การพยากรณ์ยอดขายเพื่อเพิ่มประสิทธิภาพการบริการลูกค้าของธุรกิจเครื่องดื่ม ด้วยการเรียนรู้ของเครื่อง

SALES FORECASTING TO ENHANCE CUSTOMER SERVICE EFFICIENCY IN THE BEVERAGE INDUSTRY USING MACHINE LEARNING

ธราเทพ สุดวิไล

บัณฑิตวิทยาลัย มหาวิทยาลัยศรีนครินทรวิโรฒ

2567

การพยากรณ์ยอดขายเพื่อเพิ่มประสิทธิภาพการบริการลูกค้าของธุรกิจเครื่องดื่ม ด้วยการเรียนรู้ของเครื่อง

ธราเทพ สุดวิไล

สารนิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตร
วิทยาศาสตรมหาบัณฑิต สาขาวิชาวิทยาการข้อมูล
คณะวิทยาศาสตร์ มหาวิทยาลัยศรีนครินทรวิโรฒ
ปีการศึกษา 2567
ลิขสิทธิ์ของมหาวิทยาลัยศรีนครินทรวิโรฒ

SALES FORECASTING TO ENHANCE CUSTOMER SERVICE EFFICIENCY IN THE BEVERAGE INDUSTRY USING MACHINE LEARNING

THARATHEP SUDVILAI

An Master's Project Submitted in Partial Fulfillment of the Requirements

for the Degree of MASTER OF SCIENCE

(Data Science)

Faculty of Science, Srinakharinwirot University

2024

Copyright of Srinakharinwirot University

สารนิพนธ์

เรื่อง

การพยากรณ์ยอดขายเพื่อเพิ่มประสิทธิภาพการบริการลูกค้าของธุรกิจเครื่องดื่ม

ด้วยการเรียนรู้ของเครื่อง

ของ

ธราเทพ สุดวิไล

ได้รับอนุมัติจากบัณฑิตวิทยาลัยให้นับเป็นส่วนหนึ่งของการศึกษาตามหลักสูตร

ปริญญาวิทยาศาสตรมหาบัณฑิต สาขาวิชาวิทยาการข้อมูล

ของมหาวิทยาลัยศรีนครินทรวิโรฒ

(รองศาสตราจารย์ นายแพทย์ฉัตรชัย เอกปัญญาสกุล)

คณบดีบัณฑิตวิทยาลัย

คณะกรรมการสอบปากเปล่าสารนิพนธ์

..... ที่ปรึกษาหลัก

(อาจารย์ ดร.ศุภร คนธภักดี)

..... ประธาน

(ผู้ช่วยศาสตราจารย์ ดร.จันตรี ผลประเสริฐ)

..... กรรมการ

(ผู้ช่วยศาสตราจารย์ ดร.ศิริสรรพ เหล่าหะเกียรติ)

ชื่อเรื่อง	การพยากรณ์ยอดขายเพื่อเพิ่มประสิทธิภาพการบริการลูกค้าของธุรกิจเครื่องดื่มด้วยการเรียนรู้ของเครื่อง
ผู้วิจัย	ธราเทพ สุดวิไล
ปริญญา	วิทยาศาสตรมหาบัณฑิต
ปีการศึกษา	2567
อาจารย์ที่ปรึกษา	อาจารย์ ดร. ศุภร คนธภักดี

การพยากรณ์ยอดขายมีความสำคัญอย่างยิ่งในอุตสาหกรรมเครื่องดื่มและอุตสาหกรรมผลิตภัณฑ์อื่น ๆ เนื่องจากช่วยในการบริหารจัดการสินค้าคงคลังอย่างมีประสิทธิภาพ ตอบสนองความต้องการของผู้บริโภค ลดต้นทุนด้านการผลิตและการขนส่ง รวมถึงสนับสนุนการวางแผนและการตัดสินใจเชิงกลยุทธ์ ซึ่งส่งผลต่อความสามารถในการแข่งขันและการปรับตัวต่อการเปลี่ยนแปลงของตลาดอย่างทันท่วงที งานวิจัยนี้มีวัตถุประสงค์เพื่อพัฒนาระบบพยากรณ์ยอดขายในธุรกิจเครื่องดื่มโดยใช้เทคโนโลยีการเรียนรู้ของเครื่อง (Machine Learning) เพื่อเพิ่มความแม่นยำในการคาดการณ์ยอดขาย โดยใช้ข้อมูลยอดขายระหว่างเวลา 08:00 น. ถึง 14:00 น. ในการพยากรณ์ยอดขายช่วงเวลา 17:30 น. เทคนิคที่ใช้ในการทดลองประกอบด้วย Random Forest, Gradient Boosting, XGBoost และ Linear Regression พร้อมทั้งเปรียบเทียบกับวิธีดั้งเดิม เช่น ค่าเฉลี่ยเคลื่อนที่ (Moving Average) ผลการทดลองแสดงให้เห็นว่า Gradient Boosting และ Linear Regression ให้ผลลัพธ์ที่แม่นยำสูงกว่าวิธีอื่น ๆ รวมถึง Lasso Regression, Ridge Regression, Elastic Net Regression และ Moving Average แบบดั้งเดิม ผลลัพธ์ของงานวิจัยนี้สามารถนำไปประยุกต์ใช้ในด้านการวางแผนการผลิต การจัดการสินค้าคงคลัง การขนส่งสินค้า และการบริหารจัดการธุรกิจโดยรวม เพื่อช่วยลดต้นทุน เพิ่มประสิทธิภาพในการดำเนินงาน และตอบสนองความต้องการของลูกค้าได้อย่างมีประสิทธิภาพ

คำสำคัญ : การพยากรณ์ยอดขาย อุตสาหกรรมเครื่องดื่ม การเรียนรู้ของเครื่อง การวางแผนการผลิต การตัดสินใจเชิงกลยุทธ์ ความแม่นยำในการคาดการณ์

Title	SALES FORECASTING TO ENHANCE CUSTOMER SERVICE EFFICIENCY IN THE BEVERAGE INDUSTRY USING MACHINE LEARNING
Author	THARATHEP SUDVILAI
Degree	MASTER OF SCIENCE
Academic Year	2024
Thesis Advisor	dr. Subhorn Khonthapagdee

Sales forecasting plays a vital role in the beverage industry and other product-based sectors, as it supports effective inventory management, meets customer demand, reduces production and transportation costs, and facilitates strategic business planning and decision-making. Accurate forecasting enhances market competitiveness and enables businesses to respond swiftly to market fluctuations. This study aimed to develop a sales forecasting system for the beverage industry by using machine learning (ML) techniques to improve prediction accuracy. The system uses sales data collected between 8:00 AM and 2:00 PM to forecast sales at 5:30 PM. Several ML models were implemented and evaluated, including Random Forest, Gradient Boosting, XGBoost, and Linear Regression, and were compared with traditional methods such as the Moving Average. Experimental results indicated that Gradient Boosting and Linear Regression models achieved higher forecasting accuracy than other techniques, including Lasso Regression, Ridge Regression, Elastic Net Regression, and the conventional Moving Average approach. The findings of this study can be applied to enhance various business functions in the beverage industry, such as production planning, inventory control, logistics management, and overall operational efficiency. This helps reduce costs, improve efficiency, and better meet consumer needs.

Keyword : Sales forecasting Beverage industry Machine learning Production planning Strategic decision-making Forecast accuracy

กิตติกรรมประกาศ

ในการทำสารนิพนธ์ในครั้งนี้ จะไม่สามารถทำสำเร็จได้หากไม่ได้รับคำแนะนำ ชี้แนะตลอดจนคอยให้กำลังใจและคำปรึกษาจาก อ.ดร. ศุภร คนธภักดี ที่เป็นอาจารย์ที่ปรึกษาในการทำจนทำให้สารนิพนธ์เล่มนี้สำเร็จได้ตามที่วางแผนไว้

นอกจากนี้ข้าพเจ้าต้องขอขอบพระคุณอาจารย์ทุกท่าน ในภาควิชาวิทยาการข้อมูล คณะวิทยาศาสตร์ มหาวิทยาลัยศรีนครินทรวิโรฒ ที่ได้ถ่ายทอดความรู้ในแต่ละวิชาที่ได้เรียนมาตั้งแต่ภาคการศึกษาที่ 1 จนถึงภาคการศึกษาสุดท้าย ทำให้ข้าพเจ้ามีองค์ความรู้นำมาประยุกต์ใช้ในการทำสารนิพนธ์เล่มนี้ รวมถึงต้องขอขอบพระคุณคณะกรรมการสอบสารนิพนธ์ ที่ได้ให้คำแนะนำ ตลอดจนชี้แนะจนทำให้สารนิพนธ์เล่มนี้เสร็จออกมาสมบูรณ์ได้

สุดท้ายนี้ต้องขอขอบคุณครอบครัว เพื่อนๆทั้งเพื่อนบดินทร เพื่อนปริญญาตรีเศรษฐศาสตร์ มศว และที่สำคัญเพื่อนปริญญาโท วิทยาการข้อมูลทุกคน โดยเฉพาะ มายด์ ต้น ภู ครีม บอลและเฟิร์นที่คอยช่วยเหลือและให้กำลังใจกันมาตลอดเกือบ 2 ปี รวมทั้งพี่ๆน้องๆที่ทำงาน ที่คอยให้กำลังใจและช่วยเหลือในช่วงที่ต้องเตรียมสอบและต้องทำเล่มตลอดมา ท้ายที่สุดนี้ข้าพเจ้าหวังเป็นอย่างยิ่งว่าสารนิพนธ์เล่มนี้คงมอบประโยชน์ให้แก่ผู้พบอ่านได้ไม่มากก็น้อย

ธราเทพ สุดวิไล

สารบัญ

หน้า

สารบัญตาราง

หน้า

สารบัญรูปภาพ

หน้า

บทที่ 1
บทนำ

1.1 ภูมิหลัง

การพยากรณ์ยอดขายเป็นหนึ่งในกระบวนการที่มีประโยชน์และจำเป็นกระบวนการหนึ่งในการบริหารจัดการธุรกิจในสมัยปัจจุบันนี้ โดยเฉพาะอย่างยิ่งในยุคที่ข้อมูลมีความจำเป็นอย่างมากในเชิงกลยุทธ์และการตัดสินใจของธุรกิจ ที่หากสามารถคาดคะเนยอดขายได้อย่างแม่นยำ จะทำให้สามารถสนองความต้องการของตลาดและลูกค้าได้ตรงกับความต้องการและรวดเร็ว รวมทั้งยังจะทำให้มีการจัดการสต็อกสินค้าที่มีความเหมาะสมมากขึ้น ช่วยทำให้ผู้ประกอบการสามารถลดต้นทุนการผลิตและสามารถให้บริการที่ดีกว่าแก่ลูกค้าได้ ในทางตรงกันข้ามนั้น หากการพยากรณ์ยอดขายไม่แม่นยำจะส่งผลให้เกิดความความเสียโอกาสในหลายๆด้าน เช่น การเกิดสินค้าคงคลังเกินความจำเป็นหรือการขาดแคลนสินค้าในช่วงที่สินค้านั้นมีความต้องการสูง ซึ่งอาจนำไปสู่การสูญเสียโอกาสทางการขายและลดความเชื่อถือในภาพลักษณ์ของบริษัทแก่ลูกค้า และยิ่งไปกว่านั้นในอีกมุมหนึ่งของการดำเนินงานที่สำหรับของธุรการเกี่ยวกับการค้าขายนั่นคือ ในด้านการจัดการด้านการขนส่งสินค้า ยิ่งปัจจุบันที่การแข่งขันสูงส่งผลให้แต่ละบริษัทได้มีการใช้กลยุทธ์ในการส่งสินค้าได้รวดเร็วทันใจเพื่อให้เกิดความพอใจและตอบสนองความต้องการแก่ลูกค้าหรือผู้ซื้อให้ได้ดีที่สุด หากไม่มีการวางแผนหรือเตรียมการที่ดีก็อาจจะเกิดปัญหาด้านการขนส่งสินค้าได้เช่นกัน

ในบริบทของงานวิจัยนี้จึงได้เล็งเห็นถึงความสำคัญของการพยากรณ์ยอดขายรวมที่ช่วงเวลา 17:30 น. เนื่องจากการที่สามารถพยากรณ์ยอดขายรวมที่เวลา 17:30 น. ได้ตั้งแต่ช่วงเวลา 14:00 น. นั้นจะทำให้สามารถเตรียมการเพื่อให้พร้อมรับกับยอดขายที่ต้องวางแผนจัดส่งในช่วงค่ำได้ เพื่อเพิ่มประสิทธิภาพและประสิทธิผลของงาน ซึ่งในปัจจุบันนั้นการที่จะทราบยอดขายรวมนั้นจะรู้ในช่วงเวลา 17:30 น. ของแต่ละวัน และคนวางแผนขนส่งสินค้าจะต้องวางแผนในการจัดส่งยอดขายนั้นในช่วง 17:30 น. ซึ่งเป็นเวลาที่ต่อเนื่องกันทำให้บางครั้งเกิดรถที่ใช้ขนส่งหรือพนักงานส่งสินค้าไม่เพียงพอต่อยอดขายที่ต้องวางแผนจัดส่งในพรุ่งนี้ ส่งผลให้ส่งสินค้าไม่ได้ตามเวลาที่ลูกค้าต้องการ ซึ่งหากสามารถรู้ปริมาณยอดขายก่อนเวลาปิดยอดการขายในแต่ละวันตั้งแต่ช่วงเวลาประมาณ 14:00 น. จะทำให้มีเวลาตั้งแต่ช่วง 14:00 น.- 17:30 น. ในการแก้ไขปัญหาหรือเตรียมการ เพื่อให้สามารถตอบสนองความต้องการของลูกค้าได้ 100% แต่การพยากรณ์ยอดขายที่แม่นยำในช่วงเวลานี้ถือเป็นความท้าทาย เนื่องจากต้องพิจารณาข้อมูล

จากช่วงเวลาก่อนหน้าและเหตุอื่นๆที่มีผลกระทบอย่างมากต่อความผันผวนของยอดขาย เช่น สภาพอากาศ กิจกรรมพิเศษ หรือเทศกาลต่างๆ

ในปัจจุบัน การนำเทคโนโลยี Machine Learning มาช่วยในการพยากรณ์ยอดขายเป็นที่แพร่หลายอย่างมาก เนื่องจากเทคนิคดังกล่าวสามารถช่วยในการประมวลผลข้อมูลขนาดใหญ่และสามารถตรวจจับแนวโน้มและรูปแบบที่ซับซ้อนได้อย่างมีประสิทธิภาพ การใช้เทคนิคเหล่านี้ช่วยให้แบบจำลองสามารถเรียนรู้จากข้อมูลย้อนหลังในระยะยาว สามารถนำเสนอผลการพยากรณ์ที่มีความแม่นยำที่ดีขึ้น เมื่อเทียบกับการใช้วิธีการพยากรณ์แบบดั้งเดิมที่มักอาศัยเพียงการวิเคราะห์เชิงสถิติหรือสมการคณิตศาสตร์อย่างเดียว

แนวทางการวิจัยนี้มีความสำคัญอย่างยิ่งในเชิงปฏิบัติ เพราะช่วยให้ธุรกิจสามารถจัดการการวางแผนการดำเนินงาน การผลิต และการจัดการสต็อกสินค้าของบริษัทได้อย่างมีความเหมาะสมมากขึ้น ยังช่วยให้ตอบสนองความต้องการของลูกค้าได้ดียิ่งขึ้น ลดปัญหาการขาดแคลนสินค้าในช่วงที่มีความต้องการสูง ลดความเสี่ยงที่อาจเกิดจากการจัดเก็บสินค้ามากเกินไป และลดความเลี่ยงในปัญหาด้านการขนส่งที่อาจไม่เพียงพอต่อยอดการสั่งซื้อของลูกค้า ทั้งนี้จะช่วยเสริมสร้างศักยภาพในการบริหารและจัดการของธุรกิจในตลาดที่มีการแข่งขันสูงและเปลี่ยนแปลงอย่างรวดเร็ว

การศึกษาและพัฒนาระบบพยากรณ์ยอดขายนี้ยังมีประโยชน์ในระยะยาว โดยสามารถนำไปใช้กับธุรกิจในหลากหลายอุตสาหกรรม ทั้งธุรกิจค้าปลีก การผลิต หรือการบริการ เพื่อให้การวางแผนการทำงานเป็นไปอย่างมีประสิทธิภาพ รวมถึงเป็นการพัฒนาศักยภาพของธุรกิจในการใช้เทคโนโลยีขั้นสูงในการบริหารจัดการ ซึ่งสอดคล้องกับทิศทางของโลกที่กำลังมุ่งสู่ยุคดิจิทัลและการใช้ข้อมูลในการขับเคลื่อนธุรกิจ

ด้วยเหตุนี้การวิจัยเรื่องการพยากรณ์ยอดขายที่เวลา 17:30 น. โดยใช้ข้อมูลตั้งแต่เวลา 8:00 น. ถึง 14:00 น. และการนำเทคนิค Machine Learning มาประยุกต์ใช้เพื่อให้ได้ผลลัพธ์ที่แม่นยำที่สุด จึงเป็นสิ่งที่มีความสำคัญและเป็นประโยชน์อย่างยิ่ง ไม่เพียงแต่ในเชิงทฤษฎีแต่ยังในเชิงปฏิบัติสำหรับการดำเนินธุรกิจในยุคปัจจุบันและอนาคต

1.2 **ความมุ่งหมายของงานวิจัย**

งานวิจัยครั้งนี้ผู้วิจัยได้ตั้งความมุ่งหมายไว้ดังนี้

1.2.1 เพื่อพัฒนาระบบการพยากรณ์ยอดขายรวมที่เวลาช่วง 17:30 น. โดยใช้ข้อมูลยอดขายที่เข้ามาในแต่ละชั่วโมงตั้งแต่เวลา 8:00 น. ถึง 14:00 น. ในการพยากรณ์ เพื่อให้ได้การพยากรณ์ที่แม่นยำและสามารถนำไปใช้ในการทำงานจริงของภาคธุรกิจในการบริหารจัดการการ

ส่งมอบสินค้าให้ลูกค้าได้อย่างดีที่สุด ครบถ้วนและตรงเวลา และยังสามารถใช้ในการตัดสินใจเกี่ยวกับการจัดการวางแผนการผลิต เพื่อให้สินค้าคงคลังอยู่ในจุดที่เหมาะสมและเพียงพอต่อการขายตามความต้องการของลูกค้า

1.2.2 เพื่อประยุกต์ใช้เทคนิคของ Machine Learning เช่น Random Forest, XGBoost, Linear Regressionและเทคนิคอื่นๆ ในการสร้างแบบจำลองการพยากรณ์ยอดขาย โดยมุ่งหวังที่จะเปรียบเทียบประสิทธิภาพของแต่ละแบบจำลองที่ศึกษา เพื่อระบุว่าแบบจำลองใดที่เหมาะสมและมีความใกล้เคียงในการพยากรณ์ยอดขายมากที่สุด

1.2.3 เพื่อศึกษาและวิเคราะห์ปัจจัยที่ส่งผลต่อยอดขาย โดยพิจารณาถึงตัวแปรต่างๆ เช่น ช่วงเวลาที่ยอดขายเข้ามาในแต่ละวัน, วันหยุดสุดสัปดาห์, วันหยุดนักขัตฤกษ์ หรือปัจจัยอื่นๆ ที่อาจมีผลต่อยอดขาย เพื่อให้สามารถที่จะทราบถึงพฤติกรรมการซื้อของลูกค้า และนำผลการวิเคราะห์นี้มาใช้ในการพัฒนาแผนงานให้ตอบสนองตลาดมากขึ้น

1.2.4 เพื่อสร้างเทคนิคหรือวิธีในการสนับสนุนการตัดสินใจ ที่ช่วยให้ธุรกิจสามารถจัดการสินค้าคงคลัง วางแผนการผลิตและวางแผนขนส่งสินค้าได้อย่างเหมาะสม ลดปัญหาการมีสินค้าที่มากเกินไปหรืออาจน้อยเกินไป และเพิ่มพึงพอใจของลูกค้าในการตอบสนองความต้องการได้อย่างทันท่วงที

1.2.5 เพื่อเพิ่มประสิทธิภาพและจุดแข็งขององค์กร โดยการนำผลการพยากรณ์ยอดขายที่ได้มาปรับปรุงการวางแผนและกลยุทธ์การขาย เพื่อเพิ่มยอดขายและรักษารวมถึงเพิ่มระดับความพอใจของลูกค้า ซึ่งจะเป็นปัจจัยสำคัญในการเติบโตของธุรกิจอย่างยั่งยืนและยาวนาน

1.3 ความสำคัญของงานวิจัย

การพยากรณ์ยอดขายมีความสำคัญอย่างมากต่อการดำเนินธุรกิจในยุคปัจจุบันที่มีการแข่งขันสูงและความไม่แน่นอนของตลาดที่เพิ่มขึ้นอย่างต่อเนื่อง หากธุรกิจสามารถพยากรณ์ยอดขายได้อย่างแม่นยำ จะช่วยให้สามารถตัดสินใจเกี่ยวกับการจัดการวางแผนการผลิต เพื่อให้สินค้าคงคลังอยู่ในจุดที่เหมาะสมอย่างมีประสิทธิภาพ ลดการสูญเสียจากการสต็อกสินค้ามากเกินไปหรือขาดแคลนสินค้าในช่วงที่มีความต้องการต่างจากปกติ การบริหารจัดการที่ดีจะนำไปสู่การมีต้นทุนที่ถูก เพิ่มผลกำไรให้องค์กร และมีระดับความพึงพอใจของลูกค้าที่ดี ทั้งยังสามารถช่วยให้ผู้ประกอบการสามารถวางแผนรองรับการขนส่งสินค้าได้อย่างมีประสิทธิภาพมากขึ้น สามารถตอบสนองความต้องการลูกค้าได้ทันถ่วงที ลดปัญหาการส่งสินค้าล่าช้าที่อาจเกิดขึ้นได้ หากไม่สามารถพยากรณ์ยอดขายได้อย่างแม่นยำ ธุรกิจอาจประสบปัญหาการสต็อกสินค้าขาดแคลน ซึ่งจะทำให้สูญเสียโอกาสในการขาย นอกจากนี้ การสต็อกสินค้าเกินความจำเป็นยังทำให้

เกิดการสูญเสียค่าใช้จ่ายในการจัดเก็บ ซึ่งส่งผลกระทบต่อต้นทุนของบริบัทและความสามารถในการลงทุนในโอกาสใหม่ ๆ ของธุรกิจได้ และหากมียอดขายเข้ามาเป็นจำนวนมากแต่ภาคการขนส่งสินค้าไม่สามารถรองรับความต้องการ เกิดการส่งสินค้าล่าช้าและไม่ครบถ้วน ก็จะส่งผลต่อความเชื่อมั่นของลูกค้าที่อาจลดลงได้หรือเปลี่ยนใจไปซื้อสินค้ากับคู่แข่งที่สามารถตอบสนองได้รวดเร็วกว่า

งานวิจัยนี้จึงมุ่งไปที่การพัฒนาแบบจำลองที่ใช้ในการพยากรณ์ยอดขายที่สามารถนำมาใช้ได้จริงในการทำงาน ช่วยให้ธุรกิจปรับตัวและตัดสินใจได้รวดเร็วขึ้น การนำเทคโนโลยี Machine Learning มาใช้ในการทำนายยอดขายจะเป็นการเพิ่มความถูกต้องในการคาดการณ์ให้ดีขึ้น ส่งเสริมให้ธุรกิจมีข้อมูลเชิงลึกที่ใช้ในการกำหนดแผนงานและการตัดสินใจเชิงกลยุทธ์อย่างมีประสิทธิภาพ

หากการวิจัยนี้ประสบความสำเร็จ จะส่งผลให้ธุรกิจสามารถลดความเสี่ยงจากการคาดการณ์ยอดขายที่ผิดพลาด เกิดโอกาสและเสริมการแข่งขันของธุรกิจ ลดต้นทุนที่ไม่จำเป็น และสร้างความมั่นคงในการเติบโตของธุรกิจได้ในระยะยาว ทั้งนี้ยังเป็นการเสริมสร้างความสามารถในการใช้ข้อมูลเพื่อการตัดสินใจเชิงกลยุทธ์ ซึ่งเป็นทักษะที่สำคัญสำหรับธุรกิจในยุคดิจิทัล
การวิจัยนี้จึงมีความสำคัญต่อการพัฒนาแนวทางการบริหารจัดการธุรกิจอย่างมีประสิทธิภาพ และสร้างความมั่นใจให้กับธุรกิจในการตอบสนองต่อความเปลี่ยนแปลงของตลาด ซึ่งจะนำไปสู่การเพิ่มขีดความสามารถในการแข่งขันในยุคปัจจุบันอย่างแท้จริง

1.4 ขอบเขตของการวิจัย

1.4.1 ประชากรที่ใช้ในการวิจัย

ประชากรที่ใช้ในการวิจัยนี้คือข้อมูลยอดขายสินค้ารายชั่วโมงของบริษัทหรือธุรกิจที่เกี่ยวข้องกับการจัดจำหน่ายสินค้า ซึ่งมีการจัดเก็บข้อมูลยอดขายในแต่ละชั่วโมงของแต่ละวัน เป็นระยะเวลาไม่น้อยกว่า 1 ปี เพื่อให้ได้ข้อมูลที่ครอบคลุมและสามารถนำมาใช้ในการพยากรณ์ยอดขายได้อย่างมีประสิทธิภาพ

1.4.2 กลุ่มตัวอย่างที่ใช้ในการวิจัย

กลุ่มตัวอย่างที่นำมาใช้ในการวิจัยคือข้อมูลยอดขายรายชั่วโมงที่ได้จากฐานข้อมูลของธุรกิจ โดยคัดเลือกข้อมูลที่ครอบคลุมระยะเวลาการขายตั้งแต่ช่วง 8:00 น. จนถึง 17:30 น. ในแต่ละวัน โดยข้อมูลที่ใช้จะครอบคลุมข้อมูลยอดขายในช่วงวันธรรมดาและวันหยุด เพื่อให้ได้กลุ่มข้อมูลที่หลากหลายและสามารถนำไปใช้ในการพยากรณ์ยอดขายได้อย่างแม่นยำ และเลือกเฉพาะ

คลังสินค้าขนาดใหญ่ที่อยู่ในภูมิภาคกรุงเทพ 1 คลังสินค้า เนื่องจากเป็นคลังสินค้าขนาดใหญ่และมีปัญหามากที่สุด

1.4.3 ตัวแปรที่ศึกษา

ตัวแปรที่นำมาใช้ในการวิจัยคือข้อมูลยอดขายรายชั่วโมงที่ได้จากฐานข้อมูลของธุรกิจ โดยแบ่งรายละเอียดของตัวแปรดังนี้

1.4.3.1 ตัวแปรต้น มีรายละเอียดดังนี้

ข้อมูลยอดขายที่เข้ามาในแต่ละชั่วโมง ตั้งแต่เวลา 8:00 น. ถึง 14:00 น. ของแต่ละวัน ซึ่งจะถูกใช้เป็นปัจจัยนำเข้าของแบบจำลองการพยากรณ์

1.4.3.2 ตัวแปรตาม มีรายละเอียดดังนี้

ยอดขายรวมที่เวลา 17:30 น. ซึ่งจะเป็นตัวแปรที่เราต้องการพยากรณ์ในงานวิจัยนี้

1.4.3.3 ตัวแปรร่วม มีรายละเอียดดังนี้

ปัจจัยที่อาจมีผลต่อยอดขาย เช่น วันหยุดนักขัตฤกษ์, ช่วงวันทำงานในสัปดาห์สุดท้ายของเดือนซึ่งเป็นช่วงเร่งปิดยอดขายประจำเดือน, วันก่อนวันหยุดนักขัตฤกษ์ที่ร้านค้าอาจมีการสต็อกสินค้าไว้รองรับนักท่องเที่ยว หรือวันหลังวันหยุดนักขัตฤกษ์ที่สินค้าอาจจะหมดเร็วจากการบริการนักท่องเที่ยงในช่วงวันหยุด เป็นต้น

1.5 นิยามศัพท์เฉพาะ

1.5.1 การพยากรณ์ยอดขาย (Sales Forecasting) ณ ช่วงเวลา 17:00 น.

หมายถึง กระบวนการคาดการณ์ยอดขายที่จะเกิดขึ้นในอนาคตโดยใช้ข้อมูลยอดขายที่ลูกค้าสั่งซื้อเข้ามาในระบบตั้งแต่ช่วงเช้าจนถึงเวลา 17:30 น.

1.5.2 ยอดขายรายชั่วโมง

หมายถึง ยอดขายที่ได้รับจากคำสั่งซื้อของลูกค้าตามชั่วโมง เช่น ยอดขายชั่วโมง ที่ 7 (Hour 7) คือยอดขายที่เข้ามาในระบบตั้งแต่ช่วงเวลา 07:00 น. – 07:59 น.

1.5.3 วันหยุดนักขัตฤกษ์ (Holidays)

หมายถึง วันหยุดนักขัตฤกษ์ตามประกาศราชการ ภายใต้สมติฐานว่าคนอาจไปเที่ยวแล้วส่งผลให้สินค้าไม่เพียงพอ จึงต้องมีการสั่งสินค้ามากกว่าปกติ

1.5.4 วันก่อนวันหยุดนักขัตฤกษ์ (Before Holiday)

หมายถึง วันที่ก่อนที่จะเป็นวันหยุดนักขัตฤกษ์ 1 วัน ภายใต้สมติฐานว่าอาจมีผลต่อยอดขาย เนื่องจากอาจมีการกักตุนสินค้าไว้ขายในวันหยุด

1.5.5 วันหลังวันหยุดนักขัตฤกษ์ (After Holiday)

หมายถึง วันที่หลังวันหยุดนักขัตฤกษ์ 1 วัน ภายใต้สมติฐานว่าอาจมีผลต่อยอดขาย เนื่องจากสินค้าอาจขายดีในช่วงวันหยุดทำให้มีความต้องการมากกว่าปกติ

1.5.6 สัปดาห์สุดท้าย (Last Week)

หมายถึง วันที่ในช่วงสัปดาห์สุดท้ายของแต่ละเดือน ภายใต้สมติฐานว่า เป็นช่วงที่หน่วยงานขายเร่งปิดยอดขายให้ได้ตามเป้าหมายเดือน อาจมีผลให้ยอดขายสูงกว่าปกติ

1.5.7 ยอดขายช่วงเวลา 8AM – 12:PM น.

หมายถึง ยอดขายรวมที่เกิดขึ้นในระหว่างช่วงเวลา 08:00 – 12:59 น.

1.5.8 อัตราส่วนระหว่างยอดขายช่วงเช้ากับช่วงเที่ยง (Morning_Midday_Ratio)

หมายถึง การนำยอดขายในช่วงเช้า คือ 8:00 น. – 10:59 น. มาหาผลรวมกัน แล้วหารด้วยผลรวมของยอดขายในช่วงเที่ยง คือ ยอดขายที่เวลา 11:00 น. – 13:59 น.

1.59 ความแปรปรวนของยอดขาย (Hour_SD)

หมายถึง การคำนวณความแปรปรวนของยอดขายในแต่ละชั่วโมงของแต่ละวันออกมา เพื่อดูว่าความแปรปรวนที่เกิดขึ้นจะมีผลต่อเป้าหมายที่จะพยากรณ์หรือไม่

1.6 กรอบแนวคิดงานวิจัย

งานวิจัยนี้เน้นศึกษาการพยากรณ์ยอดขายรายวัน ณ เวลา 17:30 น. โดยใช้ข้อมูลยอดขายที่เก็บรวบรวมในแต่ละชั่วโมงตั้งแต่เวลา 8:00 น. ถึง 14:00 น. รวมถึงปัจจัยอื่นๆที่เกี่ยวข้องต่าง ๆ ด้วยการใช้เทคนิคการวิเคราะห์ข้อมูลที่ทันสมัยอย่าง Machine Learning เพื่อให้ผลลัพธ์ที่ได้มีความแม่นยำมากที่สุด โดยมีองค์ประกอบหลักของกรอบแนวคิด ดังนี้

1.6.1 ข้อมูลนำเข้า (Input Data)

1.6.1.1 ข้อมูลยอดขายรายชั่วโมง: ข้อมูลยอดขายสะสมของแต่ละชั่วโมงตั้งแต่ 8:00 น. ถึง 14:00 น.

1.6.1.2 ข้อมูลยอดขายสะสมช่วงเวลา 8:00 น. – 12:59 น. ซึ่งเป็นช่วงที่ยอดขายเข้ามาในระบบประมาณ 50%แล้ว อาจสามารถประเมินผลยอดรวมในช่วง 17:30 น. ได้

1.6.1.3 ปัจจัยที่อาจมีผลต่อยอดขาย เช่น วันหยุดนักขัตฤกษ์, ช่วงวันทำงานในสัปดาห์สุดท้ายของเดือนซึ่งเป็นช่วงเร่งปิดยอดขายประจำเดือน, วันก่อนวันหยุดนักขัตฤกษ์ที่ร้านค้าอาจมีการสต็อกสินค้าไว้รองรับนักท่องเที่ยว หรือวันหลังวันหยุดนักขัตฤกษ์ที่สินค้าอาจจะหมดเร็วจากการบริการนักท่องเที่ยวในช่วงวันหยุด

ข้อมูลเหล่านี้จะถูกนำมาวิเคราะห์เพื่อหาปัจจัยสำคัญที่จะมีผลต่อยอดขาย ณ เวลา 17:30 น.

1.6.2 การประมวลผลข้อมูล (Data Processing)

1.6.2.1 ทำความสะอาดข้อมูล (Data Cleaning): จัดการกับข้อมูลที่หายไป (Missing Data), ข้อมูลผิดพลาด (Outliers) และปรับรูปแบบข้อมูลให้อยู่ในลักษณะที่พร้อมสำหรับการสร้างแบบจำลอง

1.6.2.2 การแปลงข้อมูล (Data Transformation): ปรับรูปแบบข้อมูลให้อยู่ในมาตรฐานเดียวกัน เช่น การแปลงข้อมูลให้เป็นตัวเลข (Numeric Transformation) หรือการใช้การแปลงข้อมูลแบบ One-Hot Encoding

1.6.2.3 การเลือกคุณลักษณะ (Feature Selection): เลือกข้อมูลที่สำคัญที่มีผลต่อการพยากรณ์ เช่น ยอดขายในชั่วโมงก่อนหน้า ปัจจัยวันและเวลา

1.6.3 การสร้างแบบจำลองการพยากรณ์ (Model Building) ทำการเลือกแบบจำลองที่เหมาะสมสำหรับการพยากรณ์ยอดขาย

1.6.3.1 Random Forest ใช้การรวมหลาย Decision Trees เพื่อเพิ่มความแม่นยำ ป้องกัน overfitting ใช้ได้กับข้อมูลที่ไม่เป็นเชิงเส้น

1.6.3.2 XGBoost แบบจำลอง Gradient Boosting ที่มีประสิทธิภาพสูง ใช้สำหรับข้อมูลที่ซับซ้อนและต้องการการทำนายที่แม่นยำ

1.6.3.3 Gradient Boosting การรวมหลาย Decision Trees ที่เรียนรู้จากข้อผิดพลาดของแบบจำลองก่อนหน้า ใช้ได้ดีในข้อมูลที่ไม่เป็นเชิงเส้น

1.6.3.4 Linear Regression ใช้สำหรับข้อมูลที่มีความสัมพันธ์เชิงเส้นตรงระหว่างตัวแปรต้นและตัวแปรตาม

1.6.3.5 Ridge Regression ปรับ Linear Regression โดยเพิ่มค่าลงโทษ (Penalty term)เพื่อป้องกัน overfitting เหมาะกับข้อมูลที่มี multicollinearity

1.6.3.6 Lasso Regression แบบจำลองเชิงเส้นที่เพิ่มค่าลงโทษ L1 ทำให้เลือกตัวแปรอัตโนมัติโดยลดค่าสัมประสิทธิ์บางตัวเป็นศูนย์

1.6.3.7 Elastic Net Regression ผสมระหว่าง L1 และ L2 ช่วยทั้งป้องกัน overfitting และเลือกตัวแปร เหมาะกับข้อมูลที่มีตัวแปรจำนวนมากและมีความสัมพันธ์กัน

1.6.4 การฝึกฝนและทดสอบแบบจำลอง (Training & Testing)

1.6.4.1 แบ่งข้อมูลเป็นชุดการฝึก (Training set) และชุดการทดสอบ (Test set) โดยอาจแบ่งตามสัดส่วน 80:20 เพื่อให้แบบจำลองสามารถเรียนรู้จากข้อมูลและทดสอบความแม่นยำได้

1.6.4.2 ปรับพารามิเตอร์ของแบบจำลอง (Hyperparameter Tuning) เพื่อเพิ่มประสิทธิภาพในการพยากรณ์

1.6.5 การประเมินประสิทธิภาพของแบบจำลอง (Model Evaluation)

1.6.5.1 ประเมินผลด้วยการวัดความแม่นยำ เช่น ค่า Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), r2 เพื่อเปรียบเทียบประสิทธิภาพของแต่ละแบบจำลอง ว่าแบบจำลองใดมีความแม่นยำและประสิทธิภาพดีที่สุด

1.6.5.2 วิเคราะห์ผลการทำนายและประเมินว่าผลลัพธ์สามารถนำไปใช้ได้จริงหรือไม่ และเปรียบเทียบผลการทำนายของแบบจำลองต่าง ๆ เพื่อเลือกแบบจำลองที่เหมาะสมที่สุดสำหรับการพยากรณ์ยอดขายรวมที่ช่วงเวลา 17:30 น.

1.6.6 การนำผลลัพธ์ไปใช้ (Implementation) นำแบบจำลองที่ผ่านการประเมินแล้วมาใช้ในการทำนายยอดขายรายวัน เพื่อช่วยในการวางแผนการสั่งซื้อสินค้า จัดการสต็อก และวางแผนกลยุทธ์การขาย

1.7 **สมมติฐานในการวิจัย**

1.7.1 ข้อมูลยอดขายรายชั่วโมงที่เข้ามาตั้งแต่เวลา 8:00 น. ถึง 14:00 น. สามารถใช้ในการพยากรณ์ยอดขาย ณ เวลา 17:30 น. ได้อย่างแม่นยำ โดยผ่านการวิเคราะห์และสร้างแบบจำลองด้วยเทคนิค Machine Learning

1.7.2 การใช้เทคนิค Machine Learning (เช่น XGBoost, Linear Regression) จะให้ผลลัพธ์การพยากรณ์ยอดขายที่แม่นยำกว่าเทคนิคทางสถิติแบบดั้งเดิม (เช่น Moving Average)

1.7.3 ปัจจัยต่าง ๆ เช่น วันหยุดนักขัตฤกษ์, ช่วงวันทำงานในสัปดาห์สุดท้ายของเดือนซึ่งเป็นช่วงเร่งปิดยอดขายประจำเดือน, วันก่อนวันหยุดนักขัตฤกษ์ที่ร้านค้าอาจมีการสต็อกสินค้าไว้รองรับนักท่องเที่ยว หรือวันหลังวันหยุดนักขัตฤกษ์ที่สินค้าอาจจะหมดเร็วจากการบริการนักท่องเที่ยงในช่วงวันหยุด ซึ่งอาจมีผลต่อยอดขาย และสามารถนำไปเพิ่มความแม่นยำให้กับแบบจำลองพยากรณ์ได้

บทที่ 2
ทบทวนวรรณกรรม

การวิจัยนี้ ผู้วิจัยได้ศึกษาเอกสารและงานวิจัยที่เกี่ยวข้องในการทำวิจัย โดยได้นำเสนอตามหัวข้อดังต่อไปนี้

1.การพยากรณ์ยอดขาย (Sale Forecast)

2.แบบจำลองที่ใช้ในงานวิจัยและการประเมินผล

3.งานวิจัยที่เกี่ยวข้อง

2.1 การพยากรณ์ยอดขาย (Sale Forecast)

การพยากรณ์ยอดขาย (Sales Forecasting) เป็นกระบวนการที่สำคัญในการวางแผนและการจัดการธุรกิจ โดยเฉพาะในอุตสาหกรรมที่มีการแข่งขันสูง เช่น อุตสาหกรรมเครื่องดื่ม การพยากรณ์ที่แม่นยำช่วยให้ธุรกิจสามารถตัดสินใจได้อย่างมีข้อมูล รองรับการผลิต การจัดการสินค้าคงคลัง การบริหารจัดการทรัพยากรมนุษย์และที่สำคัญยังช่วยทำให้สามารถวางแผนเตรียมการเรื่องการจัดการการขนส่งได้อย่างมีประสิทธิภาพมากขึ้น ในบทนี้จะพูดถึงหลักการและเทคนิคที่ใช้ในการพยากรณ์ยอดขาย รวมถึงแนวโน้มล่าสุดในเทคโนโลยีการวิเคราะห์ข้อมูล

2.1.1 ความสำคัญของการพยากรณ์ยอดขาย

การพยากรณ์ยอดขายมีความสำคัญต่อธุรกิจในหลายด้าน เช่น

2.1.1.1 การวางแผนขนส่งหรือการวางแผนกระจายสินค้า จะช่วยให้ธุรกิจสามารถต่อสนองต่อความต้องการของลูกค้าได้ตรงเวลา และทันถ่วงที เพื่อโอกาสในการขายสินค้าได้ดียิ่งขึ้น

2.1.1.2 การจัดการสินค้าคงคลัง จะช่วยลดความเสี่ยงจากการมีสินค้าคงคลังมากเกินไปหรือน้อยเกินไป

2.1.1.3 การจัดการทรัพยากรมนุษย์ จะช่วยในการวางแผนกำลังคนให้เหมาะสมกับความต้องการ

2.1.1.4 การปรับกลยุทธ์การตลาด จะช่วยให้ธุรกิจสามารถออกโปรโมชั่นและการตลาดได้อย่างเหมาะสม

2.1.1.5 การวางแผนการผลิต จะช่วยให้ธุรกิจสามารถผลิตสินค้าได้ตรงตามความต้องการของตลาด

2.1.2 เทคนิคการพยากรณ์ยอดขาย

การพยากรณ์ยอดขายจะใช้เทคนิคการเรียนรู้ของเครื่อง (Machine Learning) เช่น Random Forest, XGBoost, Gradient Boosting, Linear Regression ที่ใช้ในการสร้างแบบจำลองที่สามารถเรียนรู้จากข้อมูลที่มีความซับซ้อนสูง โดยเฉพาะข้อมูลที่มีหลายปัจจัยที่มีผลต่อยอดขาย

2.1.3 การใช้ข้อมูลในการพยากรณ์

การพยากรณ์ยอดขายมักจะใช้ข้อมูลจากหลายแหล่ง เช่น

2.1.3.1 ข้อมูลยอดขายในอดีต เป็นการใช้ข้อมูลการขายจากช่วงเวลาที่ผ่านมาเพื่อวิเคราะห์แนวโน้ม

2.1.3.2 ข้อมูลโปรโมชั่น เป็นการศึกษาผลกระทบของโปรโมชั่นต่อยอดขาย

2.1.3.3 ข้อมูลเชิงเศรษฐกิจ เช่น ราคาน้ำมัน อัตราแลกเปลี่ยนเงินตรา และข้อมูลทางเศรษฐกิจอื่นๆ ที่อาจส่งผลต่อความต้องการสินค้า

2.1.3.4 ข้อมูลสภาพอากาศ ซึ่งอาจมีผลกระทบต่อการขายในอุตสาหกรรมบางประเภท เช่น เครื่องดื่มที่ใช้ในฤดูร้อน

การพยากรณ์ยอดขายเป็นเครื่องมือสำคัญในการตัดสินใจทางธุรกิจที่ช่วยให้ธุรกิจสามารถเติบโตและปรับตัวเข้ากับการเปลี่ยนแปลงในตลาด การใช้เทคนิคที่หลากหลายทั้งเชิงสถิติและการเรียนรู้ของเครื่องสามารถเพิ่มความแม่นยำในการพยากรณ์ได้อย่างมาก ซึ่งจะส่งผลดีต่อการวางแผนและการดำเนินงานของธุรกิจในอนาคต ซึ่งในการทำวิจัยครั้งนี้ผู้วิจัยได้ใช้ข้อมูลยอดขายที่มีการเก็บเป็นฐานข้อมูลรายชั่วโมงแยกตามแต่ละคลังสินค้ามาใช้เป็นข้อมูลในการพยากรณ์

2.2 แบบจำลองที่ใช้ในงานวิจัยและการประเมินผล

ในงานวิจัยนี้ได้ใช้แบบจำลองหลายประเภทเพื่อพยากรณ์ยอดขาย โดยเฉพาะ ซึ่งมีลักษณะและการทำงานที่แตกต่างกัน ดังนี้

2.2.1 Random Forest

เป็นหนึ่งในเทคนิคการเรียนรู้ของเครื่องที่ใช้การสร้างชุดของต้นไม้การตัดสินใจ (decision trees) หลายๆ ต้นแล้วนำมารวมผลเพื่อให้ได้ผลลัพธ์ที่ดีที่สุดในการทำนายค่าอย่างต่อเนื่อง เช่น ในปัญหาการทำนายค่าของตัวแปรที่เป็นจำนวน (regression problem) โดย Random Forest จะช่วยลดความผิดพลาดที่เกิดจากการใช้ต้นไม้การตัดสินใจเพียงต้นเดียว โดยการพิจารณาผลลัพธ์จากต้นไม้การตัดสินใจหลายๆ ต้นและเฉลี่ยผลลัพธ์จากต้นไม้ทั้งหมด (ensemble learning)

สมการหลักสำหรับ Random Forest จะประกอบด้วยการทำนายจากหลายๆ ต้นไม้ (decision trees) และค่าเฉลี่ยของผลลัพธ์จากแต่ละต้นไม้จะเป็นผลลัพธ์สุดท้ายของแบบจำลอง

$$\hat{y} = \frac{1}{M}\sum_{m=1}^{M} f_m(x)$$

โดยที่

$\hat{y}$ คือ ผลลัพธ์ที่ทำนาย

M คือ จำนวนต้นไม้ในป่า (จำนวนต้นไม้ทั้งหมด)

$f_m(x)$ คือ ผลลัพธ์ที่ได้จากต้นไม้ที่ m เมื่อใช้ข้อมูล x เป็น input

กระบวนการทำงานของ Random Forest

1. การสร้างต้นไม้: แบบจำลองจะสร้างต้นไม้การตัดสินใจหลายๆ ต้น (เช่น 100 ต้น, 200 ต้น, หรือ 500 ต้น ตามที่กำหนด) โดยแต่ละต้นไม้จะถูกฝึกจากชุดข้อมูลที่สุ่มเลือก (Bootstrapping)

2. การเลือกคุณลักษณะ (Feature selection): ในแต่ละการแบ่ง (split) ของแต่ละต้นไม้ แบบจำลองจะเลือกคุณลักษณะบางตัวจากชุดของคุณลักษณะทั้งหมดโดยไม่ต้องใช้ทั้งหมดที่มี ซึ่งจะช่วยลดความซับซ้อนและเพิ่มความแม่นยำในการทำนาย

3. การรวมผล: หลังจากที่ฝึกต้นไม้เสร็จ แบบจำลองจะทำการรวมผลลัพธ์จากทุกต้นไม้ โดยใช้การเฉลี่ยผลลัพธ์จากแต่ละต้นไม้ (ในกรณีของ regression)

ข้อดีของ Random Forest

1. มีความสามารถในการลด overfitting เนื่องจากใช้หลายๆ ต้นไม้และไม่พึ่งพาต้นไม้เพียงต้นเดียว

2. สามารถจัดการกับข้อมูลที่มีหลายมิติและข้อมูลที่สูญหายได้ดี

3. สามารถใช้ได้กับทั้งปัญหาการทำนายค่าตัวแปรต่อเนื่อง (regression) และการทำนายค่าตัวแปรจำแนก (classification)

2.2.2 XGBoost (Extreme Gradient Boosting)

เป็นหนึ่งในแบบจำลองการเรียนรู้ของเครื่องที่ใช้เทคนิคการเสริมกำลัง (boosting) โดยหลักการของ XGBoost คือการใช้ชุดของต้นไม้การตัดสินใจ (decision trees) ที่ถูกฝึกมาแบบซ้ำๆ เพื่อแก้ไขข้อผิดพลาดจากแบบจำลองก่อนหน้านี้ ซึ่งช่วยเพิ่มประสิทธิภาพในการทำนาย โดย

XGBoost ใช้กรอบการทำงานแบบ gradient boosting ที่เน้นการปรับปรุงแบบจำลองทีละขั้นตอน เพื่อลดความผิดพลาดในแต่ละรอบ

สมการหลักของ XGBoost ในการทำนายผล (prediction) คือ

$$\hat{y}_i = \sum_{k=1}^{K} \alpha_k f_k(x_i)$$

โดยที่

$\hat{y}_i$	คือ ค่าที่ทำนายสำหรับตัวอย่างที่ i
K	คือ จำนวนของต้นไม้ที่ใช้ในแบบจำลอง
α_k	คือ น้ำหนักของต้นไม้ k
$f_k(x_i)$	คือ ผลลัพธ์จากต้นไม้ที่ k สำหรับตัวอย่างที่ i
x_i	คือ ข้อมูลอินพุตสำหรับตัวอย่างที่ i

การทำงานของ XGBoost

1. Boosting: XGBoost ใช้การเรียนรู้แบบ Boosting ซึ่งหมายถึงการฝึกแบบจำลองหลายๆ ตัวอย่างในลำดับที่สอดคล้องกัน และในแต่ละรอบจะเรียนรู้จากข้อผิดพลาดที่เกิดขึ้นจากแบบจำลองในรอบก่อนหน้า

2. Gradient Boosting การอัพเดตค่าในแต่ละรอบจะทำโดยการใช้ Gradient Descent ในการปรับปรุงค่า prediction ของต้นไม้การตัดสินใจที่มีอยู่

3. Regularization XGBoost ใช้เทคนิค Regularization (L1 และ L2) เพื่อป้องกันไม่ให้แบบจำลอง overfitting และเพิ่มความแม่นยำโดยการควบคุมค่าพารามิเตอร์ต่างๆ

ข้อดีของ XGBoost

1. ประสิทธิภาพสูง สามารถจัดการกับข้อมูลขนาดใหญ่และทำงานได้ดีในการแข่งขันด้านการทำนาย

2. ปรับแต่งได้ง่าย มีตัวเลือกในการปรับแต่งพารามิเตอร์จำนวนมาก เช่น การ regularize แบบจำลองเพื่อหลีกเลี่ยง overfitting

3. การจัดการกับข้อมูลที่หายไป XGBoost สามารถจัดการกับข้อมูลที่หายไปได้โดยไม่จำเป็นต้องทำการแทนค่าก่อน

2.2.3 Gradient Boosting

เป็นหนึ่งในเทคนิคการเรียนรู้ของเครื่องที่มีประสิทธิภาพสูงในการแก้ไขปัญหาการทำนายตัวเลขหรือการจัดกลุ่ม โดยจะใช้ต้นไม้การตัดสินใจ (decision trees) ที่เป็นต้นไม้ตื้นๆ (shallow

trees) หลายๆ ต้นในการฝึกเรียนรู้ทีละรอบ เรียกว่า boosting ซึ่งการเรียนรู้แต่ละรอบจะใช้ข้อผิดพลาดจากรอบก่อนหน้าในการปรับปรุงแบบจำลอง โดยใช้กรอบการทำงานแบบ gradient boosting

สมการของการทำนายค่าผลลัพธ์ (prediction) ของ Gradient Boosting สามารถเขียนได้ดังนี้

$$\widehat{y_i} = \sum_{k=1}^{K} f_k(x_i)$$

โดยที่

$\widehat{y_i}$ คือ ค่าที่ทำนายสำหรับตัวอย่างที่ i

K คือ จำนวนของรอบในการฝึก (หรือจำนวนต้นไม้ที่ใช้ในแบบจำลอง)

$f_k(x_i)$ คือ ผลลัพธ์จากต้นไม้ที่ k สำหรับตัวอย่างที่ i

x_i คือ ข้อมูลอินพุตของตัวอย่างที่ i

การทำงานของ Gradient Boosting

1. Gradient Boosting ใช้เทคนิค Boosting โดยการฝึกแบบจำลองหลายๆ ตัวในลำดับที่สอดคล้องกัน โดยจะทำการฝึกแบบจำลองใหม่ในแต่ละรอบที่สามารถแก้ไขข้อผิดพลาดที่เกิดจากการทำนายในรอบก่อนหน้า

2. Gradient Descent: การปรับปรุงแบบจำลองในแต่ละรอบจะทำโดยการใช้ Gradient Descent เพื่อปรับค่าของแบบจำลองในแต่ละรอบ

3. การใช้หลายๆ ต้นไม้การตัดสินใจ (Decision Trees)โดย Gradient Boosting ใช้ต้นไม้การตัดสินใจหลายๆ ต้น ซึ่งทุกต้นจะถูกฝึกมาเพื่อลดข้อผิดพลาดของแบบจำลอง

4. การ Regularize เพื่อไม่ให้แบบจำลองเกิดการ overfitting, Gradient Boosting ใช้การ regularization ผ่านทางเทคนิคต่างๆ เช่น การเพิ่มค่า learning rate หรือ max depth เพื่อควบคุมความลึกของต้นไม้

สมการสำหรับการอัพเดต (Update Equation) : การอัพเดตในแต่ละรอบของ Gradient Boosting สามารถเขียนได้เป็นสมการดังนี้

$$f_k(x_i) = -\eta \cdot \nabla_{f_{k-1}} L\big(y_i, f_{k-1}(x_i)\big)$$

โดยที่

η คือ Learning rate ที่ใช้ในการอัพเดต (มักจะมีค่าอยู่ในช่วง 0 ถึง 1)

$\nabla_{f_{k-1}}$ คือ Gradientหรืออนุพันธ์ของฟังก์ชันการสูญเสีย (Loss function)ในรอบก่อนหน้า

$L\left(y_i, f_{k-1}(x_i)\right)$ คือ ฟังก์ชันการสูญเสียของการทำนายจากแบบจำลองในรอบก่อนหน้า

ข้อดีของ Gradient Boosting

1. ประสิทธิภาพสูง สามารถทำงานได้ดีในหลากหลายประเภทของข้อมูล และมีความแม่นยำสูงในการทำนาย

2. สามารถจัดการกับข้อมูลที่หายไปได้ โดย Gradient Boosting สามารถจัดการกับข้อมูลที่หายไปโดยไม่จำเป็นต้องทำการแทนค่าก่อน

3. ปรับแต่งได้ง่าย โดยแบบจำลองสามารถปรับแต่งพารามิเตอร์ต่างๆ เช่น learning rate, n_estimators (จำนวนรอบในการฝึก), max_depth เพื่อให้ได้ผลลัพธ์ที่ดีที่สุด

2.2.4 Linear Regression

เป็นแบบจำลองทางคณิตศาสตร์ที่ใช้สำหรับการพยากรณ์ที่มีความสัมพันธ์เชิงเส้นตรง (Linear Relationship) ระหว่างตัวแปรต้นและตัวแปรตาม แบบจำลองนี้เหมาะสำหรับข้อมูลที่มีความสัมพันธ์แบบเชิงเส้นและสามารถใช้งานได้ง่าย เนื่องจากมีการคำนวณที่เร็วและเข้าใจง่าย แต่ไม่เหมาะสำหรับข้อมูลที่มีความซับซ้อนหรือไม่เป็นเชิงเส้น

2.2.4.1 Simple Linear Regression

Simple Linear Regression (SLR) เป็นแบบจำลองทางสถิติที่ใช้ในการทำนายค่า (predict) ผลลัพธ์จากตัวแปรอิสระ (independent variable หรือ feature) โดยเชื่อมโยงตัวแปรทั้งสองด้วยสมการเชิงเส้น (linear equation). แบบจำลองนี้ถือเป็นหนึ่งในเทคนิคที่ง่ายที่สุดในการทำการทำนายค่าผลลัพธ์จากชุดข้อมูล

สมการของ Simple Linear Regression สามารถเขียนได้ในรูปของสมการเชิงเส้นดังนี้

$$y = \beta_0 + \beta_1 x + \epsilon$$

โดยที่

y คือ ค่าที่ทำนาย (target variable)

β_0 คือ ค่า intercept หรือค่าคงที่ (bias) ของเส้น

β_1 คือ ค่าความชัน(slope) ของเส้น หรือการเปลี่ยนแปลงของ y เมื่อ x เปลี่ยนแปลง

x คือ ตัวแปรอิสระ (independent variable)

$\in$ คือ ความคลาดเคลื่อนหรือข้อผิดพลาดของแบบจำลอง (error term)

การอธิบายสมการ

β_0 คือค่าที่เส้นตรงจะตัดแกน y ที่ค่าของ x=0 ค่านี้บ่งชี้ค่าคงที่ของผลลัพธ์เมื่อไม่มีการเปลี่ยนแปลงในตัวแปรอิสระ

β_1 คือ เป็นตัวแทนของความสัมพันธ์ระหว่าง x และ y กล่าวคือ ถ้า β_1 เป็นบวก y จะเพิ่มขึ้นเมื่อ xเพิ่มขึ้น หาก β_1 เป็นลบ y จะลดลงเมื่อ x เพิ่มขึ้น

การทำงานของ Simple Linear Regression

1. การปรับค่าพารามิเตอร์: ในการฝึกแบบจำลอง Simple Linear Regression แบบจำลองจะพยายามหาค่า β_0 และ β_1 ที่เหมาะสมที่สุดเพื่อให้ค่า MSE (Mean Squared Error) ต่ำสุด ซึ่งจะช่วยให้แบบจำลองสามารถทำนายค่า y ได้ดีที่สุดจากค่าของ x

2. แบบจำลองนี้เหมาะสำหรับการวิเคราะห์ความสัมพันธ์ระหว่างตัวแปรสองตัวที่สามารถแสดงได้ด้วยเส้นตรง

การคำนวณค่าพารามิเตอร์ β_0 และ β_1 คำนวณได้จากสูตร ดังนี้

$$\beta_1 = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2}$$

$$\beta_0 = \bar{y} - \beta_1 \bar{x}$$

โดยที่

$\bar{x}$ และ $\bar{y}$ คือค่าเฉลี่ยของ x และ y ตามลำดับ

ข้อดีของ Simple Linear Regression

1. เข้าใจง่าย: เป็นแบบจำลองที่ง่ายที่สุดและเข้าใจได้ง่าย โดยการแสดงความสัมพันธ์ระหว่างตัวแปรอิสระและตัวแปรที่ต้องการทำนาย

2. คำนวณเร็ว: ด้วยการคำนวณที่ไม่ซับซ้อนทำให้แบบจำลองนี้สามารถฝึกฝนได้เร็ว

2.2.4.2 Polynomial Regression

เป็นการขยาย Simple Linear Regression โดยเพิ่มพหุนาม (Polynomial Features) เข้าไปในตัวแปรอิสระ (Independent Variable) เพื่อให้สามารถจับความสัมพันธ์ที่ไม่เป็นเส้นตรงได้ เช่น ข้อมูลที่มีแนวโน้มเป็นโค้ง (Curve) หรือมีลักษณะสูงขึ้นและลดลง

ความแตกต่างระหว่าง Linear Regression และ Polynomial Regression

Linear Regression: ใช้สมการเส้นตรง

$$y = \beta_0 + \beta_1 x + \in$$

Polynomial Regression: ใช้สมการพหุนาม เช่น สำหรับ degree = 2

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \in$$

หรือถ้าเป็น degree = 3

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \in$$

โดยที่

$\beta_0, \beta_1, \beta_2, \beta_3$ เป็นค่าพารามิเตอร์ของแบบจำลองที่ต้องเรียนรู้

ข้อดีของ Polynomial Regression

1. สามารถจับความสัมพันธ์ที่ไม่เป็นเส้นตรงได้ดี
2. ยืดหยุ่นกว่า Linear Regression ปกติ

2.2.4.3 Ridge Regression

เป็นการใช้เทคนิค Regularization เพื่อควบคุมค่าของค่าสัมประสิทธิ์ (coefficients) ในแบบจำลองการถดถอยเชิงเส้น (Linear Regression) โดยการเพิ่มค่าปรับลงในฟังก์ชันค่าความสูญเสีย (loss function) ซึ่งช่วยลดปัญหาการ overfitting เมื่อมีจำนวนฟีเจอร์ (features) มากๆ

สมการของ Ridge Regression

$$\hat{y} = X\beta + \in$$

โดยที่:

$\hat{y}$	คือ ค่าที่ทำนาย
X	คือ เมทริกซ์ของฟีเจอร์
β	คือ เวกเตอร์ของค่าสัมประสิทธิ์
$\in$	คือ ค่าผิดพลาด (error)

ฟังก์ชันค่าเสียหาย (Loss function) ใน Ridge Regression คือ

$$L(\beta) = \sum_{i=1}^{n} (y_i - x_i\beta)^2 + \alpha \sum_{j=1}^{p} \beta_j^2$$

โดยที่

y_i คือ ค่าจริงของข้อมูลที่ i=1

x_i คือ เวกเตอร์ของฟีเจอร์ที่ i=1

β_j คือ ค่าสัมประสิทธิ์ที่ต้องการหาค่าที่เหมาะสม

α คือ ค่าปรับ (regularization parameter) ที่ควบคุมความแรงของการ regularization โดยปกติ α จะมีค่ามากกว่า 0

Ridge Regression เป็นการปรับค่าของ β ใน Linear Regression โดยการเพิ่มเทอม $\sum_{j=1}^{p} \beta_j^2$ เข้าไปในฟังก์ชันค่าเสียหาย เพื่อควบคุมค่าของค่าสัมประสิทธิ์ β และช่วยลดการ overfitting ซึ่งเกิดขึ้นเมื่อแบบจำลองเรียนรู้ข้อมูลจนมากเกินไปและไม่สามารถทำนายข้อมูลใหม่ได้ดี

การเลือกค่า α ที่เหมาะสมจะช่วยให้แบบจำลองสามารถทำนายได้ดีทั้งในชุดข้อมูลฝึกและชุดข้อมูลทดสอบ โดย α ที่ใหญ่จะทำให้การ regularization แข็งแกร่งขึ้นและค่าค่าสัมประสิทธิ์ β จะมีค่าน้อยลง

ข้อดีของ Ridge Regression

1. ลดปัญหาการ overfitting: Ridge Regression ช่วยควบคุมการเรียนรู้ของแบบจำลองโดยการลดขนาดของค่าสัมประสิทธิ์ β ซึ่งช่วยป้องกันไม่ให้แบบจำลองเรียนรู้ความผิดพลาดจากข้อมูลฝึกที่ไม่ใช่รูปแบบทั่วไป (noise) หรือข้อมูลที่มีลักษณะเฉพาะ (outliers) ซึ่งมักจะทำให้แบบจำลอง overfit และทำนายได้ไม่ดีเมื่อทดสอบกับข้อมูลใหม่

2. ใช้ได้ดีในกรณีที่มีจำนวนฟีเจอร์มาก: เมื่อมีฟีเจอร์จำนวนมากในชุดข้อมูล (เช่น การวิเคราะห์ข้อมูลจากหลายๆ ตัวแปร) Ridge Regression จะช่วยป้องกันการที่แบบจำลองจะมีค่าสัมประสิทธิ์ที่สูงเกินไป ซึ่งอาจจะส่งผลให้แบบจำลองมีความซับซ้อนเกินความจำเป็น โดยการเพิ่มการปรับ α เข้าไปในสมการทำให้แบบจำลองสามารถจัดการกับข้อมูลที่มีจำนวนฟีเจอร์มากได้ดีขึ้น

3. ง่ายต่อการคำนวณและตีความ: Ridge Regression ทำให้แบบจำลองมีความเรียบง่ายมากขึ้น โดยไม่ต้องเลือกฟีเจอร์อย่างพิถีพิถัน ช่วยให้คำนวณและตีความได้ง่ายกว่าเมื่อเปรียบเทียบกับการทำ feature selection ด้วยวิธีอื่นๆ

2.2.4.4 Ridge Regression

Lasso Regression (Least Absolute Shrinkage and Selection Operator) เป็นการทำ L1 regularization บน Linear Regression ซึ่งเป็นวิธีการที่ใช้ในการควบคุมความซับซ้อนของแบบจำลองโดยการบีบให้ค่าสัมประสิทธิ์ของบางตัวแปรมีค่าศูนย์ เพื่อให้สามารถเลือกฟีเจอร์ที่สำคัญได้อย่างมีประสิทธิภาพ

สมการของ Lasso Regression

$$\hat{y} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p$$

โดยที่

$\hat{y}$ คือ ค่าพยากรณ์ของแบบจำลอง

β_0 คือ ค่าคงที่ (intercept)

$\beta_1, \beta_2, \cdots, \beta_p$ คือ สัมประสิทธิ์ที่แบบจำลองจะเรียนรู้

$X_1, X_2, \cdots, X_p$ คือ ฟีเจอร์ของข้อมูลที่ใช้ในการพยากรณ์

ใน Lasso Regression, การ regularization หรือการควบคุมการเติบโตของค่าสัมประสิทธิ์จะถูกทำโดยการเพิ่ม penalty term ที่เป็นค่าผลรวมของค่าสัมประสิทธิ์ทั้งหมดที่มีค่าคูณกับค่าคงที่ λ

$$\text{Loss Function} = \sum_{i=1}^{n} (y_i - \hat{y_i})^2 + \lambda \sum_{j=1}^{p} |\beta_j|$$

โดยที่

$\sum_{i=1}^{n} (y_i - \hat{y_i})^2$ คือ Residual Sum of Squares (RSS) หรือความผิดพลาดจากการทำนาย

$\lambda \sum_{j=1}^{p} |\beta_j|$ คือ L1 Regularization Term ซึ่งจะเพิ่มค่าผลรวมของค่าสัมประสิทธิ์ที่มีค่าคูณกับ λ

การทำงานของ Lasso:

L1 Regularization: ช่วยให้บางค่าสัมประสิทธิ์ β_j มีค่าเป็นศูนย์ ทำให้ฟีเจอร์ที่มีค่าสัมประสิทธิ์ศูนย์จะถูกละทิ้งจากแบบจำลอง ซึ่งช่วยในการเลือกฟีเจอร์ที่สำคัญ (feature selection)

การควบคุมความซับซ้อน: ค่าของ λ จะช่วยควบคุมว่าการ regularization จะมีผลมากน้อยแค่ไหน ถ้า λ มีค่ามาก ฟีเจอร์ที่ไม่สำคัญจะถูกลดให้เหลือศูนย์

การใช้กับข้อมูลที่มีหลายฟีเจอร์: Lasso Regression เหมาะสำหรับข้อมูลที่มีฟีเจอร์จำนวนมาก เนื่องจากมันสามารถเลือกฟีเจอร์ที่สำคัญได้และลดความซับซ้อนของแบบจำลอง

ข้อดี ของ Lasso Regression

1.Feature Selection: สามารถเลือกฟีเจอร์ที่สำคัญและละทิ้งฟีเจอร์ที่ไม่สำคัญได้

2.ลด Overfitting: การใช้ L1 Regularization ช่วยลด overfitting โดยการลดขนาดของค่าสัมประสิทธิ์

3.สามารถใช้ได้กับข้อมูลที่มีจำนวนฟีเจอร์มาก: เป็นตัวเลือกที่ดีเมื่อทำงานกับข้อมูลที่มีหลายฟีเจอร์

2.2.4.5 Elastic Net Regression

เป็นการรวมข้อดีของ Ridge Regression (L2 Regularization) และ Lasso Regression (L1 Regularization) เพื่อทำให้สามารถจัดการกับข้อมูลที่มีฟีเจอร์มากหรือมีความสัมพันธ์กันระหว่างฟีเจอร์ได้ดีขึ้น

สมการของ Elastic Net Regression

$$\hat{y} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p$$

โดยที่

$\hat{y}$ คือ ค่าพยากรณ์

β_0 คือ ค่า intercept

$\beta_1, \beta_2, \cdots, \beta_p$ คือ ค่าสัมประสิทธิ์ที่ต้องการเรียนรู้

$X_1, X_2, \cdots, X_{1p}$ คือ ฟีเจอร์ของข้อมูล

สำหรับ Elastic Net, การ regularization ใช้ค่าของทั้ง L1 (Lasso) และ L2 (Ridge) และจะถูกควบคุมด้วยพารามิเตอร์ α และ $i = 1$ ratio

$$\text{Loss Function} = \sum_{i=1}^{n} (y_i - \hat{y_i})^2 + \lambda \left[\alpha \sum_{j=1}^{p} |\beta_j| + \frac{1-\alpha}{2} \sum_{j=1}^{p} \beta_j^2 \right]$$

โดยที่

$\sum_{i=1}^{n} (y_i - \hat{y_i})^2$ คือ Residual Sum of Squares (RSS)

λ คือ ค่าปรับการ regularization ที่ควบคุมความซับซ้อน

$\sum_{j=1}^{p}|\beta_j|$ คือ L1 Regularization (Lasso Term)

$\sum_{j=1}^{p}\beta_j^2$ คือ L2 Regularization (Ridge Term)

α คือ พารามิเตอร์ที่กำหนดอัตราส่วนระหว่าง L1 และ L2 การ regularization

ถ้า α =1, Elastic Net จะกลายเป็น **Lasso Regression**

ถ้า α =0, Elastic Net จะกลายเป็น **Ridge Regression**

ค่าของ α ที่อยู่ระหว่าง 0 และ 1 จะสร้างสมการที่ผสมผสานทั้ง L1 และ L2 regularization

การทำงานของ Elastic Net

1.L1 Regularization (Lasso) ช่วยทำให้บางค่าสัมประสิทธิ์ β_j เป็นศูนย์ ทำให้ฟีเจอร์ที่ไม่สำคัญถูกละทิ้ง

2.L2 Regularization (Ridge) ช่วยควบคุมการเติบโตของค่าสัมประสิทธิ์โดยการทำให้ค่าสัมประสิทธิ์ไม่ใหญ่มากเกินไป

3.การใช้ทั้งสองแบบ ช่วยให้ Elastic Net มีประสิทธิภาพในการทำงานกับข้อมูลที่มีฟีเจอร์หลายตัวที่สัมพันธ์กัน

ข้อดีของ Elastic Net

1.จัดการกับฟีเจอร์ที่มีความสัมพันธ์กัน: Elastic Net สามารถจัดการกับฟีเจอร์ที่มีความสัมพันธ์กันสูงได้ดีกว่า Lasso เพราะจะมีการใช้ L2 regularization ควบคู่ไปกับ L1

2.ทำให้ฟีเจอร์ที่ไม่สำคัญถูกคัดออก: Elastic Net สามารถเลือกฟีเจอร์ที่สำคัญได้เหมือน Lasso โดยการทำให้บางค่าสัมประสิทธิ์เป็นศูนย์

3.เหมาะสำหรับข้อมูลที่มีฟีเจอร์จำนวนมาก: Elastic Net สามารถใช้งานได้ดีในกรณีที่ข้อมูลมีฟีเจอร์หลายตัว (High-dimensional data)

2.2.5 **ค่าเฉลี่ย** (Arithmetic Average)

คือ วิธีการหาค่ากลางของชุดข้อมูลโดยการรวมค่าทั้งหมดแล้วหารด้วยจำนวนข้อมูลทั้งหมด ซึ่งใช้เพื่อหาค่าที่เป็นตัวแทนของชุดข้อมูลทั้งหมดที่ให้ข้อมูลโดยรวม

สมการของค่าเฉลี่ย (Arithmetic Average)

$$\text{ค่าเฉลี่ย } (\mu) = \frac{x_1 + x_2 + \cdots + x_n}{n}$$

โดยที่

$x_1 + x_2 + \cdots + x_n$ คือค่าของข้อมูลในชุดข้อมูล

n คือจำนวนข้อมูลทั้งหมดในชุดข้อมูล

2.2.6 การประมวลผลหรือวัดประสิทธิภาพของแบบจำลอง

การประมวลผลหรือวัดประสิทธิภาพของแบบจำลอง (Model Evaluation) เป็นขั้นตอนสำคัญในการทดสอบว่าแบบจำลองที่พัฒนาขึ้นสามารถทำงานได้ดีเพียงใด และสามารถนำไปใช้งานจริงได้หรือไม่ โดยการประเมินผลมีหลายวิธี ขึ้นอยู่กับประเภทของแบบจำลองที่ใช้งาน (เช่น แบบจำลองการจำแนกประเภทหรือการพยากรณ์) และประเภทของข้อมูล

2.2.6.1 Mean Absolute Error (MAE)

ค่าเฉลี่ยของความแตกต่างสัมบูรณ์ระหว่างค่าที่แบบจำลองทำนายกับค่าจริง MAE จะให้ค่าผลลัพธ์ที่เป็นบวกเสมอ และจะไม่ให้ข้อมูลเกี่ยวกับทิศทางของความผิดพลาด (บวกหรือลบ) แต่เพียงแค่ให้ขนาดของความผิดพลาด และค่าที่น้อยหมายถึงแบบจำลองทำนายได้ดีใกล้เคียงกับค่าจริง

สมการ

$$MAE = \frac{1}{n}\sum_{i=1}^{n}\left|y_{\text{actual}} - y_{\text{predicted}}\right|$$

โดยที่

n คือ จำนวนข้อมูลทั้งหมด

y_{actual} คือ ค่าจริง

$y_{\text{predicted}}$ คือ ค่าที่แบบจำลองทำนาย

2.2.6.2 Mean Squared Error (MSE)

ค่าเฉลี่ยของกำลังสองของความแตกต่างระหว่างค่าที่แบบจำลองทำนายกับค่าจริง MSE ให้ความสำคัญกับความผิดพลาดที่ใหญ่กว่ามากกว่าความผิดพลาดที่เล็ก (เพราะการยกกำลังสอง) ดังนั้นมันจะให้ค่าผลลัพธ์ที่สูงขึ้นสำหรับข้อผิดพลาดที่มีค่ามาก โดยMSE จะบอกถึงความแม่นยำของแบบจำลอง และความผิดพลาดที่เกิดขึ้นในรูปแบบของค่ากำลังสอง

สมการ

$$MSE = \frac{1}{n}\sum_{i=1}^{n}\left(y_{\text{actual}} - y_{\text{predicted}}\right)^2$$

โดยที่

n คือ จำนวนข้อมูล

y_{actual} คือ ค่าจริง

$y_{\text{predicted}}$ คือ ค่าที่แบบจำลองทำนาย

2.2.6.3 R^2 (Coefficient of Determination / R-Squared / R^2 Score)

R^2 หรือ "Coefficient of Determination" ใช้เพื่อประเมินว่าแบบจำลองการทำนายสามารถอธิบายความแปรผันของข้อมูลได้ดีแค่ไหน โดยค่า R^2 จะมีค่าระหว่าง 0 ถึง 1

สมการ

$$R^2 = 1 - \frac{\sum_{i=1}^{n}\left(y_{\text{actual}} - y_{\text{predicted}}\right)^2}{\sum_{i=1}^{n}(y_{\text{actual}} - \bar{y})^2}$$

โดยที่

$\bar{y}$ คือ ค่าเฉลี่ยของค่าจริง

y_{actual} คือ ค่าจริง

$y_{\text{predicted}}$ คือ ค่าที่แบบจำลองทำนาย

การแปลผลค่า R^2

$R^2=1$ หมายถึง แบบจำลองสามารถทำนายได้ตรงกับค่าจริง 100%

$R^2=0$ หมายถึง แบบจำลองไม่มีความสามารถในการพยากรณ์เลย (เทียบได้กับการใช้ค่าเฉลี่ยของข้อมูล)

$R^2<0$ หมายถึง แบบจำลองแย่กว่าการใช้ค่าเฉลี่ยในการพยากรณ์

2.3 งานวิจัยที่เกี่ยวข้อง

ในงานวิจัยนี้ได้ทำการทบทวนวรรณกรรมและศึกษางานวิจัยที่เน้นการอธิบายการพยากรณ์ในรูปแบบของแบบจำลองต่าง ๆ เพื่อสำรวจวิธีการและแนวทางในการวิจัยที่เกี่ยวข้อง โดยมีงานวิจัยที่เกี่ยวข้อง ตัวอย่างเช่น

2.3.1 **บทความวิจัยเรื่อง** Machine-Learning Models for Sales Time Series Forecasting (Pavlyshenko, 2019)

งานวิจัยนี้มุ่งเน้นการใช้แบบจำลองการเรียนรู้ของเครื่องสำหรับการพยากรณ์ยอดขายในรูปแบบข้อมูลอนุกรมเวลา (time series forecasting) โดยศึกษาและวิเคราะห์การใช้วิธีการ generalization ของการเรียนรู้ของเครื่องเพื่อการพยากรณ์ยอดขายในกรณีที่มีข้อมูลย้อนหลังเพียงเล็กน้อย เช่น การเปิดตัวผลิตภัณฑ์ใหม่หรือร้านค้าใหม่ โดยมีจุดประสงค์เพื่อสำรวจและศึกษาแนวทางในการปรับปรุงความแม่นยำในการพยากรณ์ยอดขาย โดยใช้การเรียนรู้ของเครื่อง

การรวมแบบจำลองหลายประเภท (stacking) และวิธีการที่หลากหลายสำหรับการพยากรณ์ยอดขาย โดยเฉพาะในกรณีที่มีข้อมูลย้อนหลังไม่มาก

แบบจำลองที่ใช้ในการวิจัยรวมถึง Random Forest ซึ่งเป็นแบบจำลองการถดถอยที่ใช้ฟีเจอร์ต่างๆ เช่น โปรโมชัน, วันในสัปดาห์, และระยะทางจากร้านคู่แข่ง นอกจากนี้ยังใช้ Lasso Regression และ Neural Networks สำหรับการรวมผลการพยากรณ์จากแบบจำลองหลายตัว รวมถึง Stacking Models ซึ่งเป็นการรวมผลลัพธ์จากแบบจำลองหลายตัวเพื่อปรับปรุงความแม่นยำของการพยากรณ์

กระบวนการดำเนินงานเริ่มต้นด้วยการใช้ข้อมูลยอดขายจากการแข่งขัน “Rossmann Store Sales” ใน Kaggle ทำการทดสอบแบบจำลองหลายตัว เช่น Random Forest และการพยากรณ์ผ่านการรวมแบบจำลอง โดยใช้เทคนิคการจัดการข้อมูลอนุกรมเวลาและการจัดแบ่งข้อมูลสำหรับการทดสอบและการประเมินผล โดยมีการคำนวณความผิดพลาดของแบบจำลองโดยใช้ Mean Absolute Error (MAE) และการใช้การ generalization เพื่อคาดการณ์ยอดขายในกรณีที่มีข้อมูลย้อนหลังน้อย

ผลการศึกษาแสดงให้เห็นว่า การใช้แบบจำลองการเรียนรู้ของเครื่องสามารถปรับปรุงความแม่นยำของการพยากรณ์ยอดขายได้ โดยเฉพาะในกรณีที่ข้อมูลมีแนวโน้มซับซ้อน และการใช้วิธีการ stacking สามารถช่วยปรับปรุงผลลัพธ์ได้ดีขึ้น การใช้แบบจำลองที่รวมผลการพยากรณ์จากหลายแบบจำลองยังสามารถเพิ่มความแม่นยำในการทำนายบนชุดข้อมูลทดสอบได้อีกด้วย

2.3.2 **บทความวิจัยเรื่อง** Time series forecast of sales volume based on XGBoost (Zhang et al., 2021)

งานวิจัยนี้มุ่งเน้นการพยากรณ์ยอดขายในอนาคตโดยใช้ข้อมูลอนุกรมเวลาและแบบจำลอง XGBoost ซึ่งถูกปรับใช้เพื่อคาดการณ์ยอดขายในอุตสาหกรรมค้าปลีก โดยเพิ่มเติมปัจจัยสิ่งแวดล้อม เช่น สภาพอากาศและคุณภาพอากาศ เพื่อปรับปรุงความแม่นยำของการพยากรณ์ จุดประสงค์ของงานวิจัยนี้คือการพัฒนาระบบการพยากรณ์ยอดขายที่จะช่วยให้การจัดการพนักงานในช่วงเวลาต่างๆ มีประสิทธิภาพมากขึ้น เช่น การจัดการพนักงานให้เหมาะสมกับช่วงเวลาที่ยุ่งหรือเวลาที่ไม่ยุ่ง โดยใช้ข้อมูลจากธุรกรรมและข้อมูลสภาพอากาศ

แบบจำลองที่ใช้ในงานวิจัยประกอบด้วย XGBoost ซึ่งเป็นอัลกอริธึมการเรียนรู้ของเครื่องที่ใช้เทคนิค Gradient Boosting นอกจากนี้ยังมีการเปรียบเทียบผลลัพธ์กับแบบจำลองอื่นๆ ได้แก่ ARIMA, LSTM, Prophet, และ GBDT กระบวนการดำเนินงานเริ่มจากการรวบรวมข้อมูลยอดขายจากสองร้านค้าในปี 2019 รวมทั้งข้อมูลสภาพอากาศของกรุงปักกิ่ง โดยใช้วิศวกรรมฟีเจอร์เพื่อเลือกฟีเจอร์ที่สำคัญ เช่น อุณหภูมิ, คุณภาพอากาศ และยอดขายย้อนหลัง 7 วัน ข้อมูลจะถูกปรับ

ในช่วงเวลา 15 นาที เพื่อให้เหมาะสมกับการพยากรณ์ยอดขาย จากนั้นประเมินประสิทธิภาพของแบบจำลองโดยใช้ RMSE และ MAE เพื่อเปรียบเทียบความแม่นยำของแบบจำลองต่างๆ

ผลการศึกษาพบว่าแบบจำลอง XGBoost ทำงานได้ดีที่สุดเมื่อเทียบกับแบบจำลองอื่นๆ เช่น GBDT และ ARIMA โดยให้ผลลัพธ์ที่มีความแม่นยำสูงที่สุด โดยใช้จำนวนการวนซ้ำน้อยกว่า GBDT ถึงหนึ่งในสาม ขณะที่สามารถให้ผลการพยากรณ์ที่ดีกว่า นอกจากนี้ การเพิ่มปัจจัยภายนอก เช่น สภาพอากาศ ช่วยปรับปรุงความแม่นยำของการพยากรณ์ยอดขายได้อย่างมีนัยสำคัญ

2.3.3 **บทความวิจัยเรื่อง** Research on Sales Forecast Based on XGBoost-LSTM Algorithm Model (He & Zeng, 2021)

งานวิจัยนี้นำเสนอการพยากรณ์ยอดขายโดยใช้แบบจำลองการรวมระหว่าง XGBoost และ LSTM สำหรับการพยากรณ์อนุกรมเวลา (Time Series Forecasting) เพื่อเปรียบเทียบประสิทธิภาพกับแบบจำลองพยากรณ์แบบดั้งเดิม โดยเน้นการนำข้อมูลยอดขายจากการแข่งขัน Kaggle มาใช้ในการทดลองเพื่อหาความแม่นยำในการพยากรณ์ยอดขายของเครือซุปเปอร์มาร์เก็ต จุดประสงค์ของงานวิจัยนี้คือการพัฒนาแบบจำลองการพยากรณ์ยอดขายที่มีความแม่นยำสูงขึ้น โดยการผสมผสานจุดเด่นของแบบจำลอง XGBoost และ LSTM เพื่อช่วยให้ธุรกิจสามารถวางแผนกลยุทธ์การตลาดและตัดสินใจด้านการขายได้อย่างมีประสิทธิภาพมากขึ้น

แบบจำลองที่ใช้ในงานวิจัยนี้ประกอบด้วย LSTM (Long Short-Term Memory) ซึ่งช่วยในการเรียนรู้จากข้อมูลอนุกรมเวลาและแก้ปัญหาการสูญเสียหรือการระเบิดของ Gradient เมื่อทำการฝึกแบบจำลอง และ XGBoost ซึ่งเป็นอัลกอริธึมการเรียนรู้แบบ Gradient Boosting ที่ใช้ในการเพิ่มประสิทธิภาพการพยากรณ์

กระบวนการดำเนินงานเริ่มจากการใช้ข้อมูลยอดขายรายวันของซุปเปอร์มาร์เก็ตจากการแข่งขัน Kaggle ระหว่างปี 2017 ถึง 2019 รวมทั้งข้อมูลจากร้านค้าปลีก 78 แห่งในกรุงปักกิ่ง โดยแบ่งข้อมูลเป็นชุดฝึก (Training set) 70% และชุดทดสอบ (Test set) 30% โดยใช้การตรวจสอบไขว้ (Cross-validation) 10% เพื่อค้นหาพารามิเตอร์ที่เหมาะสมที่สุด การฝึกแบบจำลอง LSTM จะถูกทำก่อน จากนั้นจะใช้ผลลัพธ์ที่ได้เพิ่มเข้าไปในชุดข้อมูลใหม่เพื่อฝึกแบบจำลอง XGBoost

ผลการทดลองแสดงให้เห็นว่าแบบจำลอง XGBoost-LSTM มีประสิทธิภาพสูงกว่าการใช้แบบจำลองแบบเดี่ยว โดยแบบจำลองที่ใช้การรวม XGBoost และ LSTM สามารถเพิ่มความแม่นยำได้ดีกว่าแบบจำลอง XGBoost เพียงอย่างเดียว นอกจากนี้ยังช่วยให้ธุรกิจสามารถใช้ผลลัพธ์นี้ในการวางแผนการตลาดได้อย่างมีประสิทธิภาพ การทดลองใช้ชุดข้อมูลจาก 78 กลุ่ม

ยอดขาย 1,096 อนุกรมเวลา (Time Series) แสดงให้เห็นถึงความสามารถในการพยากรณ์ที่ดีกว่าแบบจำลองดั้งเดิม

2.3.4 บทความวิจัยเรื่อง Short-term Load Forecasting of Industrial Customers Based on SVMD and XGBoost (Wang et al., 2021)

งานวิจัยนี้นำเสนอวิธีการพยากรณ์โหลดไฟฟ้าระยะสั้นสำหรับลูกค้าอุตสาหกรรม โดยใช้วิธีการแยกข้อมูลแบบปรับตัว (SVMD) และแบบจำลองการพยากรณ์ XGBoost เพื่อมุ่งเน้นการพยากรณ์โหลดไฟฟ้าซึ่งมีความผันผวนและไม่แน่นอนมากกว่าระบบโหลดไฟฟ้าทั่วไป จุดประสงค์ของงานวิจัยนี้คือการพัฒนาแบบจำลองการพยากรณ์ที่มีความแม่นยำสูงสำหรับลูกค้าอุตสาหกรรม โดยผสมผสานการแยกข้อมูลด้วย SVMD และแบบจำลอง XGBoost ที่สามารถปรับค่าพารามิเตอร์ได้โดยใช้ Bayesian Optimization Algorithm (BOA)

แบบจำลองที่ใช้ในงานวิจัยประกอบด้วย SVMD (Variational Mode Decomposition with Sample Entropy) ซึ่งใช้สำหรับการแยกข้อมูลแบบปรับตัว เพื่อลดความซับซ้อนของข้อมูลโหลดไฟฟ้าและแบ่งข้อมูลออกเป็นส่วนที่มีแนวโน้มและส่วนที่มีความผันผวน ขณะเดียวกัน XGBoost ถูกใช้ในการพยากรณ์ข้อมูลที่ถูกแยกออกมา โดยมีการปรับพารามิเตอร์ผ่าน BOA

กระบวนการดำเนินงานเริ่มจากการรวบรวมข้อมูลโหลดไฟฟ้ารายวันจากลูกค้าอุตสาหกรรมในประเทศจีนและไอร์แลนด์ หลังจากนั้นใช้การแยกข้อมูลด้วย SVMD เพื่อแบ่งข้อมูลโหลดออกเป็นชุดย่อย จากนั้นใช้ XGBoost ในการพยากรณ์ข้อมูลชุดย่อย และใช้ Line Regression (LR) สำหรับแนวโน้มของข้อมูลหลัก สุดท้ายทำการประเมินความแม่นยำของแบบจำลองโดยใช้ดัชนี Mean Square Error (MSE), Mean Absolute Error (MAE), และ Mean Absolute Percentage Error (MAPE)

ผลการศึกษาพบว่าแบบจำลอง SVMD-XGBoost มีความแม่นยำสูงกว่าการใช้แบบจำลองพยากรณ์อื่นๆ โดยแสดงให้เห็นถึงค่าความผิดพลาดที่น้อยกว่า แบบจำลองที่ใช้วิธีการแยกข้อมูลด้วย SVMD ช่วยลดปัญหาเรื่องการแบ่งข้อมูลที่ไม่ดีและทำให้ผลการพยากรณ์แม่นยำขึ้น โดย MSE ของแบบจำลองที่นำเสนออยู่ที่ 52.98 kW, MAE อยู่ที่ 4.60 kW, และ MAPE อยู่ที่ 3.63% นอกจากนี้ เมื่อเปรียบเทียบกับแบบจำลองพยากรณ์อื่นๆ แบบจำลองนี้ยังให้ผลการพยากรณ์ที่แม่นยำกว่าด้วยค่าความผิดพลาดที่ต่ำกว่าแบบจำลอง เช่น LSTM, GBDT, และ CNN

2.3.5 บทความวิจัยเรื่อง Demand Forecasting in Wholesale Alcohol Distribution: An Ensemble Approach (Arora et al., 2020)

งานวิจัยนี้มุ่งเน้นการพยากรณ์ความต้องการในอุตสาหกรรมการจัดจำหน่ายแอลกอฮอล์โดยใช้แนวทางการผสมผสาน (Ensemble Approach) ของแบบจำลองการพยากรณ์เพื่อเพิ่ม

ความแม่นยำในการพยากรณ์ความต้องการสินค้าแอลกอฮอล์ที่ขายผ่านผู้จัดจำหน่าย จุดประสงค์หลักคือการปรับปรุงการจัดการสินค้าคงคลัง กระแสเงินสด และการบริการลูกค้า โดยพัฒนาแบบจำลองการพยากรณ์ที่มีประสิทธิภาพสูงในการจัดจำหน่ายแอลกอฮอล์ ซึ่งใช้วิธีผสมผสานแบบจำลองต่างๆ เพื่อให้สามารถปรับแต่งแบบจำลองให้เหมาะสมกับสินค้าที่แตกต่างกันและลดความผิดพลาดในการพยากรณ์

แบบจำลองที่ใช้ในการวิจัยประกอบด้วย S-ARIMA (Seasonal ARIMA) สำหรับการพยากรณ์ข้อมูลอนุกรมเวลา, Vector Auto-Regression (VAR) ซึ่งเป็นแบบจำลองเชิงสถิติสำหรับการพยากรณ์ที่รวมข้อมูลจากหลายตัวแปร, LSTM (Long Short-Term Memory) ซึ่งเป็นแบบจำลองการเรียนรู้เชิงลึกสำหรับการพยากรณ์ข้อมูลอนุกรมเวลา, และ Ensemble Model ที่ผสมผสานแบบจำลองทั้งสามแบบเข้าด้วยกันเพื่อเพิ่มความแม่นยำในการพยากรณ์

กระบวนการดำเนินงานเริ่มต้นด้วยการใช้ข้อมูลยอดขายรายเดือนของแอลกอฮอล์จากผู้จัดจำหน่ายในช่วงปี 2013-2019 ซึ่งรวมถึงผลิตภัณฑ์ยอดนิยมสองชนิด: Taaka Vodka และ Jack Daniel's Whiskey จากนั้นทำการทดสอบและประเมินประสิทธิภาพของแต่ละแบบจำลอง เช่น S-ARIMA, VAR, และ LSTM โดยใช้การทดสอบ cross-validation แบบ rolling window และพัฒนาแบบจำลองผสมผสาน (Ensemble) โดยให้น้ำหนักแก่แบบจำลองที่มีความแม่นยำสูงสุดเพื่อให้ได้ผลลัพธ์ที่ดีกว่าการใช้แบบจำลองเดี่ยว

ผลการทดลองแสดงให้เห็นว่าแบบจำลองผสมผสาน (Ensemble) สามารถลดความผิดพลาดในการพยากรณ์ได้อย่างมีนัยสำคัญ สำหรับ Taaka Vodka แบบจำลอง XGBoost มีค่า ASE ที่ 3,633.82 ลดลง 71% จากแบบจำลอง Naive, ส่วน LSTM มีค่า ASE ที่ 2,960.52 ลดลง 76% สำหรับ Jack Daniel's Whiskey แบบจำลอง Ensemble มีค่า ASE ที่ 6,413.11 ลดลง 56% จากแบบจำลอง Naive โดยการใช้ LSTM และ Ensemble Model มีประสิทธิภาพสูงสุดในการพยากรณ์ยอดขายแอลกอฮอล์

2.3.6 บทความวิจัยเรื่อง Sales Forecasting for Retail Chains (Jain et al., 2020)

งานวิจัยนี้นำเสนอการใช้เทคนิคเหมืองข้อมูล (Data Mining) สำหรับการพยากรณ์ยอดขายในเครือข่ายค้าปลีก โดยใช้ XGBoost เป็นอัลกอริธึมหลักในการพัฒนาแบบจำลองพยากรณ์ยอดขาย ซึ่งได้ถูกนำมาใช้กับข้อมูลจากร้านขายยาในยุโรป โดยวิเคราะห์ปัจจัยที่ส่งผลต่อยอดขาย เช่น ข้อมูลการส่งเสริมการขาย, คู่แข่ง, วันหยุด, และฤดูกาล จุดประสงค์ของงานวิจัยคือการพัฒนาแบบจำลองพยากรณ์ยอดขายที่แม่นยำเพื่อช่วยผู้จัดการร้านในการวางแผนการทำงานและเพิ่มประสิทธิภาพในการจัดการสินค้าและพนักงาน

แบบจำลองที่ใช้ประกอบด้วย XGBoost, Linear Regression, และ Random Forest Regression ซึ่งมีการทดสอบแบบจำลองต่างๆ โดยใช้ข้อมูลยอดขายรายวันจากเครือร้าน Rossmann ที่มีร้านค้ากว่า 3,000 แห่งในเยอรมนี ระหว่างปี 2013 ถึง 2015 โดยใช้วิธีการ Rolling Window Cross-validation และวิเคราะห์ความสำคัญของฟีเจอร์ต่างๆ ที่มีผลต่อการพยากรณ์

ผลการศึกษาพบว่า XGBoost มีประสิทธิภาพสูงที่สุดเมื่อเปรียบเทียบกับแบบจำลองอื่นๆ โดยมีค่า RMPSE (Root Mean Squared Percentage Error) ของแต่ละแบบจำลองดังนี้: Mean of each DayOfWeek Model = 0.18968, Linear Regression = 0.15672, Random Forest Regression = 0.13198, และ XGBoost = 0.10532 ซึ่งแสดงให้เห็นว่าแบบจำลอง XGBoost สามารถพยากรณ์ยอดขายได้อย่างแม่นยำที่สุด โดยมีความผิดพลาดลดลงอย่างชัดเจนเมื่อเปรียบเทียบกับ Linear Regression และ Random Fores

2.3.7 บทความวิจัยเรื่อง A Comparative Study of Demand Forecasting Models for a Multi-Channel Retail Company: A Novel Hybrid Machine Learning Approach (Mitra et al., 2022)

งานวิจัยนี้เปรียบเทียบประสิทธิภาพของแบบจำลองการพยากรณ์ความต้องการสินค้าหลายรูปแบบ โดยใช้ข้อมูลการขายรายสัปดาห์จากบริษัทค้าปลีกในสหรัฐอเมริกา รวมถึงปัจจัยต่างๆ เช่น ขนาดของร้านค้า อุณหภูมิของภูมิภาค และราคาน้ำมัน โดยนำเสนอแบบจำลองผสมผสานใหม่ (RF-XGBoost-LR) เพื่อเปรียบเทียบกับแบบจำลองการพยากรณ์ที่ใช้กันอยู่ จุดประสงค์คือการพัฒนาแบบจำลองการพยากรณ์ที่มีความแม่นยำสูงขึ้น โดยใช้วิธีการผสมผสานจุดเด่นของ Random Forest, XGBoost, และ Linear Regression เพื่อเพิ่มประสิทธิภาพในการพยากรณ์

กระบวนการดำเนินงานเริ่มต้นด้วยการใช้ข้อมูลยอดขายรายสัปดาห์จาก 45 ร้านค้าในช่วง 3 ปี ทดสอบแบบจำลองต่างๆ เช่น RF, XGBoost, Gradient Boosting, AdaBoost, และ ANN โดยใช้ตัวชี้วัด MAE, MSE, และ R^2 Score สร้างแบบจำลองไฮบริดโดยฝึก RF และ XGBoost แยกกันและใช้ผลลัพธ์ในการฝึก Linear Regression

ผลการศึกษาพบว่าแบบจำลอง RF-XGBoost-LR มีประสิทธิภาพสูงสุด โดยมีค่า MAE ที่ 1202.42, MSE ที่ 3428795.89, และ R^2 ที่ 0.852 ซึ่งดีกว่าแบบจำลองอื่นๆ เช่น Random Forest (MAE = 1254.67), XGBoost (MAE = 1289.12), Gradient Boosting (MAE = 1301.85), AdaBoost (MAE = 1384.93), และ ANN (MAE = 1412.35) โดยผลลัพธ์แสดงให้เห็นว่าแบบจำลองไฮบริดมีความแม่นยำในการพยากรณ์ยอดขายสูงสุดและประสิทธิภาพในการจัดการสินค้าคงคลังที่ดีขึ้น

2.3.8 บทความวิจัยเรื่อง Developing and Preliminary Testing of a Machine Learning-Based Platform for Sales Forecasting Using a Gradient Boosting Approach (Panarese et al., 2022)

งานวิจัยนี้นำเสนอการพัฒนาและทดสอบเบื้องต้นของแพลตฟอร์มการพยากรณ์ยอดขาย โดยใช้เทคนิคการเพิ่มประสิทธิภาพแบบ Gradient Boosting เพื่อปรับปรุงความแม่นยำในการพยากรณ์ในบริบทการค้าหลายช่องทาง โดยเปรียบเทียบสองแบบจำลองหลัก ได้แก่ Gradient Boosting Regressor (GBR) และ XGBoost Regressor (XGBR) จุดประสงค์คือการพัฒนาแพลตฟอร์มสำหรับการพยากรณ์ยอดขายในบริษัทที่มีหลายสาขา เพื่อปรับปรุงกระบวนการจัดการสินค้าคงคลังและเพิ่มประสิทธิภาพการดำเนินงาน

กระบวนการดำเนินงานเริ่มจากการใช้ข้อมูลการขายประจำวันจากหลายสาขา แบ่งข้อมูลเป็นชุดฝึก (80%) และชุดทดสอบ (20%) จากนั้นฝึกแบบจำลอง GBR และ XGBR โดยใช้ตัวแปรอินพุต เช่น รหัสสินค้าและปริมาณที่สั่งซื้อ พร้อมปรับพารามิเตอร์ผ่าน GridSearchCV และทดสอบแบบจำลองด้วยการคาดการณ์ยอดขายของสินค้าที่ขายอยู่และสินค้าที่เปิดตัวใหม่

ผลการศึกษาพบว่า XGBR ให้ผลลัพธ์ที่แม่นยำกว่ากว่า GBR โดยมีค่า MSE สำหรับสินค้าที่ขายอยู่ในตลาดคือ 189.3 (XGBR) เทียบกับ 225.6 (GBR) และ MAE ที่ 10.41 (XGBR) เทียบกับ 11.32 (GBR) สำหรับสินค้าที่เปิดตัวใหม่ ค่า MSE ของ XGBR อยู่ที่ 213.7 ซึ่งแสดงให้เห็นว่าแบบจำลอง XGBoost มีความแม่นยำสูงสุดในการพยากรณ์ยอดขายเมื่อเปรียบเทียบกับ GBR และมีการพยากรณ์ที่แม่นยำสูงขึ้นเมื่อฝึกจากข้อมูลที่แชร์ระหว่างสองสาขา

2.3.9 บทความวิจัยเรื่อง Boosting Learning Algorithm for Stock Price Forecasting (Wang & Bai, 2018)

งานวิจัยนี้นำเสนอการใช้แบบจำลองการเรียนรู้แบบเพิ่มประสิทธิภาพ (Boosting) โดยใช้โครงข่ายประสาทเทียม (ANN) เพื่อพยากรณ์ราคาหุ้นในตลาดหุ้น โดยมุ่งเน้นการแก้ไขปัญหาความซับซ้อนและความไม่แน่นอนของตลาดหุ้นด้วยการใช้แบบจำลอง Boosting-ANN ซึ่งรวมตัวพยากรณ์ที่อ่อนแอหลายตัว (Weak Predictors) เพื่อสร้างตัวพยากรณ์ที่แข็งแกร่ง จุดประสงค์คือการพัฒนาแบบจำลองพยากรณ์ราคาหุ้นที่มีความแม่นยำสูงขึ้นและแสดงให้เห็นว่าแบบจำลอง Boosting-ANN ทำงานได้ดีกว่าแบบจำลอง ANN แบบเดี่ยว

กระบวนการดำเนินงานเริ่มจากการรวบรวมข้อมูลราคาหุ้นจากตลาดหุ้นเซี่ยงไฮ้ตั้งแต่เดือนมกราคม 2011 ถึงมีนาคม 2016 พร้อมข้อมูลจากตลาดต่างประเทศ เช่น S&P 500, NASDAQ และ DJIA รวมถึงอัตราแลกเปลี่ยนสกุลเงิน USD โดยเลือกใช้ตัวแปรทางเทคนิค เช่น ราคาปิด ราคาสูงสุด ราคาต่ำสุด และดัชนี MACD เพื่อใช้เป็นข้อมูลอินพุต จากนั้นฝึกแบบจำลอง

ANN หลายตัวเป็น Weak Learners และสร้างแบบจำลอง Boosting-ANN เพื่อเพิ่มความแม่นยำในการพยากรณ์

ผลการทดลองแสดงให้เห็นว่าแบบจำลอง Boosting-ANN มีประสิทธิภาพสูงกว่าแบบจำลอง ANN แบบเดี่ยว โดยค่า MAE ของ Boosting-ANN อยู่ที่ 99.9720 เทียบกับ 117.6427 ของ ANN, ค่า MSE อยู่ที่ 18,279.32 เทียบกับ 25,086.48, และ RMSE อยู่ที่ 135.2010 เทียบกับ 158.3871 แบบจำลอง Boosting-ANN ยังให้ค่า MAPE ที่ต่ำที่สุดที่ 3.05% ซึ่งแสดงให้เห็นถึงความแม่นยำสูงในการพยากรณ์ราคาหุ้นเมื่อเปรียบเทียบกับแบบจำลองอื่นๆ

2.3.10 **บทความวิจัยเรื่อง** Time Series Analysis and Forecasting of Solar Generation in Spain Using eXtreme Gradient Boosting: A Machine Learning Approach (Saigustia & Pijarski, 2023)

งานวิจัยนี้มุ่งเน้นการวิเคราะห์และพยากรณ์การผลิตพลังงานแสงอาทิตย์ในสเปนโดยใช้เทคนิคการเรียนรู้ของเครื่อง eXtreme Gradient Boosting (XGBoost) โดยเน้นการใช้ข้อมูลอนุกรมเวลาในการวิเคราะห์การผลิตพลังงานระหว่างปี 2015 ถึง 2018 โดยไม่รวมข้อมูลสภาพอากาศ จุดประสงค์ของงานวิจัยคือการพัฒนาแบบจำลองการพยากรณ์ที่มีความแม่นยำสูงเพื่อช่วยในการบริหารจัดการโครงข่ายพลังงานไฟฟ้า

กระบวนการดำเนินงานเริ่มต้นด้วยการรวบรวมข้อมูลการผลิตพลังงานแสงอาทิตย์รายชั่วโมงในช่วงเวลาดังกล่าวและทำการจัดเตรียมข้อมูลโดยลบค่าผิดปกติ จากนั้นแบ่งข้อมูลออกเป็นชุดฝึก (80%) และชุดทดสอบ (20%) โดยชุดทดสอบใช้ข้อมูลจากปี 2018 ซึ่งได้ทำการวิเคราะห์ข้อมูลอนุกรมเวลา รวมถึงรูปแบบการผลิตรายวันและความแปรปรวนเชิงภูมิศาสตร์ ก่อนจะใช้แบบจำลอง XGBoost ในการพยากรณ์การผลิตพลังงานแสงอาทิตย์

ผลการศึกษาพบว่าแบบจำลอง XGBoost มีความแม่นยำสูง โดยมีค่า RMSE ที่ 11.042, MAE ที่ 5.621, R^2 ที่ 0.999, และ MAPE ที่ 0.0463 ผลลัพธ์นี้แสดงให้เห็นว่าแบบจำลองสามารถพยากรณ์การผลิตพลังงานแสงอาทิตย์ได้อย่างแม่นยำและมีประสิทธิภาพสูงในการจับรูปแบบเชิงเวลา ซึ่งมีความสำคัญต่อการวางแผนการใช้พลังงานและการจัดการโครงข่ายไฟฟ้าอย่างมีประสิทธิภาพ

2.3.11 **บทความวิจัยเรื่อง** Forecasting Floods Using Extreme Gradient Boosting – A New Approach (Venkatesan & Mahindrakar, 2019)

งานวิจัยนี้มุ่งเน้นการพัฒนาแบบจำลองพยากรณ์น้ำท่วมระยะสั้นโดยใช้เทคนิค Extreme Gradient Boosting (XGBoost) ซึ่งเป็นเทคนิคการเรียนรู้ของเครื่องที่มีประสิทธิภาพสูง โดยทำการทดลองในพื้นที่ Kolar Basin ในรัฐมัธย Pradesh, อินเดีย โดยใช้ข้อมูลฝนและการระบายน้ำราย

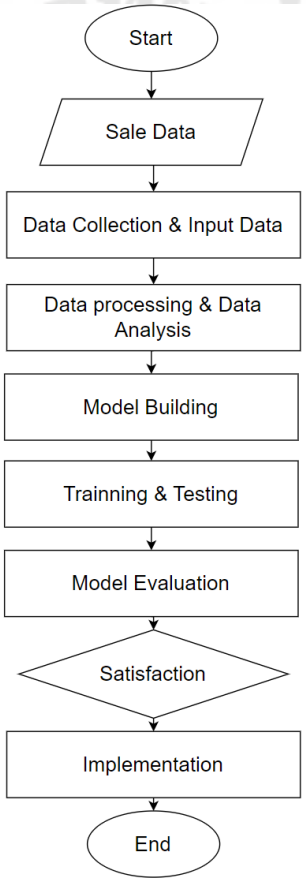
ชั่วโมงระหว่างปี 1987–1989 สำหรับการพยากรณ์น้ำท่วมล่วงหน้า 5 ชั่วโมง จุดประสงค์คือการพัฒนาแบบจำลองที่มีความแม่นยำสูงและเปรียบเทียบกับแบบจำลองที่มีอยู่ เช่น Random Forest (RF) และ Support Vector Machine (SVM)

กระบวนการดำเนินงานเริ่มจากการใช้ข้อมูลการไหลของน้ำและปริมาณฝน แบ่งข้อมูลเป็นชุดฝึก (75%) และชุดทดสอบ (25%) จากนั้นฝึกแบบจำลอง XGBoost พร้อมข้อมูลปัจจัยอินพุต เช่น ปริมาณฝนที่ตกและการไหลของน้ำ และเปรียบเทียบประสิทธิภาพของ XGBoost กับ RF และ SVM โดยใช้ตัวชี้วัด RMSE, NSE, R^2, และ PBIAS ผลการศึกษาพบว่าแบบจำลอง XGBoost มีประสิทธิภาพสูงสุด โดยมีค่า RMSE ที่ 19.77 (validation) ที่ระยะเวลา 1 ชั่วโมง และ 18.90 ที่ระยะเวลา 5 ชั่วโมง ขณะที่ Random Forest มีค่า RMSE ที่ 80.99 และ 114.21 ตามลำดับ สำหรับ SVM ค่า RMSE อยู่ที่ 37.94 และ 78.43 ซึ่งแสดงให้เห็นว่า XGBoost สามารถพยากรณ์น้ำท่วมได้แม่นยำกว่าในทั้งช่วงการสอบเทียบและการทดสอบ โดยเฉพาะในช่วงเวลาพยากรณ์ล่วงหน้า 1 ถึง 5 ชั่วโมง

บทที่ 3
วิธีดำเนินการวิจัย

ในการวิจัยครั้งนี้ผู้วิจัยได้ดำเนินการตามขั้นตอนดังนี้

1. การออกแบบกระบวนการดำเนินการวิจัย
2. การนำเข้าข้อมูล (Input Data)
3. การประมวลผลข้อมูล (Data Processing)
4. การสร้างแบบจำลอง (Model Building)
5. การประเมินประสิทธิภาพแบบจำลอง (Model Evaluation)



ภาพประกอบ 1 แสดงการออกแบบกระบวนการดำเนินการวิจัย

3.1 การออกแบบกระบวนการดำเนินการวิจัย

จากภาพประกอบ 1 ภาพนี้แสดงกระบวนการวิเคราะห์ข้อมูลยอดขาย เริ่มต้นด้วยการรวบรวมและนำเข้าข้อมูลยอดขาย จากนั้นเข้าสู่ขั้นตอนการประมวลผลและวิเคราะห์ข้อมูลเบื้องต้น เพื่อนำไปสร้างแบบจำลอง (Model) เมื่อได้แบบจำลองแล้ว จะทำการฝึกฝนและทดสอบเพื่อประเมินประสิทธิภาพของแบบจำลอง หากผลลัพธ์ที่ได้ยังไม่เป็นที่พอใจ จะกลับไปปรับปรุงและทดสอบซ้ำจนกว่าจะได้ผลลัพธ์ที่เหมาะสม เมื่อผลลัพธ์เป็นที่น่าพอใจแล้ว จึงนำแบบจำลองนั้นไปใช้งานจริงและปิดกระบวนการด้วยการสรุปผลที่ได้

3.2 การนำเข้าข้อมูล

งานวิจัยนี้ได้ใช้ข้อมูลยอดขายจากของบริษัทเครื่องดื่มรายหนึ่งในประเทศไทย โดยข้อมูลที่ใช้เป็นข้อมูลยอดขายในช่วงวันที่ 1 กรกฎาคม 2566 จนถึง 30 มิถุนายน 2567 โดยข้อมูลประกอบไปด้วย 14 คอลัมน์ 185,507 แถว รายละเอียดตัวอย่างข้อมูลตามตารางที่ 1

ตาราง 1 แสดงคอลัมน์และตัวอย่างข้อมูลที่ใช้ในการทำวิจัย

ชื่อคอลัมน์	คำอธิบายข้อมูล	ตัวอย่างข้อมูล
WH_Code	รหัสของคลังสินค้า	9311 / 9314
Delivery_Date	วันที่ส่งสินค้า	03/07/2023
Hour	ช่วงของข้อมูลที่มีการสั่งซื้อ ยึดตามชั่วโมง	7 (07:00 – 07:59)
Sale_Quantity	จำนวนยอดขายที่ลูกค้าสั่งซื้อเข้ามา หน่วยเป็นลัง	550
Delivery_Quantity	จำนวนยอดขายที่สามารถส่งมอบได้ หน่วยเป็นลัง	520
Net_Value	มูลค่าสินค้าที่จัดส่งให้ลูกค้า หน่วยเป็นบาท	42728
Month	เดือนที่จัดส่งสินค้า	กรกฎาคม
Week	สัปดาห์ที่จัดส่งสินค้า	27
Day	ชื่อวันที่จัดส่งสินค้า	Mon
Region	ภูมิภาคที่ตั้งของคลังที่จัดส่งสินค้า	BKK
Last_Week	ช่วงวันในสัปดาห์สุดท้ายของเดือนหรือวันปกติ	Last_Week / Normal
Holidays	วันหยุดนักขัตฤกษ์หรือวันปกติ	Holiday / Normal
Before holidays	วันก่อนวันหยุดนักขัตฤกษ์หรือวันปกติ	Before holidays / Normal
After holidays	วันหลังวันหยุดนักขัตฤกษ์หรือวันปกติ	After holidays / Normal

ในการทำวิจัยนี้ผู้วิจัยได้ใช้ภาษาไพทอน (Python) สำหรับการสร้างแบบจำลอง Machine Learning และประมวลผลในการทำวิจัย โดยทำผ่าน Google Colab (ซึ่งย่อมาจาก Collaboratory) ที่เป็นเครื่องมือในการเขียน Code Python แบบออนไลน์บนเว็บเบราว์เซอร์ได้แบบไม่มีค่าใช้จ่ายและสะดวกง่ายต่อการใช้งาน การใช้งานจะเริ่มจาก Import Library หรือการนำเอาโค้ดสำเร็จรูปหรือฟังก์ชันที่มีอยู่ในไลบรารี มาใช้งานเพื่อที่จะได้ไม่ต้องเขียนโค้ดใหม่ทั้งหมด ทำให้ทำงานได้ง่ายขึ้น เสร็จแล้วจึง Import Data หรือ นำเข้าไฟล์ข้อมูล ที่จะใช้ในการทำวิจัย ทั้งนี้เพื่อให้สะดวกในการทำผู้วิจัยได้เก็บข้อมูลไว้ใน Google drive และ set path สำหรับการเข้าถึงข้อมูลไว้เพื่อประหยัดเวลาและลดความซับซ้อน จะได้ไม่ต้องระบุเส้นทางเต็ม ๆ ทุกครั้งเวลาจะเรียกใช้ข้อมูล

3.3 การประมวลผลข้อมูล

หลังจากที่นำเข้าข้อมูลแล้ว ในลำดับต่อมาจะเป็นการประมวลผล โดยลำดับแรกจะเป็นการตรวจสอบข้อมูลเบื้องต้นตามภาพประกอบ 2 และภาพประกอบ 3 เพื่อดูรายละเอียดข้อมูลเบื้องต้น

	WH_Code	Delivery_Date	Hour	Sale_Quantity	Delivery_Quantity	Net_Value	Month	Week	Day	Region	Last_Week	holidays	Before holidays	After holidays
0	9311	01/07/2023	7	948	948	99,906	July	26	Sat	BKK	Normal	Normal	Normal	Normal
1	9311	01/07/2023	8	7,635	7,616	1,447,336	July	26	Sat	BKK	Normal	Normal	Normal	Normal
2	9311	01/07/2023	9	11,860	11,855	1,979,219	July	26	Sat	BKK	Normal	Normal	Normal	Normal
3	9311	01/07/2023	10	3,828	3,828	734,497	July	26	Sat	BKK	Normal	Normal	Normal	Normal
4	9311	01/07/2023	11	4,344	4,190	844,955	July	26	Sat	BKK	Normal	Normal	Normal	Normal

ภาพประกอบ 2 Preview DataFrame เพื่อดูข้อมูล 5 แถวแรก

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 185507 entries, 0 to 185506
Data columns (total 14 columns):
 #   Column             Non-Null Count   Dtype 
---  ------             --------------   ----- 
 0   WH_Code            185507 non-null  int64 
 1   Delivery_Date      185507 non-null  object
 2   Hour               185507 non-null  int64 
 3   Sale_Quantity      185507 non-null  object
 4   Delivery_Quantity  185507 non-null  object
 5   Net_Value          185507 non-null  object
 6   Month              185507 non-null  object
 7   Week               185507 non-null  int64 
 8   Day                185507 non-null  object
 9   Region             185507 non-null  object
 10  Last_Week          185507 non-null  object
 11  holidays           185507 non-null  object
 12  Before holidays    185507 non-null  object
 13  After holidays     185507 non-null  object
dtypes: int64(3), object(11)
```

ภาพประกอบ 3 แสดงข้อมูลเบื้องต้นเกี่ยวกับ DataFrame

หลังจากนั้นแปลงประเภทข้อมูลให้เหมาะสมเพื่อให้เหมาะกับการนำข้อมูลไปใช้ในการทำแบบจำลองต่อไป โดยได้เปลี่ยนประเภทข้อมูลแปลงคอลัมน์ตัวเลขให้เป็นประเภทตัวเลข และจัดการกับค่าที่มีเครื่องหมายจุลภาค แปลงข้อมูลที่ไม่ได้ใช้ในด้านการคำนวนจากประเภท int เปลี่ยนเป็น object รวมทั้งมีการลบช่องว่างที่หัวคอลัมน์ บางครั้งชื่อคอลัมน์อาจมีช่องว่างที่ไม่จำเป็น จะได้ข้อมูลใหม่ที่มีการเปลี่ยนประเภทให้เหมาะสมต่อการใช้งานตามภาพประกอบ 4

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 185507 entries, 0 to 185506
Data columns (total 14 columns):
 #   Column             Non-Null Count   Dtype
---  ------             --------------   -----
 0   WH_Code            185507 non-null  object
 1   Delivery_Date      185507 non-null  datetime64[ns]
 2   Hour               185507 non-null  int64
 3   Sale_Quantity      185507 non-null  float64
 4   Delivery_Quantity  185507 non-null  float64
 5   Net_Value          185507 non-null  float64
 6   Month              185507 non-null  object
 7   Week               185507 non-null  object
 8   Day                185507 non-null  object
 9   Region             185507 non-null  object
 10  Last_Week          185507 non-null  object
 11  holidays           185507 non-null  object
 12  Before holidays    185507 non-null  object
 13  After holidays     185507 non-null  object
dtypes: datetime64[ns](1), float64(3), int64(1), object(9)
memory usage: 19.8+ MB
```

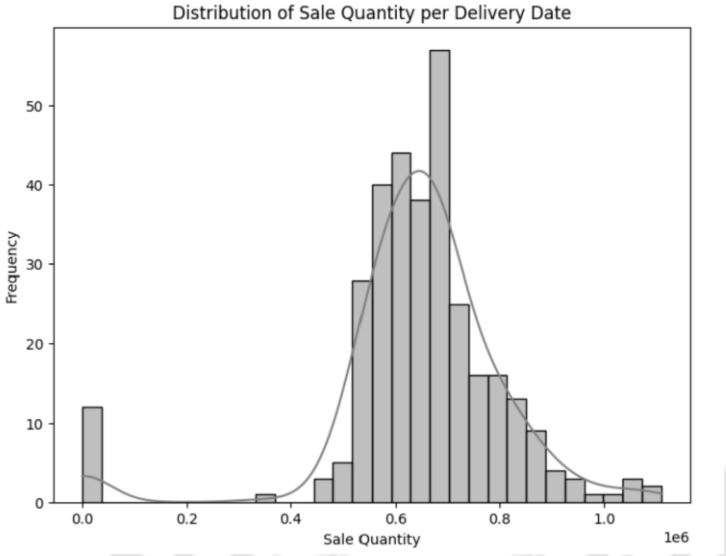
ภาพประกอบ 4 แสดงข้อมูลหลังจากมีการแปลงประเภทข้อมูลให้เหมาะสม

หลังจากแปลงประเภทของข้อมูลแล้ว ตรวจสอบว่ามีค่าว่างหรือไม่ตามภาพประกอบ5

```
Missing values in each column:
WH_Code             0
Delivery_Date       0
Hour                0
Sale_Quantity       0
Delivery_Quantity   0
Net_Value           0
Month               0
Week                0
Day                 0
Region              0
Last_Week           0
holidays            0
Before holidays     0
After holidays      0
dtype: int64
```

ภาพประกอบ 5 แสดงผลจากการตรวจสอบค่าว่างของข้อมูล

ซึ่งจากผลพบว่าข้อมูลชุดนี้ไม่มีค่าว่าง หลังจากนั้นจึงนำข้อมูลไปทำการ EDA เพื่อหาข้อมูลเชิงลึก ตามภาพประกอบที่ 6



ภาพประกอบ 6 แสดงการกระจายของปริมาณการขาย (Sale Quantity)

โดยภาพประกอบ 6 สรุปได้ว่า การกระจายของยอดขาย(Sale Quantity) มีแนวโน้มเป็นแจกแจงแบบเบ้ขวา (Right-skewed Distribution) ซึ่งข้อมูลที่มีการเบ้ขวา (Right-skewed) แสดงว่ายอดขายบางวันผิดปกติ มี(Outliers) ช่วงที่มียอดขายมากที่สุดอยู่ระหว่าง 500,000 - 700,000 หน่วย และจะเห็นว่ามีบางวันที่ยอดน้อยไปเลย โดยปกติยอดจะไม่ต่ำกว่า 400,000 หน่วย และมียอดที่มากกว่าปกติ โดยปกติจะไม่เกิน 1,000,000 หน่วย

หลังจากนั้นจึงได้ทำการตรวจสอบข้อมูลเพื่อดูข้อมูลในส่วนที่ยอดขายน้อยปิดปกติและมากผิดปกติ ตามภาพประกอบที่ 7 และ 8 ตามลำดับ

	Delivery_Date	Sale_Quantity
42	2023-08-20	20.0
109	2023-11-05	300.0
116	2023-11-12	12.0
135	2023-12-03	468.0
160	2023-12-31	25195.0
161	2024-01-03	348269.0
165	2024-01-07	42.0
172	2024-01-14	80.0
197	2024-02-11	1.0
216	2024-03-03	21.0
259	2024-04-21	710.0
266	2024-04-28	12.0
296	2024-06-02	12.0

'จำนวนวันที่มี Sale_Quantity น้อยกว่า 400,000: 13 วัน'

ภาพประกอบ 7แสดงวันที่ยอดขายน้อยผิดปกติ

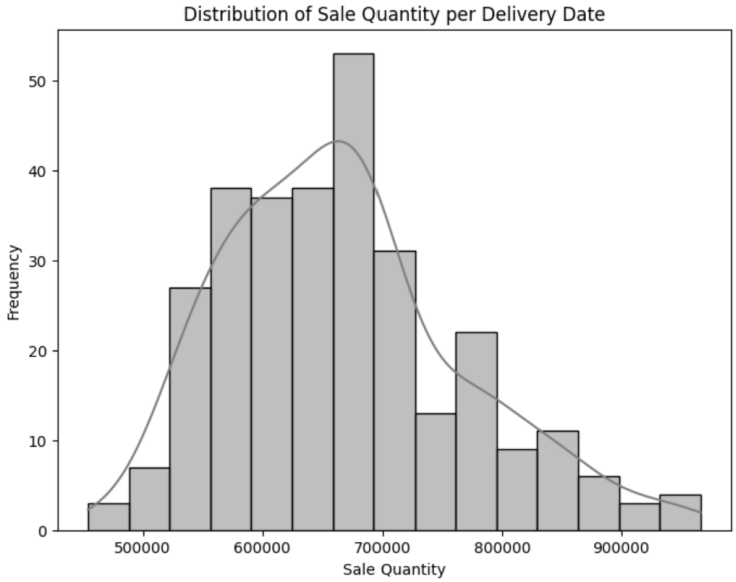
จากภาพประกอบ 7 พบว่า มี13 วันที่มียอดขายน้อยผิดปกติ คือน้อยกว่า 400,000 หน่วยต่อวัน จึงได้ทำการลบข้อมูลที่น้อยกว่าปกติออกจาก DataFrame เพื่อไม่ให้การนำข้อมูลไปใช้อาจส่งผลคลาดเคลื่อนได้

	Delivery_Date	Sale_Quantity
238	2024-04-09	1095717.0
239	2024-04-10	1062381.5
240	2024-04-11	1041307.0
256	2024-04-30	1043145.0
258	2024-05-03	1109506.0
261	2024-05-07	1021515.0

จำนวนวันที่มี Sale_Quantity มากกว่า 1,000,000: 6 วัน

ภาพประกอบ 8 แสดงวันที่ยอดขายมากผิดปกติ

จากภาพประกอบ 8 พบว่า มี 6 วันที่มียอดขายน้อยผิดปกติ คือมากกว่า 1,000,000 หน่วยต่อวัน จึงได้ทำการลบข้อมูลที่มากกว่าปกติออกจาก DataFrame หลังจากนั้นตรวจสอบการแจกแจงของปริมาณการขายอีกครั้งตามภาพประกอบ 9



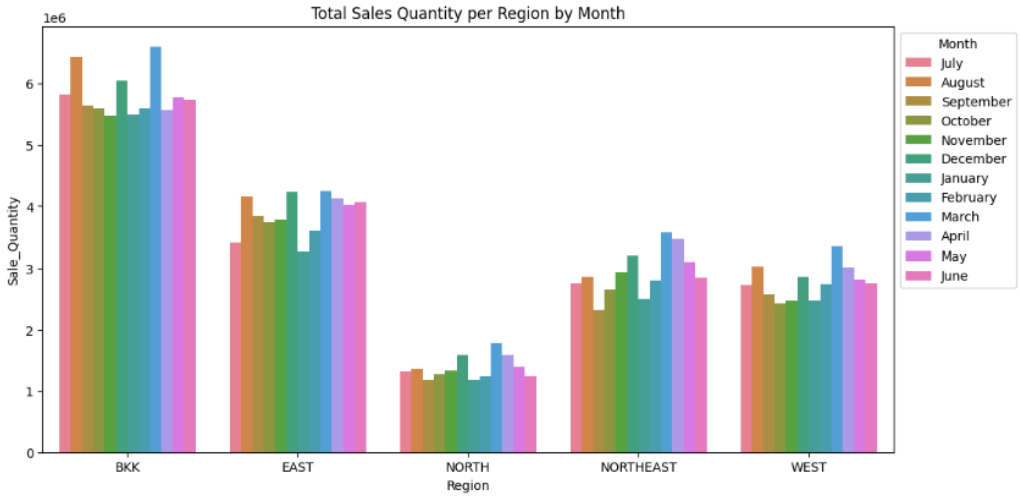
ภาพประกอบ 9 แสดงการกระจายของปริมาณการขาย (Sale Quantity)หลังลบข้อมูลที่ผิดปกติ

หลังจากนั้นได้ทำการวิเคราะห์ความสัมพันธ์ระหว่างตัวแปร โดยวิเคราะห์ความสัมพันธ์ระหว่างยอดขาย (Sale Quantity) และยอดส่งสินค้า (Delivery Quantity) ตามภาพประกอบ 10

ภาพประกอบ 10 แสดงความสัมพันธ์ระหว่าง ยอดขาย (Sale Quantity) กับ ยอดส่งมอบสินค้า (Delivery Quantity) ในแต่ละวัน (Delivery_Date)

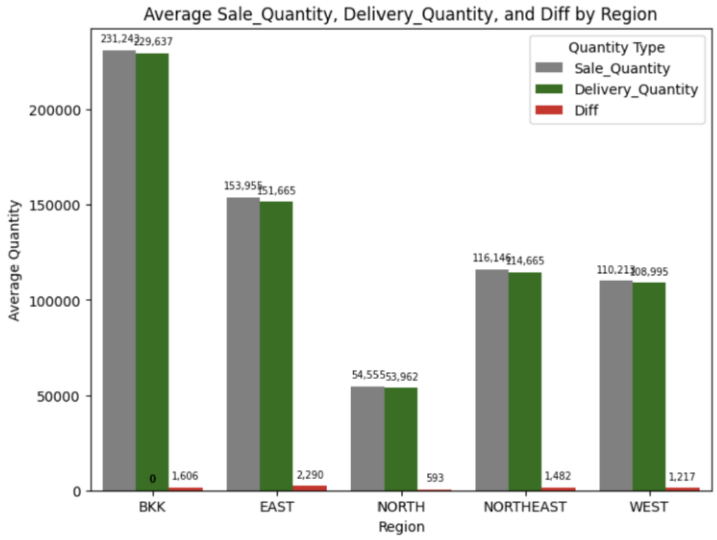
จากภาพประกอบที่ 10 พบว่าจุดข้อมูลเรียงตัวเกือบเป็นเส้นตรง ทำให้เห็นได้ว่า ยอดขาย (Sale Quantity) และยอดส่งมอบสินค้า (Delivery Quantity) มีความสัมพันธ์เชิงบวกสูง หมายความว่า เมื่อปริมาณยอดการขาย (Sale Quantity) เพิ่มขึ้น ปริมาณยอดการส่งมอบสินค้า (Delivery Quantity) ก็เพิ่มขึ้นด้วย จุดข้อมูลกระจุกตัวแน่นมาก และพบว่าส่วนใหญ่ของข้อมูลอยู่ในแนวเดียวกัน แสดงว่าการส่งสินค้าส่วนใหญ่ใกล้เคียงกับยอดขาย แต่อย่างไรก็ตามก็พบว่ามีบางวันที่ปริมาณยอดการส่งมอบสินค้าต่ำกว่าปริมาณยอดขายที่ลูกค้าสั่งซื้อเข้ามา

ต่อมาได้ทำการวิเคราะห์ความสัมพันธ์ระหว่างยอดขาย (Sale Quantity) กับภูมิภาค (Region) ตามภาพประกอบ 11 พบว่า ภูมิภาค กรุงเทพฯ(Region BKK) มีปริมาณยอดขายสูงสุดทุกเดือน ยอดขายมีความสม่ำเสมอ แต่มีบางเดือนที่มียอดขายสูงกว่าเดือนอื่นๆ เช่น สิงหาคม และ มีนาคม และ ภูมิภาคเหนือ (Region NORTH) มีปริมาณยอดขายต่ำกว่าภูมิภาคอื่น ยอดขายมีค่าน้อยที่สุดในทุกเดือนเมื่อเทียบกับภูมิภาคอื่น แต่ยังมีแนวโน้มเพิ่มขึ้นในบางเดือน



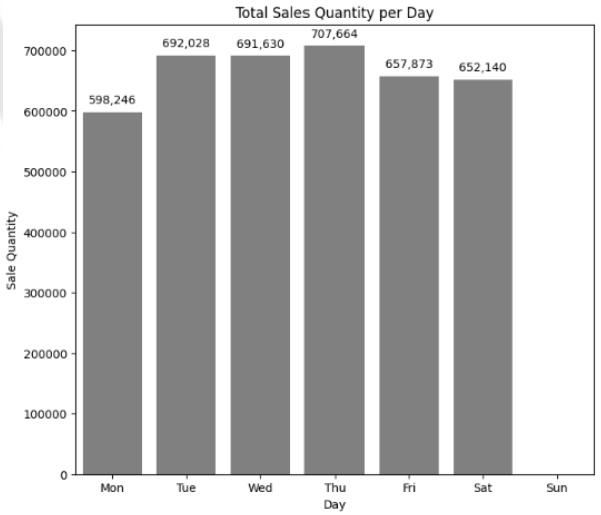
ภาพประกอบ 11 แสดงการแจกแจงปริมาณการขายตามภูมิภาค

เมื่อวิเคราะห์ต่อเพื่อดูปริมาณยอดขาย ยอดส่งมอบสินค้า และความต่างระหว่าง 2 ค่า พบว่า ในแต่ละภูมิภาคยังมีสินค้าที่ไม่สามารถส่งมอบให้ลูกค้าได้ตามที่ต้องการ



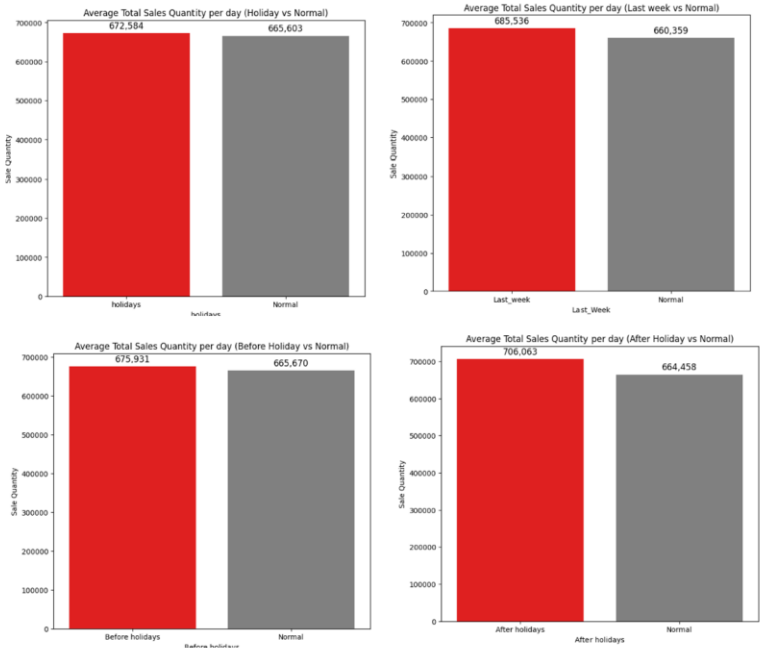
ภาพประกอบ 12 แสดงจำนวนยอดขาย,ยอดส่งมอบสินค้าและผลต่างของทั้งสองยอดเฉลี่ยรายวัน

เมื่อวิเคราะห์ยอดขายกับวันในสัปดาห์ พบว่ายอดขายสูงสุดในสัปดาห์เกิดขึ้นในวันพฤหัสบดี (Thu) และยอดขายต่ำสุดในวันจันทร์ (Mon) โดยรวมแล้ว ยอดขายในวันทำงาน (วันจันทร์ถึงวันศุกร์) สูงกว่าช่วงวันเสาร์เล็กน้อย ตามภาพประกอบที่ 13



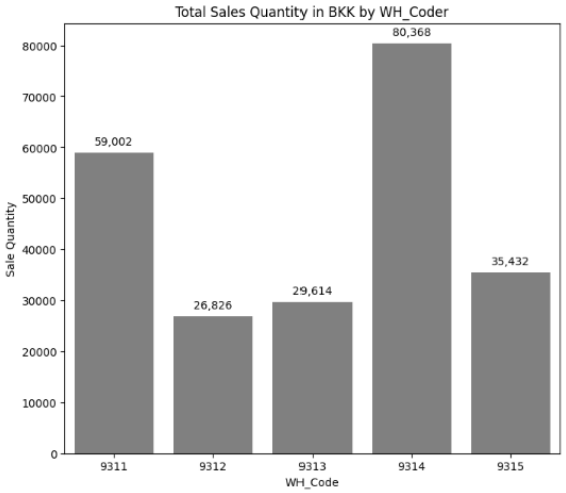
ภาพประกอบ 13 แสดงค่าเฉลี่ยยอดขายต่อวันในสัปดาห์

วิเคราะห์ต่อในมุมวันปกติ (Normal) เทียบกับวันหยุดนักขัตฤกษ์ (Holidays), วันในช่วงสัปดาห์สุดท้ายของเดือน (Last week), วันก่อนวันหยุดนักขัตฤกษ์ (Before holidays) และวันหลังวันหยุดนักขัตฤกษ์ (After holiday) พบว่า โดยรวมแล้วทุกวันต่างมียอดขายที่สูงกว่าช่วง วันปกติ (Normal) โดยวันหลังวันหยุด (After holiday) ที่มียอดค่าเฉลี่ยในวันนี้สูงกว่า วันปกติ(Normal) อย่างเห็นได้ชัด อาจเกิดจากพฤติกรรมของลูกค้าที่ กลับมาซื้อของหลังจากวันหยุดหรือในช่วงวันหยุดแล้วร้านค้าขายดีกว่าปกติทำให้สินค้าหมด จึงต้องมีการสั่งสินค้าในวันหลังวันหยุด ตามภาพประกอบ 14

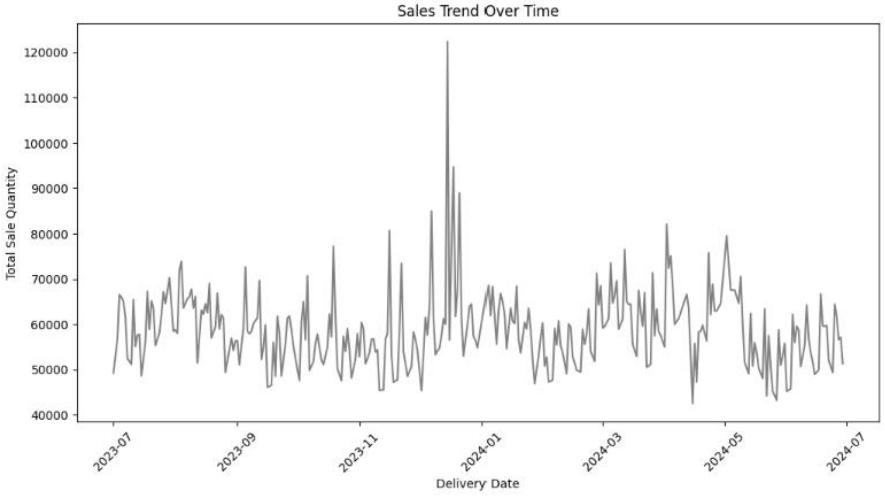


ภาพประกอบ 14 เปรียบเทียบยอดขายเฉลี่ยต่อวันระหว่างวันปกติกับวันหยุดนักขัตฤกษ์ (Holidays), วันในช่วงสัปดาห์สุดท้ายของเดือน (Last week), วันก่อนวันหยุดนักขัตฤกษ์ (Before holidays) และวันหลังวันหยุดนักขัตฤกษ์ (After holiday)

ทั้งนี้ในการทำวิจัยครั้งนี้ผู้วิจัยได้เลือกคลังสินค้าในภูมิภาคกรุงเทพ (Region BKK) จำนวน 1คลังสินค้า คือ คลังสินค้ารหัส 9311 (WH_Code 9311)มาทำแบบจำลองเนื่องจากถึงแม้ว่าจะมียอดขายสูงเป็นลำดับที่ 2 ของภูมิภาค ตามภาพประกอบที่ 15 แต่มีปัญหาภายในและบ่อยครั้งที่เกิดปัญหา ไม่สามารถส่งสินค้าได้ตามยอดขายจึงได้สนใจนำคลังสินค้านี้มาวิเคราะห์และทำแบบจำลองต่อไป



ภาพประกอบ 15 ยอดขาย(Sale Quantity) โดยเฉลี่ยต่อวันของคลังสินค้าในภูมิภาคกรุงเทพ

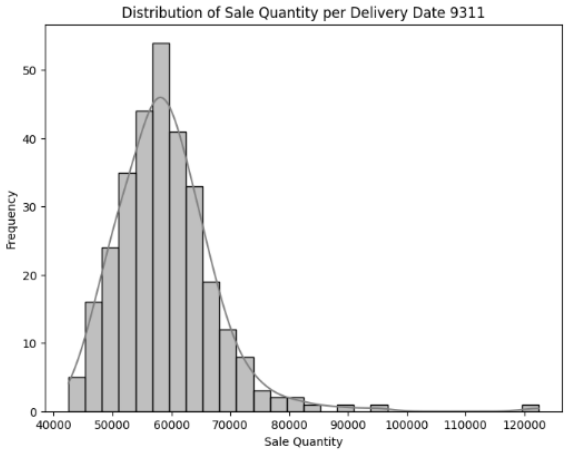


ภาพประกอบ 16 ยอดขาย(Sale Quantity) โดยเฉลี่ยต่อวันของคลังสินค้ารหัส 9311

จากภาพประกอบ 16 จะเห็นได้ว่ายอดค่าของคลังสินค้ารหัส 9311 ยอดขายมีการแปรปรวนสูง โดยเฉพาะในบางช่วงเวลา เช่นในช่วง เดือนตุลาคม 2023 (2023-10) และ มกราคม 2024 (2024-01) ซึ่งยอดขายมีการพุ่งสูงขึ้นอย่างมากในช่วงนั้น

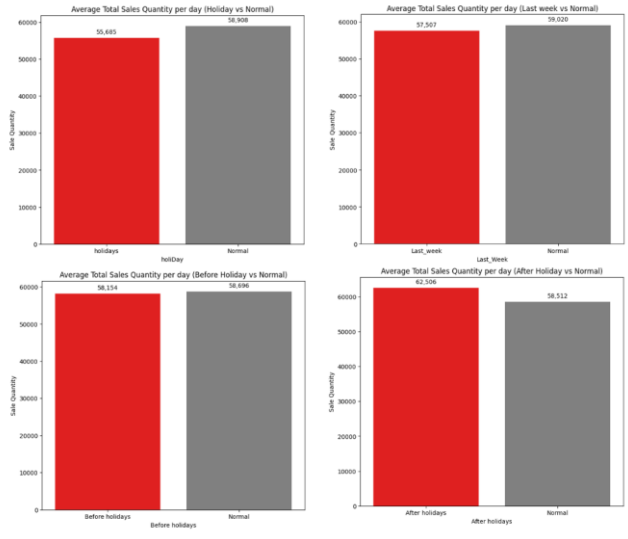
หลังจากนั้นได้ดูกระจายตัวของยอดขายตามภาพประกอบ 17 พบว่าการกระจายของยอดขายที่ เบ้ไปทางขวา ซึ่งแสดงให้เห็นว่าข้อมูลส่วนใหญ่กระจุกอยู่ในช่วงยอดขายที่ไม่สูงมาก แต่ก็มีบางครั้งที่ยอดขายพุ่งสูงขึ้นอย่างผิดปกติ ทำให้กราฟนี้แสดงลักษณะการกระจายแบบ Skewed Right ซึ่งหมายความว่า ยอดขายส่วนใหญ่จะกระจุกตัวอยู่ในช่วงที่ค่อนข้างต่ำ แต่มีการ

กระจายไปทางขวามากขึ้น (มีค่าตัวเลขสูงในบางจุด) โดยมีบางช่วงที่มียอดขายสูงผิดปกติ (outliers)



ภาพประกอบ 17 แสดงการกระจายของปริมาณการขาย (Sale Quantity) คลังสินค้า รหัส 9311

ทั้งนี้โดยปกติแล้วยอดขายเฉลี่ยต่อวันของคลังสินค้า รหัส 9311 จะไม่ต่ำกว่า 40,000 หน่วย และมาสูงกว่า 90,000 หน่วย จึงได้ทำการตัดข้อมูลที่น้อยกว่าและสูงกว่าค่าปกติออกจากข้อมูล



ภาพประกอบ 18 เปรียบเทียบยอดขายเฉลี่ยต่อวันระหว่างวันปกติกับวันหยุดนักขัตฤกษ์ (Holidays), วันในช่วงสัปดาห์สุดท้ายของเดือน (Last week), วันก่อนวันหยุดนักขัตฤกษ์ (Before holidays) และวันหลังวันหยุดนักขัตฤกษ์ (After holiday) ของคลังสินค้า รหัส 9311

จากภาพประกอบที่ 18 พบว่ายอดขายเฉลี่ยต่อวันของคลังสินค้า รหัส 9311 ของวันหลังวันหยุดนักขัตฤกษ์ (After holiday) สูงกว่าวันปกติ (Normal) อย่างเห็นได้ชัด ในขณะที่วันอื่นๆเมื่อเทียบกับวันปกตินั้น พบว่าวันปกติมีค่าเฉลี่ยของยอดขายสูงกว่า

3.4 การสร้างแบบจำลอง

หลังจากทำ EDA เสร็จแล้ว ผู้วิจัยได้แปลงข้อมูลเชิงคุณภาพ (categorical data) ให้เป็นข้อมูลเชิงปริมาณ (numeric data) ตามภาพประกอบที่ 19 เพื่อใช้ในการวิเคราะห์ข้อมูลและทำแบบจำลอง ซึ่งมักจะต้องการข้อมูลในรูปแบบตัวเลข (0 และ 1) เพื่อการประมวลผลที่ง่ายขึ้น เช่น ในการฝึกฝนแบบจำลองที่ต้องการข้อมูลที่เป็นตัวเลข โดยแปลงค่าในคอลัมน์ Last_Week, holidays, Before holidays, After holidays ให้เป็นตัวเลข 0 กับ 1 โดยกำหนดให้วันปกติจะมีค่าเป็น 0 ส่วนวันหยุดนักขัตฤกษ์ (Holidays), วันในช่วงสัปดาห์สุดท้ายของเดือน (Last week), วันก่อนวันหยุดนักขัตฤกษ์ (Before holidays) และวันหลังวันหยุดนักขัตฤกษ์ (After holiday) ให้มีค่าเป็น 1

	WH_Code	Delivery_Date	Hour	Sale_Quantity	Delivery_Quantity	Net_Value	Month	Week	Day	Region	Last_Week	holidays	Before holidays	After holidays
0	9311	2023-07-01	7	948.0	948.0	99906.0	July	26	Sat	BKK	0	0	0	0
1	9311	2023-07-01	8	7635.0	7616.0	1447336.0	July	26	Sat	BKK	0	0	0	0
2	9311	2023-07-01	9	11860.0	11855.0	1979219.0	July	26	Sat	BKK	0	0	0	0
3	9311	2023-07-01	10	3828.0	3828.0	734497.0	July	26	Sat	BKK	0	0	0	0
4	9311	2023-07-01	11	4344.0	4190.0	844955.0	July	26	Sat	BKK	0	0	0	0

ภาพประกอบ 19 DataFrame ของยอดขาย คลังสินค้า รหัส9311 ที่มีการแปลงข้อมูลเชิงคุณภาพ (categorical data) ให้เป็น ข้อมูลเชิงปริมาณ (numeric data)

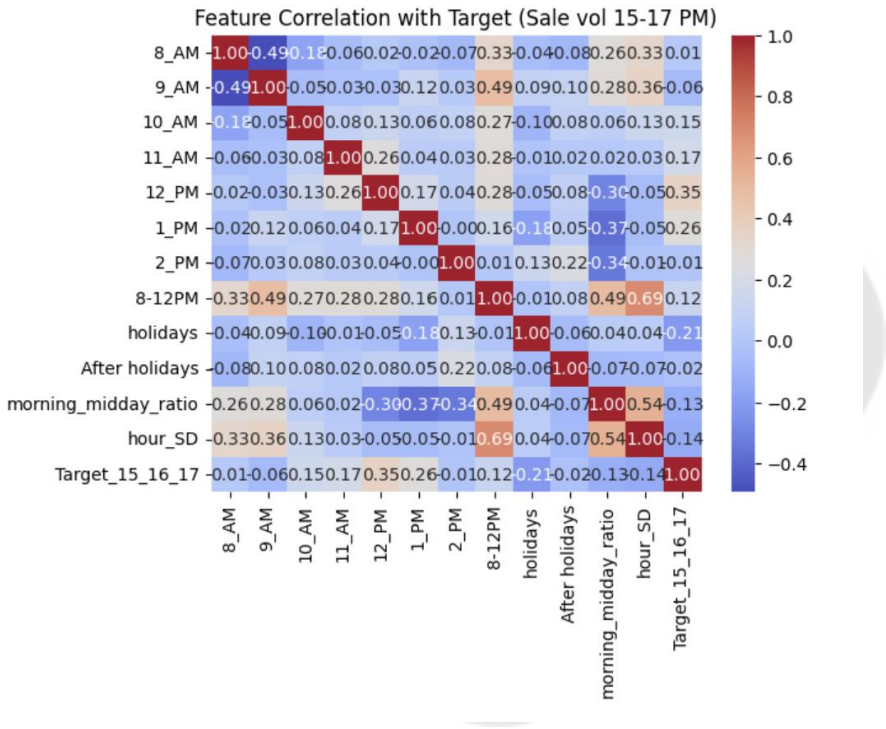
หลังจากนั้นได้สร้างฟีเจอร์ใหม่เพิ่มอีก 3 ฟีเจอร์ โดยฟีเจอร์แรก คือ ยอดขายเช้าถึงเที่ยง (8-12 PM) มาจากการนำยอดขายตั้งแต่ช่วงเวลา แปดโมงจนถึงเที่ยง มาหาผลรวมเป็นอีกฟีเจอร์ เนื่องจากประเมินว่ายอดขายรวมของช่วงเวลาเหล่านั้นจะส่งผลต่อยอดขายในช่วงเป้าหมายที่จะพยากรณ์ คือ ยอดขายรวมของช่วงเวลา บ่ายสามถึงห้าโมงเย็น ต่อมาฟีเจอร์ที่สอง อัตราส่วนระหว่างยอดขายช่วงเช้ากับช่วงเที่ยง (Morning_Midday_Ratio) และฟีเจอร์ที่สาม ความแปรปรวนของยอดขาย (Hour_SD) เพื่อดูว่าความแปรปรวนที่เกิดขึ้นจะมีผลต่อเป้าหมายที่จะพยากรณ์หรือไม่

หลังจากนั้นตัดข้อมูลคอลัมน์ที่ไม่ใช้ออกให้เหลือแต่ข้อมูลที่จะนำไปทำแบบจำลองต่อไป ได้ข้อมูล Dataframe ตามภาพประกอบ 20

8_AM	9_AM	10_AM	11_AM	12_PM	1_PM	2_PM	8-12PM	holidays	After holidays	morning_midday_ratio	hour_SD	Target_15_16_17
7635	11860	3828	4344	2335	3904	3604	30002	0	0	2.810830	3295.43	10704.0
4452	11448	3150	4251	3249	5396	7014	26550	0	0	1.488026	2912.60	14653.0
12963	3516	3587	5115	4727	5592	6317	29908	0	0	1.513645	3244.02	22206.0
4314	19800	6482	4739	3476	6683	2738	38811	0	0	2.739784	5874.16	15889.0
9197	10029	4219	5208	4235	4756	3975	32888	0	0	2.209857	2549.36	20984.0

ภาพประกอบ 20 ตัวอย่างข้อมูลในDataframe ที่จะใช้ทำแบบจำลองในการพยากรณ์

เมื่อได้ข้อมูลที่จะนำมาทำแบบจำลองได้ ผู้วิจัยได้ทำการหาความสัมพันธ์ (correlation) ระหว่างฟีเจอร์ต่างๆ กับ ยอดขายเป้าหมาย (Target_15_16_17) ตามภาพประกอบที่ 21



ภาพประกอบ 21 แสดง Heatmap ของ ความสัมพันธ์ (correlation) ระหว่างฟีเจอร์ต่างๆ กับ ยอดขายเป้าหมาย (Target_15_16_17)

จากภาพประกอบที่ 21 จะเห็นได้ว่า เวลาช่วงเที่ยง (12_PM) กับ Target_15_16_17 ความสัมพันธ์ 0.35 แสดงว่ามีความสัมพันธ์กันที่ดีจะส่งผลในเชิงบวกต่อยอดขายเป้าหมาย ในขณะที่วันหยุด (holidays) มีความสัมพันธ์ที่ต่ำกับ Target_15_16_17 อยู่ที่ -0.21 ซึ่งแสดงว่าวันหยุด ไม่ได้มีผลมากกับยอดขายเป้าหมาย

หลังจากนั้นได้แบ่งข้อมูลออกเป็น ชุดฝึกสอน (Training Data) และชุดทดสอบ (Testing Data) โดยใช้ฟังก์ชัน train_test_split และใช้ test_size=0.3 ซึ่งหมายความว่า 30% ของข้อมูลจะถูกใช้เป็นข้อมูลทดสอบ (Testing Data) และที่เหลืออีก 70% จะถูกใช้เป็นข้อมูลฝึกสอน (Training Data) ซึ่งหลังจากการแบ่งข้อมูลจะได้ ขนาดของชุดข้อมูลฝึกสอน (Training Data) ที่มี 210 ตัวอย่าง (rows) และ ขนาดของชุดข้อมูลทดสอบ (Testing Data) ที่มี 90 ตัวอย่าง (rows) ตามภาพประกอบที่ 22

```
# split train test

import sklearn
from sklearn.model_selection import train_test_split

X= df.drop('Target_15_16_17',axis=1)
y= df['Target_15_16_17']

X_train, X_test,y_train,y_test = train_test_split(X,y,test_size=0.3,random_state=42)
```

```
X_train.shape
```

```
(210, 12)
```

```
print(X_train.shape)
print(X_test.shape)
print(y_train.shape)
print(y_test.shape)
```

```
(210, 12)
(90, 12)
(210,)
(90,)
```

ภาพประกอบ 22 แสดงแบ่งข้อมูลออกเป็น ชุดฝึกสอน (Training Data) และชุดทดสอบ (Testing Data)

ในการศึกษาครั้งนี้ ผู้วิจัยได้วางแผนทำการทดลองการใช้แบบจำลอง ออกเป็น 6 การทดลอง โดยมีรายละเอียดดังนี้

การทดลองที่ 1 ใช้ความรู้จากยอดขายช่วงเวลาแปดโมงจนถึงบ่ายสอง เพื่อทำนายยอดขายช่วงบ่ายสามจนถึงห้าโมงเย็น โดยการเพิ่มฟีเจอร์ยอดขายเช้าถึงเที่ยง (8-12 PM) วันหยุดนักขัตฤกษ์ (Holiday) และ ฟีเจอร์วันหลังวันหยุดนักขัตฤกษ์ (After holiday)

```
X_train_1.head()
```

	8_AM	9_AM	10_AM	11_AM	12_PM	1_PM	2_PM	8-12PM	holidays	After holidays
194	10301	4441	7672	4399	3436	5398	3308	30249	0	0
101	7723	9585	2933	2649	3095	2916	4163	25985	0	0
68	11198	3179	4418	3664	3001	3614	2254	25460	0	0
224	13100	8169	5483	5221	5452	8259	3470	37425	0	0
37	3456	17610	7083	4638	3470	4589	5403	36257	0	0

Next steps: Generate code with X_train_1 | View recommended plots | New interactive sheet

```
[ ] X_test_1.head()
```

	8_AM	9_AM	10_AM	11_AM	12_PM	1_PM	2_PM	8-12PM	holidays	After holidays
203	17101	3809	4583	6368	2951	4066	2281	34812	0	0
266	13870	2660	2901	3152	2955	3579	1818	25538	1	0
152	13612	5805	3946	4677	3682	4146	3720	31722	1	0
9	5934	14379	4641	4043	3082	5073	4724	32079	0	0
233	10106	11500	3952	4378	3939	7957	2552	33875	0	0

Next steps: Generate code with X_test_1 | View recommended plots | New interactive sheet

ภาพประกอบ 23 แสดงข้อมูล X_train และ X_test ของการทดลองที่ 1

การทดลองที่ 2 ใช้ความรู้จากยอดขายช่วงเวลาแปดโมงจนถึงบ่ายสอง เพื่อทำนายยอดขายช่วงบ่ายสามจนถึงห้าโมงเย็น โดยการเพิ่มฟีเจอร์ ยอดขายเช้าถึงเที่ยง (8-12 PM) อัตราส่วนระหว่างยอดขายช่วงเช้ากับช่วงเที่ยง (Morning_Midday_Ratio) และ ความแปรปรวนของยอดขาย (Hour_SD)

```
[ ] X_train_2.head()
```

	8_AM	9_AM	10_AM	11_AM	12_PM	1_PM	2_PM	8-12PM	morning_midday_ratio	hour_SD
194	10301	4441	7672	4399	3436	5398	3308	30249	2.208285	2554.30
101	7723	9585	2933	2649	3095	2916	4163	25985	2.249853	2780.10
68	11198	3179	4418	3664	3001	3614	2254	25460	2.532304	3038.26
224	13100	8169	5483	5221	5452	8259	3470	37425	1.860951	3174.94
37	3456	17610	7083	4638	3470	4589	5403	36257	2.435522	5007.68

Next steps: Generate code with X_train_2 | View recommended plots | New interactive sheet

```
[ ] X_test_2.head()
```

	8_AM	9_AM	10_AM	11_AM	12_PM	1_PM	2_PM	8-12PM	morning_midday_ratio	hour_SD
203	17101	3809	4583	6368	2951	4066	2281	34812	3.426651	5114.40
266	13870	2660	2901	3152	2955	3579	1818	25538	2.703903	4201.97
152	13612	5805	3946	4677	3682	4146	3720	31722	2.428126	3585.33
9	5934	14379	4641	4043	3082	5073	4724	32079	2.251495	3805.45
233	10106	11500	3952	4378	3939	7957	2552	33875	2.071982	3490.20

Next steps: Generate code with X_test_2 | View recommended plots | New interactive sheet

ภาพประกอบ 24 แสดงข้อมูล X_train และ X_test ของการทดลองที่ 2

การทดลองที่ 3 ใช้ความรู้จากยอดขายช่วงเวลาแปดโมงจนถึงบ่ายสอง เพื่อทำนายยอดขายช่วงบ่ายสามจนถึงห้าโมงเย็น โดยรวมทุก feature จากการทดลองที่ 1 และ 2

```
[ ] X_train_3.head()
```

	8_AM	9_AM	10_AM	11_AM	12_PM	1_PM	2_PM	8-12PM	holidays	After holidays	morning_midday_ratio	hour_SD
194	10301	4441	7672	4399	3436	5398	3308	30249	0	0	2.208285	2554.30
101	7723	9585	2933	2649	3095	2916	4163	25985	0	0	2.249853	2780.10
68	11198	3179	4418	3664	3001	3614	2254	25460	0	0	2.532304	3038.26
224	13100	8169	5483	5221	5452	8259	3470	37425	0	0	1.860951	3174.94
37	3456	17610	7083	4638	3470	4589	5403	36257	0	0	2.435522	5007.68

Next steps: Generate code with X_train_3 | View recommended plots | New interactive sheet

```
X_test_3.head()
```

	8_AM	9_AM	10_AM	11_AM	12_PM	1_PM	2_PM	8-12PM	holidays	After holidays	morning_midday_ratio	hour_SD
203	17101	3809	4583	6368	2951	4066	2281	34812	0	0	3.426651	5114.40
266	13870	2660	2901	3152	2955	3579	1818	25538	1	0	2.703903	4201.97
152	13612	5805	3946	4677	3682	4146	3720	31722	1	0	2.428126	3585.33
9	5934	14379	4641	4043	3082	5073	4724	32079	0	0	2.251495	3805.45
233	10106	11500	3952	4378	3939	7957	2552	33875	0	0	2.071982	3490.20

Next steps: Generate code with X_test_3 | View recommended plots | New interactive sheet

ภาพประกอบ 25 แสดงข้อมูล X_train และ X_test ของการทดลองที่ 3

การทดลองที่ 4 เลือกฟีเจอร์ที่สำคัญ (Feature Importance) 4 ฟีเจอร์ มาเป็นฟีเจอร์ที่ใช้ในการทำนาย

การทดลองที่ 5 นำแบบจำลองที่มีผลการทดลองที่ดีมาปรับพารามิเตอร์เพิ่ม

การทดลองที่ 6 ทดลองโดยใช้วิธีการแบบดั้งเดิม (traditional approach)

ทั้งนี้ในส่วนของการทดลองที่ 1-4 ทางผู้วิจัยได้ใช้แบบจำลองทั้งหมด 7 แบบจำลองในการทดลอง เพื่อหาเปรียบเทียบผลการทดลองที่ดีที่สุด โดยแบบจำลองที่ได้นำมาใช้ในการทำมีดังนี้

1. แบบจำลอง Random Forest Regressor
2. แบบจำลอง XGBoost Regressor
3. แบบจำลอง Gradient Boosting
4. แบบจำลอง Linear Regression
5. แบบจำลอง Ridge Regression
6. แบบจำลอง Lasso Regression
7. แบบจำลอง Elastic Net Regression

แบบจำลองที่ 5 หลังจากได้ผลการทดลอง ทางผู้วิจัยได้เลือกผลการทดลองที่ดีลำดับที่ 2 มาปรับค่าพารามิเตอร์เพื่อทำการทดลองอีกครั้ง และการทดลองสุดท้ายการทดลองที่ 6 ทดลองโดยใช้วิธีการแบบดั้งเดิม โดยได้ใช้วิธีที่ผู้วิจัยได้ทำอยู่ในปัจจุบันคือ การ Moving Average ใน Microsoft Excel

3.5 การประเมินประสิทธิภาพ

ในการวิจัยนี้ผู้วิจัย ได้ใช้ค่า R^2 (R-squared), MAE (Mean Absolute Error), และ MSE (Mean Squared Error) ในการประเมินประสิทธิภาพของแบบจำลองการทำนาย เพื่อเปรียบเทียบผลการทดลองทั้ง 6 การทดลอง ว่าแบบจำลองไหนจะมีประสิทธิภาพสูงสุด

R^2 ใช้เพื่อประเมินว่าผลลัพธ์ที่ได้จากแบบจำลองสามารถอธิบายข้อมูลได้ดีแค่ไหน (ค่ามากแสดงว่าแบบจำลองดี)

MAE เป็นค่าที่บอกถึงความผิดพลาดเฉลี่ยในรูปของจำนวนที่ไม่ใช่กำลังสอง (ค่าต่ำแสดงว่าแบบจำลองแม่นยำ) ในการวิจัยนี้ MAE มีหน่วยเป็นลัง

MSE เป็นการให้ความสำคัญกับข้อผิดพลาดที่มีค่ามาก ซึ่งจะช่วยให้เห็นการผิดพลาดในกรณีที่แบบจำลองทำนายได้แย่ (ค่าต่ำแสดงว่าแบบจำลองดี) ในการวิจัยนี้ MSE มีหน่วยเป็น ลัง^2

เมื่อใช้ทั้ง 3 ค่านี้ร่วมกัน จะช่วยให้สามารถประเมินประสิทธิภาพของแบบจำลองได้อย่างมีประสิทธิภาพและครอบคลุม

บทที่ 4
ผลการดำเนินการวิจัย

ในการวิจัยเพื่อศึกษาแบบจำลองที่ช่วยในการพยากรณ์ยอดขาย ได้ใช้ข้อมูลข้อมูลยอดขายรายชั่วโมงที่ได้จากฐานข้อมูลของธุรกิจ มาเป็นข้อมูลในการทำแบบจำลองที่ใช้เทคนิคการเรียนรู้ของเครื่อง เพื่อทดลองหาแบบจำลองที่จะสามารถพยากรณ์ได้มีประสิทธิภาพสูงสุด ผู้วิจัยได้ทำการวิจัยตามกระการดำเนินการวิจัยที่ได้ออกแบบไว้ ตลอดจนได้ทำการวัดประสิทธิภาพของแบบจำลอง เพื่อให้ได้ผลบรรลุตามความมุ่งหมายของงานวิจัยที่ได้กำหนดไว้ ได้ดังนี้

1. ผลลัพธ์ของการทดลองที่ 1 ใช้ความรู้จากยอดขายช่วงเวลาแปดโมงจนถึงบ่ายสอง เพื่อทำนายยอดขายช่วงบ่ายสามจนถึงห้าโมงเย็น โดยการเพิ่มฟีเจอร์ยอดขายเช้าถึงเที่ยง (8-12 PM) วันหยุดนักขัตฤกษ์ (Holiday) และ ฟีเจอร์วันหลังวันหยุดนักขัตฤกษ์ (After holiday)

2. ผลลัพธ์ของการทดลองที่ 2 ใช้ความรู้จากยอดขายช่วงเวลาแปดโมงจนถึงบ่ายสอง เพื่อทำนายยอดขายช่วงบ่ายสามจนถึงห้าโมงเย็น โดยการเพิ่มฟีเจอร์ ยอดขายในช่วงเช้าจนถึงเที่ยง (8-12 PM) อัตราส่วนระหว่างยอดขายช่วงเช้ากับช่วงเที่ยง (morning_midday_ratio) และ ความแปรปรวนของยอดขาย (Hour_SD)

3. ผลลัพธ์การทดลองที่ 3 ใช้ความรู้จากยอดขายช่วงเวลาแปดโมงจนถึงบ่ายสอง เพื่อทำนายยอดขายช่วงบ่ายสามจนถึงห้าโมงเย็น โดยรวมทุก feature จากการทดลองที่ 1 และ 2

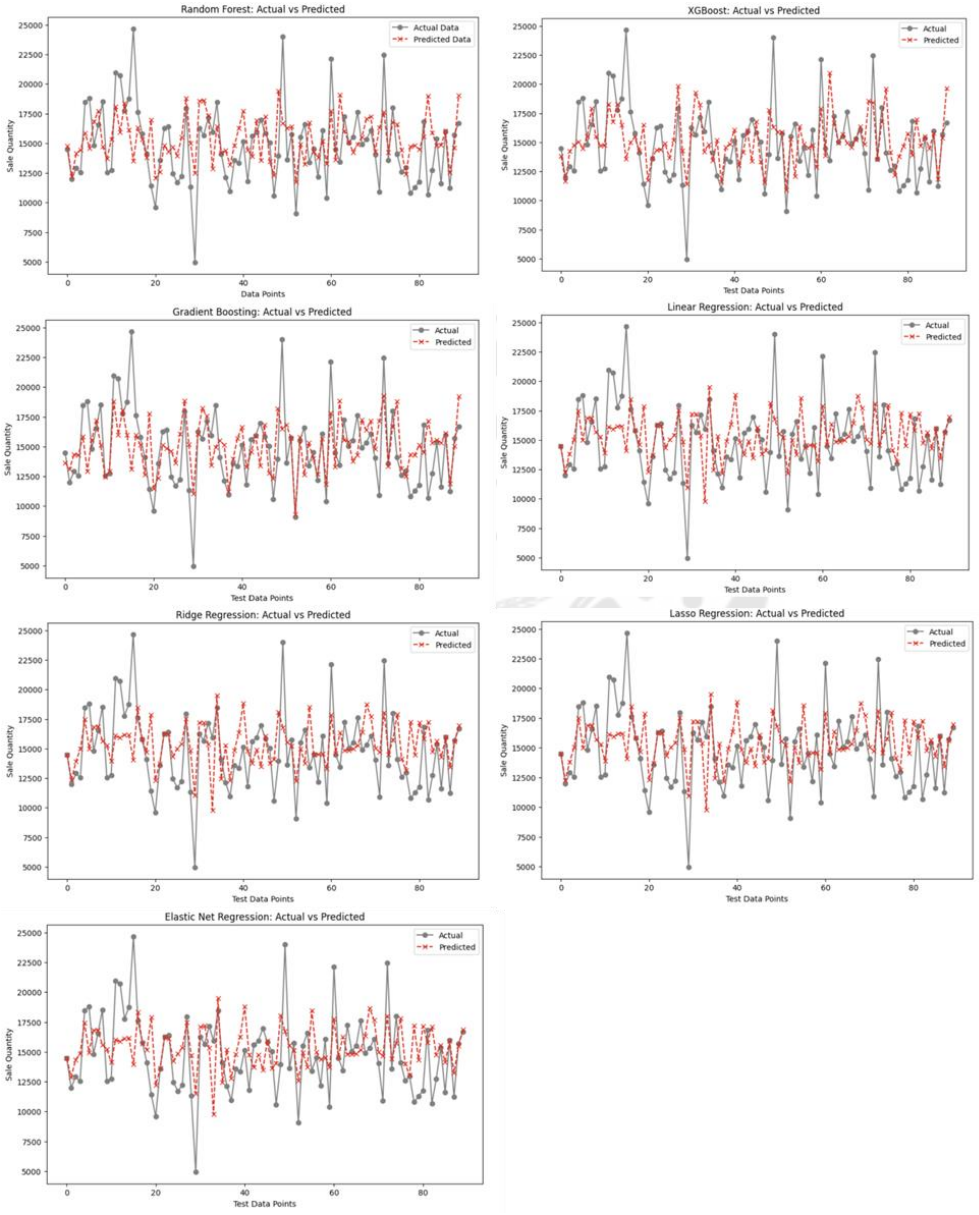
4. ผลลัพธ์การทดลองที่ 4 ที่ได้เลือกฟีเจอร์ที่สำคัญ (Feature Importance) 4 ฟีเจอร์ มาเป็นฟีเจอร์ที่ใช้ในการทำนาย

5. ผลลัพธ์การทดลองที่ 5 การนำแบบจำลองที่มีผลการทดลองที่ดีมาปรับพารามิเตอร์เพิ่ม

6. ผลลัพธ์การทดลองที่ 6 ทดลองโดยใช้วิธีการแบบดั้งเดิม (Traditional Approach)

ผลลัพธ์ของการทดลองที่ 1

จากการสร้างแบบจำลองการพยากรณ์ยอดขาย โดยใช้ความรู้จากยอดขายช่วงเวลาแปดโมงจนถึงบ่ายสอง เพื่อทำนายยอดขายช่วงบ่ายสามจนถึงห้าโมงเย็น โดยการเพิ่มฟีเจอร์ยอดขายเช้าถึงเที่ยง(8-12 PM) วันหยุดนักขัตฤกษ์ (Holiday)และฟีเจอร์วันหลังวันหยุดนักขัตฤกษ์ (After holiday) ซึ่งได้ทำการสร้างแบบจำลอง ทั้งหมด 7 แบบจำลอง ตามภาพประกอบ26 สามารถสรุปผลการทดลอง แสดงได้ตามตารางที่ 2



ภาพประกอบ 26 ผลของการแบบจำลองการทดลองที่ 1

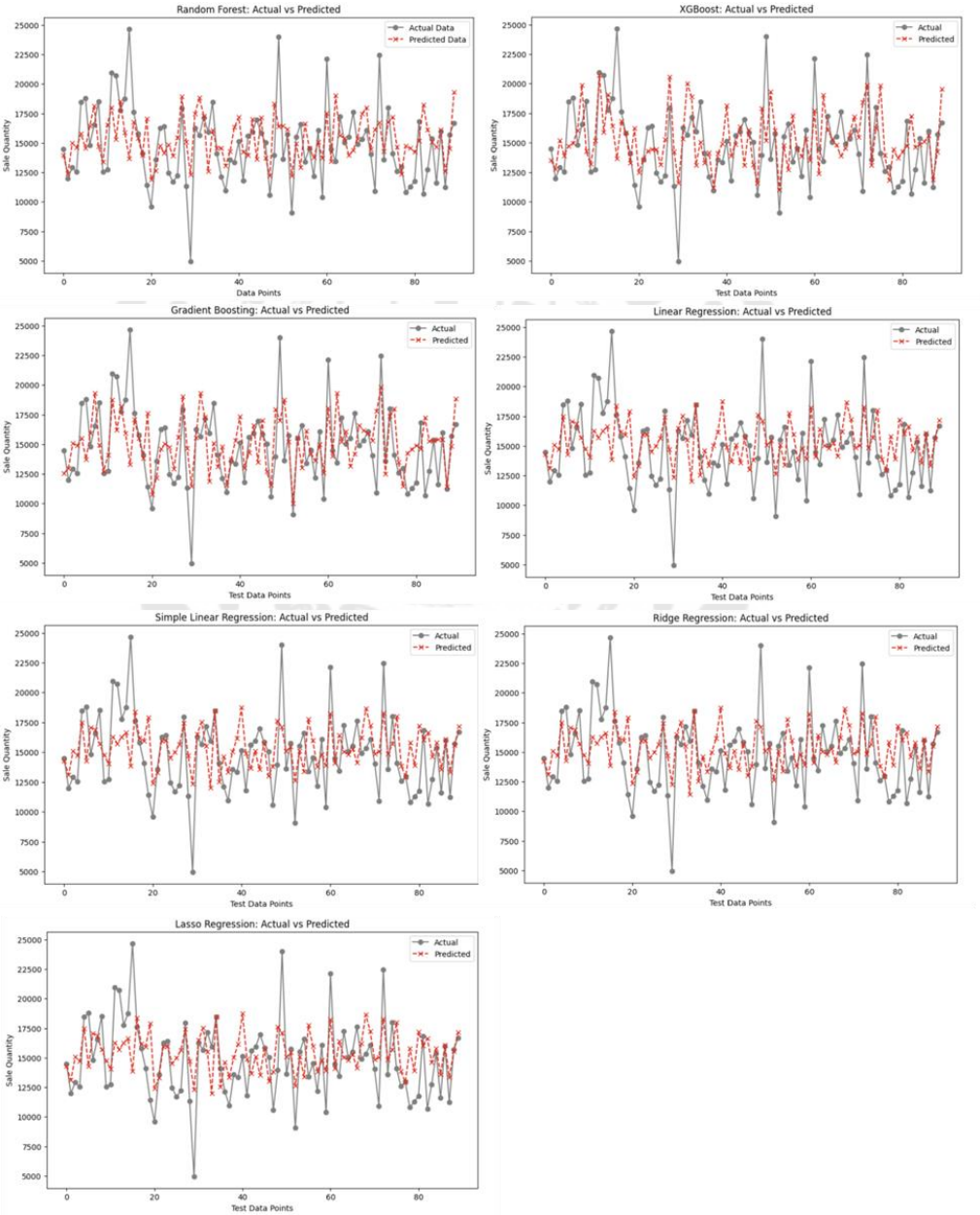
ตาราง 2 แสดงผลการประเมินประสิทธิภาพแบบจำลองของการทดลองที่ 1

แบบจำลอง	R^2	MAE (**ลัง**)	MSE (**ลัง²**)
Random Forest Regressor	0.122	2366.817	9355118.823
XGBoost Regressor	0.148	2224.748	**9072139.200**
Gradient Boosting	**0.167**	**2210.102**	9404857.126
Linear Regression	0.117	2343.100	9404857.126
Ridge Regression	0.116	2345.693	9421283.940
Lasso Regression	0.117	2343.146	9405123.417
Elastic Net Regression	0.104	2373.581	9544002.450

จากผลลัพธ์ของการประเมินประสิทธิภาพของการทดลองที่1 พบว่า แบบจำลอง Gradient Boosting เป็นแบบจำลองที่ดีที่สุด โดยมีค่า R^2 สูงที่สุด คือ 0.167 และมีค่า MAE ต่ำที่สุด คือ 2210.102 ลัง ในขณะที่แบบจำลอง XGBoost Regressor จะมีค่า MSE ต่ำที่สุด คือ 9072139.200 ลัง² เมื่อเทียบกับแบบจำลองอื่นๆ

ผลลัพธ์ของการทดลองที่ 2

จากการสร้างแบบจำลองเพื่อใช้ในการพยากรณ์ยอดขาย ใช้ความรู้จากยอดขายช่วงเวลาแปดโมงจนถึงบ่ายสอง เพื่อทำนายยอดขายช่วงบ่ายสามจนถึงห้าโมงเย็น โดยการเพิ่มฟีเจอร์ยอดขายเช้าถึงเที่ยง (8-12 PM) อัตราส่วนระหว่างยอดขายช่วงเช้ากับช่วงเที่ยง (Morning_Midday_Ratio) และ ความแปรปรวนของยอดขาย (Hour_SD)ได้ทำการสร้างแบบจำลอง ทั้งหมด 7 แบบจำลอง ตามภาพประกอบ 27 สามารถสรุปผลการทดลอง แสดงได้ตามตารางที่ 3



ภาพประกอบ 27 ผลของการแบบจำลองการทดลองที่ 2

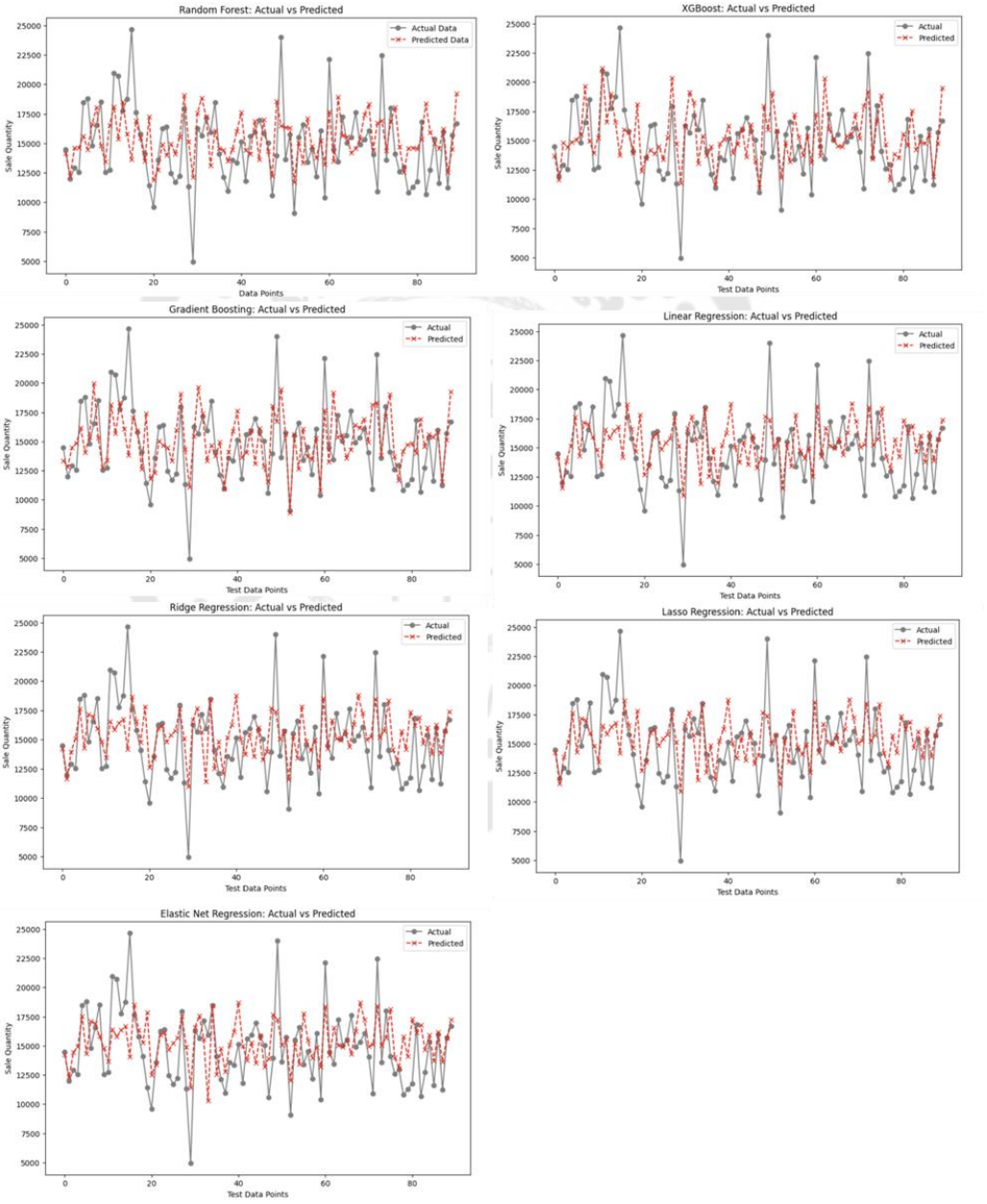
ตาราง 3 แสดงผลการประเมินประสิทธิภาพแบบจำลองของการทดลองที่ 2

แบบจำลอง	R^2	**MAE (ลัง)**	**MSE (ลัง²)**
Random Forest Regressor	0.100	2428.219	9584095.475
XGBoost Regressor	0.089	2369.292	9706091.479
Gradient Boosting	**0.148**	**2301.420**	9079399.868
Linear Regression	**0.148**	2330.550	**9073450.101**
Ridge Regression	0.144	2337.281	9121663.329
Lasso Regression	0.147	2331.150	9077401.270
Elastic Net Regression	0.132	2350.469	9240654.525

จากผลลัพธ์ของการประเมินประสิทธิภาพของการทดลองที่ 2 พบว่าแบบจำลอง Gradient Boosting และ แบบจำลอง Linear Regression เป็นแบบจำลองที่ดีที่สุด โดยแบบจำลอง Gradient Boosting มีค่า R^2 สูงที่สุด คือ 0.148 และมีค่า MAE ต่ำที่สุด คือ 2301.420 ลัง ในขณะที่แบบจำลอง Linear Regression มีค่า R^2 สูงที่สุด คือ 0.148 เท่ากันและมีค่า MSE ต่ำที่สุด คือ 9073450.101 ลัง² เมื่อเทียบกับแบบจำลองอื่นๆ

ผลลัพธ์ของการทดลองที่ 3

จากการสร้างแบบจำลองเพื่อใช้ในการพยากรณ์ยอดขาย ใช้ความรู้จากยอดขายช่วงเวลาแปดโมงจนถึงบ่ายสอง เพื่อทำนายยอดขายช่วงบ่ายสามจนถึงห้าโมงเย็น โดยรวมทุก feature จากการทดลองที่ 1 และ 2 ได้ทำการสร้างแบบจำลอง ทั้งหมด 7 แบบจำลอง ตามภาพประกอบ 28 สามารถสรุปผลการทดลอง แสดงได้ตามตารางที่ 4



ภาพประกอบ 28 ผลของการแบบจำลองการทดลองที่ 3

ตาราง 4 แสดงผลการประเมินประสิทธิภาพแบบจำลองของการทดลองที่ 3

แบบจำลอง	**R^2**	**MAE (ลัง)**	**MSE (ลัง²)**
Random Forest Regressor	0.097	2422.686	9618169.326
XGBoost Regressor	0.116	2303.275	9416514.454
Gradient Boosting	0.144	2267.319	9121946.741
Linear Regression	**0.187**	**2257.819**	**8656397.633**
Ridge Regression	0.181	2266.806	8717751.210
Lasso Regression	0.187	2258.483	8660929.253
Elastic Net Regression	0.158	2302.495	8961787.363

จากผลลัพธ์ของการประเมินประสิทธิภาพของการทดลองที่ 3 พบว่า แบบจำลอง Linear Regression เป็นแบบจำลองที่ดีที่สุด โดยมีค่า R^2 สูงที่สุด คือ 0.187 มีค่า MAE ต่ำที่สุด คือ 2257.819 ลัง และมีค่า MSE ต่ำที่สุด คือ 8656397.633 ลัง² เมื่อเทียบกับแบบจำลองอื่นๆ

จากผลการทดลองที่ 1-3 สามารถสรุปได้ว่า แบบจำลอง Linear Regression เป็นแบบจำลองที่ดีที่สุด ตามตารางที่ 5

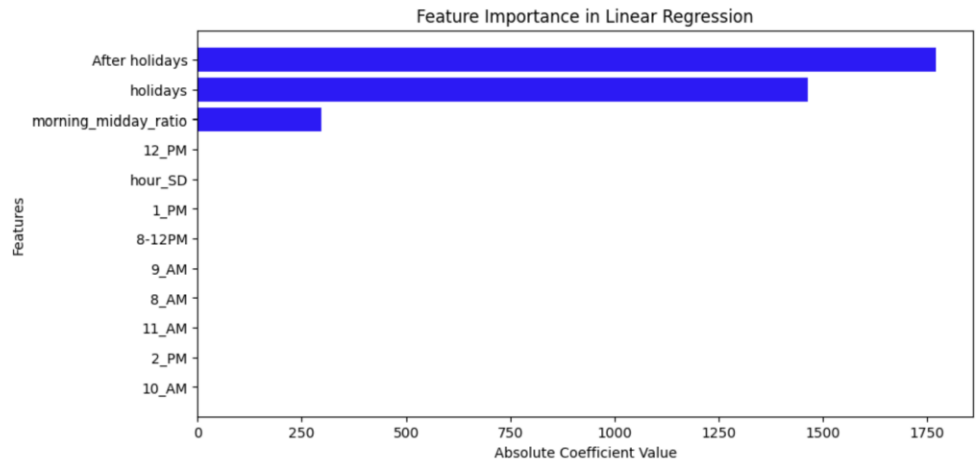
ตาราง 5 เปรียบผลการทดลองที่ 1-3 เพื่อดูผลแบบจำลองที่ดีที่สุด

การทดลอง	**R^2 สูงสุด**	**MAE ต่ำสุด**	**MSE ต่ำสุด**	**แบบจำลองที่ดีที่สุด**
1	0.167 (Gradient)	2210.102 (Gradient)	9072139.200 (XGBoost)	Gradient Boosting
2	0.148 (Gradient/ Linear)	2301.420 (Gradient)	9073450.101 (Linear)	Gradient Boosting / Linear Regression
3	0.187 (Linear)	2257.820 (Linear)	8656397.633 (Linear)	**Linear Regression**

ผลลัพธ์ของการทดลองที่ 4

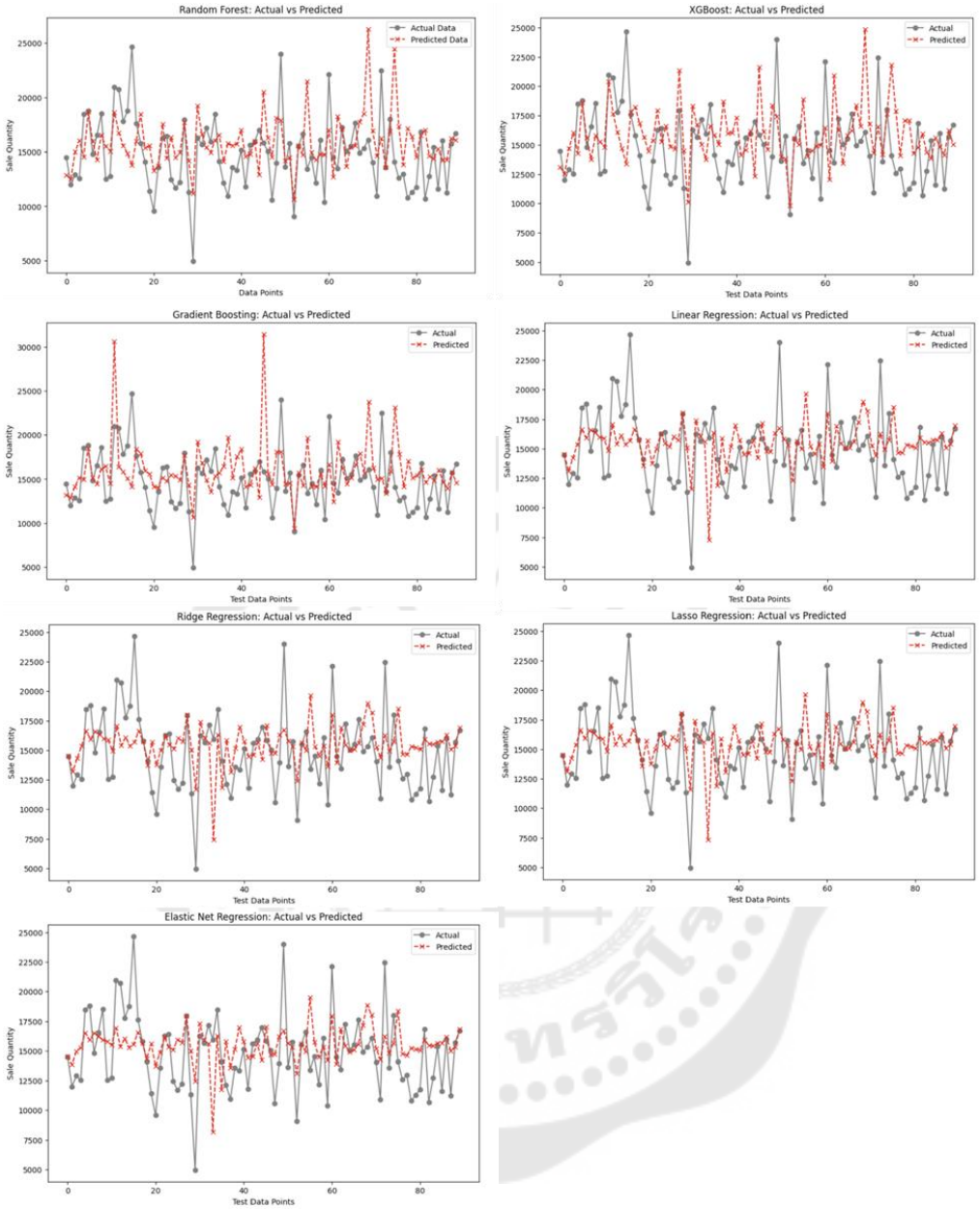
จากผลลัพธ์ผลการทดลองที่ 1-3 ที่พบว่า แบบจำลอง Linear Regression เป็นแบบจำลองที่ดีที่สุด เมื่อนำแบบจำลอง Linear Regression มาทำต่อโดยการหาฟีเจอร์ที่สำคัญ (Feature Importance) ตามภาพประกอบ 29 พบว่า ฟีเจอร์วันหลังวันหยุดนักขัตฤกษ์ (After holiday) เป็น ตัวแปรที่มีอิทธิพลมากที่สุด และเป็นค่าลบ โดยมีค่าอยู่ที่ -1771.829 แสดงว่า หลังวันหยุด ค่า Target_15_16_17 มีแนวโน้มลดลง ต่อมาฟีเจอร์วันหยุด (holidays)ก็มีผลลบเช่นกัน มีค่าอยู่ที่ -1463.900 แสดงว่า ถ้าเป็นวันหยุด ค่า Target จะก็จะลดลงด้วย นอกจากนี้พบว่าฟีเจอร์อัตราส่วนระหว่างยอดขายช่วงเช้ากับช่วงเที่ยง (morning_midday_ratio) ฟีเจอร์นี้ก็มีผลโดยมีผลเป็นบวก มีค่าอยู่ที่ 297.480 ซึ่งอาจหมายความว่า ถ้าอัตราส่วนช่วงเช้าต่อช่วงกลางวันสูง ค่า Target_15_16_17 อาจสูงขึ้น

	Feature	Coefficient
9	After holidays	-1771.829587
8	holidays	-1463.900396
10	morning_midday_ratio	297.480303
4	12_PM	0.901299
11	hour_SD	-0.566570
5	1_PM	0.548370
7	8-12PM	0.293269
1	9_AM	-0.237617
0	8_AM	-0.228468
3	11_AM	-0.155451
6	2_PM	0.090805
2	10_AM	0.013506



ภาพประกอบ 29 แสดงผลฟีเจอร์ที่สำคัญ

หลังจากได้ฟีเจอร์ที่สำคัญแล้ว ได้นำไปทดลองที่ 4 โดยใช้แบบจำลองทั้ง 7 แบบจำลองอีกครั้ง ตามภาพประกอบที่ 30 สามารถสรุปผลการทดลอง แสดงได้ตามตารางที่ 6



ภาพประกอบ 30 ผลของการแบบจำลองการทดลองที่ 4

ตาราง 6 แสดงผลการประเมินประสิทธิภาพแบบจำลองของการทดลองที่ 4

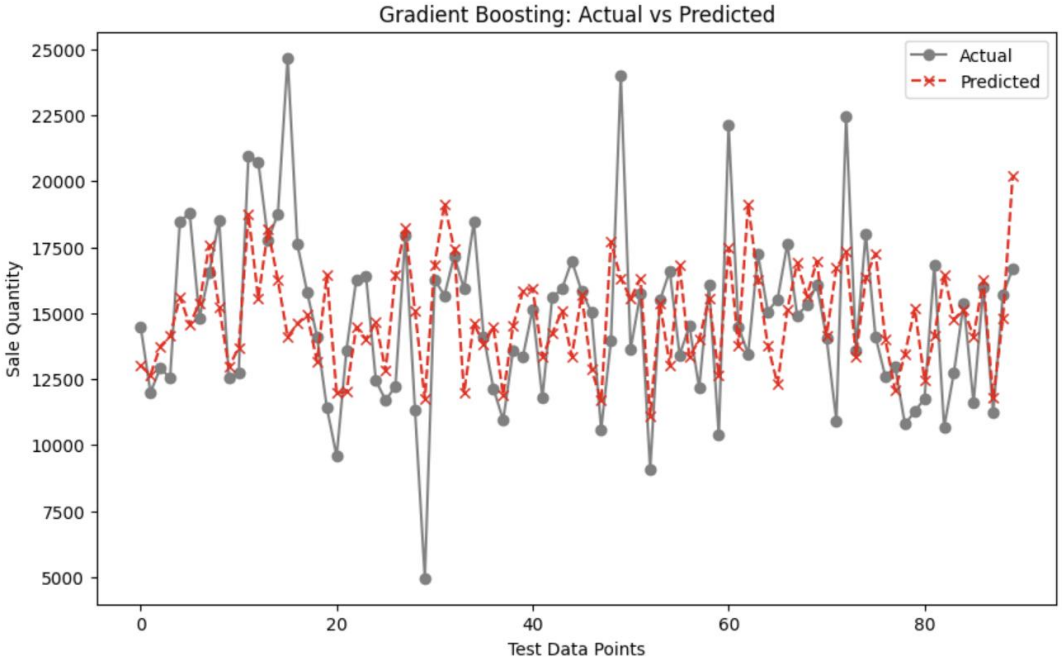
แบบจำลอง	R^2	**MAE (ลัง)**	**MSE (ลัง²)**
Random Forest Regressor	-0.195	2779.669	12730189.150
XGBoost Regressor	-0.204	2860.151	12831357.461
Gradient Boosting	-0.480	2973.175	15762549.598
Linear Regression	**0.094**	**2441.206**	9652544.083
Ridge Regression	0.092	2446.484	9668760.725
Lasso Regression	**0.094**	2441.253	**9652254.752**
Elastic Net Regression	0.082	2476.184	9776836.051

จากผลลัพธ์ของการประเมินประสิทธิภาพของการทดลองที่ 4 พบว่าแบบจำลอง Linear Regression และ แบบจำลอง Lasso Regression เป็นแบบจำลองที่ดีที่สุด โดยแบบจำลอง Linear Regression มีค่า R^2 สูงที่สุด คือ 0.094 และมีค่า MAE ต่ำที่สุด คือ 2441.206 ลัง ในขณะที่แบบจำลอง Lasso Regression มีค่า R^2 สูงที่สุด คือ 0.94 เท่ากันและมีค่า MSE ต่ำที่สุด คือ 9652254.752 ลัง² เมื่อเทียบกับแบบจำลองอื่นๆ

ทั้งนี้เมื่อเทียบกับผลการทดลอง 1-3 ยังพบว่า แบบจำลอง Linear Regression ในการทดลอง 3 ยังคงให้ผลลัพธ์ดีที่สุด ซึ่งอาจเพราะฟีเจอร์สำคัญที่เลือกมาใช้ทำการทดลองมาจากการหา Feature Importance จากแบบจำลอง Linear Regression

ผลลัพธ์ของการทดลองที่ 5

จากผลลัพธ์ผลการทดลองที่ 1-3 ที่พบว่า มีแบบจำลองที่ดีอยู่ 2 แบบจำลอง คือ แบบจำลอง Gradient Boosting และแบบจำลอง Linear Regression จึงได้นำแบบจำลอง Gradient Boosting ในการทดลองที่ 2 มาปรับค่าพารามิเตอร์เพิ่มเติม โดยได้ปรับ n_estimators จาก 100 ลองเพิ่มเป็น 500 เพื่อเพิ่มความสามารถของแบบจำลองในการเรียนรู้ข้อมูล ปรับ learning_rate (อัตราการเรียนรู้) จาก 0.1 ลดเป็น 0.05 เนื่องจากค่าที่ต่ำเกินไป อาจทำให้เรียนรู้ได้ช้า ค่าที่สูงเกินไป อาจทำให้ Overfitting ต่อมาปรับ max_depth (ความลึกของต้นไม้) จาก 3 เป็น 5 เพื่อให้แบบจำลองจับความสัมพันธ์ที่ซับซ้อนขึ้น ซึ่งหลังจากการปรับค่าตามที่กล่าวมา พบว่าผลลัพธ์ออกมาดีกว่า แบบจำลอง Linear Regression ในการทดลองที่ 3 เล็กน้อย ตามภาพประกอบที่ 31 และตารางที่ 7



ภาพประกอบ 31 แสดงผลหลังปรับพารามิเตอร์ในแบบจำลองที่ดีที่สุดในการทดลอง 2

ตาราง 7 แสดงผลการประเมินประสิทธิภาพแบบจำลองของการทดลองที่ 3 เทียบกับ 5

การทดลอง	R^2 สูงสุด	MAE ต่ำสุด	MSE ต่ำสุด	แบบจำลองที่ดีที่สุด
3	0.187 (Linear)	2257.820 (Linear)	8656397.633 (Linear)	Linear Regression
5	0.197 (Gradient)	2232.372 (Gradient)	8553787.026 (Gradient)	Gradient Boosting

ผลลัพธ์ของการทดลองที่ 6

หลังจากได้ทำการทดลองทั้ง 5 การทดลองแล้ว โดยใช้แบบจำลองจากการเรียนรู้ของเครื่อง (Machine Learning) ผู้วิจัยได้ทำการนำผลมาเปรียบเทียบกับวิธีที่ผู้วิจัยได้ใช้อยู่ในปัจจุบัน (Traditional Approach) โดยการทำ Moving Average ด้วยโปรแกรม Microsoft Excel ตามภาพประกอบ 32 ซึ่งได้ใช้ข้อมูลชุด Test ข้อมูลเดียวกับที่ใช้ทดสอบในแบบจำลอง สาขาสรุปผลได้ตามตารางที่ 8

	A	B	C	D	E	F	G	H	I
1	8_AM	9_AM	10_AM	11_AM	12_PM	1_PM	2_PM	Target_15_16_17	Moving Average
2	17101	3809	4583	6368	2951	4066	2281	14477	
3	13870	2660	2901	3152	2955	3579	1818	11998	
4	13612	5805	3946	4677	3682	4146	3720	12915	
5	5934	14379	4641	4043	3082	5073	4724	12548	13,130
6	10106	11500	3952	4378	3939	7957	2552	18481	12,487
7	19054	3391	2905	6348	3814	3941	2916	18787	14,648
8	11812	5373	6462	4183	4040	5354	2232	14815	16,605
9	10931	3625	5252	4816	3631	6414	3707	16540	17,361
10	7583	12071	4014	4452	3620	5317	2599	18528	16,714
11	14675	1080	6277	3929	3513	3690	3779	12534	16,628
12	16331	3601	3393	6744	3782	3566	11451	12761	15,867
13	7898	6014	3330	4844	4170	4707	3029	20967	14,608
14	16629	2985	6604	6749	3334	4753	3830	20741	15,421
15	13556	5823	4743	5599	3686	5473	3139	17774	18,156
16	8247	8104	7571	3633	3186	5579	2219	18753	19,827
17	10195	3607	4096	2983	3124	3334	5091	24677	19,089
18	3834	11792	11596	4711	4066	5751	3903	17603	20,401
19	8360	10844	6493	3839	3386	5067	4232	15781	20,344
20	14491	3878	9725	3475	3353	4029	3932	14112	19,354
21	14572	6545	6694	3807	3267	9023	2542	11417	15,832
22	10976	9665	2875	3402	2223	3101	3450	9574	13,770
23	10626	2630	2254	4323	2640	4225	3720	13602	11,701
24	16602	6661	3574	3571	3768	6639	2646	16261	11,531
25	14895	4093	6357	3792	3030	6502	5269	16405	13,146
26	11463	6561	3698	6328	3110	3644	3655	12448	15,423
27	12054	12581	4018	3285	3692	4455	4318	11696	15,038
28	11785	5598	6599	3404	3612	3840	2140	12230	13,516
29	14864	4966	5414	5438	5128	4571	3027	17938	12,125
30	7337	14918	4591	4613	2952	5124	4546	11308	13,955

X_test_traditional

ภาพประกอบ 32 การพยากรณ์ยอดขายโดยใช้วิธีแบบดั้งเดิม (Traditional Approach)

ตาราง 8 แสดงผลการประเมินประสิทธิภาพแบบจำลองของการทดลองที่ 6

แบบจำลอง	R^2	MAE (ลัง)	MSE (ลัง^2)
Traditional approach (Moving Average)	0.012	2951.452	13731030.606

บทที่ 5
สรุปผลการวิจัย อภิปรายผล และข้อเสนอแนะ

ในการวิจัยเพื่อศึกษาแบบจำลองที่ช่วยในการพยากรณ์ยอดขาย ได้ใช้ข้อมูลข้อมูลยอดขายรายชั่วโมงที่ได้จากฐานข้อมูลของธุรกิจ มาเป็นข้อมูลในการทำแบบจำลองที่ใช้เทคนิคการเรียนรู้ของเครื่อง เพื่อทดลองหาแบบจำลองที่จะสามารถพยากรณ์ได้มีประสิทธิภาพสูงสุด ผู้วิจัยได้ทำการวิจัยได้ประเมินประสิทธิภาพของแต่ละแบบจำลอง เพื่อนำมาเปรียบเทียบและสรุปผล สามารถแบ่งหัวข้อในการสรุปผลได้ ดังนี้

1.สรุปผลการวิจัย

2.อภิปรายผลการวิจัย

3.ข้อเสนอแนะ

สรุปผลการวิจัย

ในการวิจัยครั้งนี้ผู้วิจัยได้ทำการพยากรณ์ยอดขาย โดยวางแผนการทดลองไว้ทั้งหมด 6 การทดลอง ซึ่งการทดลองที่ 1-5 จะใช้แบบลำลองการเรียนรู้ของเครื่อง (Machine Learning) ในการทำ และแต่ละการทดลองจะทำทั้งหมด 7 แบบจำลองต่อ 1 การทดลอง และในส่วนของการทดลองที่ 6 จะใช้วิธีการ Moving Average ซึ่งเป็นวิธีการดั้งเดิม (Traditional Approach) ที่ผู้วิจัยได้ทำอยู่ในปัจจุบัน ผ่านการทำบนโปรแกรม Microsoft Excel โดยหลังจากทำครบทั้ง 6 การทดลองแล้ว ได้นำผลจากการทำ มาประเมินประสิทธิภาพของแบบจำลอง ด้วยการใช้ R^2 (R-squared), MAE (Mean Absolute Error), และ MSE (Mean Squared Error)

ในการประเมินประสิทธิภาพของแบบจำลองการทำนายโดยผลการทดลองพบว่า การพยากรณ์ยอดขายด้วยแบบจำลอง Gradient Boosting จากการทดลองที่ 5 ที่ใช้ความรู้จากยอดขายช่วงเวลาแปดโมงจนถึงบ่ายสอง เพื่อทำนายยอดขายช่วงบ่ายสามจนถึงห้าโมงเย็น โดยการเพิ่มฟีเจอร์ ยอดขายในช่วงเช้าจนถึงเที่ยง (8-12 PM) อัตราส่วนระหว่างยอดขายช่วงเช้ากับช่วงเที่ยง (Morning_Midday_Ratio) และ ความแปรปรวนของยอดขาย (Hour_SD) และมีการปรับค่าพารามิเตอร์ในแบบจำลองเพิ่มเติม ได้ผลลัพธ์ออกมาดีที่สุด กว่าแบบจำลองอื่นๆ และทุกการทดลอง ซึ่งดูจากค่า R^2 ที่ได้ผลลัพธ์มากที่สุด คือ 0.197 ค่า MSE ต่ำที่สุด คือ 8553787.026 ลัง^2 ในส่วนของค่า MAE ในการทดลองที่ 5 ถึงแม้ค่า MAE จะสูงกว่า การทดลองที่ 1 แต่แบบจำลองที่ให้ค่า MAE น้อยที่สุดก็ยังเป็น แบบจำลอง Gradient Boosting ตามตารางที่ 9

ตาราง 9 แสดงผลการประเมินประสิทธิภาพแบบจำลองของการทดลองที่ 1 เทียบกับ 5

การทดลอง	R^2 สูงสุด	MAE ต่ำสุด	MSE ต่ำสุด	แบบจำลองที่ดีที่สุด
1	0.167 (Gradient)	**2210.102 (Gradient)**	9072139.200 (Xgboost)	Gradient Boosting
5	**0.197 (Gradient)**	2232.372 (Gradient)	**8553787.026 (Gradient)**	**Gradient Boosting**

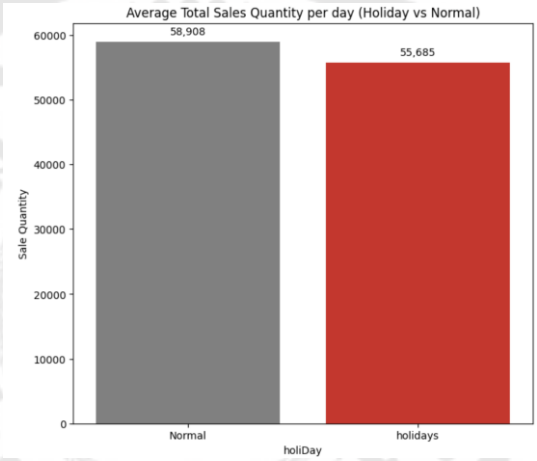
อภิปรายผลการวิจัย

การวิจัยในครั้งนี้ผู้วิจัยใช้ภาษาไพทอน (Python) สำหรับการสร้างแบบจำลอง Machine Learning และประมวลผลในการทำวิจัย โดยทำผ่าน Google Colab ที่เป็นเครื่องมือในการเขียน Code Python แบบออนไลน์บนเว็บเบราว์เซอร์ได้แบบไม่มีค่าใช้จ่ายและสะดวกง่ายต่อการใช้งาน ใช้ โดยได้ทำการทดลองทั้งหมด 6 การทดลอง และสร้างแบบจำลองในแต่ละการทดลองทั้งหมด 7 แบบจำลอง ทั้งนี้ในการทำการทดลองผู้วิจัยได้ใช้ฟีเจอร์ทั้งหมด 12 ฟีเจอร์ที่คิดว่าจะมีผลต่อยอดขายในการทำการทดลอง ตามตารางที่ 10

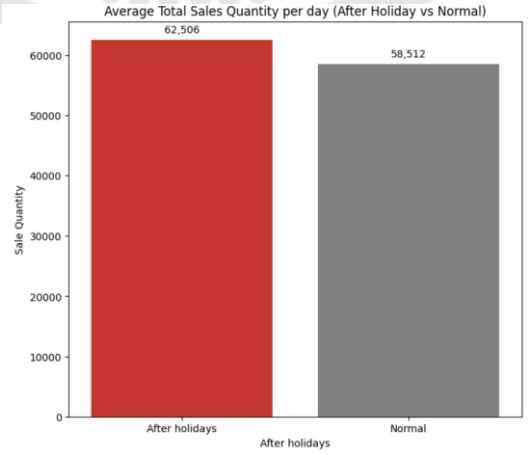
ตาราง 10 แสดงฟีเจอร์ทั้งหมดที่ใช้ในการทำการลดลอง

ลำดับ	ฟีเจอร์	คำอธิบาย
1	8_AM	ยอดขายรวมที่ช่วงเวลา 08:00-08:59
2	9_AM	ยอดขายรวมที่ช่วงเวลา 09:00-09:59
3	10_AM	ยอดขายรวมที่ช่วงเวลา 10:00-10:59
4	11_AM	ยอดขายรวมที่ช่วงเวลา 11:00-11:59
5	12_PM	ยอดขายรวมที่ช่วงเวลา 12:00-12:59
6	1_PM	ยอดขายรวมที่ช่วงเวลา 13:00-13:59
7	2_PM	ยอดขายรวมที่ช่วงเวลา 14:00-14:59
8	8-12 PM	ยอดขายรวมที่ช่วงเวลา 08:00-12:59
9	holidays	วันหยุดนักขัตฤกษ์
10	After_holidays	วันหลังวันหยุดนักขัตฤกษ์ 1 วัน
11	Morning_Midday_Ratio	อัตราส่วนยอดขายระหว่างช่วง 08:00-11:59 เทียบกับช่วง 12:00-14:59
12	Hour_SD	ค่าความแปรปรวนของยอดขายในแต่ละช่วงโมง

ซึ่งจะเห็นได้ว่าฟีเจอร์ที่ 1-8 ที่เลือกใช้จะเป็นในส่วนของยอดขายรวมในแต่ละช่วงเวลา ในส่วนฟีเจอร์ holidays หรือ วันหยุดนักขัตฤกษ์นั้นผู้วิจัยได้ตั้งต้นจากสมมติฐานที่ว่าวันหยุดนักขัตฤกษ์อาจมีผลให้ยอดขายในวันนั้นน้อยกว่าวันปกติคือร้านค้าอาจหยุดเพื่อไปเที่ยว หรือยอดขายอาจจะมากกว่าปกติจากการที่ลูกค้ามาเที่ยวแล้วมีการซื้อสินค้ามากขึ้น และฟีเจอร์ After_holidays มาจากสมมติฐานที่ว่าถ้าหากวันหยุดคนเที่ยวมากขึ้น ร้านค้าอาจขายสินค้าได้มากขึ้นกว่าเดิมทำให้วันถัดมาจะมีการสั่งสินค้าเพื่อเติมสต็อกในร้านมากกว่ากว่าวันอื่นๆ ซึ่งผลจากการทำ EDA พบว่า ในส่วนของวันหยุดยอดขายจะน้อยกว่าวันปกติ ตามภาพประกอบ 32 ในขณะที่หลังวันหยุดยอดขายกลับมากกว่าวันปกติ ตามภาพประกอบ 33



ภาพประกอบ 33 แสดงยอดขายเฉลี่ยของวันธรรมดาเทียบกับวันหยุดนักขัตฤกษ์



ภาพประกอบ 34 แสดงยอดขายเฉลี่ยของวันธรรมดาเทียบกับวันหลังหยุดนักขัตฤกษ์ 1 วัน

หลังจากทำการทดลองทั้ง 6 การทดลอง เมื่อพิจารณาผลจากการประเมินประสิทธิภาพแบบจำลองทั้งหมด พบว่า แบบจำลอง Gradient Boosting สามารถพยากรณ์ยอดขายได้ดีที่สุดจากทั้ง 7 แบบจำลองของการเรียนรู้ของเครื่องที่ผู้วิจัยได้ทำการทดลอง รวมถึงยังได้ผลลัพธ์ดีกว่าวิธีการแบบดั้งเดิมที่ผู้วิจัยใช้ ตามข้อมูลที่แสดงในตารางที่ 9

ตาราง 11 เปรียบเทียบผลการประเมินประสิทธิภาพของแบบจำลองที่ดีที่สุดในแต่ละการทดลอง

การทดลอง	**R^2 สูงสุด**	**MAE ต่ำสุด**	**MSE ต่ำสุด**	**แบบจำลองที่ดีที่สุด**
1	0.167 (Gradient)	**2210.102 (Gradient)**	9072139.200 (XGBoost)	Gradient Boosting
2	0.148 (Gradient/ Linear)	2301.420 (Gradient)	9073450.101 (Linear)	Gradient Boosting / Linear Regression
3	0.187 (Linear)	2257.820 (Linear)	8656397.633 (Linear)	Linear Regression
4	0.094 (Linear/Lasso)	2441.206 (Linear)	9652254.752 (Lasso)	Linear Regression / Lasso Regression
5	**0.197 (Gradient)**	2232.372 (Gradient)	**8553787.026 (Gradient)**	**Gradient Boosting**
6	0.012	2951.452	13731030.606	Moving Average

ซึ่งจากผลการทดลองทั้งหมดยังพบว่าทุกการทดลองประสิทธิภาพของแบบจำลองยังให้ผลของค่า R^2 ที่ต่ำอยู่ ซึ่งอาจเกิดจากข้อมูลที่ยังไม่มากเพียงพอ หรือข้อมูลมีความแปรปรวนสูงก็เป็นได้ แต่ถึงอย่างนั้นการใช้แบบจำลองด้วยการเรียนรู้ของเครื่องโดยใช้ความรู้ที่เรียนมา ก็สามารถพยากรณ์ยอดขายได้ดีกว่าวิธีการแบบดั้งเดิมที่ทำอยู่ในปัจจุบัน

นอกจากนี้เมื่อนำผลลัพธ์จากการทดลองที่ดีที่สุดคือผลลัพธ์จากแบบจำลอง Gradient Boosting ในการทดลองที่ 5 มาดูผลของการพยากรณ์เทียบกับยอดขายจริง ตามตารางที่ 12 พบว่า ยิ่งค่ายอดขายมากหรือน้อยกว่าค่าเฉลี่ยและค่าฐานนิยมข้อข้อมูลที่ใช้ทดสอบ พบว่าค่าพยากรณ์ที่ได้ออกมายิ่งมีความต่างจากค่ายอดขายจริง ซึ่งอาจสรุปได้ว่าในวันที่ยอดขายมากกว่าปกติมากหรือยอดขายน้อยกว่าปกติมาก การใช้แบบจำลองด้วยการเรียนรู้ของเครื่องอาจให้ผลลัพธ์ที่ไม่ดี

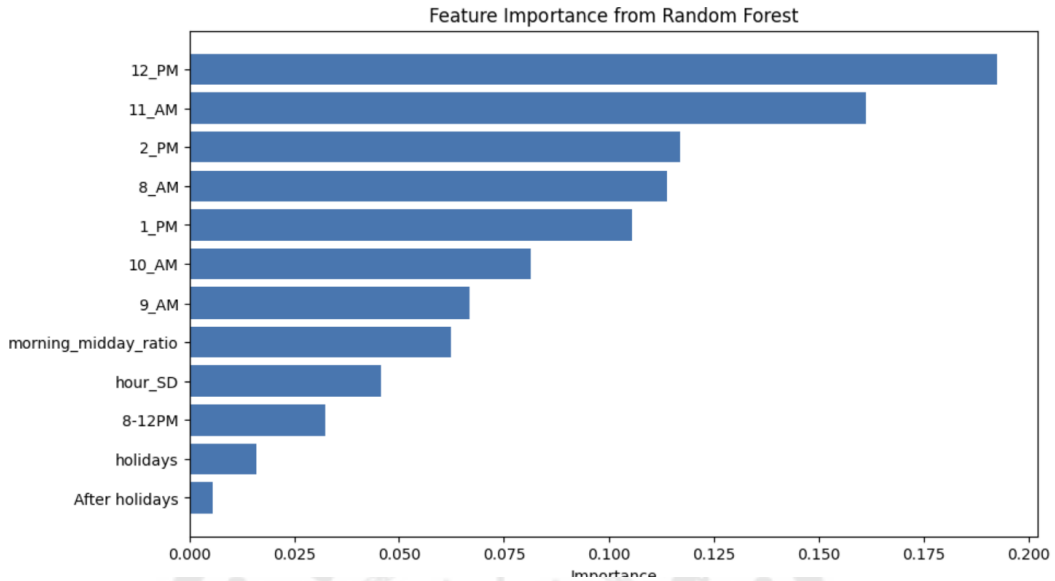
ตาราง 12 แสดงค่าความต่างระหว่างยอดขายจริงกับค่าที่พยากรณ์ได้ตามกลุ่มของยอดขาย

Range of Actual Sale	Count of Delivery Day	Average of Diff	Min of Diff	Max of Diff	Note
4000-4999	1	6,648	6,648	6,648	
9000-9999	2	2,432	2,091	2,772	
10000-10999	6	3,141	971	5,830	
11000-11999	9	2,298	596	4,931	
12000-12999	11	1,709	312	4,182	
13000-13999	8	2,560	409	5,870	
14000-14999	9	1,169	68	3,188	*Average*
15000-15999	16	1,243	6	3,961	*Mode*
16000-16999	10	2,038	516	3,616	
17000-17999	6	1,227	258	2,901	
18000-18999	6	3,047	1,692	4,210	
20000-20999	2	3,723	2,285	5,162	
22000-22999	2	4,905	4,487	5,323	
23000-23999	1	7,647	7,647	7,647	
24000-25000	1	10,521	10,521	10,521	

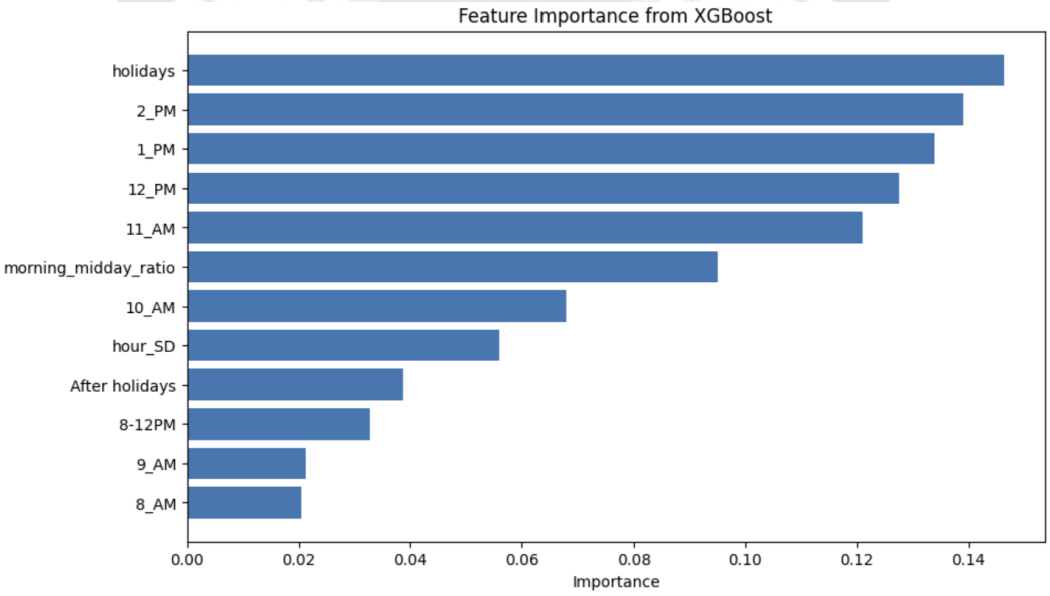
ซึ่งเมื่อดูข้อมูลเพิ่มเติมพบว่าในวันที่ยอดขายน้อยกว่าปกติคือวันหยุดนักขัตฤกษ์ในช่วงวันสงกรานต์(15-16 เมษายน) และในวันที่ยอดเยอะกว่าปกติจะเป็นในช่วงปลายเดือนมีนาคมถึงต้นเดือนเมษายน ซึ่งเป็นช่วงเดือนที่ยอดขายของบริษัทจะขายดี

และในส่วนของผลการทดลองที่ 4 เนื่องจากผู้วิจัยตั้งต้นการทดลองโดยเลือกแบบจำลอง Linear Regression เป็นแบบจำลองตั้งต้นและได้นำแบบจำลอง Linear Regression นี้มาหาค่าฟีเจอร์ที่สำคัญ (Feature Importance) โดยใช้วิธีการหา ขนาดสัมประสิทธิ์ (coefficient) ของแต่ละตัวแปร ซึ่งวิธีนี้จะใช้กับแบบจำลองที่เป็นแบบจำลองเชิงเส้น เช่น Linear Regression, Lasso, Ridge, Elastic Net จึงอาจส่งผลให้ในการทดลองที่ 4 ผลลัพธ์ที่ได้ออกมาทำให้แบบจำลอง Linear Regression ดีที่สุด ทั้งนี้ผู้วิจัยได้ดำเนินการเพิ่มเติมโดยใช้การหาค่าฟีเจอร์ที่สำคัญ (Feature Importance)ให้เหมาะสมกับลักษณะแบบจำลองประเภท Tree-based เช่น มากขึ้นตามภาพประกอบที่ 35 – 37 แล้วทำการเลือกฟีเจอร์ที่สำคัญของแต่ละแบบจำลองมาลองทำการ

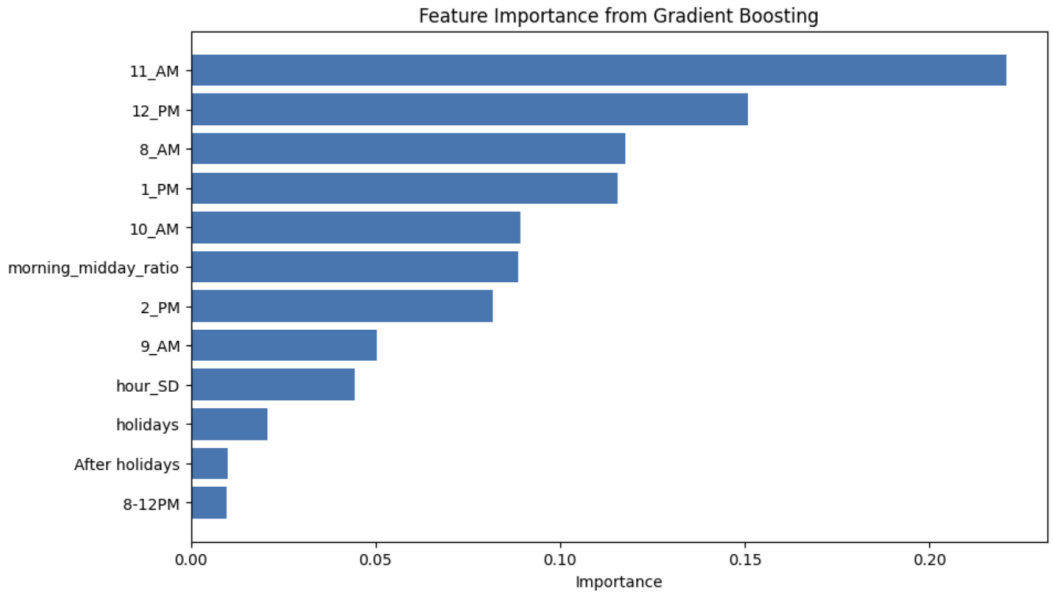
พยากรณ์อีกครั้ง ผลที่ได้แบบจำลอง Gradient Boosting ในการทดลองที่ 5 ก็ยังคงให้ผลลัพธ์ที่ดีสุดเช่นเดิม ตามตารางที่ 13



ภาพประกอบ 35 แสดงฟีเจอร์ที่สำคัญของแบบจำลอง Random Forest



ภาพประกอบ 36 แสดงฟีเจอร์ที่สำคัญของแบบจำลอง XGBoost



ภาพประกอบ 37 แสดงฟีเจอร์ที่สำคัญของแบบจำลอง Gradient Boosting

ตาราง 13 เปรียบเทียบผลการประเมินประสิทธิภาพของแบบจำลอง Gradient ในการทดลองที่ 5 เทียบกับแบบจำลอง Random, XGBoost และ Gradient ที่ใช้ผลของการหาค่าฟีเจอร์ที่สำคัญมาเป็นตัวกำหนดฟีเจอร์ที่ใช้ในแบบจำลอง

การทดลอง	**R^2 สูงสุด**	**MAE ต่ำสุด**	**MSE ต่ำสุด**
การทดลองที่ 5 Gradient	0.197	2232.372	8553787.026
Feature Importance Random	0.153	2362.194	9026993.660
Feature Importance XGBoost	0.167	2207.725	8873846.174
Feature Importance Gradient	0.149	2280.522	9063858.109

ข้อเสนอแนะ

1.ในการวิจัยครั้งนี้ เนื่องจากมีข้อจำกัดในการดึงข้อมูลที่นำมาใช้ในการทำแบบจำลองเพื่อพยากรณ์ ซึ่งทำให้ได้ข้อมูลย้อนหลังแค่ 1 ปี อาจทำให้ไม่สามารถสร้างแบบจำลองที่มีประสิทธิภาพมากกว่านี้ หรือสร้างแบบจำลองที่อาจมีฤดูกาลเข้ามาเกี่ยวข้องได้

2.อาจเลือกคลังสินค้าอื่นๆมาทำเพิ่มเติม หรือเลือกคลังสินค้าต่างภูมิภาคมาทำแบบจำลองเพิ่ม เพื่อดูผลของแบบจำลองเพิ่มเติม

3.ในการจะช่วยบริหารจัดการสินค้าคงคลัง ข้อมูลที่พยากรณ์อาจเลือกข้อมูลสินค้าหลักที่มีนัยะสำคัญหรือมีปัญหาอยู่ในปัจจุบัน มาทำการเข้าแบบจำลองแทนการใช้ข้อมูลยอดขายรวม เพื่อให้สามารถพยากรณ์สินค้าคงคลังที่ควรมีได้

4.ในการทำครั้งต่อไปอาจหาฟีเจอร์อื่นๆที่อาจมีผลต่อยอดขายเพิ่มเติมเข้ามาเป็นฟีเจอร์ เช่น อุณหภูมิ, วันที่ฝนตก, วันในสัปดาห์ หรืออาจเป็นยอดขายเฉลี่ยช่วงเที่ยงของยอดขายวันก่อนหน้า เป็นต้น

บรรณานุกรม

Arora, T., Chandna, R., Conant, S., Sadler, B., & Slater, R. (2020). Demand forecasting in wholesale alcohol distribution: An ensemble approach. *International Journal of Forecasting*, 36(3), 1083-1097.

He, W., & Zeng, Q. (2021). Research on sales forecast based on XGBoost-LSTM algorithm model. *Big Data and Business Intelligence (MLBDBI)*, 1754, 265-269.

Jain, A., Menon, M. N., & Chandra, S. (2020). Sales forecasting for retail chains. *International Journal of Data Science*, 5(2), 141-153.

Mitra, A., Jain, A., Kishore, A., & Kumar, P. (2022). A comparative study of demand forecasting models for a multi-channel retail company: A novel hybrid machine learning approach. *Procedia Computer Science*, 200, 456-463.

Panarese, A., Settanni, G., Vitti, V., & Galiano, A. (2022). Developing and preliminary testing of a machine learning-based platform for sales forecasting using a gradient boosting approach. *Computers in Industry*, 141, 103692.

Pavlyshenko, B. M. (2019). Machine-learning models for sales time series forecasting. *Data*, 4(1), 15.

Saigustia, C., & Pijarski, P. (2023). Time series analysis and forecasting of solar generation in Spain using eXtreme gradient boosting: A machine learning approach. *Energies*, 16(3), 1017.

Venkatesan, E., & Mahindrakar, A. B. (2019). Forecasting floods using extreme gradient boosting – a new approach. *Water Resources Management*, 33, 1899-1912.

Wang, C., & Bai, X. (2018). Boosting learning algorithm for stock price forecasting. *Procedia Computer Science*, 127, 166-174.

Wang, Y., Sun, S., Chen, X., Zeng, X., Kong, Y., Chen, J., & et al. (2021). Short-term load forecasting of industrial customers based on SVMD and XGBoost. *IEEE Access*, 9, 66656-66667.

Zhang, L., Bian, W., Qu, W., Tuo, L., & Wang, Y. (2021). Time series forecast of sales volume based on XGBoost. *Procedia Computer Science*, 187, 257-262.

ประวัติผู้เขียน