



KKU SCIENCE JOURNAL

Journal Home Page : <https://ph01.tci-thaijo.org/index.php/KKUSciJ>

Published by the Faculty of Science, Khon Kaen University, Thailand



การวิเคราะห์และเปรียบเทียบประสิทธิภาพของโมเดลพยากรณ์ข้อมูลอนุกรมเวลา

ในการพยากรณ์จำนวนนักท่องเที่ยวในจังหวัดเชียงใหม่

Comparative Analysis of Time Series Forecasting Models

for Predicting Tourist Arrivals in Chiang Mai

เทวา พรหมนุชานนท์¹ รุจีพันธุ์ โกษารัตน์^{2*} และ วรکار ใจดี¹

Tewa Promnuchanont¹, Rujipan Kosarat^{2*} and Worakarn Jaidee¹

¹คณะบริหารธุรกิจและศิลปศาสตร์ มหาวิทยาลัยเทคโนโลยีราชมงคลล้านนา จังหวัดเชียงใหม่ 50300

²คณะวิศวกรรมศาสตร์ มหาวิทยาลัยเทคโนโลยีราชมงคลล้านนา จังหวัดเชียงใหม่ 50300

¹Faculty of Business Administration and Liberal Arts, Rajamangala University of Technology Lanna, Chiang Mai, 50300, Thailand

²Faculty of Engineering, Rajamangala University of Technology Lanna, Chiang Mai, 50300, Thailand

บทคัดย่อ

วัตถุประสงค์ของการศึกษานี้คือการประเมินประสิทธิภาพในการพยากรณ์ของโมเดลการพยากรณ์อนุกรมเวลา (Time series) โดยใช้โมเดล 5 แบบ ได้แก่ ARIMA LSTM Prophet XGBoost และ Random forest ในการพยากรณ์จำนวนนักท่องเที่ยวที่มาเยือนจังหวัดเชียงใหม่ การศึกษาใช้ชุดข้อมูลที่ประกอบด้วยจำนวนนักท่องเที่ยว ตั้งแต่เดือนมกราคม พ.ศ. 2563 ถึงเดือนธันวาคม พ.ศ. 2566 โดยแบ่งข้อมูลออกเป็นสองชุด ได้แก่ ชุดข้อมูลการฝึกตั้งแต่เดือนมกราคม พ.ศ. 2563 ถึงเดือนธันวาคม พ.ศ. 2565 และชุดข้อมูลการทดสอบตั้งแต่เดือนมกราคม พ.ศ. 2566 ถึงเดือนธันวาคม พ.ศ. 2566 ผู้วิจัยใช้ค่าความคลาดเคลื่อนสัมบูรณ์เฉลี่ย (MAE) ค่ารากที่สองของค่าความคลาดเคลื่อนสัมบูรณ์เฉลี่ยกำลังสอง (RMSE) และค่าความคลาดเคลื่อนสัมบูรณ์เฉลี่ยแบบร้อยละ (MAPE) ในการประเมินประสิทธิภาพของโมเดลต่าง ๆ ผลการศึกษาระบุว่าโมเดล ARIMA มีความแม่นยำสูงสุด จากผลการวิเคราะห์เชิงเปรียบเทียบพบว่าโมเดล ARIMA มีความคลาดเคลื่อนในการพยากรณ์ต่ำที่สุด (MAE = 8,325.33 RMSE = 11,462.63 และ MAPE = 10.16)

ABSTRACT

The purpose of this study is to evaluate how well five time series forecasting models—ARIMA, LSTM, Prophet, XGBoost, and Random Forest—predict Chiang Mai's arrivals of tourists. The study utilized a dataset that comprised visitor counts from January 2020 to December 2023. We split the data into two sets: a training set from January 2020 to December 2022, and a test set from January 2023 to December 2023. We used the mean absolute error (MAE), root mean square error (RMSE) and mean absolute percentage (MAPE) to evaluate the models' effectiveness. The results indicate that the ARIMA model

*Corresponding Author, E-mail: rujipan@rmutl.ac.th

demonstrated the highest accuracy. The Comparative analysis indicate that the ARIMA model exhibits the lowest forecasting error metrics among the models evaluated (MAE = 8,325.33, RMSE = 11,462.63, and MAPE = 10.16).

คำสำคัญ: พยากรณ์ ข้อมูลอนุกรมเวลา ประสิทธิภาพโมเดล

Keywords: Forecasting, Time Series Data, Model Performance

บทนำ

จังหวัดเชียงใหม่ เป็นจังหวัดที่มีความสำคัญทั้งในด้านประวัติศาสตร์และวัฒนธรรม ด้วยความหลากหลายทางธรรมชาติและวัฒนธรรม เชียงใหม่เป็นจุดหมายปลายทางยอดนิยมสำหรับนักท่องเที่ยวทั้งชาวไทยและชาวต่างชาติ การท่องเที่ยวจึงมีบทบาทสำคัญในการกระตุ้นเศรษฐกิจของจังหวัด ทำให้เกิดการพัฒนาด้าน ทั้งโครงสร้างพื้นฐาน การบริการ และกิจกรรมที่น่าสนใจ อย่างไรก็ตามการเติบโตของการท่องเที่ยวก็สร้างปัญหาในด้านการอนุรักษ์สิ่งแวดล้อมและวัฒนธรรมท้องถิ่น ปัญหาเหล่านี้ทำให้เกิดความจำเป็นในการวางแผนและบริหารจัดการการท่องเที่ยวอย่างยั่งยืนเพื่อรักษาสมดุลระหว่างการพัฒนาและการอนุรักษ์ เพื่อให้การท่องเที่ยวเชียงใหม่เติบโตในระยะยาว

การพยากรณ์จำนวนนักท่องเที่ยวอย่างแม่นยำจำเป็นสำหรับการวางแผน การจัดสรรทรัพยากร และการพัฒนายุทธศาสตร์ในอุตสาหกรรมการท่องเที่ยว การพัฒนาทางด้านการพยากรณ์ข้อมูลอนุกรมเวลามีโมเดลที่หลากหลายที่สามารถใช้ในการพยากรณ์ โดยงานวิจัยนี้เปรียบเทียบประสิทธิภาพของโมเดล 5 แบบ ได้แก่ ARIMA LSTM Prophet XGBoost และ Random forest ซึ่งทั้ง 5 โมเดลนี้ เป็นโมเดลที่สามารถจัดการกับข้อมูลในรูปแบบอนุกรมเวลาได้เป็นอย่างดีเหมาะสำหรับการนำมาใช้ในการพยากรณ์จำนวนนักท่องเที่ยวในเชียงใหม่

การพยากรณ์ที่แม่นยำช่วยให้ผู้มีส่วนได้ส่วนเสียสามารถคาดการณ์ความต้องการ ปรับปรุงบริการและเพิ่มประสบการณ์ท่องเที่ยว ซึ่งโมเดลที่นำมาศึกษาในครั้งนี้ เป็นโมเดลที่มีประสิทธิภาพและเข้าถึงง่ายสำหรับการจัดการกับแนวโน้มและฤดูกาล โดยในงานวิจัยนี้ ใช้ข้อมูลจำนวนนักท่องเที่ยวในเชียงใหม่ ระหว่างปี พ.ศ. 2563 - พ.ศ.2566 จากเว็บไซต์กระทรวงการท่องเที่ยวและกีฬา (กระทรวงการท่องเที่ยวและกีฬา, 2567) ซึ่งครอบคลุมช่วงเวลาที่ได้รับผลกระทบอย่างมากจากการแพร่ระบาดของ COVID-19 ซึ่งนำมาซึ่งมีความผันผวนของข้อมูลที่ไม่เคยมีมาก่อนในแนวโน้มการท่องเที่ยว โดยใช้ชุดข้อมูลที่หลากหลายนี้ งานวิจัยจะประเมินความแม่นยำของแต่ละโมเดลโดยอิงจากเมตริกการประเมินประสิทธิภาพหลัก ได้แก่ ค่าความคลาดเคลื่อนสัมบูรณ์เฉลี่ย รากที่สองของค่าความคลาดเคลื่อนสัมบูรณ์เฉลี่ย และร้อยละของค่าความคลาดเคลื่อนสัมบูรณ์เฉลี่ย

ผลการวิจัยมีเป้าหมายเพื่อแก้ปัญหาในการหาโมเดลที่เหมาะสม และสามารถให้ผลลัพธ์ที่แม่นยำที่สุดในการพยากรณ์จำนวนนักท่องเที่ยวในบริบทของจังหวัดเชียงใหม่ ซึ่งข้อมูลจำนวนนักท่องเที่ยวในสภาพแวดล้อมที่เปลี่ยนแปลงอย่างรวดเร็ว การวิเคราะห์เปรียบเทียบไม่เพียงแต่เน้นถึงจุดแข็งและจุดอ่อนของแต่ละโมเดล แต่ยังนำเสนอคำแนะนำ เพื่อเพิ่มขีดความสามารถในการพยากรณ์ต่อไป นอกจากนี้ งานวิจัยนี้จะช่วยในการเพิ่มพูนความรู้ในด้านการพยากรณ์ข้อมูลอนุกรมเวลา และสนับสนุนการตัดสินใจอย่างมีข้อมูลและการวางแผนเชิงกลยุทธ์ เพื่อส่งเสริมการเติบโตด้านการท่องเที่ยวอย่างยั่งยืนในจังหวัดเชียงใหม่ต่อไป

วิธีการดำเนินการวิจัย

1. งานวิจัยนี้ ใช้ข้อมูลแบบอนุกรมเวลา (Time series data) ซึ่งเป็นข้อมูลที่เก็บรวบรวมตามระยะเวลาเป็นช่วง ๆ อย่างต่อเนื่องกัน โดยข้อมูลที่จะใช้ในการวิจัย จะเป็นข้อมูลจำนวนนักท่องเที่ยวในเชียงใหม่ ระหว่างปี พ.ศ. 2563 - พ.ศ.

2566 จากเว็บไซต์กระทรวงการท่องเที่ยวและกีฬา จำนวนทั้งหมด 48 เดือน ซึ่งเมื่อทำความสะอาดข้อมูล (Data cleaning) และทำการแปลงข้อมูลให้อยู่ในรูปแบบข้อมูลแบบอนุกรมเวลา

ในการวิจัยนี้ มีตัวแปรที่เกี่ยวข้องคือ

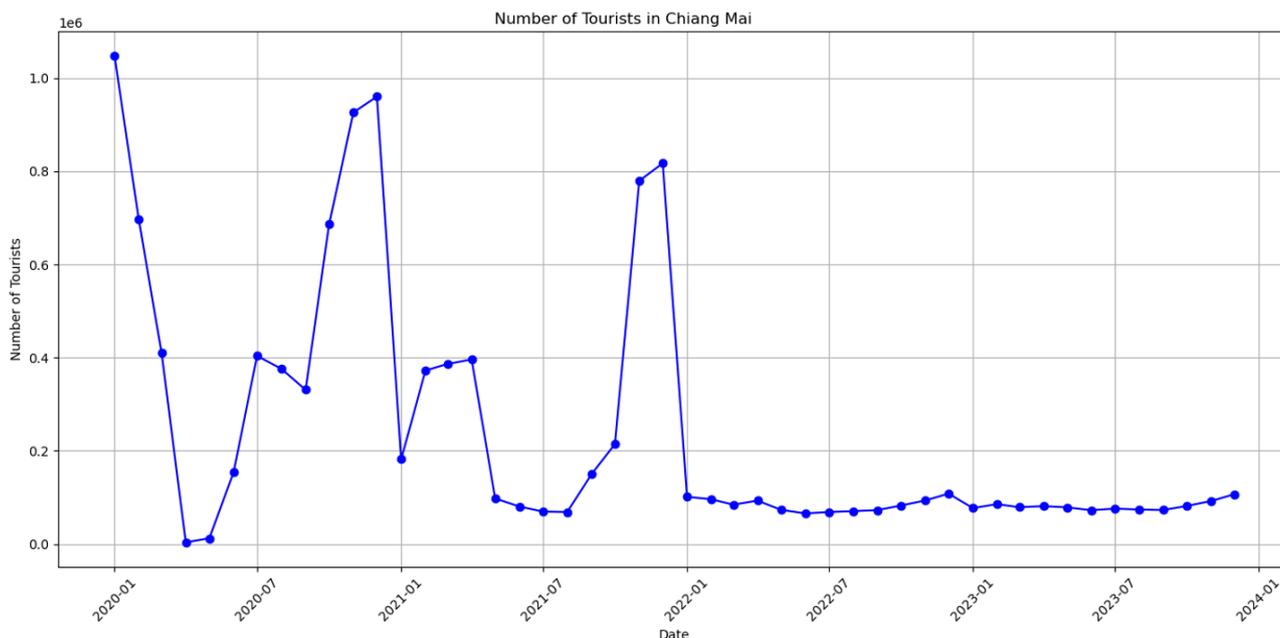
- ตัวแปร y (ตัวแปรตาม) คือ จำนวนนักท่องเที่ยวรายเดือนในจังหวัดเชียงใหม่
- ตัวแปร t (ตัวแปรเวลา) คือ เวลาเป็นรายเดือน ตั้งแต่เดือนมกราคม พ.ศ. 2563 ถึงเดือนธันวาคม พ.ศ. 2566

ข้อมูลจำนวนนักท่องเที่ยวในจังหวัดเชียงใหม่ ระหว่างปี พ.ศ. 2563 - พ.ศ. 2566 มีค่าสถิติพื้นฐานแสดงได้ดังตารางที่ 1

ตารางที่ 1 แสดงค่าทางสถิติของข้อมูลจำนวนนักท่องเที่ยวในจังหวัดเชียงใหม่ ระหว่างปี พ.ศ. 2563 - พ.ศ. 2566

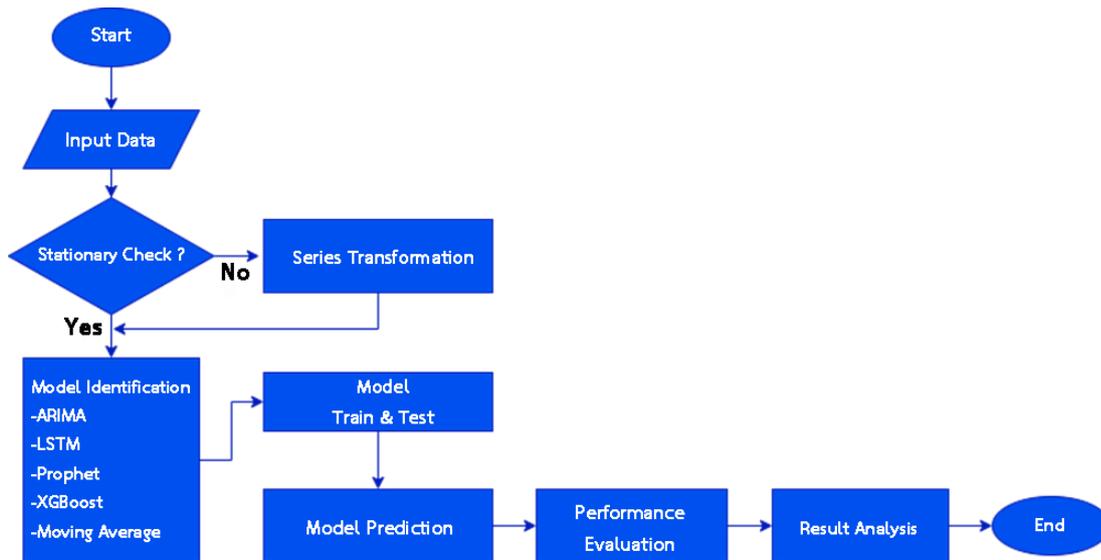
Statistic	Value
Mean	241,856.4
Median	93,307.0
Range	1,044,023.0
Min	3,523.0
Max	1,047,546.0
Standard Deviation	278,585.1

ซึ่งข้อมูลนักท่องเที่ยวในจังหวัดเชียงใหม่ ระหว่างปี พ.ศ. 2563 - พ.ศ. 2566 จะเห็นได้ว่าข้อมูลมีลักษณะเป็นข้อมูลฤดูกาล (Seasonality) โดยแสดงรูปแบบที่ซ้ำกันในช่วงเวลาที่แน่นอน สังเกตได้จากกราฟจะมีค่าสูงขึ้นในช่วงปลายปีถึงต้นปีของแต่ละปี แต่ในช่วงปี พ.ศ. 2565 - พ.ศ. 2566 นักท่องเที่ยวจะลดลงอย่างมาก เนื่องจากเกิดสถานการณ์การระบาดของไวรัสโควิด-19 ดังแสดงในกราฟรูปที่ 1



รูปที่ 1 กราฟแสดงจำนวนนักท่องเที่ยวในจังหวัดเชียงใหม่ช่วงปี พ.ศ. 2563 - พ.ศ.2566

2. เครื่องมือที่ใช้วิจัย
 - 2.1 ใช้ซอฟต์แวร์สเปรดชีต (Spread sheet) ในการรวบรวมและเตรียมข้อมูลเพื่อนำเข้าสู่โมเดล
 - 2.2 ใช้ภาษาไพธอน (Python) เป็นภาษาโปรแกรมมิ่งในการการเขียนโค้ดเพื่อสร้าง และทดสอบโมเดลพยากรณ์
 - 2.3 ใช้แอปพลิเคชัน Anaconda ในการจัดการสภาพแวดล้อม ติดตั้งแพ็คเกจ (Package) และไลบรารี (Library) ที่ใช้ในการทดสอบโมเดล
 - 2.4 ใช้แอปพลิเคชัน Jupyter notebook ในการเขียนเขียนโค้ดภาษาไพธอนประมวลผลข้อมูล และแสดงผลในรูปแบบอินเตอร์แอคทีฟ



รูปที่ 2 กระบวนการในการเปรียบเทียบประสิทธิภาพโมเดลเพื่อพยากรณ์จำนวนนักท่องเที่ยวในจังหวัดเชียงใหม่

วิธีการวิจัย

3. ในการวิจัยครั้งนี้ได้ใช้กระบวนการตามกรอบแนวคิดในการวิจัย แสดงดังรูปที่ 2
 - 3.1 นำข้อมูลที่รวบรวมมาจากเว็บไซต์ของการท่องเที่ยวแห่งประเทศไทย ซึ่งเป็นไฟล์สเปรดชีตเก็บข้อมูลรายปีจำนวน 4 ปี รวมจำนวน 4 ไฟล์ จากนั้นรวมไฟล์ให้เป็นไฟล์เดียวแล้วตัดข้อมูลที่ไม่เกี่ยวข้องออก เหลือเฉพาะข้อมูลที่จะนำไปเข้าสู่โมเดล 2 คอลัมน์ คือข้อมูลเวลาและจำนวนนักท่องเที่ยว
 - 3.2 นำข้อมูลมาตรวจสอบว่าเป็นข้อมูลคงที่ (Stationary data) หรือไม่ เนื่องจากการพยากรณ์ข้อมูลที่เป็นข้อมูลอนุกรมเวลาจำเป็นต้องใช้ข้อมูลที่คงที่ เพื่อให้ได้ผลการพยากรณ์ที่เสถียรและแม่นยำ ซึ่งในการตรวจสอบข้อมูล จะใช้การตรวจสอบยูนิตรูท (Unit Root) ด้วยวิธี Augmented Dickey-Fuller test (ADF) (บุญศรี, 2557) ซึ่งสามารถเขียนสมการทางคณิตศาสตร์ได้ดังสมการที่ 1

$$\Delta y_t = \alpha + \beta t + y_{t-1} + \sum_{i=1}^k \delta_i \Delta y_{t-i} + \epsilon_t \quad (1)$$

- โดยที่ Δy_t คือ การเปลี่ยนแปลงของตัวแปร y_t ระหว่างช่วงเวลาปัจจุบันและช่วงเวลาก่อนหน้า
- α คือ ค่าคงที่ (Constant) ในสมการเพื่อชดเชยการเบี่ยงเบนของข้อมูลจากศูนย์ และเพื่อให้สมการสามารถปรับให้เข้ากับข้อมูลจริงได้ดีขึ้น
- βt คือ เทอมที่แสดงถึงแนวโน้ม (Trend) ที่เป็นเชิงเส้น เมื่อเวลาผ่านไป
- โดย β คือ ค่าสัมประสิทธิ์
- t คือ ตัวแปรเวลา

γy_{t-1} คือ ค่าของ y_t ในช่วงเวลาก่อนหน้า ($t - 1$) คูณด้วยค่าสัมประสิทธิ์ γ

$\sum_{i=1}^k$ คือ ผลรวม (Summation) ของค่า Δy_{t-i} ตั้งแต่ $i = 1$ จนถึง $i = k$

δ_i คือ ค่าสัมประสิทธิ์ (Coefficient) สำหรับค่า Δy_{t-i} ที่เป็นค่าล่าช้า (Lagged value) ในช่วงเวลา $t - i$

Δy_{t-i} คือ การเปลี่ยนแปลงของตัวแปร y_t ในช่วงเวลาก่อนหน้า ($t - i$)

ϵ_t คือ ค่าความคลาดเคลื่อน (Error term)

ในการทดสอบ ADF มีการตั้งสมมติฐานดังนี้

- สมมติฐานหลัก (Null Hypothesis, H_0) คือ ข้อมูลมียูนิตรูท และสถานะไม่คงที่ (Non-stationary)
- สมมติฐานรอง (Alternative Hypothesis, H_1) คือ ข้อมูลไม่มียูนิตรูท และสถานะคงที่ (Stationary)

ซึ่งในการทดสอบ ADF จะเกณฑ์พิจารณา 3 ค่าดังนี้

1. ADF statistic คือ ค่าที่ได้จากการคำนวณในการทดสอบ ADF ซึ่งเป็นตัวชี้วัดหลักในการทดสอบว่าข้อมูลมีหน่วยรากหนึ่ง (Unit root) หรือไม่
2. P-value (Probability value) คือ ค่าความน่าจะเป็นที่ใช้ในการทดสอบทางสถิติ เพื่อใช้เป็นเกณฑ์ในการตัดสินใจว่าควรปฏิเสธสมมติฐานว่างหรือไม่
3. Critical values คือค่าที่กำหนดขึ้น เพื่อใช้เป็นเกณฑ์ในการตัดสินใจว่าควรปฏิเสธสมมติฐานหลักหรือไม่

โดยข้อมูลนี้จะปฏิเสธสมมติฐานหลักเมื่อค่า ADF statistic มีค่าน้อยกว่าค่า Critical values ที่ระดับนัยสำคัญทั้ง

1% 5% และ 10% และ ค่า P-value มีค่าน้อยกว่า 0.05

จากการทดสอบ ADF ได้ผลลัพธ์ดังนี้

ADF statistic เท่ากับ -4.892

P-value เท่ากับ 0.000036

Critical values

- ระดับนัยสำคัญ ที่ 1% มีค่าเท่ากับ -3.578
- ระดับนัยสำคัญ ที่ 5% มีค่าเท่ากับ -2.925
- ระดับนัยสำคัญ ที่ 10% มีค่าเท่ากับ -2.601

จากการทดสอบสมมติฐาน สรุปผลได้คือ ปฏิเสธ H_0 และยอมรับสมมติฐานรอง เนื่องจากข้อมูลไม่มียูนิตรูทและข้อมูลมีสถานะคงที่

3.3 กำหนดโมเดลที่จะใช้ทดสอบ ซึ่งประกอบด้วย 5 โมเดลคือ

3.3.1 Autoregression integrated moving average (ARIMA) (Nochai and Nochai, 2006) เป็นหนึ่งในโมเดลที่ใช้ในการวิเคราะห์ข้อมูลอนุกรมเวลา ถูกคิดค้นโดย Box and Jenkins 1970 โดยอาศัยข้อมูลในอดีตเพื่อกำหนดรูปแบบของข้อมูลและพยากรณ์ข้อมูลในอนาคต ARIMA เป็นวิธีการที่สามารถระบุและพยากรณ์แนวโน้มของข้อมูลได้ดี และยังมีคามยืดหยุ่นในการปรับเปลี่ยนพารามิเตอร์เพื่อให้เหมาะสมกับลักษณะของข้อมูลที่มีความซับซ้อน โมเดล ARIMA เป็นการรวมเอาส่วนของการถดถอยอัตโนมัติ (Auto regressive: AR) การเคลื่อนที่เฉลี่ย (Moving average: MA) และการบูรณาการ (Integration: I) เข้าด้วยกัน สมการของ ARIMA จะมีรูปแบบดังสมการที่ 2

$$y_t = c + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \dots + \phi_p y_{t-p} + \epsilon_t + \theta_1 \epsilon_{t-1} + \theta_2 \epsilon_{t-2} + \dots + \theta_q \epsilon_{t-q} \quad (2)$$

โดยที่ c คือ ค่าคงที่

y_t คือ ข้อมูลที่ผ่านการทำ Differencing แล้ว

ϕ_i คือ ค่าสัมประสิทธิ์ของส่วน AR

θ_i คือ ค่าสัมประสิทธิ์ของส่วน MA

ϵ_t คือ ค่าความคลาดเคลื่อน

3.3.2 Long short-term memory (LSTM) (Greff *et al.*, 2017) เป็นเทคนิคหนึ่งที่ถูกพัฒนาจาก Recurrent neural network (RNN) ซึ่งเทคนิค RNN นั้นเหมาะสำหรับการนำมาใช้งานกับข้อมูลที่มีลักษณะเป็นลำดับ (Sequence) หรือข้อมูลที่มีความต่อเนื่อง เช่น ข้อมูลอนุกรมเวลา ข้อมูลเสียง ข้อมูลประเภท ข้อความ ข้อมูลรูปภาพ และ วิดีทัศน์ เป็นต้น

LSTM ประกอบด้วย 3 หน่วยความจำหลักที่เรียกว่า Gate ได้แก่ Forget gate Input gate และ Output gate โดยแต่ละประตูจะทำหน้าที่ควบคุมการไหลของข้อมูลเข้าสู่และออกจาก Cell state ซึ่งเป็นหน่วยความจำหลักของ LSTM

Forget gate ทำหน้าที่ตัดสินใจว่าจะลบข้อมูลอะไรจากสถานะเซลล์ โดยใช้ฟังก์ชันซิกมอยด์ (Sigmoid function) ซึ่งผลลัพธ์จะเป็น 0 และ 1 ค่าเท่ากับ 0 หมายถึง ข้อมูลจะถูกลบ และเท่ากับ 1 หมายถึง ข้อมูลจะถูกเก็บรักษาไว้ แสดงสมการได้ดังสมการที่ 3

$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f) \quad (3)$$

โดยที่ f_t คือ Output ของ Forget gate (0 = ลืม 1 = เก็บไว้)

σ คือ ฟังก์ชันซิกมอยด์ ที่ใช้ในการจำกัดค่า Output ให้อยู่ในช่วง 0 ถึง 1

W_f คือ น้ำหนักของ Forget gate

h_{t-1} คือ ข้อมูลจากการคำนวณในเดือนก่อนหน้า ที่โมเดลจำได้และจะใช้เป็นส่วนหนึ่งในการคำนวณเพื่อช่วยในการพยากรณ์ในเดือนปัจจุบัน

x_t คือ ข้อมูลจำนวนนักท่องเที่ยวในเดือนปัจจุบันที่ใช้ในการพยากรณ์

b_f คือ ค่าคงที่ที่ใช้ปรับค่าผลลัพธ์ในการตัดสินใจว่าจะเก็บหรือลบข้อมูลจากสถานะก่อนหน้า

Input gate ตัดสินใจว่าจะเพิ่มข้อมูลใหม่อะไรเข้าไปใน Cell state ประกอบด้วยสองส่วน คือ Sigmoid layer ที่ตัดสินใจว่าจะปรับปรุงค่าใด และ Tanh layer ที่สร้างค่า Candidate value ใหม่ แสดงสมการได้ดังสมการที่ 4 และสมการที่ 5

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i) \quad (4)$$

$$\tilde{C}_t = \tanh(W_c \cdot [h_{t-1}, x_t] + b_c) \quad (5)$$

โดยที่ i_t คือ น้ำหนักของ Input gate และค่า Candidate value ตามลำดับ

\tilde{C}_t คือ ค่า Candidate value ใหม่สำหรับ Cell state

W_i และ W_c คือ น้ำหนักของ Input gate และค่า Candidate value

b_i และ b_c คือ Bias ของ Input gate และค่า Candidate value

Output gate ตัดสินใจว่าจะส่งข้อมูลอะไรออกไปจาก Cell state เพื่อสร้างผลลัพธ์ ประกอบด้วย Sigmoid layer และ Tanh layer ที่ปรับค่าข้อมูลให้อยู่ในช่วงที่ต้องการ แสดงสมการดังสมการที่ 6 และ 7

$$o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o) \quad (6)$$

$$h_t = o_t \cdot \tanh(C_t) \quad (7)$$

โดยที่ o_t คือ ผลลัพธ์ของ Output gate

h_t คือ Hidden state ปัจจุบัน

W_o คือ น้ำหนักของ Output gate

b_o คือ Bias ของ Output gate

3.3.3 Prophet model (Taylor and Letham, 2018) พัฒนาโดย Facebook โดยใช้โมเดลแบบเพิ่ม (Additive model) ซึ่งสามารถประมวลผลแนวโน้มที่ไม่เชิงเส้นด้วยความถี่รายปี รายสัปดาห์ และรายวัน พร้อมทั้งคำนึงถึงวันหยุดพิเศษด้วย โมเดลนี้เหมาะกับข้อมูลอนุกรมเวลาที่มีฤดูกาลที่ชัดเจน และมีข้อมูลในอดีตหลายฤดูกาล Prophet ยังสามารถจัดการกับข้อมูลที่ขาดหายได้โดยไม่ทำให้การพยากรณ์เสียหาย นอกจากนี้ยังสามารถจัดการกับค่าผิดปกติได้อย่างมีประสิทธิภาพ

สมการพื้นฐานของ Prophet คือการรวมองค์ประกอบต่าง ๆ ได้แก่ แนวโน้ม (trend) ฤดูกาล (seasonality) และวันหยุดพิเศษ (holidays) ซึ่งสามารถเขียนเป็นสมการได้ดังสมการที่ 8

$$y(t) = g(t) + s(t) + h(t) + \epsilon_t \quad (8)$$

โดยที่ $g(t)$ คือ ฟังก์ชันแนวโน้ม (Trend function) ที่สามารถเป็นแบบเส้นตรงหรือแบบ Logistic growth

$$g(t) = (k + a(t)^T \delta) \cdot t + (m + a(t)^T \gamma) \quad (9)$$

โดยที่ k คือ อัตราการเติบโต

m คือ ค่าคงที่ (Intercept)

$a(t)$ คือ เวกเตอร์ตัวชี้วัดสำหรับช่วงเวลาที่มีการเปลี่ยนแปลง

δ คือ การปรับเปลี่ยนอัตราการเติบโตที่กำหนด

โดยที่ $s(t)$ คือ ฟังก์ชันฤดูกาล (Seasonality function)

$$s(t) = \sum_{n=1}^N \left[a_n \cos\left(\frac{2\pi n t}{P}\right) + b_n \sin\left(\frac{2\pi n t}{P}\right) \right] \quad (10)$$

โดยที่ P คือ ช่วงเวลาปกติ (เช่น $P = 365.25$ วัน สำหรับข้อมูลรายปี หรือ $P = 7$ สำหรับข้อมูลรายสัปดาห์)

a_n และ b_n คือ สัมประสิทธิ์ของฟังก์ชันฤดูกาล

โดยที่ $h(t)$ คือ ฟังก์ชันผลกระทบวันหยุด (Holiday effects function)

$$h(t) = \sum_{j=1}^L [d_j \cdot \text{indicator}(t \in \text{holiday}_j)] \quad (11)$$

โดยที่ L คือ จำนวนของวันหยุดที่กำหนด

d_j คือ ผลกระทบของวันหยุด j

โดยที่ ϵ_t คือ ค่าความคลาดเคลื่อน

3.3.4 XGBoost model (Chen and Guestrin, 2016) เป็นโมเดลที่ถูกนำมาใช้สำหรับการวิเคราะห์ การถดถอยและการจำแนกข้อมูล และถือเป็นหนึ่งในอัลกอริทึมที่มีประสิทธิภาพดีที่สุดสำหรับการเรียนรู้แบบมีผู้สอน ที่พัฒนา และเพิ่มประสิทธิภาพจากเทคนิคที่รู้จักกันในชื่อ Gradient boosting machines (GBM) หลักการทำงานของเทคนิคนี้ จะพิจารณาชุดข้อมูล (Data set: DS) ที่มีจำนวนตัวอย่างข้อมูล m และ n โดยที่ $DS = \{(x_i, y_i) : i = 1 \dots n, x_i \in \mathbb{R}^m, y_i \in \mathbb{R}\}$ โดยให้ y_i เป็นคำตอบของโมเดลต้นไม้ตัดสินใจ (Decision trees) ร่วมกันดังสมการดังสมการที่ 12 (Pecho et al., 2024)

$$\hat{A}_i = \phi(X_i) = \sum_{k=1}^K f_k(X_i), f_k \in T \quad (12)$$

โดยที่ K คือ จำนวนต้นไม้ตัดสินใจ

f_k คือ ฟังก์ชันการทำนายที่ได้จากต้นไม้ตัดสินใจที่ k

โดยต้องทำการลดจำนวนการสูญเสีย (Loss) เข้าไปในสมการ

$$l(\phi) = \sum_i l(y_i, \hat{A}_i) + \sum_k \Omega(f_k) \quad (13)$$

โดยที่ l คือ ฟังก์ชันการสูญเสีย ที่แสดงความแตกต่างระหว่างค่าที่คาดหวัง y_i และ y_i ผลลัพธ์ที่เกิดขึ้นจริง

$\Omega(f_k)$ คือ เป็นฟังก์ชัน Regularization ที่ช่วยควบคุมความซับซ้อนของโมเดล เพื่อป้องกันปัญหา over-fitting

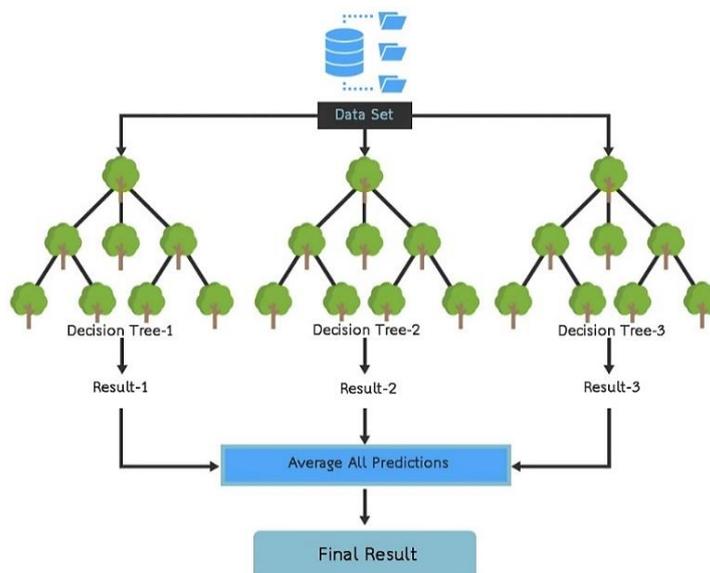
$$\Omega(f_k) = \gamma T + \frac{1}{2} \lambda \|W\|^2 \quad (14)$$

โดยที่ γ และ λ คือ ไฮเปอร์พารามิเตอร์ที่ต้องกำหนดค่า

W คือ ค่าน้ำหนักของแต่โหนดใบ (Leaf)

T คือ จำนวนโหนดของต้นไม้ตัดสินใจ

3.3.5 Random forest (Breiman, 2001) เป็นโมเดลที่ใช้อัลกอริทึมการเรียนรู้ของเครื่อง (Machine learning) โดยใช้วิธีการที่รวมการพยากรณ์ของแผนผังการตัดสินใจหลาย ๆ แบบ เพื่อให้การพยากรณ์แม่นยำและมีประสิทธิภาพมากขึ้น จากการสร้างโมเดล ต้นไม้ตัดสินใจหลาย ๆ โมเดลย่อย แต่ละโมเดลจะได้รับชุดข้อมูลไม่เหมือนกัน และทำหน้าที่พยากรณ์ให้แต่ละต้นไม้ตัดสินใจของตนเอง และคำนวณผลพยากรณ์ด้วยการหาค่าเฉลี่ยของผลการพยากรณ์จากทุกโมเดลย่อย ดังแสดงดังรูปที่ 3



รูปที่ 3 แสดงการทำงานของ Random forest

Random forest เป็นโมเดลที่ใช้ในการแก้ปัญหาการพยากรณ์ทั้งแบบการจัดหมวดหมู่ (Classification) และการถดถอย (Regression) โดยการใช้ Random forest เพื่อการพยากรณ์ข้อมูลอนุกรมเวลา จำเป็นต้องแปลงข้อมูลอนุกรมเวลาให้เหมาะสมกับการพยากรณ์แบบถดถอยก่อน สมการสำหรับการถดถอยแสดงดังสมการที่ 15

$$\hat{y} = \frac{1}{n} \sum_{i=1}^n T_i(x) \quad (15)$$

โดยที่ $T_i(x)$ คือ ผลลัพธ์ของต้นไม้การตัดสินใจที่ i

n คือ จำนวนต้นไม้ตัดสินใจ

3.4 นำข้อมูลทั้งหมด 48 เดือน แบ่งเป็น 2 ส่วน เพื่อเรียนรู้และทดสอบ โดยใช้ข้อมูล 36 เดือนในการเรียนรู้ (Train) และใช้ข้อมูล 12 เดือนในการทดสอบ (Test)

3.5 นำโมเดล ทั้ง 5 โมเดลไปพยากรณ์ โดยใช้ข้อมูลและการแบ่งข้อมูลเหมือนกันทั้ง 5 โมเดล

3.6 ประเมินประสิทธิภาพของโมเดล โดยใช้วิธีการประเมินประสิทธิภาพ 3 วิธี

3.6.1 ค่าความคลาดเคลื่อนสัมบูรณ์เฉลี่ย (Mean absolute error: MAE) (Vandepu, 2019; Botchkarev, 2019) เป็นการวัดค่าเฉลี่ยของความแตกต่างในแนวตั้งระหว่างค่าที่คาดการณ์กับค่าจริง คำนวณจากผลต่างแบบสัมบูรณ์ระหว่างค่าที่คาดการณ์กับค่าจริง แล้วหารด้วยจำนวนข้อมูลทั้งหมดดังสมการที่ 16

$$MAE = \frac{1}{n} \sum_{i=1}^n |A_i - F_t| \quad (16)$$

โดยที่ n คือ จำนวนข้อมูลทั้งหมด

A_t คือ ค่าจริงในช่วงเวลา t

F_t คือ ค่าที่พยากรณ์ได้ในช่วงเวลา t

3.6.2 รากที่สองของค่าความคลาดเคลื่อนสัมบูรณ์เฉลี่ย (Root mean square Error: RMSE) (Vandepu, 2019; Botchkarev, 2019) เป็นการวัดค่าเฉลี่ยของความแตกต่างที่ถูกยกกำลังสองระหว่างค่าที่คาดการณ์กับค่าจริง แล้วนำไปถอดรากที่สอง คำนวณได้จากการยกกำลังสองผลต่างระหว่างค่าที่คาดการณ์กับค่าจริง หารด้วยจำนวนข้อมูล แล้วถอดรากที่สองดังสมการที่ 17

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (A_i - F_t)^2} \quad (17)$$

โดยที่ n คือ จำนวนข้อมูลทั้งหมด

A_t คือ ค่าจริงในช่วงเวลา t

F_t คือ ค่าที่พยากรณ์ได้ในช่วงเวลา t

3.6.3 ร้อยละของค่าความคลาดเคลื่อนสัมบูรณ์เฉลี่ย (Mean absolute percentage error: MAPE) (Vandepu, 2019; Botchkarev, 2019) เป็นการวัดค่าเฉลี่ยของความแตกต่างในแนวตั้งระหว่างค่าที่คาดการณ์กับค่าจริงในรูปแบบร้อยละ โดยคำนวณจากผลต่างแบบสัมบูรณ์ระหว่างค่าที่คาดการณ์กับค่าจริง แล้วหารด้วยค่าจริง และคูณด้วย 100 เพื่อให้อยู่ในรูปแบบร้อยละ ดังสมการที่ 18

$$MAPE = \frac{1}{n} \sum_{i=1}^n \left| \frac{A_i - F_t}{A_t} \right| \times 100 \quad (18)$$

โดยที่ n คือ จำนวนข้อมูลทั้งหมด

A_t คือ ค่าจริงในช่วงเวลา t

F_t คือ ค่าที่พยากรณ์ได้ในช่วงเวลา t

3.7 หลังจากได้ผลลัพธ์จากการวิจัยแล้ว นำผลลัพธ์ไปวิเคราะห์ในขั้นตอนสุดท้าย

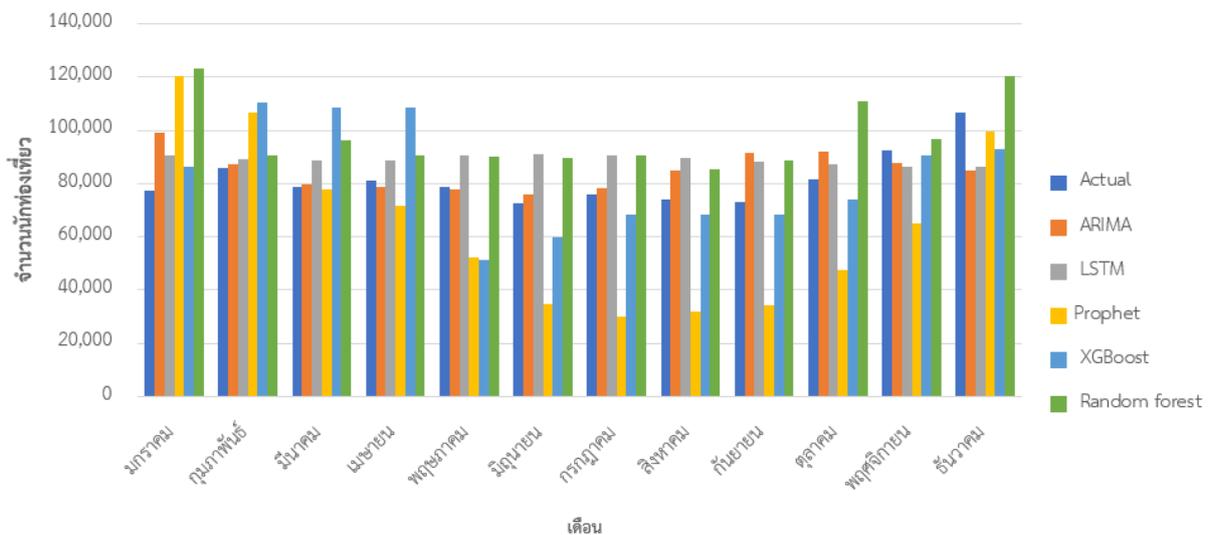
ผลการวิจัยและวิจารณ์ผล

จากการนำทั้ง 5 โมเดลประกอบด้วย ARIMA LSTM Prophet XGBoost และ Random forest ไปใช้ในการพยากรณ์จำนวนนักท่องเที่ยวในเชียงใหม่ โดยได้พยากรณ์จำนวนนักท่องเที่ยวในจังหวัดเชียงใหม่ ในปี พ.ศ. 2566 ตั้งแต่เดือนมกราคมจนถึงธันวาคมเป็นจำนวน 12 เดือน ได้ผลการพยากรณ์จำนวนนักท่องเที่ยวเทียบกับจำนวนจริงของนักท่องเที่ยวออกมาแสดงดังตารางที่ 2

ตารางที่ 2 แสดงผลพยากรณ์ของแต่ละโมเดลเทียบกับค่าจริง (Actual)

Month	Actual	ARIMA	LSTM	Prophet	XGBoost	Random forest
มกราคม	77,106	98,923	90,721	120,294	86,407	123,152
กุมภาพันธ์	85,610	87,282	88,929	106,568	110,282	90,425
มีนาคม	78,912	79,776	88,434	77,753	108,305	96,374
เมษายน	81,290	78,521	88,861	71,518	108,305	90,520
พฤษภาคม	78,735	77,994	90,765	52,363	51,113	89,847
มิถุนายน	72,419	75,998	91,132	34,966	59,973	89,416
กรกฎาคม	75,879	78,362	90,478	30,075	68,265	90,348
สิงหาคม	73,931	84,774	89,495	31,732	68,265	85,147
กันยายน	72,826	91,547	88,331	34,376	68,265	88,778
ตุลาคม	81,678	91,767	87,219	47,746	74,200	111,036
พฤศจิกายน	92,261	87,862	86,451	65,247	90,757	96,654
ธันวาคม	106,813	84,886	86,273	99,746	92,946	120,277

เมื่อนำผลการพยากรณ์จากตารางที่ 2 มาแสดงในรูปแบบกราฟจะเห็นได้ว่าเส้นกราฟของโมเดล ARIMA มีความใกล้เคียงกับค่าจริงมากที่สุด ดังแสดงในรูปที่ 4



รูปที่ 4 เปรียบเทียบค่าที่พยากรณ์ได้ในแต่ละโมเดลเทียบกับค่าจริง

เมื่อทำการประเมินประสิทธิภาพของโมเดลทั้ง 3 วิธี คือ คำนวณหาค่าความคลาดเคลื่อนสัมบูรณ์เฉลี่ย รากที่สองของค่าความคลาดเคลื่อนสัมบูรณ์เฉลี่ย และร้อยละของค่าความคลาดเคลื่อนสัมบูรณ์เฉลี่ย ออกมาก็พบว่า โมเดล ARIMA มีความคลาดเคลื่อนต่ำที่สุด จากการทดสอบประสิทธิภาพทั้ง 3 วิธี โดยแสดงผลการประเมินประสิทธิภาพของโมเดลทั้ง 5 โมเดล ดังตารางที่ 3

โดยโมเดล ARIMA ที่ใช้ทำการทดสอบ และได้ผลการประเมินประสิทธิภาพออกมาที่มีความแม่นยำสูงสุด ใช้พารามิเตอร์ ARIMA (p,d,q) เท่ากับ ARIMA (8, 1, 0)

p = 8 แสดงถึงจำนวนค่าก่อนหน้า (Lagged) ที่ใช้ในส่วนของ Autoregressive (AR)

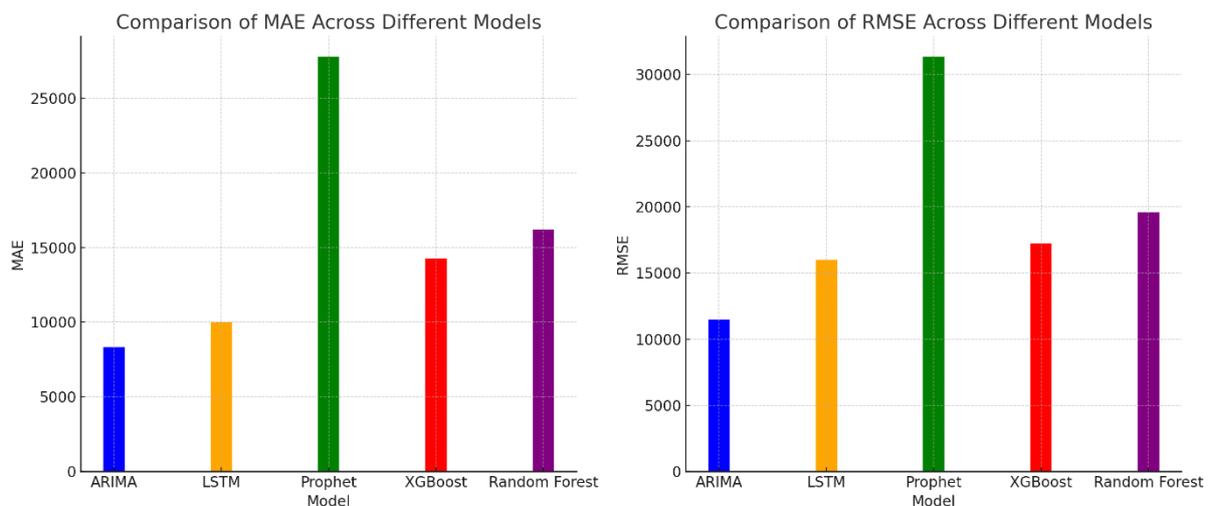
d = 1 บ่งชี้ว่ามีการทำ First-order differencing เพื่อให้อนุกรมเวลาที่มีความนิ่ง (Stationary)

q = 0 แสดงว่าไม่มีการใช้ส่วนของ Moving Average (MA) ในแบบจำลองนี้

ตารางที่ 3 ผลการประเมินประสิทธิภาพด้วยค่าความคลาดเคลื่อนสัมบูรณ์เฉลี่ย รากที่สองของค่าความคลาดเคลื่อนสัมบูรณ์เฉลี่ย และร้อยละของค่าความคลาดเคลื่อนสัมบูรณ์เฉลี่ยโมเดลทั้ง 5 โมเดล

โมเดล	MAE	RMSE	MAPE
ARIMA	8,325.33	11,462.63	10.16 %
LSTM	9,973.01	15,991.98	11.15 %
Prophet	27,780.50	3,1341.67	35.57%
XGBoost	14,261.57	17,243.82	17.61%
Random forest	16,209.42	19,574.49	20.49%

เมื่อนำค่าความคลาดเคลื่อนสัมบูรณ์เฉลี่ยกับรากที่สองของค่าความคลาดเคลื่อนสัมบูรณ์เฉลี่ยแต่ละโมเดล มาเปรียบเทียบกับในรูปแบบกราฟ จะเห็นความแตกต่างได้อย่างชัดเจน ดังรูปที่ 5



รูปที่ 5 แสดงกราฟเปรียบเทียบค่า ค่าความคลาดเคลื่อนสัมบูรณ์เฉลี่ยกับรากที่สองของค่าความคลาดเคลื่อนสัมบูรณ์เฉลี่ย

ผลการวิจัยชี้ให้เห็นว่าโมเดล ARIMA มีประสิทธิภาพสูงสุดในการพยากรณ์โดยมีค่าความคลาดเคลื่อนสัมบูรณ์เฉลี่ย รากที่สองของค่าความคลาดเคลื่อนสัมบูรณ์เฉลี่ย และร้อยละของค่าความคลาดเคลื่อนสัมบูรณ์เฉลี่ย พบว่า โมเดล ARIMA มีความคลาดเคลื่อนต่ำที่สุด โดยค่าความคลาดเคลื่อนสัมบูรณ์เฉลี่ยของโมเดล ARIMA เท่ากับ 8,325.33 เมื่อเทียบกับขนาดของข้อมูลของจำนวนนักท่องเที่ยวในจังหวัดเชียงใหม่ ที่มีค่าสูงสุดของข้อมูลทั้งหมดเท่ากับ 1,047,546 คน และต่ำสุดที่ 3,523 คน ซึ่งจากข้อมูลนี้ แสดงให้เห็นว่าข้อมูลมีความผันผวนสูงมาก เนื่องจากข้อมูลครอบคลุมช่วงเวลาที่ได้รับผลกระทบอย่างมากจากสถานการณ์การแพร่ระบาดของ COVID-19 ในช่วงปี พ.ศ. 2563 - 2564 แต่เมื่อพิจารณาเป็นร้อยละเทียบกับค่าสูงสุด ค่าความคลาดเคลื่อนจะมีค่าเพียงร้อยละ 0.79 ของค่าสูงสุด และคิดเป็นร้อยละ 3.44 ของค่าเฉลี่ย

ในส่วนของการที่สองของค่าความคลาดเคลื่อนสัมบูรณ์เฉลี่ย มีค่าเท่ากับ 11,462.63 เทียบกับจำนวนสูงสุด คิดเป็นร้อยละอยู่ที่ 1.09 ของค่าสูงสุด และเมื่อเทียบกับค่าเฉลี่ยแล้วจะมีคิดเป็น 4.73 ของค่าเฉลี่ย และค่าร้อยละของค่าความคลาดเคลื่อนสัมบูรณ์เฉลี่ยมีค่าอยู่ที่ 10.16 ซึ่งเมื่อพิจารณาจากค่าความคลาดเคลื่อนทั้ง 3 ค่าแล้ว ถือว่าโมเดล ARIMA มีความคลาดเคลื่อนในการพยากรณ์อยู่ในระดับต่ำ ซึ่งสอดคล้องกับงานวิจัยของ Fan *et al.* (2021) ได้ใช้โมเดลในการคาดการณ์การผลิตหุ้มน้ำมันและประสิทธิภาพ ถือเป็นสิ่งสำคัญในการยืดวงจรชีวิตของหุ้มน้ำมันและปรับปรุงการฟื้นตัวของอ่างเก็บน้ำ ซึ่งผลการทดลองระบุว่าโมเดล ARIMA มีประสิทธิภาพที่ดีกว่า LSTM ในช่วงที่มีการผลิตลดลงอย่างต่อเนื่อง และ Verma *et al.* (2021) การพยากรณ์ข้อมูลอนุกรมเวลาได้รับความสนใจอย่างมากในยุคปัจจุบัน การประยุกต์ใช้ในทุกอุตสาหกรรม รวมถึงการดูแลสุขภาพ เศรษฐศาสตร์ ห่วงโซ่อุปทาน และอื่น ๆ ทำให้เป็นกลยุทธ์ที่นำไปใช้สำหรับโมเดลธุรกิจหรือแอปพลิเคชันการวิจัยใด ๆ ที่เกี่ยวข้องกับการคาดการณ์ มีเทคนิคทางสถิติที่ได้รับการยอมรับและเป็นแบบแผนหลายอย่างที่ถูกนำมาใช้เพื่อคาดการณ์ โดยพื้นฐานแล้วโมเดล ARIMA ได้รับการยอมรับว่าทำงานด้วยความแม่นยำสูงสุดในการทำนายการเปลี่ยนแปลงถัดไปในอนุกรมเวลา การวิจัยนี้มีจุดมุ่งหมายเพื่อเปรียบเทียบโดยใช้ตัวชี้วัดประสิทธิภาพต่าง ๆ เช่น ค่าความคลาดเคลื่อนสัมบูรณ์เฉลี่ย รากที่สองของค่าความคลาดเคลื่อนสัมบูรณ์เฉลี่ย และร้อยละของค่าความคลาดเคลื่อนสัมบูรณ์เฉลี่ย สำหรับแต่ละโมเดลเหล่านี้ โมเดลดังกล่าวได้รับการทดสอบเพื่อระบุดัชนีคุณภาพอากาศของกรุงเทพฯ เมืองหลวงของอินเดีย งานวิจัยของนิตกรและจารี (2565) ได้ทำการเปรียบเทียบประสิทธิภาพของโมเดล ARIMA และเทคนิคถดถอยการเรียนรู้ของเครื่องในการพยากรณ์ราคาปิดของหุ้น ผลการวิจัยพบว่าแบบจำลองจากโมเดล ARIMA มีผลการพยากรณ์ที่ใกล้เคียงค่าจริงมากที่สุด โดยมีค่าความคลาดเคลื่อนสัมบูรณ์เฉลี่ย รากที่สองของค่าความคลาดเคลื่อนสัมบูรณ์เฉลี่ยสำหรับการพยากรณ์ ราคาเปิดตลาด ราคาสูงสุด ราคาต่ำสุด ราคาปิดตลาด ปริมาณซื้อขายในแต่ละวัน และมูลค่าตลาดรวมต่ำที่สุด รวมถึงงานวิจัยของวรารัตนา (2565) การพยากรณ์มูลค่าการส่งออกของไทย มีวัตถุประสงค์เพื่อสร้างตัวแบบพยากรณ์มูลค่าการส่งออกของไทยด้วยวิธีการทางสถิติ ทั้งหมด 7 วิธี ผลการวิจัยพบว่าแบบจำลองจากเทคนิค Box-Jenkins Method ที่มีตัวแบบ ARIMA (p,d,q) เป็นวิธีที่เหมาะสมมากที่สุด

ผลการศึกษาแสดงให้เห็นถึงความแตกต่างอย่างมีนัยสำคัญในประสิทธิภาพของโมเดลต่าง ๆ ที่ใช้ในการพยากรณ์จำนวนนักท่องเที่ยวในจังหวัดเชียงใหม่ ซึ่งโมเดล ARIMA แสดงให้เห็นถึงความแม่นยำสูงสุด โดยมีค่าความคลาดเคลื่อนต่ำที่สุดในบรรดาโมเดลทั้งหมดที่ทำการทดสอบ ส่วนโมเดลอื่น ๆ ได้แก่ LSTM XGBoost และ Random forest แม้จะมีค่าความคลาดเคลื่อนสูงกว่า ARIMA แต่ยังคงอยู่ในระดับที่ยอมรับได้ โมเดลเหล่านี้แสดงให้เห็นถึงความสามารถที่น่าพอใจในการจัดการกับข้อมูลที่มีความซับซ้อน และมีรูปแบบตามฤดูกาล ซึ่งเป็นลักษณะเฉพาะของข้อมูลการท่องเที่ยวในจังหวัดเชียงใหม่ อย่างไรก็ตามโมเดล Prophet แสดงผลลัพธ์ที่แตกต่างอย่างมีนัยสำคัญ โดยมีค่าความคลาดเคลื่อนสูงกว่าโมเดลอื่น ๆ อย่างเห็นได้ชัด ผลลัพธ์นี้บ่งชี้ว่า โมเดล Prophet อาจไม่เหมาะสมสำหรับการพยากรณ์ข้อมูลนักท่องเที่ยวในบริบทเฉพาะของจังหวัดเชียงใหม่

สรุปผลการวิจัย

การวิจัยนี้มีวัตถุประสงค์เพื่อเปรียบเทียบประสิทธิภาพของโมเดลการพยากรณ์ข้อมูลอนุกรมเวลา 5 โมเดล ได้แก่ ARIMA LSTM Prophet XGBoost และ Random forest ในการพยากรณ์จำนวนนักท่องเที่ยวในจังหวัดเชียงใหม่ โดยใช้ข้อมูลตั้งแต่เดือนมกราคม พ.ศ. 2563 ถึงเดือนธันวาคม พ.ศ. 2566

จากผลการวิจัยนี้ สามารถสรุปได้ว่าโมเดล ARIMA เป็นตัวเลือกที่เหมาะสมที่สุดสำหรับการพยากรณ์จำนวนนักท่องเที่ยวในจังหวัดเชียงใหม่ เนื่องจากสามารถจับแนวโน้มและฤดูกาลในข้อมูลอนุกรมเวลาได้ดี อย่างไรก็ตามการปรับปรุงและการปรับแต่งโมเดล อาจช่วยเพิ่มความแม่นยำในการพยากรณ์ได้ในอนาคต โดยงานวิจัยนี้มีส่วนช่วยให้ผู้มีส่วนได้ส่วนเสีย

ในอุตสาหกรรมการท่องเที่ยวของจังหวัดเชียงใหม่ สามารถตัดสินใจอย่างมีข้อมูลและวางแผนกลยุทธ์เพื่อการเติบโตอย่างยั่งยืน และการจัดการกับความไม่แน่นอนในอนาคตได้ดียิ่งขึ้น

ข้อเสนอแนะในการวิจัยครั้งต่อไป

1. ควรเพิ่มข้อมูลจำนวนนักท่องเที่ยวที่จะใช้ในการเรียนรู้ให้มากขึ้น เพื่อให้โมเดลมีความสามารถในการเรียนรู้ แนวโน้มและรูปแบบได้ดียิ่งขึ้น
2. เพิ่มโมเดลในการพยากรณ์แบบอื่น ๆ เพื่อที่จะได้โมเดลที่มีความสามารถพยากรณ์ได้แม่นยำมากที่สุด ในการพยากรณ์ข้อมูลจำนวนนักท่องเที่ยว เช่น LLTime และ TimesFM
3. ใช้ปัจจัยภายนอกเพิ่มเติม เช่น ข้อมูลเศรษฐกิจ สภาพอากาศ หรือเหตุการณ์สำคัญต่าง ๆ ที่ส่งผลต่อจำนวนนักท่องเที่ยว อาจช่วยปรับปรุงความแม่นยำของการพยากรณ์ในงานวิจัยต่อไป

เอกสารอ้างอิง

- กระทรวงการท่องเที่ยวและกีฬา. (2567). ข้อมูลสถิตินักท่องเที่ยว (Tourism Statistic). แหล่งข้อมูล: www.mots.go.th/more_news_new.php?cid=411. ค้นเมื่อวันที่ 4 เมษายน 2567.
- นิติกร จันทาญ และจारी ทองคำ. (2565). การเปรียบเทียบประสิทธิภาพของเทคนิคอริมาและเทคนิคถดถอยการเรียนรู้ของเครื่องในการพยากรณ์ราคาบิตคอยน์. วารสารวิทยาศาสตร์และเทคโนโลยี มหาวิทยาลัยอุบลราชธานี. 24 (1): 62 - 75.
- บุญศรี ไศภณัฐยานนท์. (2557). ความสัมพันธ์ระหว่างอัตราผลตอบแทนของตลาดหลักทรัพย์ อัตราดอกเบี้ย และอัตราแลกเปลี่ยนเงินตราต่างประเทศ ของกลุ่มประเทศอาเซียน AEC. สารนิพนธ์ปริญญาบัณฑิต, มหาวิทยาลัยมหิดล. กรุงเทพฯ. 59 หน้า.
- วรางคณา เรียนสุทธิ. (2565). การพยากรณ์มูลค่าการส่งออกพริกไทย. วารสารวิจัยและส่งเสริมวิชาการเกษตร สำนักวิจัยและส่งเสริมวิชาการเกษตร มหาวิทยาลัยแม่โจ้. 39(3): 156 - 168.
- Breiman, L. (2001). Random Forest. *Machine Learning* 45: 5 - 32. doi: 10.1023/A:1010933404324.
- Botchkarev, A. (2019). Performance metrics (error measures) in machine learning regression, forecasting and prognostics: Properties and typology. *Interdisciplinary Journal of Information, Knowledge, and Management* 14: 45 - 79.
- Chen, T. and Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, San Francisco. 785 - 794.
- Fan, D., Sun, H., Yao, J., Zhang, K., Yan, X. and Sun, Z. (2021). Well production forecasting based on ARIMA-LSTM model considering manual operations. *Energy* 220: 119708. doi 10.1016/j.energy.2020.119708.
- Greff, K., Srivastava, R.K., Koutnik, J., Steunebrink, B.R. and Schmidhuber, J. (2017). LSTM: A Search Space Odyssey. *IEEE Transactions on Neural Networks and Learning Systems* 28(10): 1 - 12. doi 10.1109/TNNLS.2016.2582924.
- Nochai, R. and Nochai, T. (2006). ARIMA model for forecasting oil palm price. In: *Proceedings of the 2nd IMT-GT Regional Conference on Mathematics, Statistics and applications*, Penang. 13 - 15.

- Pecho, R.D.C., Vijaya, K., Sharma, N., Pal, H. and Jose, B.K. (2024). An approach for crop yield prediction using hybrid XGBoost, SVM and C4. 5 classifier algorithms. *Engineering & Applied Science Research* 51(3): 300 - 312.
- Taylor, S.J. and Letham, B. (2018). Forecasting at scale. *The American Statistician* 72(1): 37 - 45. doi: 10.7287/peerj.preprints.3190v2.
- Vandeput, N. (2019). Forecast KPI: RMSE, MAE, MAPE & Bias. *Towards Data Science*. Source: <https://towardsdatascience.com/forecast-kpi-rmse-mae-mape-bias-cdc5703d242d>. Retrieved from March 2024.
- Verma, P., Reddy, S.V., Raha, L. and Datta, D. (2021). Comparison of time-series forecasting models. In: 2021 International Conference on Intelligent Technologies (CONIT). Hubli, India. 82 - 88. doi 10.1109/CONIT51480.2021.9498451.

