์ข้อมูลมีหลากหลายเทคนิคเปรียบเทียบสำหรับสร้างแบบจำลองทำให้ผู้ใช้งานสามารถเลือกใช้เทคนิคการพยากรณ์ให้เหมาะสมกับจุดประสงค์การใช้งานของตนเองได้
3. ระบบใช้งานง่ายไม่ซับซ้อนและไม่จำเป็นต้องมีความรู้เกี่ยวกับเทคนิคระดับเชี่ยวชาญก็สามารถนำระบบนี้ไปใช้งานให้เกิดประโยชน์ได้

ทฤษฎี และผลงานที่เกี่ยวข้อง

1. เทคนิคการวิเคราะห์การถดถอย

1.1 การวิเคราะห์ความถดถอยพหุคูณ (Multiple Regression Analysis) เป็นการวิเคราะห์ความสัมพันธ์ระหว่างตัวแปรอิสระตั้งแต่ 2 ตัวขึ้นไปกับผลของตัวแปรตามที่เกิดขึ้น

1.2 ต้นไม้ตัดสินใจ (Decision Tree) เป็นวิธีการเรียนรู้ของเครื่อง (Machine Learning) โดยเทคนิคการให้ผลลัพธ์จะออกมาในรูปแบบลักษณะของโครงสร้างต้นไม้สร้างแบบจำลองที่คาดการณ์ค่าของตัวแปรตามโดยการเรียนรู้กฎการตัดสินใจอย่างง่ายที่อนุมานจากคุณลักษณะข้อมูลโดยจำลองแผนผังการตัดสินใจ



ภาพที่ 1: ลักษณะของต้นไม้ตัดสินใจในรูปแบบของ ต้นไม้ไบนารี (Binary tree)

ที่มา: <https://gdcdoder.com/decision-tree-regressor-explained-in-depth/>

1.3 ป่าสุ่ม (Random Forest) (ชนัท, 2561) ป่าสุ่มเป็นวิธีที่ใช้พื้นฐานพัฒนามาจากต้นไม้ตัดสินใจ (Decision Tree) ในทำนายในรูปแบบชุดของต้นไม้ตัดสินใจหลายๆ ต้น (Ensemble of Decision Trees) โดยสร้างจากการสุ่มข้อมูลตัวอย่างแบบเลือกแล้วใส่กลับ (random sampling with replacement) เพื่อนำมาสร้างเป็นแบบจำลองต้นไม้โดยแต่ละต้นมีลักษณะที่ไม่ซ้ำกันซึ่งผลจากการทำนายของต้นไม้แต่ละต้นจะทำการโหวตเลือกผลการทำนายที่ได้รับการโหวตมากที่สุดวิธีการนี้เรียกว่า Bagging หรือ Bootstrapping

2. การพยากรณ์ด้วยเทคนิค Box-Jenkins

2.1 Autoregressive integrated moving average (ARIMA) เป็นเทคนิคพยากรณ์วิเคราะห์ข้อมูลอนุกรมเวลาในอดีตและค่าความคลาดเคลื่อนทั้งในปัจจุบันและในอดีตมีลักษณะการเกิดที่เป็นไปตามกฎความน่าจะเป็นซึ่งการวิเคราะห์ห้อนุกรมเวลาโดยวิธีนี้ลักษณะของอนุกรมเวลาต้องเป็นอนุกรมเวลาที่มีคุณสมบัติ Stationary (มีความหยุดนิ่ง)

(ผศ.ดร. เฉลิมพล, 2562) แบบจำลอง ARIMA ประกอบด้วยองค์ประกอบ 3 ส่วน คือ

1. AR(p) คือ Y_t ที่ถูกกำหนดด้วย Autoregressive process
โดย $Y_t = Y_{t-1}, Y_{t-2}, \dots, Y_p$
2. I(d) คือ อันดับความหยุดนิ่งของ Y_t
โดย $d = 0, 1, 2, \dots, d$
3. MA(q) คือ Y_t ที่ถูกกำหนดด้วย Moving average process
โดย $Y_t = \varepsilon_{t-1}, \varepsilon_{t-2}, \dots, \varepsilon_q$

2.2 Seasonal Autoregressive integrated moving average (SARIMA) เป็นเทคนิคพยากรณ์วิเคราะห์ข้อมูลอนุกรมเวลาถูกพัฒนาขึ้นเพื่อรองรับการทำงานที่คำนึงถึงความผันแปรตามฤดูกาลซึ่งเป็นส่วนประกอบที่มีความสำคัญดังนั้นรูปแบบจะแสดงเป็น SARIMA(p,d,q)x(P,D,Q)_s

3. เปรียบเทียบค่าความคลาดเคลื่อนในการประเมินผลการพยากรณ์

เทียบความแตกต่างระหว่างข้อมูลจริงกับค่าพยากรณ์จากค่าความคลาดเคลื่อนที่เกิดขึ้นสำหรับงานวิจัยนี้จะเลือกนำตัวชี้วัดในการประเมินคือค่าเฉลี่ยของค่าสัมบูรณ์เปอร์เซ็นต์ความคลาดเคลื่อน (Mean Absolute Percentage Error, MAPE)

$$MAPE = \frac{1}{n} \sum_{t=1}^n \left| \frac{Y_t - \hat{Y}_t}{Y_t} \right| \times 100$$

n คือ จำนวนตัวอย่าง, Y_t คือ ข้อมูลจริง ณ เวลา t, $\hat{Y}_t$ คือ ค่าพยากรณ์ของข้อมูล ณ เวลา t

ระเบียบวิธีวิจัย

แนวทางการวิจัยนำหลักการ Cross-Industry Standard Process For Data Mining (CRISP-DM) มาประยุกต์ใช้และนำข้อมูลอนุกรมเวลาที่แตกต่างกันมาทำการทดลองใช้กับระบบที่พัฒนาขึ้นมา

เป็น 3 ประเภทของข้อมูล 1) ข้อมูลยอดขายของร้านกาแฟ 2) ข้อมูลจำนวนผู้ใช้บริการโครงการรถไฟฟ้ามหานคร สายฉลองรัชธรรม 3) ข้อมูลรายงานสถานการณ์ COVID-19 ระลอก 3

1. การทำความเข้าใจกับธุรกิจ (Business Understanding)

ทำความเข้าใจธุรกิจที่มีลักษณะแตกต่างกันในแต่ละธุรกิจสามารถใช้เครื่องมือที่ใช้ในการช่วยให้เราเข้าใจธุรกิจมากขึ้น เช่น กรณีศึกษาข้อมูลยอดขายของร้านกาแฟที่มาทำการพยากรณ์จะใช้โมเดลธุรกิจพื้นฐาน (Business Model Canvas)

2. การทำความเข้าใจกับข้อมูล (Data Understanding)

รวบรวมข้อมูลที่เกี่ยวข้องกับการสร้างแบบจำลองพยากรณ์

2.1 กรณีศึกษาเป็นประเภทข้อมูลยอดขายของร้านกาแฟ เลือกใช้ข้อมูลยอดขายรายวันจากเครื่อง POS (Point of sale system) ของร้านค้าเพราะมีความน่าเชื่อถือและมีความเหมาะสมกับการนำข้อมูลมาใช้งาน

2.2 กรณีศึกษาเป็นประเภทข้อมูลจำนวนผู้ใช้บริการโครงการรถไฟฟ้ามหานคร สายฉลองรัชธรรม เลือกใช้ข้อมูลจากองค์กรการไฟฟ้าขนส่งมวลชนแห่งประเทศไทยเพราะมีความน่าเชื่อถือและมีการปรับปรุงข้อมูลให้อัปเดตอยู่เสมอ

2.3 กรณีศึกษาเป็นประเภทข้อมูลรายงานสถานการณ์ COVID-19 ระลอก 3 เลือกใช้ข้อมูลจาก API ของกระทรวงสาธารณสุขเพราะมีความน่าเชื่อถือและ มีการอัปเดตข้อมูลทุกวัน

3. การเตรียมข้อมูล (Data Preparation)

3.1 การคัดเลือกข้อมูล (Data Selection)

เลือกข้อมูลที่มีลักษณะเป็นตัวเลขมาใช้เนื่องจากการวิเคราะห์อนุกรมเวลาด้วยเทคนิคทางสถิติและการเรียนรู้ของเครื่องตามรายละเอียดดังภาพที่ 2

	รูปแบบที่ 1: เทคนิค ARIMA, SARIMA		รูปแบบที่ 2: ใช้กับเทคนิค Multiple Linear Regression, Decision Tree Random Forest	
	Attirbute	Target	Attirbute	Target
ฟิลด์บังคับ	Date	Total	Date, ID	Total
กรณีศึกษาเป็นประเภทข้อมูล ยอดขายของร้านกาแฟ	เวลา (Date),	ยอดขาย (Total)	เวลา (Date), รหัสสินค้า (ID), อุณหภูมิ (Temperature), วันหยุด (Holiday), มีการจัดโปรโมชัน (Promotion),	ยอดขาย (Total)
กรณีศึกษาเป็นประเภทข้อมูล จำนวนผู้ใช้บริการ โครงการ รถไฟฟ้าในนคร สายฉลองรัช ธรรม	เวลา (Date),	จำนวน ผู้โดยสาร รวม (Total)		
กรณีศึกษาเป็นประเภทข้อมูล รายงานสถานการณ์ COVID-19 ระลอก 3			วันแถลง (Date), รหัสประเทศ (ID), จำนวนผู้ป่วยรายใหม่ (new_case), จำนวนผู้ป่วยรักษาหายรายใหม่ (new_recovered)	จำนวนผู้ป่วย ตายรายใหม่ (Total)

ภาพที่ 2: การคัดเลือกข้อมูลสำหรับการสร้างแบบจำลองสำหรับกรณีศึกษาและเทคนิคต่าง ๆ

3.2 การกลั่นกรองข้อมูล (Data Cleaning)

ในกรณีศึกษาข้อมูลยอดขายของร้านกาแฟที่มีการพบข้อมูลที่หายไป (missing value) คือ เป็นช่วงเวลาที่มีการปิดร้านค้างจึงได้มีการแทนที่ข้อมูลด้วยค่าเฉลี่ยของยอดขายภายในเดือนนั้นๆ มาทำการแทนที่และมีการกรองข้อมูลที่เป็นค่าผิดปกติก่อนที่จะนำไปสร้างแบบจำลองการพยากรณ์คือ รายการที่ไม่มียอดขายจริงเช่น ชื่อสินค้าคือ “รายการไม่รับช้อนส้อมพลาสติก” “ใช้สิทธิโครงการรัฐ” “ลด 50% สูงสุด 100.- เพียงใส่โค้ด "KBANK" ในหน้าชำระเงิน” หรือ “ของแถม” เป็นต้น

3.3 การแปลงรูปแบบข้อมูล (Data Transformation)

สร้างข้อมูลให้อยู่ในรูปแบบที่พร้อมนำไปใช้ในการวิเคราะห์คือ 1) กรณีศึกษาเป็นประเภทข้อมูลยอดขายของร้านกาแฟเพิ่มแอตทริบิวต์ที่อาจจะเพิ่มประสิทธิภาพในการวิเคราะห์เพิ่มเติม เช่น วันหยุด (อ้างอิงวันหยุดตามประเพณีของสถาบันการเงิน) มีการจัดโปรโมชันหรือไม่ อุณหภูมิ (องศาเซลเซียส) แปลงข้อมูลให้เป็น Numerical (ตัวเลข) เช่น วันหยุด (Holiday) การจัดโปรโมชัน

(Promotion) ให้อยู่ในรูปแบบของ {0, 1} และแปลงข้อมูลวันที่ออกเป็นวันที่เท่าไหระของสัปดาห์ (วันจันทร์-อาทิตย์ 0-6) เดือน ปี วันที่ 2) กรณีศึกษาเป็นประเภทข้อมูลรายงานสถานการณ์ COVID-19 ระลอก 3 เพิ่มแอตทริบิวต์ ID เป็นรหัสประเทศเพื่อนำเข้าระบบ (ฟิลด์บังคับ) กรณีนี้จะใส่เป็นรหัส ID เดียวกัน และแปลงข้อมูลวันที่ที่มีรูปแบบเหมือนกรณีที่ 1

ข้อมูลหลังทำการแปลงข้อมูลในกรณีศึกษาต่างๆสรุปออกมาได้ดังภาพที่ 3

	รูปแบบที่ 1: เทคนิค ARIMA, SARIMA	รูปแบบที่ 2: ใช้กับเทคนิค Multiple Linear Regression, Decision Tree Random Forest	
การแบ่งไฟล์ก่อนนำเข้าแบบจำลอง และพยากรณ์	คอลัมภ์ข้อมูลในไฟล์ train	คอลัมภ์ข้อมูลในไฟล์ train	คอลัมภ์ข้อมูลในไฟล์ test
กรณีศึกษาเป็นประเภทข้อมูลยอดขาย ของร้านกาแฟ	เวลา (Date), ยอดขาย (Total)	เวลา (Date), รหัสสินค้า (ID), อุณหภูมิ (Temperature), วันหยุด (Holiday), มีการจัดโปรโมชัน (Promotion), ยอดขาย (Total)	เวลา (Date), รหัสสินค้า (ID), อุณหภูมิ (Temperature), วันหยุด (Holiday), มีการจัดโปรโมชัน (Promotion), ยอดขาย (Total)
กรณีศึกษาเป็นประเภทข้อมูลจำนวน ผู้ใช้บริการโครงการรถไฟฟ้าฟ้ามหานคร สายฉลองรัชธรรม	เวลา (Date), จำนวนผู้โดยสารรวม (Total)		
กรณีศึกษาเป็นประเภทข้อมูลรายงาน สถานการณ์ COVID-19 ระลอก 3		วันแถลง (Date), รหัสประเทศ (ID), จำนวนผู้ป่วยรายใหม่ (new_case), จำนวนผู้ป่วยรักษาหายรายใหม่ (new_recovered), จำนวนผู้ป่วยตายรายใหม่ (Total)	วันแถลง (Date), รหัสประเทศ (ID), จำนวนผู้ป่วยรายใหม่ (new_case), จำนวนผู้ป่วยรักษาหายรายใหม่ (new_recovered)

ภาพที่ 3: คอลัมภ์ในข้อมูลไฟล์ train, test สำหรับการสร้างแบบจำลองสำหรับกรณีศึกษาและเทคนิค
ต่าง ๆ

4. การสร้างแบบจำลอง (Modeling)

ขั้นตอนนี้จะใช้ข้อมูลจากข้อ 3 มาทำการสร้างแบบจำลองโดยการแบ่งออกเป็น 2 กลุ่ม คือ

4.1 อนุกรมเวลา (Time series)

ข้อมูลจะถูกนำเข้ามาเพื่อสร้างแบบจำลองด้วยวิธี ARIMA หรือ SARIMA โดยแบ่งข้อมูลจำนวน 80% เป็นข้อมูลสำหรับสร้างแบบจำลองและข้อมูลจำนวน 20% เป็นข้อมูลใช้เป็นข้อมูลทดสอบแบบจำลองที่สร้างขึ้นได้ ใช้วิธีการ Split Test

พิจารณารูปแบบของ ARIMA กรณีที่เป็น Non Stationary series ที่มีแนวโน้มและมีความจำเป็นต้องทำ Regular Differencing หาค่า (d) อันดับความหุคหนึ่ง (จำนวนครั้งที่หาผลต่างเพื่อปรับอนุกรมเวลาให้เป็นอนุกรมเวลาที่ Stationary) ด้วยวิธี ADF test หา p, q ใช้วิธีชยับตัวเลขไปที่ละขั้นหาแบบจำลองโอกาสที่เป็นไปได้ และพิจารณาค่า AIC น้อยที่สุดเป็นแบบจำลองในการพยากรณ์

พิจารณารูปแบบของ SARIMA กรณีที่เป็น Non Stationary series ที่มีแนวโน้มและมีอิทธิพลของฤดูกาลมาเกี่ยวข้องมีความจำเป็นต้องทำ Seasonal Differencing หาค่า (D) อันดับความหุคหนึ่ง (จำนวนครั้งที่หาผลต่างเพื่อปรับอนุกรมเวลาให้เป็นอนุกรมเวลาที่ Stationary และไม่มีความเป็นฤดูกาล) ด้วยวิธี CH test หา p, q และ P, Q ใช้วิธีชยับตัวเลขไปที่ละขั้นหาแบบจำลองโอกาสที่เป็นไปได้ และแทนค่า s ด้วยคาบของฤดูกาลรายวันแทนด้วย 7 รายเดือนแทนด้วย 12 งานวิจัยนี้จะมีข้อมูลที่นำมาทดลองเป็นลักษณะรายวันและรายเดือน ซึ่งค่า s จะถูกปรับตามคาบของฤดูกาลที่ใช้และพิจารณาค่า AIC น้อยที่สุดเป็นแบบจำลองในการพยากรณ์

```
Performing stepwise search to minimize aic
ARIMA(0,1,0)(0,1,1)[12] : AIC=734.923, Time=0.04 sec
ARIMA(0,1,0)(0,1,0)[12] : AIC=747.992, Time=0.01 sec
ARIMA(1,1,0)(1,1,0)[12] : AIC=745.017, Time=0.03 sec
ARIMA(0,1,1)(0,1,1)[12] : AIC=738.959, Time=0.05 sec
ARIMA(0,1,0)(1,1,1)[12] : AIC=736.904, Time=0.05 sec
ARIMA(0,1,0)(0,1,2)[12] : AIC=736.450, Time=0.12 sec
ARIMA(0,1,0)(1,1,0)[12] : AIC=747.011, Time=0.03 sec
ARIMA(0,1,0)(1,1,2)[12] : AIC=737.523, Time=0.11 sec
ARIMA(1,1,0)(0,1,1)[12] : AIC=738.079, Time=0.05 sec
ARIMA(1,1,1)(0,1,1)[12] : AIC=740.918, Time=0.08 sec
ARIMA(0,1,0)(0,1,1)[12] intercept : AIC=737.099, Time=0.04 sec
Best model: ARIMA(0,1,0)(0,1,1)[12]
Total fit time: 0.616 seconds
```

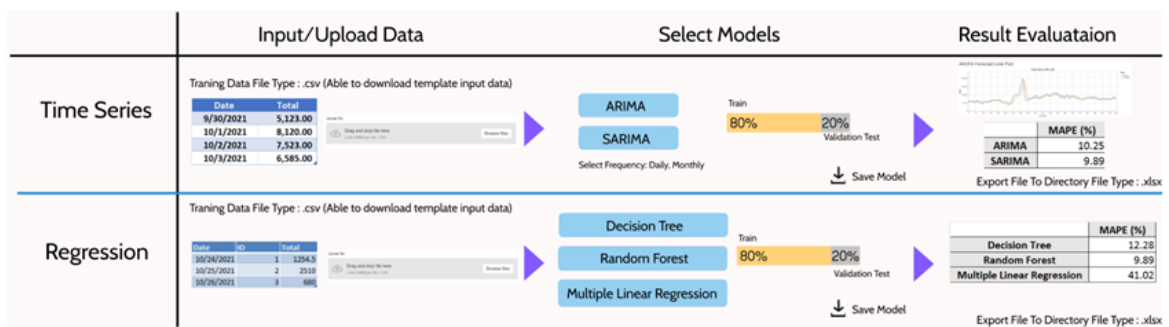
ภาพที่ 4: ตัวอย่างการหารูปแบบ ARIMA(p,d,q) SARIMA (P,D,Q)T หรือ ARIMA(p,d,q)x(P,D,Q)_s จากการพิจารณาค่า AIC

4.2 สถิติข้อมูล/การวิเคราะห์เชิงถดถอย (Statistic/Regression Analysis)

ข้อมูลจะถูกนำเข้าเพื่อสร้างแบบจำลองด้วยวิธี Decision Tree, Random Forest และ Multiple Linear Regression โดยแบ่งข้อมูลจำนวน 80% เป็นข้อมูลสำหรับสร้างแบบจำลองและข้อมูลจำนวน 20% ใช้เป็นข้อมูลทดสอบโมเดลที่สร้างขึ้นได้ ใช้วิธีการ Split Test แบบสุ่มข้อมูลก่อนแบ่ง

สรุปภาพรวมขั้นตอนการสร้างแบบจำลองดังภาพที่ 5

Build A New Predictive Model



ภาพที่ 5: ขั้นตอนการทำงานของระบบเพื่อสร้างแบบจำลอง

5. การประเมินผล (Evaluation)

หลังจากที่แบ่งข้อมูลออกเป็น 2 ส่วนและใช้แบบจำลองที่สร้างได้จากข้อมูลจำนวน 80% มาทำการพยากรณ์ข้อมูลที่อยู่ในข้อมูลจำนวน 20% ที่เหลือนั้น และใช้ค่าเฉลี่ยของค่าสัมบูรณ์เปอร์เซ็นต์ความคลาดเคลื่อน (MAPE) แสดงค่าความคลาดเคลื่อนตามเปรียบเทียบกับเทคนิควิธีต่าง ๆ สามารถ เลือกจัดเก็บแบบจำลองที่สนใจผ่านการ download ในรูปแบบของ .pkl ได้หรือ export ไฟล์ข้อมูลเปรียบเทียบพยากรณ์ในรูปแบบของ .xlsx

6. การปฏิบัติตามผลเสนอแนะ (Deployment)

6.1 เลือกใช้แบบจำลองที่ทำการสร้างล่าสุดในระบบ

รูปแบบการพยากรณ์นี้ระบบจะทำการจัดเก็บแบบจำลองล่าสุดอัตโนมัติผู้ใช้งานสามารถเลือกวิธีการพยากรณ์และทำการระบุเวลาในอนาคตที่ต้องการพยากรณ์สำหรับเทคนิค ARIMA, SARIMA หรือ upload ข้อมูลสำหรับทดสอบสำหรับเทคนิค Decision Tree, Random Forest, Multiple Linear Regression ที่ต้องการพยากรณ์ข้อมูลในอนาคตระบบจะแสดงผลลัพธ์ออกมาในรูปแบบของกราฟและ/หรือตาราง สามารถ export ให้อยู่ในรูปแบบของ .xlsx ไฟล์ ให้สำหรับนำไปใช้งาน

6.2 เลือกใช้แบบจำลองที่ได้จัดเก็บไว้มาใช้พยากรณ์ข้อมูลใหม่

รูปแบบการพยากรณ์นี้ผู้ใช้งานสามารถ upload แบบจำลองที่เคยจัดเก็บไว้ นำมาพยากรณ์ข้อมูลได้โดยการ upload แบบจำลอง .pkl เข้าระบบและทำการระบุคาบเวลาในอนาคตที่ต้องการพยากรณ์สำหรับเทคนิค ARIMA, SARIMA หรือ upload ข้อมูลสำหรับทดสอบสำหรับเทคนิค Decision Tree, Random Forest, Multiple Linear Regression ที่ต้องการพยากรณ์ข้อมูลในอนาคตระบบจะแสดงผลลัพธ์ออกมาในรูปแบบของกราฟและ/หรือตาราง สามารถ export ให้อยู่ในรูปแบบของ .xlsx ไฟล์ ให้สำหรับนำไปใช้งาน

ผลการศึกษา

เปรียบเทียบค่า MAPE ของแบบจำลองทางอนุกรมเวลาด้วยเทคนิคและกรณีศึกษาต่างๆ ดังภาพที่ 6

1. ผลการวัดประสิทธิภาพความถูกต้องของแบบจำลอง

เทคนิค	ค่าเฉลี่ยของค่าสัมบูรณ์เปอร์เซ็นต์ความคลาดเคลื่อน (Mean Absolute Percentage Error, MAPE)				
	รูปแบบที่ 1: เทคนิค ARIMA, SARIMA		รูปแบบที่ 2: ใช้กับเทคนิค Multiple Linear Regression, Decision Tree Random Forest		
	ARIMA	SARIMA	Decision Tree	Random Forest	Multiple Linear Regression
กรณีศึกษาเป็นประเภท ข้อมูลยอดขายของร้าน กาแฟ	14.22	14.22	57.54	54.51	156.86
กรณีศึกษาเป็นประเภท ข้อมูลจำนวนผู้ใช้บริการ โครงการรถไฟฟ้าฟ้ามหานคร สายฉลองรัชธรรม	35.15	25.90			
กรณีศึกษาเป็นประเภท ข้อมูลรายงานสถานการณ์ COVID-19 ระลอก 3			22.45	18.78	28.08

ภาพที่ 6: เปรียบเทียบค่าเฉลี่ยของค่าสัมบูรณ์เปอร์เซ็นต์ความคลาดเคลื่อนของการพยากรณ์ด้วยกรณีศึกษาและเทคนิคในรูปแบบต่าง ๆ

2. การวัดผลประสิทธิภาพความถูกต้องจากการนำแบบจำลองพยากรณ์ในอนาคตเทียบกับข้อมูลจริง

ขั้นตอนนี้จะเป็นการเลือกแบบจำลองที่เหมาะสมมาใช้พยากรณ์ข้อมูลอนาคต

2.1 กรณีศึกษาประเภทข้อมูลยอดขายของร้านกาแฟ เลือกใช้เทคนิค ARIMA จากเทคนิค ARIMA, SARIMA สำหรับการสร้างแบบจำลองและพยากรณ์ข้อมูลยอดขายรายวันกรณีนี้แทนค่าด้วยคาบเวลาเท่ากับ 7 วัน (วันที่ 1-7 กันยายน พ.ศ. 2564) เมื่อหาค่า MAPE ของข้อมูลพยากรณ์เทียบกับข้อมูลจริง 7 วัน ได้เท่ากับ 39.23% เลือกใช้เทคนิค Random Forest จากเทคนิค Decision Tree, Random Forest, Multiple Regression สำหรับพยากรณ์ข้อมูลยอดขายรายวันแบ่งตามรหัสสินค้ากรณีนี้พยากรณ์ข้อมูลยอดขายรวมรายรหัสสินค้าของวันที่ 1 กันยายน พ.ศ. 2564 เมื่อหาค่า MAPE ของข้อมูลพยากรณ์เทียบกับข้อมูลจริงจากตัวอย่างสินค้า 7 ตัวอย่างได้เท่ากับ 35.18%

2.2 กรณีศึกษาประเภทข้อมูลจำนวนผู้ใช้บริการโครงการรถไฟฟ้าฟ้ามหานคร สายฉลองรัชธรรม เลือกใช้เทคนิค SARIMA จากเทคนิค ARIMA, SARIMA สำหรับการสร้างแบบจำลองและพยากรณ์ข้อมูลจำนวนผู้ใช้บริการกรณีนี้แทนค่าด้วยคาบเวลาเท่ากับ 7 เดือน (เดือน มกราคม ถึง กรกฎาคม พ.ศ. 2564) เมื่อหาค่า MAPE ของข้อมูลพยากรณ์เทียบกับข้อมูลจริง 7 เดือนได้เท่ากับ เท่ากับ 77.30%

2.3 กรณีศึกษาเป็นประเภทข้อมูลรายงานสถานการณ์ COVID-19 ระลอก 3 เลือกใช้เทคนิค Random Forest จากเทคนิค (Decision Tree, Random Forest, Multiple Regression) สำหรับพยากรณ์ข้อมูลจำนวนผู้เสียชีวิตรายใหม่กรณีนี้พยากรณ์ข้อมูลจำนวนผู้เสียชีวิตของวันที่ วันที่ 1-6 ตุลาคม 2564 เมื่อหาค่า MAPE ของข้อมูลพยากรณ์เทียบกับข้อมูลจริงได้เท่ากับ 37.69%

สรุปและอภิปรายผลการวิจัย

ผลการทดลองพิจารณาประสิทธิภาพของแบบจำลองเปรียบเทียบกับเทคนิค ARIMA กับ SARIMA เมื่อนำข้อมูลกรณีศึกษาของยอดขายรวม (รายวัน) ของร้านกาแฟและจำนวนผู้ใช้บริการ (รายเดือน) โครงการรถไฟฟ้าฟ้ามหานคร สายฉลองรัชธรรมพบว่าหากข้อมูลมีการเคลื่อนไหวแนวโน้มที่ไม่เสถียรและไม่มีความสอดคล้องตามฤดูกาลผลของการพยากรณ์จะให้ค่า MAPE ใกล้เคียงกันหรือแทบไม่ต่างกันเลยเนื่องจากพารามิเตอร์ที่นำผลของเกณฑ์ในการคัดเลือกแบบจำลอง AIC ค่าที่ดีที่สุดสำหรับเทคนิคทั้งสองข้างต้นมีค่าเดียวกันในทางตรงกันข้ามหากข้อมูลมีแนวโน้มไม่สอดคล้องตามฤดูกาลผลของการพยากรณ์ของเทคนิค SARIMA จะให้ผลดีกว่าโดยให้ค่า MAPE ต่ำกว่าเทคนิค ARIMA

เมื่อพิจารณาประสิทธิภาพของแบบจำลองเปรียบเทียบกับเทคนิค Decision Tree, Random Forest และ Multiple Linear Regression เมื่อนำข้อมูลกรณีศึกษาของยอดขายรวมของแต่ละรายสินค้า (รายวัน) ของร้านกาแฟและจำนวนผู้ป่วยและผู้เสียชีวิตจากสถานการณ์ COVID-19 ระลอก 3 (รายวัน) พบว่าเทคนิค Random Forest พยากรณ์ยอดขายตามรายการสินค้าและจำนวนผู้เสียชีวิตรายใหม่ของสถานการณ์ COVID-19 ให้ผลค่าความคลื่อนไปในทิศทางเดียวกันโดยให้ค่า MAPE ต่ำที่สุดสอดคล้องกับผลการศึกษา (Abdulwahed Salam and Abdelaziz El Hibaoui, 2018) ซึ่งศึกษาเปรียบเทียบวิธีพยากรณ์ 5 วิธี Linear regression, Decision tree, Random forest, Feedforward Neural network และ Supper vector ของการพยากรณ์ปริมาณการใช้ไฟฟ้าของเมืองเตโตนวนประเทศโมร็อกโกพบว่าวิธี Random Forest เป็นวิธีที่เหมาะสมที่สุด

ข้อเสนอแนะ

1. ข้อมูลที่นำมาศึกษาวิจัยควรมีความสมบูรณ์ถูกต้องและมีความน่าเชื่อถือ
2. ควรแบ่งข้อมูลของคาบเวลามาสร้างแบบจำลองให้เหมาะสมตามจุดประสงค์ของการพยากรณ์
3. ควรมีการพิจารณาประเมินผลที่ได้จากการพิจารณาอย่าสม่ำเสมอปรับเปลี่ยนลดตัวแปรอิสระเพิ่มเติมที่อาจช่วยเพิ่มประสิทธิภาพของการสร้างแบบจำลองมากขึ้นหรือลองนำเทคนิคใหม่ ๆ ที่ถูกพัฒนามาปรับใช้ให้มีประสิทธิภาพมากขึ้น

บรรณานุกรม

ภาษาไทย

ชนัท จระณะสมบูรณ์. (2561). การทำนายการซื้อซ้ำของผู้ซื้อโดยใช้เทคนิคการเรียนรู้ของเครื่องจักร.

ปริญญาณพนธ์ วิทยาศาสตร์มหาบัณฑิต สาขาวิชาเทคโนโลยีสารสนเทศ
มหาวิทยาลัยศรีนครินทรวิโรฒ.

วันหยุดตามประเพณีของสถาบันการเงิน ประจำปี. (2564). สืบค้น 14 มิถุนายน 2564, จาก

<https://www.bot.or.th/Thai/FinancialInstitutions/FIholiday/Pages/2021.aspx>

ปริมาณผู้โดยสารรถไฟฟ้ามหานคร. (2564). สืบค้น 5 ตุลาคม 2564, จาก

<https://data.go.th/dataset/mrta-crmk>

สถานการณ์ผู้ติดเชื้อ COVID-19 อัปเดตรายวัน รูปแบบ API(Json/CSV Data Format). (2564). สืบค้น 28 กันยายน 2564, จาก <https://covid19.ddc.moph.go.th/>
ผศ.ดร. เจริมพล. (2562). การพยากรณ์ทางอนุกรมเวลา (*Time Series Forecasting*). สืบค้น 30 กันยายน 2564, จาก <https://cj007blog.files.wordpress.com/2020/04/07-time-series-forecasting.pdf>

ภาษาต่างประเทศ

Georgios D. (2019). *Decision Tree Regressor explained in depth*. Retrieved September 19, 2021, from <https://gdcoder.com/decision-tree-regressor-explained-in-depth/>.
A Salam, A El Hibaoui. (2018). Comparison of Machine Learning Algorithms for the Power Consumption Prediction : - Case Study of Tetouan city-. *2018 6th International Renewable and Sustainable Energy Conference (IRSEC)*. pp 1210.
Weather Underground. (2020). *Bangkok, Thailand Weather History*. Retrieved June 14,2021 from <https://www.wunderground.com/history/monthly/th/bangkok>.