**Market Basket Analysis by Association Rule Approach**

**using Various of Machine Learning Algorithm**

**Introduction:**

ในช่วงเริ่มแรกของการค้าขายนั้น ตลาดทั่วไปจะตั้งเป้าไปที่การผลิตสินค้าออกมาเป็นจำนวนมาก และจำหนายสินค้าออกไปให้ได้อย่างรวดเร็ว เราเรียกยุคเหล่านี้ว่า Product Era และ Sales Era . เมื่อเวลาผ่านไปก็ได้กำเนิดเป็น marketing era ที่เริ่มมีการกล่าวถึงกลยุทธ์ในการขาย แล้วจึงเกิดการตระหนักถึงลูกค้าที่เป็นอีกปัจจัยในการทำการขาย จึงเกิดขึ้นมาเป็นกลยุทธ์ที่ชื่อว่า CRM ในที่สุด. CRM ไม่ได้เป็นเพียงแต่การติดต่อกับลูกค้าเพียงเท่านั้น แต่รวมถึงการวิเคราะห์พฤติกรรมของพวกเขาเพื่อที่จะนำไปวางแผนการตลาดต่อไป โดยในปัจจุบัน เทคโนโลยีต่างๆ เช่น ซอฟต์แวร์ลูกค้าสัมพันธ์ หรือ ระบบ data analytic ได้เข้ามามีส่วนช่วยอำนวยความสะดวกในเรื่องของการเก็บและวิเคราะห์ข้อมูล ทำให้การวางแผน strategy สำหรับองค์กรหรือผู้ประกอบการเป็นไปได้อย่างง่าย ทันการณ์ อีกทั้งยังช่วยลดค่าใช้จ่ายและข้อผิดพลาดในส่วนของทรัพยากรบุคคลลงอีกด้วย.

data analytic นั้นมีความสามารถในการศึกษาพฤษติกรรมของมนุษย์จากการจดจำ pattern ต่างๆ แล้วทำนาย action ถัดไปที่อาจเกิดขึ้นในอนาคตได้. เมื่อเดือนกุมภาพันธ์ ปี 2021 ที่ผ่านมา นักพัฒนาชาวเกาหลีใต้ seunghyun choi และทีมได้ตีพิมพ์บทความที่กล่าวถึงการใช้ sequential pattern strategy สำหรับ classify เหตุการณ์ต่างๆจากการ monitor พฤติกรรมของมนุษย์ในพื้นที่จำกัด และทำนายเหตุหารณ์ที่อาจจะเกิดขึ้นจาก object detection algorithm โดย algoritm สามารถจำลอง pattern ของเหตุการณ์ที่อาจเกิดขึ้นต่อจากกันซึ่งอ้างอิงจากข้อมูลที่รวบรวมได้. ยังมีอีกหลายการศึกษาที่กล่าวถึงการใช้เทคโนโลยีในการประยุคต์ใช้กับการทำนายพฤติกรรมของลูกค้า เช่น ในปี 2013 Laura Badea จาก Bucharest University ได้ใช้ Artificial Neural Networks (ANN) สำหรับตรวจสอบลักษณะของปัจเจคบุคคลที่มีโอกาสที่จำทำรายการฝากเงินจากข้อมูลส่วนบุคคล เช่น เชื้อชาติ เพศ การศึกษา อายุ หรือ รายได้ โดยผลลัพธ์แสดงให้เห็นว่าอัลกอริทึมสามารถตรวจจับกลุ่มตัวอย่างที่มีโอกาสฝากและไม่ฝากมีความแม่นยำของโมเดลอยู่ระหว่าง 73-78 % แม้อัตราการตัวจับจะอยู่ในระดับสูง แต่เทคโนโลยีดังกล่าวก็ยังมีเรื่อง่ที่ต้องนำมาพิจารณาเพิ่มเติมเช่น time consuming หรือ chance ที่จะเกิดสถารการณ์ overfitting ที่เพิ่มขึ้นเมื่อเปรียบเทียบกับ algorithm อีกชนิดหนึ่ง

หนึ่งในวิธีการที่ simple และมีการถูกนำมาใช้ใน CRM อย่างแพร่หลายคือการหาความสัมพันธ์ของสินค้าแต่ละชนิดว่ามีโอกาสที่จะถูกซื้อมากเท่าไหร่เมื่อพิจารณาจากสินค้าที่มีอยู่ในตะกร้าปัจจุบัน เรียกวิธีการนี้ว่า market basket analytic โดยใช้ association rule algorithm ด้วยวิธีนี้จะช่วยให้ทางองค์กรสามารถคาดเดาพฤติกรรมการซื้อของลูกค้าเพื่อวางแผนกลยุทธ์ทางการตลาดได้. Market basket analytic นั้นสามารถนำไปประยุคต์ใช้ผ่าน tool และวิธีวิเคราะห์ที่หลากหลาย ในสถานการณ์ที่ต่างกันไป โดยจะให้ผลลัพธ์ที่มีประสิทธิภาพที่ไม่เหมือนกันซะทีเดียวในข้อมูลแต่ละชุด. การศึกษานี้มุ่งเน้นไปที่การวิเคราะห์ association rule ที่ได้มาจาก apriori algorithm และ frequent pattern growth algorithm ใน development tool หรือ programming language ที่ต่างกันไป ซึ่งประกอบไปด้วย Weka, Python, Rstudio, และ Matlab และนำผลลัพธ์มาวิเคราะห์เพื่อเปรียบเทีบประสิทธิภาพของวิธีการต่างๆ ก่อนที่จะนำไปแสดงผล และอภิปรายผลลัพธ์ในขั้นตอนสุดท้าย

การศึกษานี้ตั้งเป้าไปที่การแสดงผลลัพธ์ด้านประสิทธิภาพ ได้แก่ ระยะเวลาที่ใช้ และความแม่นยำ ของการวิเคราะห์ Association rule จาก data analytic tool ประเภทต่างๆ เพื่อให้เกิดความเข้าใจในการเลือกใช้ tools ให้เหมาะสนกับสถานการณ์

**Literature Review:**

CRM เกิดขึ้นจากแนวคิด Relationship marketing ในช่วงปลายศตวรรษที่ 20 ที่มอง customer เป็นส่วนหนึ่งในการยกระดับการตลาดด้วย โดย prof Philip Kotler จาก Northwestern universities ได้กล่าวเกียวกับการตลาดรูปแบบใหม่นี้ว่า บริษัททั้งหลายจะต้องเปลี่ยนจาก short-term transaction-oriented goal ที่เน้นไปที่การขายให้ได้มากๆ เป็น long-term relationship-building goal ที่เน้นการสร้างความสัมพันธ์ระยะยาวให้กับผู้บริโภค (<https://bit.ly/3nKzE70>). CRM นั้นถูกนิยามว่าเป็น infrastructure ที่ต้องทำความเข้าใจและเพิ่มคุณค่าให้กับ customer เพื่อจูงใจให้ valuable customer ยังคงความ loyalty กับ brand และกลับมาซื้อผลิตภัณฑ์ใหม่อีกครั้ง (<https://bit.ly/3Ih3VSy>) .

ในบัจจุบัน เทคโนโลยีได้เข้ามามีบทบาทกับการเก็บและวิเคราะห์ข้อมูลเพื่อนำไปสร้างความสัมพันธ์กับผู้บริโภค. การนำ analytic tools มาใช้ เพื่อเข้าถึงข้อมูลที่ซ่อนอยู่ของลูกค้าและปรับปรุง CRM strategy ถูกพูดกันอย่างกว้างขวางทั้งในแง่มุมของการวิจัย. Data mining เป็นหนึ่งใน approach ที่มีบทบาทกับการวิเคราะห์ข้อมูล, อ้างอิงจาก (**DM for CRM**) ประโยชน์ของ data mining ประกอบด้วย (1) การหา hidden pattern จากข้อมูลที่มีอยู่ (2) การนำ pattern จากข้อมูลที่มีอยู่มาทำนายเหตุการณ์ในอนาคต และ (3) การนำ extracted pattern ไปหา unusual data elements. ซึ่งจากบทความยังกล่าวถึง approach สำหรับใช้ในสถานการณ์ต่างๆด้วย เช่น fraud detection สำหรับระบบ banking, customer churn analysis สำหรับระบบ telecommunication, or customer segmentation, เขายังมีการพูดถึง approach ที่เกี่ยวกับ retail marketing ซึ่งเป็นหัวข้อที่เราได้ศึกษาอีกด้วยว่า data mining สามารถนำมาใช้ในเรื่องของการทำ basket analysis, sale forecasting, database marketing, and merchandise planning and allocation. หัวข้อการศึกษานี้มุ้งเน้นไปที่การใช้ data mining tools เพื่อวิเคราะห์ association rules ด้วย market basket analysis (MBA) จากข้อมูล transaction ของห้างสรรพสินค้า.

MBA เป็นวิธีการวิเคราะห์รูปแบบของการเลือกซื้อสินค้าชนิดหนึ่งที่มีผลกับการเลือกซื้อสินค้าอีกชนิด โดยความสัมพันธ์ระหว่างสินค้าเหล่านี้จะถูกนำเสนอในรูปแบบของ association rule (**book**) โดยจะนิยามกลุ่มของสินค้านั้นว่า itemset และวิเคราะห์ความแม่นยำจากค่า confidence. การศึกษาของ Roshan Gangurde **(ANN market Basket analysis)** ได้นำ Artificial Neural Network มาใช้ในการทำ MBA โดยเขาใช้ Feed Forward Neural Network (FFNN) ในการ train ข้อมูล transaction ด้วยภาษา Java, hidden layer และ output node ที่เขาใช้จะมีจำนวนเท่ากับผลรวมของค่า combination จากจำนวนของ input ทั้งหมดที่เป็นไปได้ โดย order is not important and repetition is not allowed. การศึกษานี้ใช้ตัวแปรสำหรับทดลองจำนวนทั้งหมด 4 ตัว ได้แก่ X1, X2, X3, และ X4 โดยกำหนด weight ให้ตัวแปรสองตัวมีค่าเท่ากับ 1 และที่เหลือมีค่าเท่ากับ 0 และ -5 ตามลำดับ ผลลัพธ์ของการวิเคราะห์แสดงให้เห็นถึงความสัมพันธ์ระหว่าง 4 กลุ่ม ที่มีความเป็นไปได้ว่าจะปรากฎอยู่ในตะกร้าเดียวกัน โดยมีการสรุปว่า อัลกอริทึมของพวกเขาสามารถค้นหา association rule จาก transaction data อีกทั้งยังสามารถรับมือกับ data cleaning กับ continuously challenge ได้. อย่างไรก็ตาม การศึกษานี้ไม่ได้ระบุผลลัพธ์เชิงประสิทธิภาพของเวลาที่ใช้ในการวิเคราะห์ข้อมูลหากนำไปประยุคต์ใช้กับสถานการณ์จริง. อีกการศึกษาของ Agus และ Triyoso ก็ได้กล่าวถึงการใช้ Neural Network เพื่อหา association rule เช่นกัน. แม้ในการศึกษานี้จะมีการระบุ epoch ทีใช้, จำนวนของ data, และ ประสิทธิภาพขอโมเดล แต่ก็ไม่ได้กล่าวถึงผลลัพธ์เชิงเวลาที่ใช้ในการทำงานแต่อย่างใด .

สำหรับวิธีการอื่นๆนอกเหนือจาก neural network ที่ได้ยกตัวอย่างไปก่อนหน้า ในการศึกษาของ (**MBA Algorithm**) ได้มีการกล่าวถึง Algorithms ต่างๆ ที่ใช้ในการหา association rule. โดยสำหรับการศึกษานี้จะยกมาทดลองด้วยกัน 2 algorithm ได้แก่ Apriori Algorithm ซึ่งเป็นอัลกอริทึมพื้นฐานที่ใช้ในการหา association ระหว่างสินค้า เนื่องจากมีความง่ายในการ implement, และ frequent pattern growth (FP-Growth) algorithm ที่เป็นแนวคิดที่ถูกปรับปรุงมาจาก Apriori. การศึกษาเกี่ยวกับ customer churn predicting (**FP growth**) ได้มีการประยุคต์ใช้อัลกริทึ่มทั้งสองในการหา frequent itemset บทความได้กล่าวถึง อัลกอริทึ่มเหล่านี้ว่า Apriori นั้นมีประสิทธิภาพมากว่าในแง่มุมของการ generate candidate แต่อาจจะเกิดปัญหา bottlenecks จากการประมวลผล complex data และการแสกน database หลายๆครั้ง ในขณะที่ FP Growth นั้นสามารถจัดการปัญหาเหล่านี้ได้ดีกว่า. ในตอนท้าย พวกเขาได้แสดงผลลัพธ์ของการทำนาย customer churn ด้วยการหา association rule แต่ไม่ได้กล่าวถึงผลลัพธ์ในเรื่องเวลาในการประมวลผลเช่นกัน.

**Methodology:**

การศึกษานี้มีจุดประสงค์ในการเปรียบเทียบความแตกต่างเชิงคุณภาพ และ execution time ระหว่าง Apriori Algorithm และ FP-Growth Algorithm โดยใช้ข้อมูล Sale transaction ของร้านค้า ในแง่มุมของข้อมูลที่ใช้ จะใช้ ข้อมูล online retail จาก UCI Machine Learning Repository (**ref**) จำนวน 25900 transaction เพื่อจำลองชุดข้อมูลขนาดใหญ่ และ ข้อมูล grocery transaction จาก arules library (**ref**) จำนวน 9835 transaction เพื่อจำลองข้อมูลขนาดเล็ก

Executer ที่ใช้ในการสร้างโมเดลประกอบไปด้วย tools ที่ต่างกันจำนวน 4 ชนิด ได้แก่

1. Rstudio – IDE สำหรับ execute R language ซึ่งเป็นภาษาที่เหมาะสำหรับการคำนวณทางสถิติ
2. Jupyter lab – IDE สำหรับ execute python language ซึ่งเป็นภาษาที่พบบ่อยในการทำ data mining
3. Matlab – โปรแกรมสำหรับวิเคราะห์ข้อมูลเชิงคณิตศาสตร์ ถูกพัฒนาสมาจากภาษา C และ C++
4. Weka – Tool สำเร็จรูปที่รวบรวม machine learninig algorithm ที่หลากหลาย

เพื่อให้ประสิทธิภาพในการทดสอบโมเดลออกมาใกล้เตียงกันที่สุด Device ที่ใช้ทดสอบจึงเป็นอุปกรณ์ชนิดเดียว โดยเป็น laptop ที่มี \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

ผลลัพธ์ที่ได้จากการวิเคราะห์จะเป็น pos-processed data จำนวนทั้งหมด 8 ชุด ประกอบไปด้วย ข้อมูลขนาดใหญ่ จาก 4 tools จำนวน 4 ชุด และ ข้อมูลขนาดเล็ก จาก 4 tools จำนวน 4 ชุด โดยผลลัพธ์แต่ละชุดจะประกอบด้วยข้อมูลประสิทธิภาพการวิเคราะห์ (Confidence) และ เวลาที่ใช้ในการทำงาน (time cost). ข้อมูลทั้งหมดจะถูกอภิปรายและเปรียบเทียบความแตกต่างของแต่ละการทดสอบ ซึ่งจะถูกสรุปและแสดงผลเชิงเปรียบเทียบในรูปแบบกราฟและตารางในส่วนท้ายของรายงาน

**Reference:**