การศึกษาประสิทธิภาพการพยากรณ์ความสนใจในผลิตภัณฑ์จากบัญชีทางการธุรกิจของแอปพลิเคชันไลน์ด้วยเทคนิคการเรียนรู้ของเครื่อง

THE STUDY ON THE EFFICIENCY OF INTEREST PREDICTION IN BUSINESS ACCOUNT PRODUCTS OF THE LINE APPLICATION USING MACHINE LEARNING TECHNIQUES

ธัญลักษณ์ เอียดเจริญ[1]
เอกสิทธิ์ พัชรวงศ์ศักดา[2]

บทคัดย่อ

บัญชีทางการของแอปพลิเคชันไลน์ สามารถสร้างฐานลูกค้าผู้ติดตามให้เพิ่มมากขึ้น โดยสื่อสาร บรอดแคสต์ และ ยังส่งข้อมูลกิจกรรม การขาย การตลาด พร้อมทั้งโปรโมชั่นพิเศษต่าง ๆ ที่จะสื่อถึงลูกค้าหลายๆรูปแบบได้ โดยธุรกิจหนึ่งมีข้อมูลในการบรอดแคส ซึ่งออกมาเป็นเอกสารรายงานประจำวันทุกวัน เพื่อดูสถิติความสนใจของลูกค้าที่เข้ามาใช้งานบัญชีของธุรกิจเท่านั้น ดังนั้นจึงเป็นที่มาของงานวิจัยครั้งนี้ซึ่งมีวัตถุประสงค์ เพื่อใช้รายงานประจำวันเพื่อดูสถิติของกลุ่มเป้าหมาย มาใช้อย่างเหมาะสมและใช้ประโยชน์ให้มากขึ้น โดยเสนอเปรียบเทียบเทคนิคการเรียนรู้ของเครื่องที่มีประสิทธิภาพมากที่สุด ที่เกี่ยวกับการคาดการณ์การมีส่วนร่วมของกลุ่มเป้าหมาย โดยใช้การศึกษาเครื่องมือ “ฟีเจอร์ ออโต้ โมเดล” บน RapidMiner Studio ในการหาแบบจำลอง และ วัดประสิทธิภาพแบบจำลองจาก Confusion Matrix โดยคำนวณเปรียบเทียบ ค่าความถูกต้อง, ค่าความระลึก, ค่าความแม่นยำและ ค่าความถ่วงดุล

ผลการวิจัยครั้งนี้พบว่า 1)พบว่าเทคนิค Gradient Boosted Trees จาก “Feature Auto Model” บน RapidMiner ให้ผลคะแนนของแบบจำลองนี้มีผลลัพธ์ดีมากที่สุด 2) ค่าร้อยละความถูกต้องจากการนำแบบจำลองพยากรณ์ในอนาคตเทียบกับข้อมูลจริง ซึ่งข้อมูลจริงมีค่า การเข้าร่วมแคมเปญ 38.93% และ การไม่เข้าร่วมแคมเปญ 61.07% พบว่าสองแบบจำลองมีค่าใกล้เคียง คือแบบจำลอง Naive Bayes ร้อยละของข้อมูลพยากรณ์ การเข้าร่วมแคมเปญ 31.33% และ การไม่เข้าร่วมแคมเปญ 68.67% และ Gradient Boosted Trees ร้อยละของข้อมูลพยากรณ์ การเข้าร่วมแคมเปญ 34.24% และ การไม่เข้าร่วมแคมเปญ 65.67% 3) แบบจำลอง Gradient Boosted Trees มีค่าการคำนวณประสิทธิภาพสูงที่สุดคือ ค่าความแม่นยำ 80.24%, ค่าความถูกต้องของข้อมูล 89.51% และ ค่าความถ่วงดุล 85.62% แต่

---

[1] นักศึกษาหลักสูตรวิศวกรรมศาสตรมหาบัณฑิต สาขาวิศวกรรมข้อมูลขนาดใหญ่

[2] ที่ปรึกษาสารนิพนธ์

ค่าความแม่นยำ 91.77% ซึ่งมีค่าน้อยกว่า แบบจำลอง Decision Tree ที่มีค่าความแม่นยำ 92.03% จากผลลัพธ์ทั้งสามข้อข้างต้นสามารถอภิปรายผลการวิจัยได้ว่า แบบจำลอง Gradient Boosted Trees มีประสิทธิภาพในการคาดการณ์ว่าลูกค้าคนใดบ้างที่จะมีส่วนร่วมในการปล่อยแคมเปญได้ดีที่สุด และ ข้อมูลเหมาะสม ในการปล่อยแคมเปญครั้งต่อไปในอนาคต

คำสำคัญ : บัญชีทางการของแอปพลิเคชันไลน์, วัดประสิทธิภาพแบบจำลอง, ฟีเจอร์ ออโต้ โมเดล

ABSTRACT

LINE Official Account can be built more customers by communicating, broadcasting, and sending sales activity. A daily report of business has after sending data to seen statistics of interest only. Therefore, it was the origin of this research whose objective was to use daily reports to see statistics of target groups appropriately and more usefully. It was a comparison of the most effective machine learning techniques related to the propensity to engagement. Used the "Feature Auto Model" tool on RapidMiner Studio to find techniques and estimate performance by Confusion Matrix.

The results found that: 1) Technique Gradient Boosted Trees by the "Feature Auto Model" on RapidMiner delivered the best score. 2) Percentage embracing future forecast models compared to actual data. Actual data had values of 38.93% campaign engaged and 61.07% campaign no-engaged. Technique Naive Bayes percentage of forecast data 31.33% campaign engaged and 68.67% campaign no-engaged and Technique Gradient Boosted Trees percentage of forecast data 34.24% campaign engaged and 65.67% campaign no-engaged. These two techniques were the closest percentages to the actual results. 3) Technique Gradient Boosted Trees has the highest efficiency calculation value of 80.24% Recall, 89.51% Accuracy, and 85.62% F-Measure but 91.77% Precision which was less than Technique Decision Tree with a Precision of 92.03% Results can be discussed: From the three results above, it can be concluded that Technique Gradient Boosted Trees was effective in predicting which engaged the campaign launch best and suitable to the data provided. It was used to create forecast models as much as possible to predict which customers will engage in future campaign launches.

Keywords: LINE Official Account, Estimate Performance, Feature Auto Model

บทนำ

บทความจาก Line Corp website Line official กล่าวว่า ในปี 2564 แอปพลิเคชันไลน์สามารถเข้าถึงคนไทยได้มากกว่า 47 ล้านคน และ ครอบคลุมผู้ใช้ทั่วประเทศไทย จากยอดการใช้งานแอปพลิเคชันไลน์ที่เพิ่มขึ้นเป็น 5 ล้านบัญชี (ณ เดือน ส.ค. 2564) มีสถิติการเติบโตเพิ่มขึ้นถึง 25% ภายในปีเดียว ด้วยเหตุนี้เองผู้ประกอบการจึงเลือกใช้งานแพลตฟอร์มนี้ นำเสนองานวิจัยชิ้นนี้ เป็นการนำข้อมูลของข้อความตัวอักษร ในรูปแบบประโยคการสื่อสาร ซึ่งเป็นรูปแบบที่มีการสื่อสาร และ บรอดแคสต์ ที่ธุรกิจส่งออกไปแบบระบุกลุ่มเป้าหมาย ซึ่งการได้มาของข้อมูลนั้น เริ่มต้นโดยลูกค้าจะต้องมีการเลือกกดที่ข้อมูลที่ส่งไป เพื่อดูข้อมูลเพิ่มเติมจาก แคมเปญ ข่าวสาร โปรโมชั่นต่าง ๆ ที่ส่งออกไปให้กับลูกค้า หลังจากนั้นหนึ่งวัน จะมีการดำเนินการรวบรวมสถิติความสนใจของการเลือกกดข้อมูล ซึ่งอยู่ในรูปแบบข้อความตัวอักษร จากของกลุ่มเป้าหมายเหล่านี้ ออกมาเป็นเอกสารรายงานประจำวันทุกวัน เพื่อสำหรับดูสถิติ ความสนใจของลูกค้าของธุรกิจ ซางข้อมูลที่ได้มาจึงที่มีขนาดใหญ่มาก เพราะ ผู้ติดตามบัญชีทางการของแอปพลิเคชันไลน์ของธุรกิจเป็นผู้ติดตามจำนวนมาก หลายกลุ่มเป้าหมาย จากข้อมูลขนาดใหญ่เหล่านี้ โดยใช้กระบวนการทาง Data Engineering เพื่อจัดการข้อมูลสำหรับใช้วิเคราะห์ เปรียบเทียบเทคนิคของแบบจำลองที่พยากรณ์ข้อมูล และ นำเสนอผลลัพธ์ด้วยวิธีการที่เหมาะสมเพื่อแสดงผลการเปรียบเทียบต่างๆ ให้เกิดประโยชน์เพิ่มขึ้น และ สามารถนำไปใช้สำหรับวิเคราะห์ในการออกแบบวางแผนการบรอดแคสต์โฆษณาในอนาคต

วัตถุประสงค์ของงานวิจัย

1. เพื่อนำข้อมูลข้อมูลรายงานประจำวันจากผลการสื่อสารบรอดแคสต์แบบระบุกลุ่มเป้าหมาย ที่ใช้สำหรับดูแค่สถิติของกลุ่มเป้าหมายที่ติดตามอยู่บนแพลตฟอร์มออนไลน์ว่ามีส่วนร่วมกับการสื่อสารกับธุรกิจในช่องทางบัญชีทางการของแอปพลิเคชันไลน์อยางเดียว เพื่อให้สามารถใช้ข้อมูลขนาดใหญ่ที่มีนั้นให้เกิดประโยชน์เพิ่มขึ้น

2. จัดทำการเสนอ พร้อมทั้งเปรียบเทียบ แบบจำลองจากเทคนิคการเรียนรู้ของเครื่องที่มีส่งผลให้มีประสิทธิภาพสูงที่สุด ที่เกี่ยวกับการคาดการณ์การมีส่วนร่วมของลูกค้าเป้าหมาย

3. เพื่อเป็นแนวทางสำหรับการคาดการณ์ว่าลูกค้าคนใดบ้างที่จะมีส่วนร่วมในการปล่อยแคมเปญ โดยมีความเหมาะสมกับข้อมูลที่นำมาใช้ในการสร้างแบบจำลอง สำหรับการพยากรณ์มากที่สุด

ขอบเขตงานวิจัย

1. ข้อมูลกรณีศึกษาในงานวิจัย คือ ข้อมูลรายงานประจำวันจากผลการสื่อสารบรอดแคสต์แบบระบุกลุ่มเป้าหมาย จากบัญชีทางการธุรกิจของแอปพลิเคชันไลน์ เริ่มต้น เดือนกรกฎาคม 2563 ถึง ตุลาคม 2563 ทั้งหมดเป็นจำนวน 405,538 ข้อมูลตัวอย่าง

2. งานวิจัยฉบับนี้เพื่อศึกษาเครื่องมือ “Feature Auto Model” บน RapidMiner Studio ในการหาแบบจำลองในการพยากรณ์ข้อมูล โดยประเมินเปรียบเทียบโดยใช้แบบจำลองทั้งหมด 9 วิธีการ ดังนี้ การจำแนกประเภทข้อมูลแบบเบย์อย่างง่าย, ตัวแบบเชิงเส้นนัยทั่วไป, การถดถอยโลจีสติก, การจำแนกประเภทแบบเร็วโดยใช้ขอบเขตการแยกแยะที่มีขนาดกว้าง, การเรียนรู้เชิงลึก, ต้นไม้การตัดสินใจ, แบบจำลองป่าสุ่ม, ต้นไม้การเพิ่มไล่ระดับการตัดสินใจ, ซัพพอร์ตเวกเตอร์แมทชีน

3. การเลือกแบบจำลองโดยการวัดประสิทธิภาพแบบจำลองจาก Confusion Matrix ซึ่งคำนวณค่าทั้ง 4 แบบคำนวณดังนี้ ค่าความถูกต้อง (Accuracy), ค่าความระลึก (Recall), ค่าความแม่นยำ (precision) และ ค่าความถ่วงดุล (F-measure)

ทฤษฎี และผลงานที่เกี่ยวข้อง

งานวิจัยนี้เน้นศึกษากระบวนการ ในการจัดการข้อมูลเพื่อให้สามารถนำมาวิเคราะห์ และ เปรียบเทียบหาเทคนิคที่สามารถพยากรณ์ได้เหมาะสมกับข้อมูล ซึ่งการศึกษาค้นคว้าหลักการ แนวคิด ทฤษฎีและงานวิจัยที่เกี่ยวข้องมีดังนี้

1. กระบวนการ Cross-Industry Standard Process for Data Mining (CRISP-DM)

เป็นมาตรฐานในลักษณะของวงจรชีวิตในการทำเหมืองข้อมูล เพื่อวิเคราะห์และนำไปใช้ประโยชน์ ประกอบไปด้วย 6 ขั้นตอนดังนี้

ขั้นตอนที่ 1 ทำความเข้าใจปัญหา (Business Understanding)

ขั้นตอนที่ 2 ทำความเข้าใจข้อมูล (Data Understanding)

ขั้นตอนที่ 3 การเตรียมข้อมูล (Data Preparation

ขั้นตอนที่ 4 การสร้างแบบจำลอง (Modeling Phase)

ขั้นตอนที่ 5 การประเมินผลแบบจำลอง (Evaluation Phase)

ขั้นตอนที่ 6 การนำไปใช้งาน (Deployment Phase)

2. โปรแกรม RapidMiner Studio and Feature Auto Model

โปรแกรม RapidMiner Studio เป็นเครื่องมือที่เหมาะสำหรับการคำนวณวิเคราะห์ประมวลผลโดยมีโอเปอร์เรเตอร์มากมายให้เลือก สามารถนำมาใช้งานได้ง่าย ตั้งแต่กระบวนการเตรียมข้อมูล จนถึงขั้นตอน สร้างแบบจำลอง ให้เป็นไปต้องการ

3. เทคนิคการเรียนรู้ของเครื่อง (Machine Learning Technique)

3.1. การจำแนกประเภทข้อมูลแบบเบย์อย่างง่าย (Naive Bayes) ซึ่งเป็นเทคนิคของการแก้ปัญหาแบบจำแนกประเภทที่สามารถคาดการณ์ ผลลัพธ์ได้ โดยวิเคราะห์ความสัมพันธ์ระหว่าง ตัวแปรในการใช้สำหรับการสร้างเงื่อนไขความน่าจะเป็น สำหรับแต่ละความสัมพันธ์ เหมาะกับกรณีของเซต ตัวอย่างที่มีจำนวนมากและ คุณลักษณะของตัวอย่างไม่ขึ้นต่อกัน

3.2. ตัวแบบเชิงเส้นนัยทั่วไป (Generalized Linear Model) ซึ่งสามารถแสดงผลการวิเคราะห์ที่สามารถบ่งบอกระดับความเชื่อมั่นทางสถิติ หรือ บ่งบอกผลการประเมินคุณภาพของตัวแปรได้ด้วย แนวคิดคือ หาความสัมพันธ์ระหว่างตัวแปรต้น (Explanatory Variables) และตัวแปรตาม (Response Variable) โดยที่ตัวแปรนำมาวิเคราะห์หาความสัมพันธ์นั้นสามารถเป็นได้ทั้งข้อมูลตัวเลข หรือ รูปแบบอื่น ๆ ที่ครอบคลุมทั้ง ตัวแปรผลลัพธ์แบบต่อเนื่องและไม่ต่อเนื่อง ซึ่งอยู่ภายใต้ฟังก์ชันการแจกแจงตระกูลเอกซ์โพเนนเชียล

3.3. การถดถอยโลจิสติก (Logistic Regression) เพื่อประมาณค่า หรือ ทำนายว่าจะเกิดเหตุการณ์ขึ้น หรือ ไม่เกิดเหตุการณ์นั้นภายใต้ปัจจัยต่างๆ ซึ่งแบบจำลองโลจิสติก จะประกอบไปด้วยตัวแปรตาม หรือ ตัวแปรเกณฑ์ ที่เป็นตัวแปรแบบทวินาม (Dichotomous Variable) และ ตัวแปรอิสระ หรือตัวแปรทำนาย อาจจะมีตัวเดียวหรือหลายตัว ซึ่งเป็นได้ทั้งตัวแปรเชิงกลุ่ม (Categorical Variable) หรือ ตัวแปรแบบต่อเนื่อง (Continuous Variable)

3.4. ซัพพอร์ตเวกเตอร์แมทชีน เป็นเทคนิคที่สามารถนำมาช่วยแก้ปัญหาในการจำแนกข้อมูลได้ โดยเฉพาะกับปัญหาของข้อมูลที่มีขนาดของข้อมูลไม่ใหญ่มากนักแต่คุณลักษณะ (features) ของข้อมูลมีจำนวนมาก มันอาศัยหลักการหาสัมประสิทธิ์ของสมการ ในการสร้างเส้นแบ่งแยกกลุ่มข้อมูล

3.5. การจำแนกประเภทแบบเร็วโดยใช้ขอบเขตการแยกแยะที่มีขนาดกว้าง เป็นเทคนิคที่อ้างอิงมาจาก the linear support vector learning ของ R.E. Fan, K.W. Chang, C.J. Hsieh, X.R. Wang, และ C.J. Lin. ซึ่งเทคนิคนี้เป็นเทคนิคใหม่ที่มีการคำนวณผลลัพธ์คล้ายกับเทคนิคของวิธีซัพพอร์ตเวกเตอร์แมทชีน สามารถประมวลผลข้อมูลตัวอย่างและแอททริบิวต์(Attribute) ที่มีจำนวนมากได้

3.6. การเรียนรู้เชิงลึก ซึ่งประมวลผลข้อมูลจำนวนมากขนาดใหญ่ ที่เลียนแบบการทำงานของโครงข่ายประสาทของมนุษย์ (Neurons) ทำให้สามารถทำการตัดสินใจด้วยการขับเคลื่อนด้วยข้อมูลได้อย่างแม่นยำ

3.7. ต้นไม้การตัดสินใจ (Decision Tree) วิธีการเรียนรู้ที่ใช้สถิติจากการเรียนรู้ของเครื่อง และการทำเหมืองข้อมูลจะสังเกตการณ์แบ่งแยกข้อมูล ซึ่งเป็นกระบวนการวิเคราะห์ข้อมูลที่นักวิทยาศาสตร์ข้อมูลทั้งมือใหม่และมือเก่านิยมใช้อย่างแพร่หลาย เนื่องจากเป็นแบบจำลองที่ง่ายต่อการเข้าใจ และ แปรผลลัพธ์ออกมา

3.8. แบบจำลองป่าสุ่ม (Random Forest) หลักการทำงานของเทคนิคนี้ คือจะแบ่งข้อมูลออกเป็น ต้นไม้ตัดสินใจ (Decision Tree) หลายๆ รูปแบบจำลองย่อย ๆ โดยแต่ละแบบจำลองจะได้รับคุณลักษณะ (Feature) และข้อมูล (Data) ที่ไม่เหมือนกันทั้งหมด เพื่อทำให้ได้ต้นไม้ที่มีความหลากหลายและมีความอิสระต่อกัน

4. การวัดประสิทธิภาพของแบบจำลอง

จำเป็นจะต้องมีการวัดประสิทธิภาพของแบบจำลองดังกล่าวว่าแบบจำลองมีประสิทธิภาพเพียงพอที่จะนำมาพัฒนาต่อและสามารถนำใช้งานได้จริงหรือไม่ ตาราง Confusion Matrix เป็นหนึ่งในการวัดประสิทธิภาพประเภทหนึ่งที่เป็นที่นิยม สำหรับในงานทางด้าน Data มี 3 การคำนวณที่เป็นที่นิยมซึ่งนำมาใช้ในงานวิจัยนี้ มีการคำนวณดังนี้

- ค่าความถูกต้อง (Accuracy) หมายถึง ค่าที่วัดได้เข้าใกล้ค่าจริงมากน้อยเพียงใด ถือว่ามีความถูกต้องมาเท่านั้น

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN}$$

- ค่าความครบถ้วน (Recall) หมายถึง อัตราส่วนของการค้นพบข้อมูลที่ถูกต้องจากข้อมูลที่ถูกต้องทั้งหมด

$$Recall = \frac{TP}{TP + FN}$$

- ค่าความแม่นยำ (Precision) หมายถึง ความแม่นยำของผลทำนายโดยสนใจผลทำนาย หรือ Prediction คำนวณเป็นค่าสัดส่วนร้อยละเท่าไหร่

$$Percision = \frac{TP}{TP + FP}$$

- ค่าความถ่วงดุล (F-measure) คือ ค่าที่ได้จากการเอาค่าความแม่นยำ และ ค่าความครบถ้วน มาพิจารณาร่วมกันเฉลี่ยแบบ ค่าเฉลี่ยฮาร์มอนิก (harmonic mean)

$$F - meansure = 2 * \left(\frac{Percision * Recall}{Percision + Recall}\right)$$

โดย True Positive (TP) คือ ความถี่ของสิ่งที่ทำนาย ตรงกับสิ่งที่เกิดขึ้นจริง
True Negative (TN) คือ ความถี่ของสิ่งที่ทำนายตรงกับสิ่งที่เกิดไม่จริง
False Positive (FP) คือ ความถี่ของสิ่งที่ทำนายไม่ตรงกับสิ่งที่เกิดขึ้นจริง
False Negative (FN) คือ ความถี่ของสิ่งที่ทำนายไม่ตรงกับที่เกิดขึ้นไม่จริง

5. บทความและงานวิจัยที่เกี่ยวข้อง

สุรวัชร ศรีเปารยะ และสายชล สินสมบูรณทอง (2560) งานวิจัยการเปรียบเทียบประสิทธิภาพวิธีการจำแนกกลุ่มการ เป็นโรคไตเรื้อรัง มีวัตถุประสงค์เพื่อเปรียบเทียบประสิทธิภาพของวิธีการจำแนกกลุ่มการเป็นโรคไตเรื้อรังของโรงพยาบาลแห่งหนึ่ง โดยเลือกใช้ 7วิธี ได้แก่ วิธีความใกล้เคียงกันมากที่สุด วิธีต้นไม้ตัดสินใจวิธีโครงข่ายประสาทเทียม วิธีซัพพอร์ตเวกเตอร์แมชชีน วิธีฐานกฎวิธีการถดถอยโลจิสติก และวิธีนาอีฟ เบย์ ซึ่งผลจากการ

เปรียบเทียบประสิทธิภาพพบว่าวิธีต้นไม้ตัดสินใจ เป็นวิธีที่มีค่าความถูกต้องและค่าความคลาดเคลื่อนกำลังสองเฉลี่ยมากที่สุดคือพบ 100%และ 0.0059 ตามลำดับ ดังนั้นจึสรุปได้ว่า วิธีต้นไม้ตัดสินใจ ในการเปรียบเทียบมีประสิทธิภาพการจำแนกดีที่สุดสำหรับข้อมูลดังกล่าวที่นำมาใช้กับงานวิจัยชิ้นนี้

Oleksandr Andrieiev.(2022) อธิบายว่า Propensity Model คือแบบจำลองที่ทำให้เข้าใจว่า กลุ่มเป้าหมายหรือลูกค้า ตอบสนองอย่างไรกับการส่งข้อมูลหรือดำเนินการบางประเภทออกไป เป็นการทำนายพฤติกรรมการมีส่วนร่วมของลูกค้า ซึ่งการมีส่วนร่วมนั้นจะช่วยให้ธุรกิจ สามารถประเมินแนวโน้มของลูกค้าที่จะคาดว่าจะมีส่วนร่วมกับ สิ่งที่จะส่งออกหรือโฆษณา ไปให้กับกลุ่มคนเหล่านั้น ยกตัวอย่างเช่น เป็นความน่าจะเป็นของคะแนนความชอบที่แสดงว่าผู้เยี่ยมชมเว็บไซต์คนใดบ้างจะคลิกโฆษณา หรือ ประชาชนคนประเภทไหนอาจลงคะแนนให้พรรคที่กำหนดในการเลือกตั้ง เป็นต้น

LAURA LAHINDAH and IVAN DIRYANA SUDIRMAN.(2023) งานวิจัย CLASSIFICATION APPROACH TO PREDICT CUSTOMER DECISION BETWEEN PRODUCT BRANDS BASED ON CUSTOMER PROFILE AND TRANSACTION เป็นการเปรียบเทียบ machine learning models ดังนี้ naive bayes, linear models, deep learning, and decision trees ใช้โปรแกรม RapidMiner โดยผลลัพธ์ที่ได้ระบุว่า naive bayes สามารถทำงานได้ดีที่สุดในฐานะอัลกอริทึมการทำนาย ซึ่งมีค่า Precision เท่ากับร้อยละ 60.3 ,Recall เท่ากับร้อยละ 65.5 และ Accuracy เท่ากับร้อยละ 61.7 ซึ่งมีค่าสูงที่สุด แต่ปัจจัยด้านประชากรศาสตร์ เช่น อายุและสถานที่ ตลอดจนปัจจัยเชิงบริบท เช่น ช่วงเวลาของสัปดาห์ มีอิทธิพลอย่างมากต่อแนวโน้มการซื้อของลูกค้า

ระเบียบวิธีวิจัย

ใช้หลักการ Cross-Industry Standard Process For Data Mining (CRISPDM) เป็นกระบวนการมาตรฐานที่นิยมใช้ในการวิเคราะห์ข้อมูลสำหรับการทำเหมืองข้อมูลมาประยุกต์ใช้ ซึ่งมีรายละเอียดดังนี้

1. ทำความเข้าใจปัญหา (Business Understanding Phase)

การส่งออกข้อมูลให้ครอบคลุมของธุรกิจใหญ่ๆนั้น จะต้องใช้ ทรัพยากรและเวลาจำนวนมากใน การจัดทำโหลดเข้าและส่งออกข้อมูล โดยการโฆษณาบรอดแคสในแต่ละครั้ง ค่าใช้จ่ายในการส่งข้อมูลแบบกระจายให้ครอบคลุมทุก ๆ กลุ่มลุ่มลูกค้าที่มีนั้นจะถูกคิดจากปริมาณในการส่งออกไปเป็นรายบุคคล หากส่งมากๆก็จะจ่ายมาก โดยนำเข้าสู่กระบวนการ และ เปรียบเทียบเทคนิคต่าง ๆ ที่ใช้แบบจำลองในการพยากรณ์ข้อมูล เพื่อนำมาใช้ในการปรับปรุงการคาดการณ์ว่าลูกค้าคนใดบ้างที่จะมีส่วนร่วมในการปล่อยแคมเปญครั้งต่อไป

2. ทำความเข้าใจข้อมูล (Data Understanding Phase)

กำหนดประชากรก็คือ ลูกค้าที่เป็นสมาชิกจากบัญชีทางการธุรกิจของแอปพลิเคชันไลน์ของธุรกิจหนึ่ง และข้อมูลที่สนใจคือ การบันทึกคำตอบจากการกดคลิกสนใจแคมเปญในครั้งแรกที่ถูกส่งไปถึงลูกค้าที่เป็นสมาชิกคนนั้น ๆ ซึ่งสามารถแบ่งกลุ่มข้อมูลได้สองประเภทดังตารางต่อไปนี้

ตารางที่ 1 แบ่งประเภทกลุ่มเป้าหมายออกเป็น segment 5 กลุ่มและมีความหมาย ดังนี้

ประเภทกลุ่มเป้าหมาย	คำนิยาม
AF (Affluence)	กลุ่มลูกค้าที่มีปริมาณสินทรัพย์เยอะมากตามปริมาณที่ธุรกิจกำหนด
MOL (Money loan)	กลุ่มลูกค้าเงินกู้
Payroll	กลุ่มลูกค้าที่รับเงินค่าตอบแทนผ่านช่องทางของธุรกิจทุกเดือน
GNI (Generic)	กลุ่มลูกค้าทั่วไป
UNI (university)	กลุ่มลูกค้านักเรียนนักศึกษา

ตารางที่ 2 แบ่งประเภทของแคมเปญออกเป็น 20 ประเภท และมีความหมายดังนี้

ประเภทของแคมเปญ	คำนิยาม
PRODUCTS	แคมเปญประเภทที่เกี่ยวข้องกับผลิตภัณฑ์ของธุรกิจ
INVEST	แคมเปญประเภทที่เกี่ยวข้องกับการลงทุน
MARKET	แคมเปญประเภทที่เกี่ยวข้องกับการค้าขาย
VARIETY	แคมเปญประเภทที่เกี่ยวข้องกับการร่วมสนุกหลายๆชนิด
WORKER	แคมเปญประเภทที่เกี่ยวข้องกับการทำงาน
LEARN	แคมเปญประเภทที่เกี่ยวข้องกับการเรียนรู้
PROMOTION	แคมเปญประเภทที่เกี่ยวข้องกับโปรโมชั่นของผลิตภัณฑ์ของธุรกิจ
CAR	แคมเปญประเภทที่เกี่ยวข้องกับประกันรถยนต์
COVID	แคมเปญประเภทที่เกี่ยวข้องกับการให้ความรู้ไวรัสโควิด-19
TECH	แคมเปญประเภทที่เกี่ยวข้องกับเทคโนโลยี
INSURANCE	แคมเปญประเภทที่เกี่ยวข้องกับประกันสุขภาพ
TAX	แคมเปญประเภทที่เกี่ยวข้องกับภาษีประจำปี
GAMER	แคมเปญประเภทที่เกี่ยวข้องกับเกมส์
TRAVEL	แคมเปญประเภทที่เกี่ยวข้องกับการท่องเที่ยวธรรมชาติ
HAUNTINGTRIP	แคมเปญประเภทที่เกี่ยวข้องกับการท่องเที่ยวกินดื่ม ใช้จ่าย
FEATURE	แคมเปญประเภทที่เกี่ยวข้องกับตกแต่งบ้าน
LEARNTHINK	แคมเปญประเภทที่เกี่ยวข้องกับข่าวสารการเงิน
HOROSCOPE	แคมเปญประเภทที่เกี่ยวข้องกับการดูดวง
ACTIVITY	แคมเปญประเภทที่เกี่ยวข้องกับกิจกรรมแอดเวนเจอร์
TREE	แคมเปญประเภทที่เกี่ยวข้องกับบ้านและสวน

ตัวอย่างข้อมูลของรายงานที่นำมาใช้งานมีดังนี้

1. ข้อมูลรายงานประจำวันจากผลการสื่อสารบรอดแคสต์แบบระบุกลุ่มเป้าหมาย จากบัญชีทางการธุรกิจของแอปพลิเคชันไลน์ เริ่มต้น เดือนกรกฎาคม 2563 ถึง ตุลาคม 2563 คิดเป็นเดือนทั้งหมด 3 เดือน และคิดเป็นวันทั้งหมด 92 วัน โดยข้อมูลมีจำนวน 405,538 ข้อมูลตัวอย่าง โดยเก็บเป็นข้อมูลรายวันและรายแคมเปญเป็นไฟล์ Excel เดือนกรกฎาคม 2563 มี 101 ไฟล์ เดือนสิงหาคม 2563 มี 84 ไฟล์ และ เดือน กันยายน 2563 มี 122 ไฟล์ รวมทั้งหมด 307 ไฟล์ โดยที่คุณลักษณะที่สนใจและนำมาใช้ในการดำเนินการมีดังนี้ User_id : Line Social ID ของลูกค้า, Line_Name: Line Display ของลูกค้า, Time stamp : ช่วงเวลาที่กด campaign ครั้งแรก, Campaing_name: ชื่อแคมเปญ, Answer : #***ข้อความคำตอบของแคมเปญ และ Segment : segment ของกลุ่มลูกค้าคนนั้น ๆ

2. ข้อมูลชุดชื่อรายการบรอดแคสต์และคำตอบทั้งหมด เป็นจำนวนทั้งหมด 400 แคมเปญ โดยที่คุณลักษณะที่สนใจและนำมาใช้ในการดำเนินการมีดังนี้ category_ID : ID ของประเภทแคมเปญ, campainge_name_sum : ชื่อแคมเปญ, answer : #***ข้อความคำตอบของแคมเปญ และ category : ประเภทของแคมเปญ

3. การเตรียมข้อมูล (Data Preparation Phase)

3.1. การจัดรูปแบบข้อมูล (Data format) จากข้อมูลดิบ (raw data) ของรายงานประจำวันจากผลการสื่อสารบรอดแคสต์แบบระบุกลุ่มเป้าหมาย จากบัญชีทางการธุรกิจของแอปพลิเคชันไลน์ที่ได้รวบรวมมานั้น มีการบันทึก ชื่อแคมเปญเป็นชื่อของตาราง ดังนั้นจึงจะต้องแก้ไขจัดชื่อของตารางใหม่จาก user_id ,name ,timestamp และ ชื่อของแคมเปญ มาเป็น user_id ,line_name ,datestamp ,timestamp ,campainge_name ,answer และ segment

3.2. การเก็บรวบรวมข้อมูล (Data Collection เนื่องจากข้อมูลรายงาน first response จากการ broadcast campaign ที่มีนั้นเป็นข้อมูลที่อยู่ในรูปแบบไฟล์ Excel ที่มีทั้งหมด 307 ไฟล์ จึงต้องการรวบรวมข้อมูลให้เป็นหนึ่งเดียวกันดังนั้นจึงจำเป็นจะต้องเลือกใช้งาน Operation Loop file เพื่อวนลูป และ Operation Read Excel เพื่ออ่านข้อมูลไฟล์ที่ละไฟล์ แล้วใช้ Operation Append เพื่อรวมข้อมไฟล์ทั้งหมดแต่ละเดือน และ ใช้ Operation Append อีกครั้ง เพื่อผนวกข้อมูลให้เป็นตารางเดียวกันทั้งหมด จากนั้นใช้ Operation Store เพื่อบันทึกข้อมูลทั้งหมดลงตารางให้สามารถเข้าสู่กระบวนการวิเคราะห์และสร้างแบบจำลองในการใช้งานในโปรแกรม RapidMiner

3.3. เตรียมข้อมูลให้พร้อม (Data Preparation) ดำเนินการคัดเลือกข้อมูล ที่ใช้ประโยชน์ได้ ให้เพียงพอและพร้อมที่จะใช้งานให้ได้ตรงกับความต้องการในทันทีในการนำมาวิเคราะห์ และ เกิดความผิดพลาดลดลง

3.4. การแปลงรูปแบบข้อมูล (Data Transformation) เมื่อแปลงข้อมูลที่ต้องการนำไปวิเคราะห์เรียบร้อยแล้ว จึงทำการรวมและจัดกลุ่มข้อมูลให้เป็นมูลชุดเดียวกัน และนำเข้าแบบจำลองเพื่อทำการพยากรณ์ข้อมูลในขั้นตอนไป

4. การสร้างแบบจำลอง (Modeling Phase)

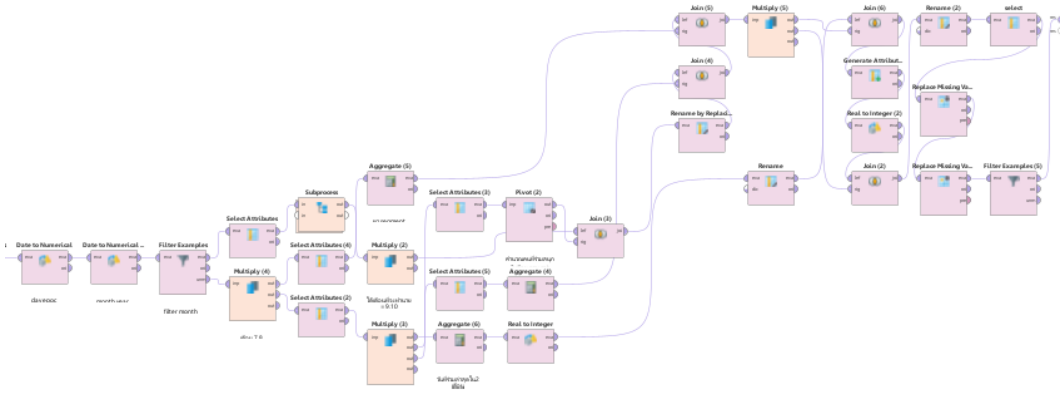
นำข้อมูลที่ เตรียมจากขั้นตอนก่อนหน้ามาทำการสร้างแบบจำลอง Propensity Model เพื่อการทำนายเฉพาะแคมเปญที่สำคัญที่สุด และ จากความถี่ของความสนใจของคมเปญ ที่รวมผลจากค่า 0 และ1 ก่อนหน้าออกมา จำแนกแยกประเภทพบว่าเป็นแคมเปญที่มีลูกค้ามีส่วนร่วมมากที่สุดเป็นอันดับหนึ่ง คือ PRODUCTS ซึ่งมีลูกค้าที่สนใจมาก

ที่สุดจำนวน 112,012 คน ขั้นตอนการสร้างแบบจำลองจะแบ่งข้อมูลออกแบบ Time Frame for Modeling และ สร้างตัวแปรใหม่เพิ่มจากข้อมูลตัวแปรเดิมที่มี (derived variable) แสดงดังตาราง

ตารางที่ 3 ตาราง Time Frame for Modeling

July	August	September and October
46	16	31
observation period	latency period	event outcome period

เมื่อวางแผน Time Frame for Modeling เรียบร้อยแล้วจึงนำข้อมูลมาดำเนินการสร้างแบบจำลอง Propensity โดยใช้ Operation ดำเนินการดังนี้



ภาพที่ 1 Operation ของแบบจำลอง Propensity ภาพรวมทั้งหมดในโปรแกรม RapidMiner

หลังจากนำข้อมูลมาดำเนินการสร้างแบบจำลอง Propensity จะดำเนินสร้างตัวแปรใหม่เพิ่มจากข้อมูลตัวแปรเดิมที่มี (derived variable) โดยจะสร้างตัวแปรใหม่ออกมาดังตารางดังนี้

ตารางที่ 4 ตารางตัวแปรใหม่เพิ่มจากข้อมูลตัวแปรเดิมที่มี

ชื่อตารางหลังเข้า Propensity Process	คำนิยาม
user_id	รหัสของลูกค้า
count(answer)	จำนวนรวมการเข้าร่วมแคมเปญทุกประเภท
Recency_Category_Product	Recency_Category_Product = End of Training date - Sub_max_date
Frequency_Category__ACTIVITY	จำนวนการเข้าร่วมแคมเปญประเภทที่เกี่ยวข้องกับกิจกรรมแอดเวนเจอร์
Frequency_Category__CAR	จำนวนการเข้าร่วมแคมเปญประเภทที่เกี่ยวข้องกับประกันรถยนต์
Frequency_Category__COVID	จำนวนการเข้าร่วมแคมเปญประเภทที่เกี่ยวข้องกับการให้ความรู้ไวรัสโควิด-19
Frequency_Category__GAMER	จำนวนการเข้าร่วมแคมเปญประเภทที่เกี่ยวข้องกับเกมส์
Frequency_Category__HAUNTINGTRIP	จำนวนการเข้าร่วมแคมเปญประเภทที่เกี่ยวข้องกับการท่องเที่ยวในเมือง
Frequency_Category__HOROSCOPE	จำนวนการเข้าร่วมแคมเปญประเภทที่เกี่ยวข้องกับการดูดวง
Frequency_Category__INSURANCE	จำนวนการเข้าร่วมแคมเปญประเภทที่เกี่ยวข้องกับประกันสุขภาพ
Frequency_Category__INVEST	จำนวนการเข้าร่วมแคมเปญประเภทที่เกี่ยวข้องกับการลงทุน
Frequency_Category__LEARN	จำนวนการเข้าร่วมแคมเปญประเภทที่เกี่ยวข้องกับการเรียนรู้
Frequency_Category__MARKET	จำนวนการเข้าร่วมแคมเปญประเภทที่เกี่ยวข้องกับการค้าขาย
Frequency_Category__PRODUCTS	จำนวนการเข้าร่วมแคมเปญประเภทที่เกี่ยวข้องกับผลิตภัณฑ์ของธุรกิจ
Frequency_Category__PROMOTION	จำนวนการเข้าร่วมแคมเปญประเภทที่เกี่ยวข้องกับโปรโมชั่นของผลิตภัณฑ์ของธุรกิจ
Frequency_Category__TAX	จำนวนการเข้าร่วมแคมเปญประเภทที่เกี่ยวข้องกับภาษีประจำปี
Frequency_Category__TECH	จำนวนการเข้าร่วมแคมเปญประเภทที่เกี่ยวข้องกับเทคโนโลยี
Frequency_Category__TRAVEL	จำนวนการเข้าร่วมแคมเปญประเภทที่เกี่ยวข้องกับการท่องเที่ยวธรรมชาติ
Frequency_Category__VARIETY	จำนวนการเข้าร่วมแคมเปญประเภทที่เกี่ยวข้องกับการร่วมสนุกหลายๆชนิด
Frequency_Category__WORKER	จำนวนการเข้าร่วมแคมเปญประเภทที่เกี่ยวข้องกับการทำงาน
Frequency_Category__FEATURE	จำนวนการเข้าร่วมแคมเปญประเภทที่เกี่ยวข้องกับตกแต่งบ้าน
Frequency_Category__LEARNTHINK	จำนวนการเข้าร่วมแคมเปญประเภทที่เกี่ยวข้องกับข่าวสารการเงิน
Frequency_Category__TREE	จำนวนการเข้าร่วมแคมเปญประเภทที่เกี่ยวข้องกับบ้านและสวน
Label	เข้าร่วมแคมเปญประเภทที่เกี่ยวข้องกับผลิตภัณฑ์ของธุรกิจหรือไม่

5. การประเมินผลแบบจำลอง (Evaluation Phase)

การประเมินผลแบบจำลองจะใช้การวัดประสิทธิภาพแบบจำลองจาก Confusion Matrix ซึ่งประสิทธิภาพของผลลัพธ์ว่ามีความน่าเชื่อถือมากน้อยเพียงใด โดยคำนวณค่าทั้ง 4 แบบคำนวณดังนี้

- วัดค่าความถูกต้องของแบบจำลองโดยใช้ค่าความถูกต้องของข้อมูล (Accuracy)
- วัดค่าสัดส่วนของข้อมูลที่ถูกต้องกับข้อมูลที่ต้องการทั้งหมด ค่าความระลึก (Recall)
- วัดค่าสัดส่วนของข้อมูลที่ถูกต้อง และ ตรงตามความต้องการส่วนด้วยข้อมูลทั้งหมดค่า ความแม่นยำ (Precision)
- วัดค่าเฉลี่ยแบบ ค่าเฉลี่ยฮาร์มอนิก (harmonic mean) โดยจะนำ “ค่าความแม่นยำ” และ “ค่าความระลึก” มาพิจารณาร่วมกัน ค่าความถ่วงดุล (F-measure)

6. การนำไปใช้งาน (Deployment Phase)

นำผลลัพธ์ของการเปรียบเทียบประสิทธิภาพ แสดงผลบน Dashboard เพื่อนำแบบจำลองที่ดีที่สุด โดยหลังจากที่วิจัย และ เปรียบเทียบค่าความคลาดเคลื่อนของแต่ละเทคนิคแล้ว แบบจำลองใดมีประสิทธิภาพมากที่สุด และมีความเหมาะสมกับความต้องการของการพยากรณ์ข้อมูลในอนาคตจะถูกแนะนำให้นำมาประยุกต์ใช้โดยนำเทคนิคที่ดีที่สุด

ผลการวิจัย

1. ผลการเปรียบเทียบประสิทธิภาพการทำนายผล

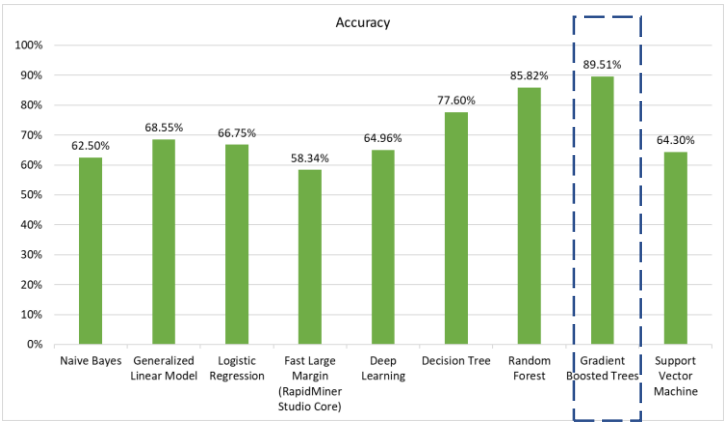
จาก “Feature Auto Model” ได้ข้อผลของประสิทธิภาพผลเปรียบเทียบค่าความถูกต้อง ของแบบจำลองโดยใช้ค่าความถูกต้องของข้อมูล, ค่าสัดส่วนของข้อมูลที่ถูกต้องกับข้อมูลที่ต้องการทั้งหมด ค่าความระลึก, ค่าสัดส่วนของข้อมูลที่ถูกต้อง ตรงตามความต้องการส่วนด้วยข้อมูลทั้งหมดค่าความแม่นยำ และ ค่าเฉลี่ยแบบ ค่าเฉลี่ยฮาร์มอนิก โดยจะนำ “ค่าความแม่นยำ” และ “ค่าความระลึก” มาพิจารณาร่วมกัน ค่าความถ่วงดุล (F-measure) ตามลำดับดังนี้

ตารางที่ 5 ผลการวิเคราะห์ประสิทธิภาพของแบบจำลอง

Model	Accuracy	Precision	Recall	F Measure
Naive Bayes	62.50%	52.27%	42.34%	46.78%
Generalized Linear Model	68.55%	65.15%	41.36%	50.59%
Logistic Regression	66.75%	62.55%	36.45%	46.04%
Fast Large Margin (RapidMiner Studio Core)	58.34%	47.10%	57.32%	51.71%

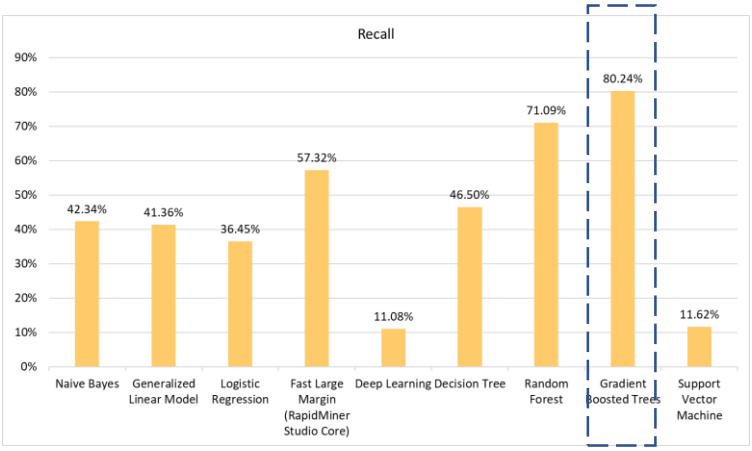
ตารางที่ 5 (ต่อ)

Model	Accuracy	Precision	Recall	F Measure
Deep Learning	64.96%	91.00%	11.08%	19.76%
Decision Tree	77.60%	92.03%	46.50%	61.78%
Random Forest	85.82%	90.45%	71.09%	79.61%
Gradient Boosted Trees	89.51%	91.77%	80.24%	85.62%
Support Vector Machine	64.30%	78.20%	11.62%	20.22%



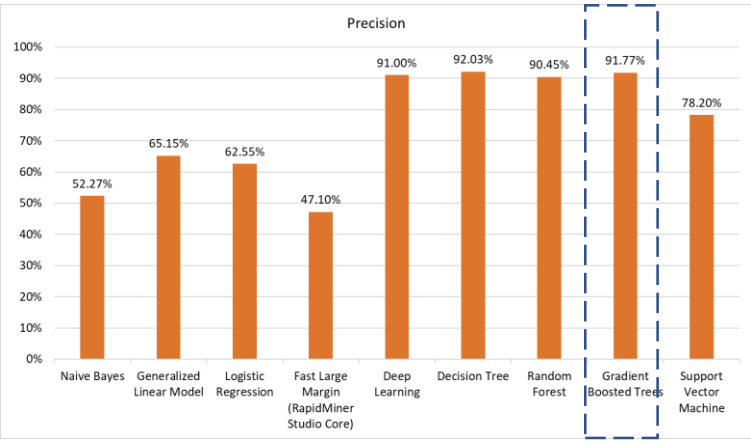
ภาพที่ 2 ผลความถูกต้องของแบบจำลองโดยใช้ค่าความถูกต้องของข้อมูล (Accuracy)

ค่าความระลึก (Recall) โดยเปรียบเทียบกับ 9 แบบจำลองสามารถแจกแจงประสิทธิภาพได้ดังนี้



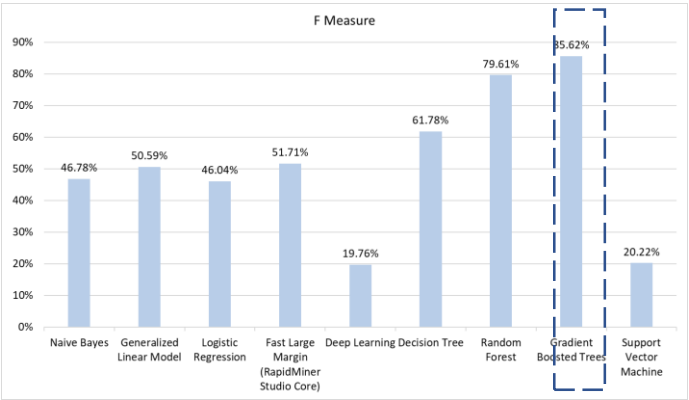
ภาพที่ 3 ผลค่าสัดส่วนของข้อมูลที่ถูกต้องกับข้อมูลที่ต้องการทั้งหมด ค่าความระลึก (Recall)

ค่าความแม่นยำ (Precision) โดยเปรียบเทียบกับ 9 แบบจำลองสามารถแจกแจงประสิทธิภาพได้ดังนี้



ภาพที่ 4 ผลค่าสัดส่วนของข้อมูลที่ถูกต้อง ตรงตามความต้องการส่วนด้วยข้อมูลทั้งหมดค่าความแม่นยำ (Precision)

ค่าความถ่วงดุล (F-measure) โดยเปรียบเทียบกับ 9 แบบจำลองสามารถแจกแจงประสิทธิภาพได้ดังนี้



ภาพที่ 5 ผลค่าเฉลี่ยแบบ ค่าเฉลี่ยฮาร์มอนิก (harmonic mean) โดยจะนำ “ค่าความแม่นยำ” และ “ค่าความระลึก” มาพิจารณาร่วมกัน ค่าความถ่วงดุล (F-measure) โดยเปรียบเทียบกับ 9 แบบจำลอง

2. ผลความถูกต้องจากการนำแบบจำลองพยากรณ์ในอนาคตเทียบกับข้อมูลจริง

ซึ่งข้อมูลจริงมีค่า Engaged(เข้าร่วมแคมเปญ) ร้อยละ 38.93 และ Not_ Engaged(ไม่เข้าร่วมแคมเปญ) ร้อยละ 61.07

2.1. วิธีการจำแนกประเภทข้อมูลแบบเบย์อย่างง่าย (Naive Bayes) โดยข้อมูลพยากรณ์ Engaged(เข้าร่วมแคมเปญ) มีที่ร้อยละ 31.33 และ Not_ Engaged(ไม่เข้าร่วมแคมเปญ) มีที่ร้อยละ 68.67

2.2. วิธีตัวแบบเชิงเส้นนัยทั่วไป (Generalized Linear Model) โดยข้อมูลพยากรณ์ Engaged(เข้าร่วมแคมเปญ) มีที่ร้อยละ 25.03 และ Not_ Engaged(ไม่เข้าร่วมแคมเปญ) มีที่ร้อยละ 74.97

2.3. วิธีการถดถอยโลจีสติก (Logistic Regression) โดยข้อมูลจริง Engaged(เข้าร่วมแคมเปญ) มีที่ร้อยละ 38.93 และ Not_ Engaged(ไม่เข้าร่วมแคมเปญ) มีที่ร้อยละ 61.07

2.4. วิธีจำแนกประเภทโดยใช้ขอบเขตที่มีขนาดกว้างที่สุดแบบเร็ว (Fast Large Margin (RapidMiner Studio Core)) โดยข้อมูลพยากรณ์ Engaged(เข้าร่วมแคมเปญ) มีที่ร้อยละ 47.22 และ Not_ Engaged(ไม่เข้าร่วมแคมเปญ) มีที่ร้อยละ 52.78

2.5. วิธีการเรียนรู้เชิงลึก (Deep Learning) โดยข้อมูลพยากรณ์ Engaged(เข้าร่วมแคมเปญ) มีที่ร้อยละ 4.74 และ Not_ Engaged(ไม่เข้าร่วมแคมเปญ) มีที่ร้อยละ 95.26

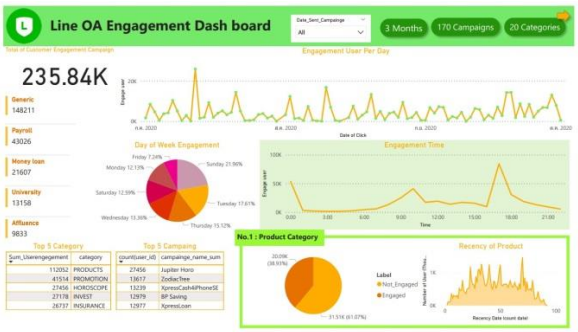
2.6. วิธีต้นไม้การตัดสินใจ (Decision Tree) โดยข้อมูลพยากรณ์ Engaged(เข้าร่วมแคมเปญ) มีที่ร้อยละ 19.69 และ Not_ Engaged(ไม่เข้าร่วมแคมเปญ) มีที่ร้อยละ 80.31

2.7. วิธีแบบจำลองป่าสุ่ม (Random Forest) โดยข้อมูลพยากรณ์ Engaged(เข้าร่วมแคมเปญ) มีที่ร้อยละ 30.73 และ Not_ Engaged(ไม่เข้าร่วมแคมเปญ) มีที่ร้อยละ 69.27

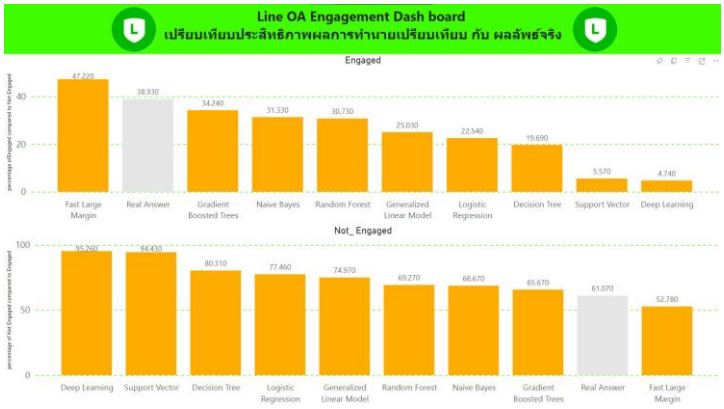
2.8. วิธีต้นไม้การเพิ่มไล่ระดับการตัดสินใจ (Gradient Boosted Trees) โดยข้อมูลพยากรณ์ Engaged(เข้าร่วมแคมเปญ) มีที่ร้อยละ 34.24 และ Not_ Engaged(ไม่เข้าร่วมแคมเปญ) มีที่ร้อยละ 65.67

2.9. วิธีซัพพอร์ตเวกเตอร์แมทชีน (Support Vector) โดยข้อมูลพยากรณ์ Engaged(เข้าร่วมแคมเปญ) มีที่ร้อยละ 5.57 และ Not_ Engaged(ไม่เข้าร่วมแคมเปญ) มีที่ร้อยละ 94.43

3. ผลของการแสดงผลเพื่อนำเสนอ



ภาพที่ 6 ตัวอย่างหน้าจอการแสดงผลการพยากรณ์ของข้อมูลสำหรับผู้ใช้งานหน้าที่ 1



ภาพที่ 7 ตัวอย่างหน้าจอการแสดงผลการพยากรณ์ของข้อมูลสำหรับผู้ใช้งานหน้าที่ 2

สรุปและอภิปรายผลการวิจัย

งานวิจัยชิ้นนี้ได้เสนอการเปรียบเทียบการนำชุดข้อมูล เข้าแบบจำลองทั้ง 9 แบบจำลอง โดยได้ทำ การวัดประสิทธิภาพของแต่ละแบบจำลอง เพื่อนำผลลัพธ์มาเปรียบเทียบและนำเสนอเทคนิคที่ดีและเหมาะสมที่สุดใช้สำหรับการพยากรณ์ข้อมูล สำหรับข้อมูลที่มีลักษณะหรือรูปแบบเดียวกันกับที่ใช้ในงานวิจัยในการใช้ วิเคราะห์เพื่อคาดการณ์ว่าลูกค้าคนใดบ้างที่คาดว่าจะมีส่วนร่วมในการปล่อยแคมเปญครั้งต่อไป สามารถสรุปผลการวิจัยได้ดังนี้

1. พบว่าเทคนิคการเรียนรู้ของเครื่องด้วยวิธีการของ ต้นไม้การเพิ่มไล่ระดับการตัดสินใจ (Gradient Boosted Trees) โดย “Feature Auto Model” บน RapidMiner ให้ผลคะแนนของแบบจำลองนี้ดีมากที่สุดใน 9 แบบจำลองข้างต้น จากประสิทธิภาพมากที่สุด และ มีความเร็วของการทำนายมากที่สุด

2. จากผลการเปรียบเทียบความถูกต้องจากการนำแบบจำลองพยากรณ์ในอนาคตเทียบกับข้อมูลจริง ซึ่งข้อมูลจริงมีค่า การเข้าร่วมแคมเปญ 38.93% และ การไม่เข้าร่วมแคมเปญ 61.07% พบว่าสองแบบจำลองมีค่าใกล้เคียงคือแบบจำลอง Naive Bayes ร้อยละของข้อมูลพยากรณ์ การเข้าร่วมแคมเปญ 31.33% และ การไม่เข้าร่วมแคมเปญ 68.67% และ Gradient Boosted Trees ร้อยละของข้อมูลพยากรณ์ การเข้าร่วมแคมเปญ 34.24% และ การไม่เข้าร่วมแคมเปญ 65.67%

ตารางที่ 15 ผลการวิเคราะห์ประสิทธิภาพแบบจำลองที่ดีที่สุดที่มีค่าประสิทธิภาพมากกว่า ร้อยละ 80 จะได้แบบจำลอง 4 เทคนิค

	Deep Learning	Decision Tree	Random Forest	Gradient Boosted Trees
Accuracy	64.96%	77.60%	85.82%	89.51%
Recall	11.08%	46.50%	71.09%	80.24%
Precision	91.00%	92.03%	90.45%	91.77%
F Measure	19.76%	61.78%	79.61%	85.62%

3. จากตารางสรุปได้ว่าจาก 4 แบบจำลองที่ดีที่สุดในการวัดประสิทธิภาพทั้ง 4 แบบคำนวณ พบว่าแบบจำลองต้นไม้การเพิ่มไล่ระดับการตัดสินใจ (Gradient Boosted Trees) มีค่าที่สูงที่สุด 3 ใน 4 แบบคำนวณคือ ค่าความแม่นยำ (Precision) ,ค่าความถูกต้องของข้อมูล (Accuracy) และ ค่าความถ่วงดุล (F-measure)

ข้อเสนอแนะ

1. ตัวแปรที่นำมาใช้สำหรับงานวิจัยชิ้นนี้เป็นเพียงส่วนหนึ่งของการพิจารณาควรเพิ่มตัวแปรอื่น ๆที่เกี่ยวข้องอีกและ ปรับปรุงการเก็บข้อมูลเพื่อให้สามารถนำเข้าแบบจำลองและผลลัพธ์ดียิ่งขึ้น

2. ควรมีการพิจารณาตัวแปร หรือปัจจัยอื่น ๆ เช่น ข้อมูลช่วงเวลาการส่งโปรโมชั่น ช่วงที่เก็บผลการกดดูแคมเปญมากกว่า 1 วัน หรือ แผนการพัฒนาสินค้าหรือผลิตภัณฑ์ของธุรกิจ และรูปแบบแนวทางของการตลาดในช่วงนั้น ๆ เป็นต้น

3. ควรมีการประเมินผลที่ได้อย่างสม่ำเสมอ ในการที่จะพิจารณาถึง ความเหมาะสม และ ลองนำเทคนิคใหม่ๆที่ถูกพัฒนามาปรับใช้เพื่อให้มีประสิทธิภาพมากยิ่งขึ้น หรือเพื่อลดเวลาของการดำเนินการของ “Feature Auto Model” สามารถดูผลประสิทธิภาพของงานวิจัยชิ้นนี้แล้วเลือกแบบจำลองให้น้อยลงเพื่อลดระยะเวลา

บรรณานุกรม

สุรวัชร ศรีเปารยะ และ สายชล สินสมบูรณทอง, “การเปรียบเทียบประสิทธิภาพวิธีการจำแนกกลุ่มการ เปนโรคไตเรื้อรัง,” ภาควิชาสถิติ คณะวิทยาศาสตร์, สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง, เขตลาดกระบัง, กรุงเทพฯ, 2560

Corporate, “บทสรุปความสำเร็จ LINE ประเทศไทย ปี 2564 การเติบโตที่พร้อมรุกหน้ายกระดับแพลตฟอร์มให้ตอบโจทย์คนไทยอย่างไม่หยุดยั้ง,” LINE Corporation [ออนไลน์]. https://linecorp.com/th/pr/news/th/2022/4073. (เข้าถึงเมื่อ 1 มกราคม 2565).

Faraway et al.,” Deep Learning In Digital Pathology,” global engage [Online]. https://www.global-engage.com/life-science/deep-learning-in-digital-pathology/. (Accessed: May. 20, 2023).

F. Pochetti, “Extreme Label Imbalance: When You Measure the Minority Class in Basis Points.,” [Online]. http://francescopochetti.com/extreme-label-imbalance-when-you-measure-the-minority-class-in-basis-points/. (Accessed: May. 24, 2023).

L. LAHINDAH,I.D. SUDIRMAN, “CLASSIFICATION APPROACH TO PREDICT CUSTOMER DECISION BETWEEN PRODUCT BRANDS BASED ON CUSTOMER PROFILE AND TRANSACTION,” Program Studi Manajemen, Sekolah Tinggi Ilmu Ekonomi Harapan Bangsa, BINUS Business School Undergraduate Program, Bina Nusantara University, Bandung Campus, Bandung, Indonesia, 2023 [Online]. Available: http://www.jatit.org/volumes/Vol101No9/12Vol101No9.pdf

O. Andrieiev, “Propensity Model: Using Data to Predict Customer Behavior,” [Online]. https://jelvix.com/blog/propensity-model. (Accessed: May. 24, 2023).

P. Gatchalee, “Confusion Matrix เครื่องมือสำคัญในการประเมินผลลัพธ์ของการทำนาย ในMachine learning,” Medium[Online]. https://medium.com/@pagongatchalee/confusion-matrix-เครื่องมือสำคัญในการประเมินผลลัพธ์ของการทำนาย-ในmachine-learning-fba6e3f9508c. (Accessed: May. 23, 2023).

RapidMiner, Inc., “Fast Large Margin(RapidMiner Studio Core),” RapidMiner Document [Online].https://docs.rapidminer.com/latest/studio/operators/modeling/predictive/support_vector_machines/fast_large_margin.html. (Accessed: May. 20, 2023).

S. Nuchitprasitchai, “หลักการประเมิน Machine Learning Model และการเลือกใช้ Precision,Recall,F1 Score,” Medium[Online]. https://medium.com/botnoi-classroom/หลักการประเมิน-machine-learning-model-และการเลือกใช้-precision-recall-f1-score-e1db39fc79dd. (Accessed: May. 23, 2023).

W. Daroontham, “เจาะลึก Random Forest !!!— Part 2 of “รู้จัก Decision Tree, Random Forest, และ XGBoost!!!”,” Medium[Online]. https://medium.com/@witchapongdaroontham/เจาะลึก-random-forest-part-2-of-รู้จัก-decision-tree-random-forest-และ-xgboost-79b9f41a1c1c. (Accessed: May. 22, 2023).