DPU DHURAKIJ PUNDIT UNIVERSITY มหาวิทยาลัยธุรกิจบัณฑิตย์

การศึกษาประสิทธิภาพการพยากรณ์ความสนใจในผลิตภัณฑ์ จากบัญชีทางการธุรกิจของแอปพลิเคชันไลน์ ด้วยเทคนิคการเรียนรู้ของเครื่อง

ธัญลักษณ์ เอียดเจริญ

สารนิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตร
วิศวกรรมศาสตรมหาบัณฑิต สาขาวิชาวิศวกรรมข้อมูลขนาดใหญ่
วิทยาลัยนวัตกรรมด้านเทคโนโลยีและวิศวกรรมศาสตร์
มหาวิทยาลัยธุรกิจบัณฑิตย์
ปีการศึกษา 2565

DPU DHURAKIJ PUNDIT UNIVERSITY มหาวิทยาลัยธุรกิจบัณฑิตย์

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

TANYALUK EADCHAROEN

A Thematic Paper Submitted in Partial Fulfillment of the
Requirements for the Degree of Master of Engineering
Department of Big Data Engineering,
College of Innovative Technology and Engineering
Dhurakij Pundit University
Academic Year 2022

DPU DHURAKIJ PUNDIT UNIVERSITY มหาวิทยาลัยธุรกิจบัณฑิตย์

ใบรับรองสารนิพนธ์

วิทยาลัยนวัตกรรมด้านเทคโนโลยีและวิศวกรรมศาสตร์ มหาวิทยาลัยธุรกิจบัณฑิตย์
วิศวกรรมศาสตรมหาบัณฑิต สาขาวิชาวิศวกรรมข้อมูลขนาดใหญ่

หัวข้อสารนิพนธ์ การศึกษาประสิทธิภาพการพยากรณ์ความสนใจในผลิตภัณฑ์ จากบัญชีทางการธุรกิจของแอพพลิเคชั่นไลน์ ด้วยเทคนิคการเรียนรู้ของเครื่อง

เสนอโดย ธัญลักษณ์ เอียดเจริญ

สาขาวิชา วิศวกรรมข้อมูลขนาดใหญ่

อาจารย์ที่ปรึกษาสารนิพนธ์ ดร.เอกสิทธิ์ พัชรวงศ์ศักดา

ได้พิจารณาเห็นชอบโดยคณะกรรมการสอบสารนิพนธ์แล้ว

(รองศาสตราจารย์ ดร.วฤษาย์ ร่มสายหยุด) ประธานกรรมการ

(ดร.เอกสิทธิ์ พัชรวงศ์ศักดา) กรรมการที่ปรึกษาสารนิพนธ์

(ผู้ช่วยศาสตราจารย์ ดร.ดวงใจ จิตคงชื่น) กรรมการ

วิทยาลัยนวัตกรรมด้านเทคโนโลยีและวิศวกรรมศาสตร์ รับรองแล้ว

(ดร.ชัยพร เชมะภาตะพันธ์) คณบดีวิทยาลัยนวัตกรรมด้านเทคโนโลยีและวิศวกรรมศาสตร์

วันที่ 31 เดือน กรกฎาคม พ.ศ. 2566

DPU DHURAKIJ PUNDIT UNIVERSITY มหาวิทยาลัยธุรกิจบัณฑิตย์

หัวข้อสารนิพนธ์	การศึกษาประสิทธิภาพการพยากรณ์ความสนใจในผลิตภัณฑ์จากบัญชีทางการธุรกิจของแอปพลิเคชันไลน์ด้วยเทคนิคการเรียนรู้ของเครื่อง
ชื่อผู้เขียน	ธัญลักษณ์ เอียดเจริญ
อาจารย์ที่ปรึกษา	ดร.เอกสิทธิ์ พัชรวงศ์ศักดา
หลักสูตร	วิศวกรรมข้อมูลขนาดใหญ่
ปีการศึกษา	2565

บทคัดย่อ

บัญชีทางการของแอปพลิเคชันไลน์ สามารถสร้างฐานลูกค้าผู้ติดตามให้เพิ่มมากขึ้น โดยสื่อสารบรอดแคสต์ และ ยังส่งข้อมูลกิจกรรม การขาย การตลาด พร้อมทั้งโปรโมชั่นพิเศษต่าง ๆ ที่จะสื่อถึงลูกค้าหลายๆ รูปแบบได้ 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 ค่าความแม่นยำ 91.77% ซึ่งมีค่าน้อยกว่า แบบจำลอง Decision Tree ที่มีค่าความแม่นยำ 92.03%

DPU DHURAKIJ PUNDIT UNIVERSITY มหาวิทยาลัยธุรกิจบัณฑิตย์

จากผลลัพธ์ทั้งสามข้อข้างต้นสามารถอภิปรายผลการวิจัยได้ว่า แบบจำลอง Gradient Boosted Trees มีประสิทธิภาพในการการคาดการณ์ว่าลูกค้าคนใดบ้างที่จะมีส่วนร่วมในการปล่อยแคมเปญได้ดีที่สุด และ ข้อมูลเหมาะสม ในการปล่อยแคมเปญครั้งต่อไปในอนาคต

คำสำคัญ: บัญชีทางการของแอปพลิเคชันไลน์, วัดประสิทธิภาพแบบจำลอง, ฟีเจอร์ ออโต้ โมเดล

อาจารย์ที่ปรึกษา

DPU DHURAKIJ PUNDIT UNIVERSITY มหาวิทยาลัยธุรกิจบัณฑิตย์

Thematic Paper Title	THE STUDY ON THE EFFICIENCY OF INTEREST PREDICTION IN BUSINESS ACCOUNT PRODUCTS OF THE LINE APPLICATION USING MACHINE LEARNING TECHNIQUES
Author	Tanyaluk Eadcharoen
Thematic Paper Advisor	Dr. Eakasit Pacharawongsakda
Program	Big Data Engineering
Academic Year	2022

ABSTRACT

LINE Official Account can build more customers by communicating, broadcasting, and sending sales activity. A daily report of business has after sent data to see statistics of interest only. Therefore, it is the origin of this research whose objective is to use daily reports to see statistics of target groups appropriately and more usefully. It is a comparison of the most effective machine learning techniques related to propensity to engagement. Using 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 gave 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 are 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 is less than Technique Decision Tree with a Precision of 92.03%

DPU DHURAKIJ PUNDIT UNIVERSITY มหาวิทยาลัยธุรกิจบัณฑิตย์

From the three results above, it can be concluded that Technique Gradient Boosted Trees is effective in predicting which engaged the campaign launch best and suitable to the data provided. It is used to create forecasting models as much as possible to predict which customers will engage in future campaign launches.

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

Advisor

DPU DHURAKIJ PUNDIT UNIVERSITY มหาวิทยาลัยธุรกิจบัณฑิตย์

กิตติกรรมประกาศ

สารนิพนธ์ฉบับนี้สำเร็จสมบูรณ์ลุล่วงไปได้ด้วยดีอันเนื่องมาจากการได้รับความเมตตากรุณา แนะนำแนวทาง พร้อมทั้งชี้แจ้ง ตรวจสอบ แก้ไขข้อบกพร่อง และ สนับสนุนช่วยเหลืออย่างดียิ่ง เพื่อให้สารนิพนธ์ฉบับนี้สำเร็จเสร็จสิ้นสมบูรณ์ โดยการให้ความช่วยเหลือของ ดร. เอกสิทธิ์ พัชรวงศ์ศักดา ซึ่งเป็นทั้งอาจารย์ที่ปรึกษาสารนิพนธ์ และ เป็นผู้ประสิทธิ์ประสาทความรู้ ดังนั้นผู้เขียนจึงระลึกถึงพระคุณ และ ซาบซึ้งใจในความเมตตากรุณามาโดยตลอด จึงขอกราบขอบพระคุณเป็นอย่างสูงไว้ ณ โอกาสนี้

ขอกราบขอบพระคุณ รศ.ดร.วฤษาย์ ร่มสายหยุด และ ผศ.ดร.ดวงใจ จิตคงชื่น ที่กรุณาให้เกียรติในการสอบสารนิพนธ์ครั้งนี้ โดยมาเป็นประธานกรรมการ และ กรรมการ ตามลำดับ ซึ่งท่านได้กรุณาตรวจแก้ไขสารนิพนธ์ฉบับนี้ ให้ถูกต้อง และ สมบูรณ์ยิ่งขึ้น

ขอขอบพระคุณ นางสาวกุลธิดา รอดบุญ ที่ช่วยดำเนินการ และ ประสานงานในเรื่องการจัดการในด้านของการประมวลผล ลงทะเบียนรายวิชาและไขข้อสงสัยทั้งหมดของทุกภาคการศึกษา ตลอดจนถึงการจัดการในการสอบสารนิพนธ์ในครั้งนี้ที่สำเร็จลุล่วงไปได้ด้วยดี และ ขอขอบพระคุณเจ้าหน้าที่บัณฑิตมหาวิทยาลัยธุรกิจบัณฑิตย์ ทุกท่านที่ได้ให้ความสะดวกด้านอำนวยการ และประสานงานในเรื่องต่าง ๆ

ขอกราบขอบพระคุณบิดา มารดา ผู้ให้กำเนิด และ ครอบครัว ที่ช่วยส่งเสริมทุกอย่างที่อยากทำและสนใจ พร้อมทั้งให้กำลังใจในการดำเนินชีวิตเสมอ ถึงแม้ว่าบิดาผู้เป็นแบบอย่างในความพยายาม และ การเรียนรู้อย่าสม่ำเสมอ จะล่วงลับไปแล้วก็ตามท่านยังคงเป็นแบบอย่างและระลึกถึงเสมอมา

สุดท้ายนี้ขอขอบพระคุณเพื่อนๆ ทุกท่านในสาขาวิศวกรรมข้อมูลขนาดใหญ่ที่สนิทหรือไม่สนิทก็ตาม ทุก ๆ ท่านที่เป็นแบบอย่าง และกำลังใจ ให้คำปรึกษา ช่วยเหลือจนงานสารนิพนธ์สำเร็จลุล่วงไปได้ด้วยดี

ธัญลักษณ์ เอียดเจริญ

DPU DHURAKIJ PUNDIT UNIVERSITY มหาวิทยาลัยธุรกิจบัณฑิตย์

สารบัญ

หน้า

บทคัดย่อภาษาไทย..... ง
บทคัดย่อภาษาอังกฤษ..... จ
กิตติกรรมประกาศ..... ฉ
สารบัญ..... ช
สารบัญตาราง..... ฌ
สารบัญภาพ..... ญ
บทที่
1. บทนำ..... 1
1.1. ที่มาและความสำคัญของปัญหา..... 1
1.2. วัตถุประสงค์ของการวิจัย..... 3
1.3. ขอบเขตงานวิจัย..... 3
1.4. ประโยชน์ที่คาดว่าจะได้รับ..... 4
2. แนวคิด ทฤษฎี และงานวิจัยที่เกี่ยวข้อง..... 5
2.1. กระบวนการ Cross-Industry Standard Process for Data Mining (CRISP-DM).. 5
2.2. โปรแกรม RapidMiner Studio and “Feature Auto Model”..... 6
2.3. เทคนิคการเรียนรู้ของเครื่อง (Machine Learning Technique)..... 8
2.4. การวัดประสิทธิภาพของแบบจำลอง..... 18
2.5. บทความและงานวิจัยที่เกี่ยวข้อง..... 19
3. ระเบียบวิธีวิจัย..... 21
3.1. ทำความเข้าใจปัญหา (Business Understanding Phase)..... 21
3.2. ทำความเข้าใจข้อมูล (Data Understanding Phase)..... 22
3.3. การเตรียมข้อมูล (Data Preparation Phase)..... 26
3.4. การสร้างแบบจำลอง (Modeling Phase)..... 31
3.5. การประเมินผลแบบจำลอง (Evaluation Phase)..... 38
3.6. การนำไปใช้งาน (Deployment Phase)..... 38

DPU DHURAKIJ PUNDIT UNIVERSITY มหาวิทยาลัยธุรกิจบัณฑิตย์

สารบัญ (ต่อ)

บทที่	หน้า
4. ผลการวิจัย	39
4.1. ผลการทำนาย และ ค่าน้ำหนักของแอตทริบิวต์	39
4.2. ผลการเปรียบเทียบประสิทธิภาพการทำนายผล	48
4.3. ผลความถูกต้องจากการนำแบบจำลองพยากรณ์ในอนาคตเทียบกับข้อมูลจริง	56
4.4. ผลของการแสดงผลเพื่อนำเสนอ	61
5. สรุปผลการวิจัย อภิปรายผล และข้อเสนอแนะ	63
5.1. สรุปผลการวิจัยและอภิปรายผล	63
5.2. ข้อเสนอแนะ	65
บรรณานุกรม	66
ประวัติผู้เขียน	70

DPU DHURAKIJ PUNDIT UNIVERSITY มหาวิทยาลัยธุรกิจบัณฑิตย์

สารบัญตาราง

ตารางที่ **หน้า**

3.1 แบ่งประเภทกลุ่มเป้าหมายออกเป็น segment 5 กลุ่มและมีความหมาย ดังนี้..... 22
3.2 แบ่งประเภทของแคมเปญออกเป็น 20 ประเภท และมีความหมายดังนี้..... 22
3.3 ตาราง Time Frame for Modeling..... 31
3.4 ตารางตัวแปรใหม่เพิ่มจากข้อมูลตัวแปรเดิมที่มี..... 33
4.1 ผลในส่วนของตาราง Confusion Matrix ของวิธีการพยากรณ์ของเทคนิคการจำแนกประเภทข้อมูลแบบเบย์อย่างง่าย (Naive Bayes)..... 40
4.2 ผลในส่วนของตาราง Confusion Matrix ของวิธีการพยากรณ์ของเทคนิคตัวแบบเชิงเส้นนัยทั่วไป (Generalized Linear Model)..... 41
4.3 ผลในส่วนของตาราง Confusion Matrix ของวิธีการพยากรณ์ของเทคนิคการถดถอยโลจีสติก (Logistic Regression)..... 42
4.4 ผลในส่วนของตาราง Confusion Matrix ของวิธีการพยากรณ์ของเทคนิคจำแนกประเภทโดยใช้ขอบเขตที่มีขนาดกว้างที่สุดแบบเร็ว (Fast Large Margin (RapidMiner Studio Core)..... 43
4.5 ผลในส่วนของตาราง Confusion Matrix ของวิธีการพยากรณ์ของเทคนิคการเรียนรู้เชิงลึก (Deep Learning)..... 44
4.6 ผลในส่วนของตาราง Confusion Matrix ของวิธีการพยากรณ์ของเทคนิคต้นไม้การตัดสินใจ (Decision Tree)..... 44
4.7 ผลในส่วนของตาราง Confusion Matrix ของวิธีการพยากรณ์ของเทคนิคแบบจำลองป่าสุ่ม (Random Forest)..... 45
4.8 ผลในส่วนของตาราง Confusion Matrix ของวิธีการพยากรณ์ของเทคนิคต้นไม้การเพิ่มไล่ระดับการตัดสินใจ (Gradient Boosted Trees)..... 46
4.9 ผลในส่วนของตาราง Confusion Matrix ของวิธีการพยากรณ์ของเทคนิคซัพพอร์ตเวกเตอร์แมทชีน (Support Vector)..... 47
4.10 ผลการวิเคราะห์ประสิทธิภาพของแบบจำลอง..... 48
5.1 ผลการวิเคราะห์ประสิทธิภาพแบบจำลองที่ดีที่สุดที่มีค่าประสิทธิภาพมากกว่า ร้อยละ 80 จะได้แบบจำลอง 4 เทคนิค..... 64

DPU DHURAKIJ PUNDIT UNIVERSITY มหาวิทยาลัยธุรกิจบัณฑิตย์

สารบัญภาพ

ภาพที่ **หน้า**

1.1 ภาพข้อมูลความสำเร็จของ Line Application ที่ใช้บัญชีทางการของแอปพลิเคชันไลน์ (LINE Official Account) ตลอดปี 2564..... 1
1.2 ภาพตัวอย่างการส่งแคมเปญของธุรกิจต่าง ๆ และผลลัพธ์หลังกดเข้าร่วมแคมเปญของธุรกิจ..... 2
2.1 กระบวนการทำงานของ CRISP-DM..... 5
2.2 สรุปขั้นตอนกระบวนการทำงานของ CRISP-DM..... 5
2.3 โปรแกรม RapidMiner Studio Version 10.0..... 7
2.4 ส่วนเสริม “Feature Auto Model” บน RapidMiner Studio Version 10.0..... 7
2.5 ตัวอย่างหลังจาก Run “Feature Auto Model” สำเร็จ ในส่วนเสริม “Feature Auto Model” บน RapidMiner Studio Version 10.0..... 8
2.6 การจำแนกประเภทข้อมูลแบบเบย์อย่างง่าย..... 8
2.7 เทคนิคตัวแบบเชิงเส้นนัยทั่วไป..... 9
2.8 เทคนิคการถดถอยโลจิสติก..... 11
2.9 หลักการวิเคราะห์การถดถอยโลจีสติก..... 11
2.10 เทคนิคซัพพอร์ตเวกเตอร์แมทชีน..... 12
2.11 เทคนิคการจำแนกประเภทแบบเร็วโดยใช้ขอบเขตการแยกแยะที่มีขนาดกว้าง(Fast Large Margin (RapidMiner Studio Core))..... 13
2.12 เทคนิคการเรียนรู้เชิงลึก..... 14
2.13 ตัวอย่าง ราก (Root Node) ,กิ่ง (Branch) และ ใบ (Leaf Node) ของต้นไม้การตัดสินใจ..... 14
2.14 เทคนิคแบบจำลองป่าสุ่ม..... 15
2.15 เทคนิคต้นไม้การเพิ่มไล่ระดับการตัดสินใจ..... 16
2.16 ตาราง Confusion Matrix..... 17
3.1 ตัวอย่างชุดข้อมูล first response จากการ broadcast campaign..... 24
3.2 ตัวอย่างชุดข้อมูล campaign category จากการ broadcast campaign..... 25
3.3 ตัวอย่างก่อนและหลังการปรับปรุงชื่อของตารางโดยใช้ Excel Application..... 26
3.4 Operation ขั้นตอนการรวมข้อมูลในโปรแกรม RapidMiner..... 27
3.5 ตัวอย่างข้อมูลหลังจากรวบรวมไฟล์ Excel 307 ไฟล์ให้อยู่ในรูปแบบเดียวกัน..... 27

DPU DHURAKIJ PUNDIT UNIVERSITY มหาวิทยาลัยธุรกิจบัณฑิตย์

สารบัญภาพ (ต่อ)

ภาพที่ **หน้า**

3.6 Operation การเตรียมข้อมูลในโปรแกรม RapidMiner..... 28
3.7 Operation Subprocess การเตรียมข้อมูลในโปรแกรม RapidMiner..... 28
3.8 Operation การรวมชุดข้อมูลในโปรแกรม RapidMiner 1..... 29
3.9 ตัวอย่างข้อมูลของการรวมชุดข้อมูลในโปรแกรม RapidMiner 1..... 29
3.10 Operation การรวมชุดข้อมูลในโปรแกรม RapidMiner 2..... 29
3.11 ตัวอย่างข้อมูลของการรวมชุดข้อมูลในโปรแกรม RapidMiner 2..... 30
3.12 Operation การแปลงรูปแบบข้อมูลในโปรแกรม RapidMiner..... 30
3.13 ตัวอย่างข้อมูลของการแปลงรูปแบบข้อมูลในโปรแกรม RapidMiner..... 30
3.14 Operation ของแบบจำลอง Propensity ในโปรแกรม RapidMiner 1..... 31
3.15 Operation ของแบบจำลอง Propensity ในโปรแกรม RapidMiner 2..... 32
3.16 Operation ของแบบจำลอง Propensity ในโปรแกรม RapidMiner 3..... 32
3.17 Operation ของแบบจำลอง Propensity ทั้งหมดในโปรแกรม RapidMiner..... 32
3.18 ตัวอย่างข้อมูลหลังดำเนินการสร้างแบบจำลอง Propensity..... 35
3.19 ขั้นตอน Prepare Target..... 35
3.20 ขั้นตอน Select Input..... 36
3.21 ขั้นตอน การเลือกแบบจำลองที่ต้องการนำมาวัดประสิทธิภาพทั้งหมด Model Types.... 37
4.1 ค่าน้ำหนักของแอตทริบิวต์ ของวิธีการจำแนกประเภทข้อมูลแบบเบย์อย่างง่าย (Naive Bayes)..... 39
4.2 ค่าน้ำหนักของแอตทริบิวต์ ของวิธีตัวแบบเชิงเส้นนัยทั่วไป (Generalized Linear Model) 40
4.3 ค่าน้ำหนักของแอตทริบิวต์ ของวิธีการพยากรณ์ของเทคนิคการถดถอยโลจีสติก (Logistic Regression)..... 41
4.4 ค่าน้ำหนักของแอตทริบิวต์ ของวิธีการพยากรณ์ของเทคนิคจำแนกประเภทโดยใช้ขอบเขตที่มีขนาดกว้างที่สุดแบบเร็ว (Fast Large Margin (RapidMiner Studio Core))..... 42
4.5 ค่าน้ำหนักของแอตทริบิวต์ ของวิธีการพยากรณ์ของเทคนิคการเรียนรู้เชิงลึก (Deep Learning) 43

DPU DHURAKIJ PUNDIT UNIVERSITY มหาวิทยาลัยธุรกิจบัณฑิตย์

สารบัญภาพ (ต่อ)

ภาพที่ **หน้า**

4.6 ค่าน้ำหนักของแอตทริบิวต์ ของวิธีการพยากรณ์ของเทคนิคต้นไม้การตัดสินใจ (Decision Tree) 44
4.7 ค่าน้ำหนักของแอตทริบิวต์ ของวิธีการพยากรณ์ของเทคนิคแบบจำลองป่าสุ่ม (Random Forest) 45
4.8 ค่าน้ำหนักของแอตทริบิวต์ ของวิธีการพยากรณ์ของเทคนิคต้นไม้การเพิ่มไล่ระดับการตัดสินใจ (Gradient Boosted Trees) 46
4.9 ค่าน้ำหนักของแอตทริบิวต์ ของวิธีการพยากรณ์ของเทคนิคซัพพอร์ตเวกเตอร์แมทชีน (Support Vector) 47
4.10 ผลประสิทธิภาพของแบบจำลองวิธีการพยากรณ์ของเทคนิควิธีการจำแนกประเภทข้อมูลแบบเบย์อย่างง่าย (Naive Bayes)..... 49
4.11 ผลประสิทธิภาพของแบบจำลองวิธีการพยากรณ์ของเทคนิคตัวแบบเชิงเส้นนัยทั่วไป (Generalized Linear Model) 49
4.12 ผลประสิทธิภาพของแบบจำลองวิธีการพยากรณ์ของเทคนิคการถดถอยโลจีสติก (Logistic Regression) 50
4.13 ผลประสิทธิภาพของแบบจำลองวิธีการพยากรณ์ของเทคนิคการจำแนกประเภทแบบเร็วโดยใช้ขอบเขตการแยกแยะที่มีขนาดกว้าง (Fast Large Margin (RapidMiner Studio Core)) 50
4.14 ผลประสิทธิภาพของแบบจำลองวิธีการพยากรณ์ของเทคนิคการเรียนรู้เชิงลึก (Deep Learning) 51
4.15 ผลประสิทธิภาพของแบบจำลองวิธีการพยากรณ์ของเทคนิคต้นไม้การตัดสินใจ (Decision Tree) 51
4.16 ผลประสิทธิภาพของแบบจำลองวิธีการพยากรณ์ของเทคนิคแบบจำลองป่าสุ่ม (Random Forest) 52
4.17 ผลประสิทธิภาพของแบบจำลองวิธีการพยากรณ์ของเทคนิคต้นไม้การเพิ่มไล่ระดับการตัดสินใจ (Gradient Boosted Trees) 52
4.18 ผลประสิทธิภาพของแบบจำลองวิธีการพยากรณ์ของเทคนิคซัพพอร์ตเวกเตอร์แมทชีน (Support Vector) 53

DPU DHURAKIJ PUNDIT UNIVERSITY มหาวิทยาลัยธุรกิจบัณฑิตย์

สารบัญภาพ (ต่อ)

ภาพที่ **หน้า**

4.19 ผลความถูกต้องของแบบจำลองโดยใช้ค่าความถูกต้องของข้อมูล (Accuracy) โดยเปรียบเทียบกับ 9 แบบจำลอง..... 53
4.20 ผลค่าสัดส่วนของข้อมูลที่ถูกต้องกับข้อมูลที่ต้องการทั้งหมด ค่าความระลึก (Recall) โดยเปรียบเทียบกับ 9 แบบจำลอง..... 54
4.21 ผลค่าสัดส่วนของข้อมูลที่ถูกต้อง ตรงตามความต้องการส่วนด้วยข้อมูลทั้งหมดค่าความแม่นยำ (Precision) โดยเปรียบเทียบกับ 9 แบบจำลอง..... 54
4.22 ผลค่าเฉลี่ยแบบ ค่าเฉลี่ยฮาร์มอนิก (harmonic mean) โดยจะนำ “ค่าความแม่นยำ” และ “ค่าความระลึก” มาพิจารณาร่วมกัน ค่าความถ่วงดุล (F-measure) โดยเปรียบเทียบกับ 9 แบบจำลอง..... 55
4.23 ผลแบบจำลองพยากรณ์ในอนาคตเทียบกับข้อมูลจริง ด้วยวิธีการจำแนกประเภทข้อมูลแบบเบย์อย่างง่าย (Naive Bayes)..... 56
4.24 ผลแบบจำลองพยากรณ์ในอนาคตเทียบกับข้อมูลจริง ด้วยวิธีตัวแบบเชิงเส้นนัยทั่วไป (Generalized Linear Model)..... 56
4.25 ผลแบบจำลองพยากรณ์ในอนาคตเทียบกับข้อมูลจริง ด้วยวิธีการถดถอยโลจีสติก (Logistic Regression) 57
4.26 ผลแบบจำลองพยากรณ์ในอนาคตเทียบกับข้อมูลจริง ด้วยวิธีจำแนกประเภทโดยใช้ขอบเขตที่มีขนาดกว้างที่สุดแบบเร็ว (Fast Large Margin (RapidMiner Studio Core))..... 57
4.27 ผลแบบจำลองพยากรณ์ในอนาคตเทียบกับข้อมูลจริง ด้วยวิธีการเรียนรู้เชิงลึก (Deep Learning)..... 58
4.28 ผลแบบจำลองพยากรณ์ในอนาคตเทียบกับข้อมูลจริง ด้วยวิธีต้นไม้การตัดสินใจ (Decision Tree) 58
4.29 ผลแบบจำลองพยากรณ์ในอนาคตเทียบกับข้อมูลจริง ด้วยวิธีแบบจำลองป่าสุ่ม (Random Forest) 59
4.30 ผลแบบจำลองพยากรณ์ในอนาคตเทียบกับข้อมูลจริง ด้วยวิธีต้นไม้การเพิ่มไล่ระดับการตัดสินใจ (Gradient Boosted Trees) 59
4.31 ผลแบบจำลองพยากรณ์ในอนาคตเทียบกับข้อมูลจริง ด้วยวิธีซัพพอร์ตเวกเตอร์แมทชีน (Support Vector) 60

DPU DHURAKIJ PUNDIT UNIVERSITY มหาวิทยาลัยธุรกิจบัณฑิตย์

สารบัญภาพ (ต่อ)

ภาพที่ **หน้า**

4.32 ตัวอย่างหน้าจอการแสดงผลการพยากรณ์ของข้อมูลสำหรับผู้ใช้งานหน้าที่ 1..... 61
4.33 ตัวอย่างหน้าจอการแสดงผลการพยากรณ์ของข้อมูลสำหรับผู้ใช้งานหน้าที่ 2..... 62
5.1 กราฟการเปรียบเทียบของ4 แบบจำลองที่ดีที่สุดในการวัดประสิทธิภาพทั้ง 4 แบบคำนวณ 65

DPU DHURAKIJ PUNDIT UNIVERSITY มหาวิทยาลัยธุรกิจบัณฑิตย์

บทที่ 1
บทนำ

1.1. ที่มาและความสำคัญของปัญหา

ปัจจุบันการส่งออกข้อมูลหาลูกค้า เข้ามามีบทบาทในการ สื่อสาร และ โฆษณา เป็นอย่างมาก ทำให้รูปแบบการใช้ชีวิตประจำวันของผู้คนมากมายในสังคมเปลี่ยนไป ซึ่งในความก้าวหน้าของเทคโนโลยีที่เพิ่มขึ้นทุกวันนั้น มีอิทธิพลกับ ผู้คนทุก ๆ เพศทุก ๆ วัย สิ่งนี้เองจำเป็นต่อธุรกิจ ไม่ว่าจะเป็นรายเล็กหรือรายใหญ่ ธุรกิจต่าง ๆ จำเป็นต้องปรับกลยุทธ์ ในการสร้างโอกาสให้เป็นที่รู้จักเยอะ ๆ และ แพร่หลายสามารถเข้าถึงทุกกลุ่มคนมากมายทั้งที่เป็นลูกค้าเก่า ลูกค้าใหม่ หรือคนที่ไม่เคยรู้จักธุรกิจเลยก็ตาม ผ่านแพลตฟอร์มที่สามารถสื่อสารไปยังลูกค้าได้หลายๆรูปแบบ เพราะ มีผู้ใช้งานหลากหลายช่วงอายุ ที่สามารถเข้าถึงและใช้งานแพลตฟอร์มดังกล่าวได้อย่างสะดวกสบาย ซึ่งแพลตฟอร์มดังกล่าวยังใช้งานอย่างง่ายและสามารถใช้งานได้ทุก ๆวันอยู่แล้ว ดังนั้นเองแพลตฟอร์มที่หลาย ๆ ธุรกิจมักเลือกใช้งานนั่นคือ บัญชีทางการของแอปพลิเคชันไลน์ (LINE Official Account) หรือ ไลน์โอเอ (Line OA) ซึ่งบัญชีทางการนั้นเป็นตัวช่วยให้สามารถสร้างฐานลูกค้าผู้ติดตามทั้งเก่าและใหม่ให้เพิ่มมากขึ้น พร้อมทั้งสามารถสื่อสาร โดยการบรอดแคสต์แบบระบุหรือไม่ระบุกลุ่มเป้าหมายได้ และ ยังส่งข้อมูลกิจกรรม การขาย การตลาด พร้อมทั้งโปรโมชั่นพิเศษต่าง ๆ ที่ต้องการสื่อสารไปยัง ลูกค้าหลายๆรูปแบบได้ โดยดำเนินการผ่านช่องทางของแอปพลิเคชันไลน์ได้อย่างง่าย ใช้ทีมงานน้อยในการดำเนินการ ซึ่งจากบทความบน Line Corp website Line official กล่าวว่า ในปี 2564 แอปพลิเคชันไลน์สามารถเข้าถึงคนไทยได้มากกว่า 47 ล้านคน และ ครอบคลุมผู้ใช้ทั่วประเทศไทย ดังนั้นจึงเป็นไปไม่ได้เลยที่ Line OA ได้ครองตำแหน่งช่องทางยอดนิยมของธุรกิจในไทยหลายๆราย เพื่อการนำเสนอสินค้าและ บริการ เพื่อพูดคุยสื่อสารกับลูกค้า จากยอดการใช้งานแอปพลิเคชันไลน์ที่เพิ่มขึ้นเป็น 5 ล้านบัญชี (ณ เดือน ส.ค. 2564) มีสถิติการเติบโตเพิ่มขึ้นถึง 25% ภายในปีเดียว ด้วยเหตุนี้เองผู้ประกอบการในหลายๆขนาดทั้งองค์กรใหญ่และเล็กจึงเลือกใช้งานแพลตฟอร์มนี้

DPU DHURAKIJ PUNDIT UNIVERSITY มหาวิทยาลัยธุรกิจบัณฑิตย์

ภาพที่ 1.1 ภาพข้อมูลความสำเร็จของบัญชีทางการของแอปพลิเคชันไลน์ ตลอดปี 2564

ที่มา: https://linecorp.com/th/pr/news/th/2022/4073

จากผลสำรวจบัญชีทางการของแอปพลิเคชันไลน์ กลุ่มอุตสาหกรรมที่นิยมใช้งานบัญชีทางการของแอปพลิเคชันไลน์ มากที่สุดใน 3 กลุ่มอุตสาหกรรม ได้แก่ อันดับหนึ่งคือ กลุ่มการเงิน ประกันและการธนาคาร (Financial, Insurance/Banking) อันดับสองคือ กลุ่มรีเทล และ อีคอมเมิร์ซ (Retail & E-commerce) และอันดับสาม กลุ่มภาครัฐ (Government & Public Sector) ตามลำดับ (ณ เดือน ส.ค. 2564) ซึ่งความนิยมที่กล่าวมาข้างต้น ขนาดของข้อมูลที่ธุรกิจ ๆ หนึ่งนั้นมีข้อมูลปริมาณมหาศาล และ ขนาดใหญ่มาก ไม่ว่าจะเป็นรูปภาพ วีดีโอ และ ข้อความตัวอักษร เป็นต้น ซึ่งข้อมูลส่วนที่ใหญ่มากที่สุดนั่นก็ คือข้อมูลของข้อความตัวอักษร ซึ่งมาเป็นรูปแบบประโยคการสื่อสาร เพราะสืบเนื่องมาจากธุรกิจนั้นจำเป็นที่จะต้องถ่ายทอดข้อมูลอย่างง่าย ๆ ให้กลุ่มลูกค้าเพื่อให้เข้าถึง โฆษณา สินค้า บริการของธุรกิจแล้ว ยังสามารถนำมารวบรวม และนำมาวิเคราะห์ความต้องการได้ ซึ่งตัวแปรที่สำคัญ ของการที่จะนำมาวิเคราะห์นั้นก็คือ เทรนด์ และ พฤติกรรมของผู้บริโภค

DPU DHURAKIJ PUNDIT UNIVERSITY มหาวิทยาลัยธุรกิจบัณฑิตย์

ภาพที่ 1.2 ภาพตัวอย่างการส่งแคมเปญของธุรกิจต่าง ๆ และ ผลลัพธ์หลังกดเข้าร่วมแคมเปญของธุรกิจ

จากจำนวนข้อมูลของข้อความขนาดใหญ่ข้างต้น จึงเป็นที่มาของการนำเสนองานวิจัยชิ้นนี้ เป็นการนำข้อมูลของข้อความตัวอักษร ในรูปแบบประโยคการสื่อสาร ซึ่งเป็นรูปแบบที่มีการสื่อสาร และ บรอดแคสต์ ที่ธุรกิจส่งออกไปแบบระบุกลุ่มเป้าหมาย ไม่ว่าจะเป็นแคมเปญ ข่าวสาร โปรโมชั่นต่าง ๆ ที่เกี่ยวข้องกับธุรกิจๆ หนึ่ง ซึ่งการได้มาของข้อมูลนั้น เริ่มต้นโดยลูกค้าจะต้องมีการเลือกกดที่ข้อมูลที่ส่งไป เพื่อดูข้อมูลเพิ่มเติมจากแคมเปญ ข่าวสาร โปรโมชั่นต่าง ๆ ที่ส่งออกไปให้กับลูกค้า หลังจากนั้นหนึ่งวัน จะมีการดำเนินการรวบรวมสถิติความสนใจของการเลือกกดข้อมูล ซึ่งอยู่ในรูปแบบข้อความตัวอักษร จากของกลุ่มเป้าหมายเหล่านี้ออกมาเป็นเอกสารรายงานประจำวันทุกวัน เพื่อสำหรับดูสถิติ ความสนใจของลูกค้า ที่เข้ามาใช้งานบัญชีทางการของแอปพลิเคชันไลน์ของธุรกิจ ดังนั้นจากเอกสารรายงานประจำวันที่กล่าวมา เมื่อได้รวบรวมข้อมูลแค่เพียงสามเดือน ข้อมูลที่ได้มาจึงที่มีขนาดใหญ่มาก เพราะ ผู้ติดตามบัญชีทางการของแอปพลิเคชันไลน์ของธุรกิจเป็นผู้ติดตามจำนวนมาก หลายกลุ่มเป้าหมาย จากข้อมูลขนาดใหญ่เหล่านี้ ทางผู้วิจัยจึงต้องการใช้กระบวนการทาง Data Engineering เพื่อจัดการข้อมูลสำหรับใช้วิเคราะห์ เปรียบเทียบเทคนิคของแบบจำลองที่พยากรณ์ข้อมูล และ นำเสนอผลลัพธ์ด้วยวิธีการที่เหมาะสมเพื่อแสดงผลการเปรียบเทียบต่าง ๆ ให้เข้าใจผลการดำเนินการได้อย่างง่าย และ สามารถสื่อสารองค์ความรู้ให้กับผู้ที่สามารถนำไปใช้ประโยชน์ต่อเพื่อแก้ปัญหา การบริหารจัดการ เพื่อให้สามารถใช้ข้อมูลขนาดใหญ่ที่มีนั้น ให้เกิดประโยชน์เพิ่มขึ้น และสามารถนำไปใช้สำหรับวิเคราะห์ในการออกแบบวางแผนการบรอดแคสต์โฆษณา ต่อไปในอนาคตเพื่อให้ตรงกับกลุ่มเป้าหมายมากยิ่งขึ้น

DPU DHURAKIJ PUNDIT UNIVERSITY มหาวิทยาลัยธุรกิจบัณฑิตย์

1.2. วัตถุประสงค์ของการวิจัย

1.2.1. เพื่อนำข้อมูลข้อมูลรายงานประจำวันจากผลการสื่อสารบรอดแคสต์แบบระบุกลุ่มเป้าหมาย ที่ใช้สำหรับดูแค่สถิติของกลุ่มเป้าหมายที่ติดตามอยู่บนแพลตฟอร์มออนไลน์ว่ามีส่วนร่วมกับการสื่อสารกับธุรกิจในช่องทางบัญชีทางการของแอปพลิเคชันไลน์อยางเดียว เพื่อให้สามารถใช้ข้อมูลขนาดใหญ่ที่มีนั้นให้เกิดประโยชน์เพิ่มขึ้น

1.2.2. จัดทำการเสนอ พร้อมทั้งเปรียบเทียบ แบบจำลองจากเทคนิคการเรียนรู้ของเครื่องที่มีส่งผลให้มีประสิทธิภาพสูงที่สุด ที่เกี่ยวกับการคาดการณ์การมีส่วนร่วมของลูกค้าเป้าหมาย

1.2.3. เพื่อเป็นแนวทางสำหรับการคาดการณ์ว่าลูกค้าคนใดบ้างที่จะมีส่วนร่วมในการปล่อยแคมเปญ โดยมีความเหมาะสมกับข้อมูลที่นำมาใช้ในการสร้างแบบจำลอง สำหรับการพยากรณ์มากที่สุด

1.3. ขอบเขตงานวิจัย

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

1.3.2. งานวิจัยฉบับนี้เพื่อศึกษาเครื่องมือ “Feature Auto Model” บน RapidMiner Studio ในการหาแบบจำลองในการพยากรณ์ข้อมูล โดยประเมินเปรียบเทียบโดยใช้แบบจำลองทั้งหมด 9 วิธีการ ดังนี้

- การจำแนกประเภทข้อมูลแบบเบย์อย่างง่าย (Naive Bayes)
- ตัวแบบเชิงเส้นนัยทั่วไป (Generalized Linear Model)
- การถดถอยโลจีสติก (Logistic Regression)
- การจำแนกประเภทแบบเร็วโดยใช้ขอบเขตการแยกแยะที่มีขนาดกว้าง(Fast Large Margin (RapidMiner Studio Core))
- การเรียนรู้เชิงลึก (Deep Learning)
- ต้นไม้การตัดสินใจ (Decision Tree)
- แบบจำลองป่าสุ่ม (Random Forest)
- ต้นไม้การเพิ่มไล่ระดับการตัดสินใจ (Gradient Boosted Trees)
- ซัพพอร์ตเวกเตอร์แมทชีน (Support Vector)

1.3.3. การเลือกแบบจำลองโดยการวัดประสิทธิภาพแบบจำลองจาก Confusion Matrix ซึ่งคำนวณค่าทั้ง 4 แบบคำนวณดังนี้

- ค่าความถูกต้อง (Accuracy) เป็นค่าของข้อมูลที่ทำนายถูกของทุก ๆ คลาส
- ค่าความระลึก (Recall) เป็นค่าอัตราส่วนของจำนวนของการทำนายถูก
- ค่าความแม่นยำ (precision) เป็นค่าที่ดูว่าสิ่งที่ทำนายออกมาแล้วนั้นสามารถทายถูกเท่าไหร่

DPU DHURAKIJ PUNDIT UNIVERSITY มหาวิทยาลัยธุรกิจบัณฑิตย์

- ค่าความถ่วงดุล (F-measure) เป็นค่าเฉลี่ย ของค่าความแม่นยำและค่าความระลึก

1.4. ประโยชน์ที่คาดว่าจะได้รับ

1.4.1. สามารถใช้ข้อมูลขนาดใหญ่ที่มีให้เกิดประโยชน์มากยิ่งขึ้น และ สามารถนำมาประยุกต์ใช้ในการแก้ปัญหา การบริหารจัดการ ในการออกแบบการบรอดแคสต์ ในอนาคตเพื่อให้ตรงกับเป้าหมาย และ ใช้เพื่อวางแผนในการลดการใช้ทรัพยากรของเครื่อง

1.4.2. เพิ่มศักยภาพในการแข่งขันของธุรกิจ เพื่อลดการส่งการบรอดแคสต์ที่ต้องเสียค่าใช้จ่ายแบบสิ้นเปลือง โดยนำผลการวิเคราะห์ ไปใช้ให้ตรงกับเป้าหมายมากยิ่งขึ้น

1.4.3. สามารถใช้เพื่อเป็นแนวทางในการ พัฒนา และ ปรับปรุงกระบวนการการพยากรณ์และ วางแผนรูปแบบของการเก็บข้อมูลเพื่อให้ใช้งานได้อย่างง่ายมากขึ้นในอนาคต

DPU DHURAKIJ PUNDIT UNIVERSITY มหาวิทยาลัยธุรกิจบัณฑิตย์

บทที่ 2
แนวคิด ทฤษฎีและงานวิจัยที่เกี่ยวข้อง

งานวิจัยแบบจำลองการพยากรณ์ความสนใจในผลิตภัณฑ์ ของบัญชีทางการธุรกิจของแอปพลิเคชันไลน์ ด้วยเทคนิคการเรียนรู้ของเครื่อง โดยเน้นศึกษากระบวนการ ในการจัดการข้อมูลเพื่อให้สามารถนำมาวิเคราะห์ และ เปรียบเทียบหาเทคนิคที่สามารถพยากรณ์ได้เหมาะสมกับข้อมูลรายงานประจำวันจากผลการสื่อสารบรอดแคสต์แบบระบุกลุ่มเป้าหมาย จากบัญชีทางการธุรกิจของแอปพลิเคชันไลน์ซึ่งการศึกษาค้นคว้าหลักการ แนวคิด ทฤษฎีและงานวิจัยที่เกี่ยวข้องดังนี้

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

2.2 โปรแกรม RapidMiner Studio and Feature Auto Model

2.3 เทคนิคการเรียนรู้ของเครื่อง (Machine Learning Technique)

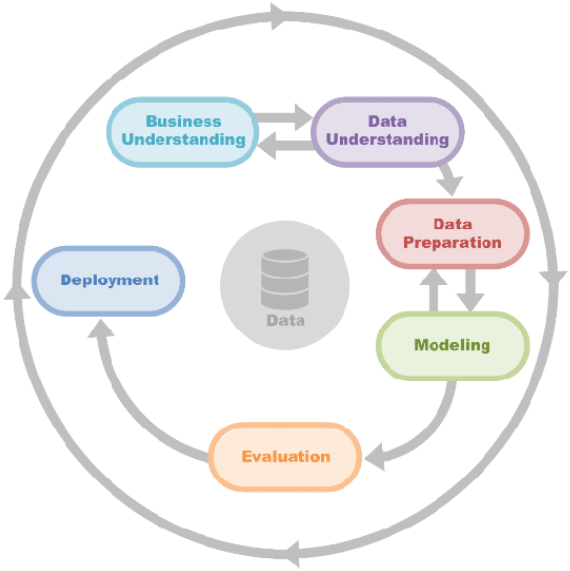
2.4 การวัดประสิทธิภาพของแบบจำลอง

2.5 บทความและงานวิจัยที่เกี่ยวข้อง

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

การทำเหมืองข้อมูล (Data Mining) เป็นกระบวนการที่ดำเนินการกับข้อมูลที่มีขนาดใหญ่ เพื่อค้นหารูปแบบ แนวทาง และความสัมพันธ์ที่ซ่อนอยู่ภายใต้ชุดข้อมูล โดยจะใช้หลักสถิติจากการเรียนรู้ของเครื่อง และ หลักคณิตศาสตร์ ในการแบ่งประเภท รูปแบบ ในการเชื่อมโยงข้อมูลที่สัมพันธ์กัน โดยพยากรณ์สิ่งที่น่าจะเกิด ทำให้ได้องค์ความรู้ใหม่ ๆ เพื่อใช้ในการตัดสินใจต่าง ๆ ในการได้ข้อมูลออกมา โดยข้อมูลที่ได้จะมีเหตุผลที่ สนับสนุนให้สามารถนำไปใช้ประโยชน์ได้ ซึ่งปกติแล้วกระบวนการที่ใช้ในการทำเหมืองข้อมูล ในการวิเคราะห์ข้อมูล และ นำไปใช้ประโยชน์คือ กระบวนการ Cross-industry standard process for data mining หรือ CRISP-DM ซึ่งมีกระบวนการด้วยกัน 6 ขั้นตอน

DPU DHURAKIJ PUNDIT UNIVERSITY มหาวิทยาลัยธุรกิจบัณฑิตย์



ภาพที่ 2.1 กระบวนการทำงานของ CRISP-DM

ที่มา: https://www.researchgate.net/figure/CRISP-DM-data-mining-framework_fig1_341627969

จากปีค.ศ. 1996 DaimlerChrysler (Daimler-Benz) SPSS (ISL) และ NCR ได้พัฒนาเครื่องมือในการทำเหมืองข้อมูล ต่อมาปีค.ศ. 1999 จึงได้มีการเสนอตัวแบบ CRISP-DM (Cross -Industry Standard Process for Data Mining) ขึ้นมาโดยเป็นมาตรฐานในลักษณะของวงจรชีวิตในการทำเหมืองข้อมูล เพื่อวิเคราะห์และนำไปใช้ประโยชน์ประกอบไปด้วย 6 ขั้นตอนดังนี้

ขั้นตอนที่ 1 ทำความเข้าใจปัญหา (Business Understanding) มุ่งเน้นไปที่การทำความเข้าใจธุรกิจ ปัญหาและวัตถุประสงค์ จากนั้นทำปัญหาให้อยู่ในรูปแบบของโจทย์เพื่อวิเคราะห์ข้อมูล และวางแผนการดำเนินงานเบื้องต้น

ขั้นตอนที่ 2 ทำความเข้าใจข้อมูล (Data Understanding) เป็นการรวบรวมข้อมูล และ ทำความเข้าใจ ตรวจสอบคุณภาพ โดยเลือกข้อมูลที่เก็บรวบรวมมาว่าจะใช้ข้อมูลใดบ้างในการวิเคราะห์

ขั้นตอนที่ 1 และ 2 สามารถวนกลับไปมาได้ เพราะ การทำความเข้าใจธุรกิจทำให้เข้าใจข้อมูลมากขึ้น และการเข้าใจข้อมูลก็ทำให้เข้าใจธุรกิจมากขึ้นด้วย

ขั้นตอนที่ 3 การเตรียมข้อมูล (Data Preparation) เป็นการทำเพื่อให้ข้อมูลดิบ ให้กลายเป็นข้อมูลที่สมบูรณ์ โดยพร้อมจะเข้าสู่แบบจำลองในขั้นตอนถัดไป เช่น ลบข้อมูลไม่ต้องการออก เพิ่มหรือสร้างตาราง และ แปลงข้อมูลในรูปแบบที่ต้องการให้เป็น เป็นต้น

ขั้นตอนที่ 4 การสร้างแบบจำลอง (Modeling Phase) ดำเนินการเลือก และ ทดสอบสร้างแบบจำลองหลายๆรูปแบบที่คาดว่าจะสามารถแก้ไขปัญหาได้ แล้วค่อยๆปรับค่าพารามิเตอร์ ในแต่ละแบบจำลองที่

DPU DHURAKIJ PUNDIT UNIVERSITY มหาวิทยาลัยธุรกิจบัณฑิตย์

เหมาะสมที่สุดใช้ในการแก้ปัญหา หากยังไม่ได้แบบจำลองที่พอใจ หรือเหมาะสม ก็สามารถกลับไปเตรียมข้อมูลให้พร้อมได้ เนื่องจากข้อมูลที่ดี ก็จะทำให้ได้ผลลัพธ์ที่ดีเพิ่มขึ้น

ขั้นตอนที่ 5 การประเมินผลแบบจำลอง (Evaluation Phase) เป็นการวัดประสิทธิภาพของแบบจำลองที่ได้จากขั้นตอนที่ 4 เพื่อวัดแบบจำลองว่ามีประสิทธิภาพหรือไม่ ซึ่งจะเลือกใช้การวัดประสิทธิภาพที่แตกต่างกันออกไปตามแต่แบบจำลองแต่ละประเภท ที่เหมาะสม

ขั้นตอนที่ 6 การนำไปใช้งาน (Deployment Phase) เป็นการนำแบบจำลองที่เหมาะสมที่สุดไปใช้งานในการวิเคราะห์ และ แก้ปัญหาตามที่ต้องการ

สรุป CRISP-DM ถือประกระบวนการที่สำคัญสำหรับโจทย์ทางธุรกิจ และ ทางการวิเคราะห์การตลาดหรืออื่นๆ ที่จะช่วยให้สามารถใช้วิเคราะห์โจทย์ที่ต้องการได้อย่างครบถ้วนตั้งแต่เริ่มต้น จนสิ้นสุด รวมไปถึงการเก็บข้อมูลความคิดเห็นของลูกค้า ด้วยนั่นเอง

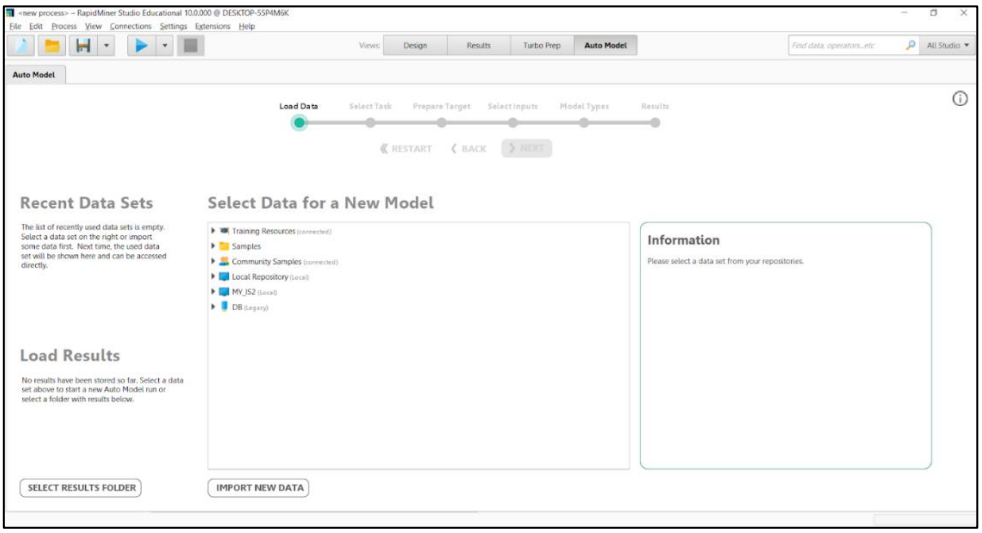
2.2. โปรแกรม RapidMiner Studio and Feature Auto Model

2.2.1. โปรแกรม RapidMiner Studio เป็นเครื่องมือที่เหมาะสำหรับการคำนวณวิเคราะห์ประมวลผล โดยมีโอเปอร์เรเตอร์มากมายให้เลือก สามารถนำมาใช้งานได้ง่าย ตั้งแต่กระบวนการเตรียมข้อมูล จนถึงขั้นตอน สร้างแบบจำลอง ให้เป็นไปต้องการ ซึ่งในงานวิจัยนี้ผู้ศึกษาวิจัยเลือกใช้ โปรแกรม RapidMiner Studio Version 10.0 ซึ่งเป็นประเภท Educational License Program ไว้ใช้สำหรับภาคการศึกษาทำให้สามารถใช้งานได้ฟรีโดยไม่จำกัดจำนวนแถว

ภาพที่ 2.3 โปรแกรม RapidMiner Studio Version 10.0

DPU DHURAKIJ PUNDIT UNIVERSITY มหาวิทยาลัยธุรกิจบัณฑิตย์

2.2.2. Feature Auto Model เป็นส่วนเสริม ๆ หนึ่งที่มีอยู่บนโปรแกรม RapidMiner โดยส่วนเสริม Auto Model ได้เริ่มเปิดตัวอย่างเป็นทางการใน RapidMiner Studio Version 8.1 ซึ่งส่วนเสริมดังกล่าวเป็นส่วนเสริม ที่ช่วยในการเปรียบเทียบ การทำนาย(Prediction) , การรวมกลุ่ม(Clustering) และ การแก้ไขปัญหาค่าผิดปกติ(Outliers) โดยใช้ Automated Machine Learning เพื่อให้สามารถสร้าง และเปรียบเทียบแบบจำลองต่าง ๆทาง Machine Learning หลายๆแบบจำลองพร้อม ๆกันเพียงไม่กี่ขั้นตอน อีกทั้งช่วยลดเวลาในการสร้างแบบจำลองหลายๆแบบจำลอง พร้อมทั้งยังแสดงข้อมูลเปรียบเทียบประสิทธิภาพจากส่วนเสริม “Feature Auto Model” บน RapidMiner Studio



ภาพที่ 2.4 ส่วนเสริม “Feature Auto Model” บน RapidMiner Studio Version 10.0

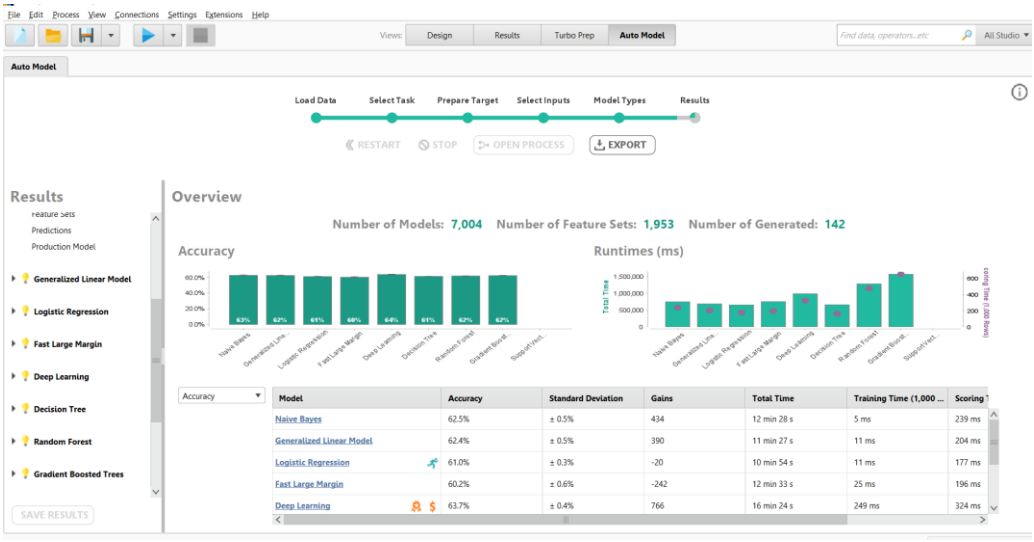
ช่องทางการเข้าใช้งาน “Feature Auto Model” บน RapidMiner Studio มีหลากหลายช่องทางดังนี้

1.หน้าต่าง New process: Welcome to RapidMiner Studio เมนู Auto Model

2.เมนู Auto Model บน Views Tab

3.เมนู Auto Model บนตาราง Example Set

DPU DHURAKIJ PUNDIT UNIVERSITY มหาวิทยาลัยธุรกิจบัณฑิตย์

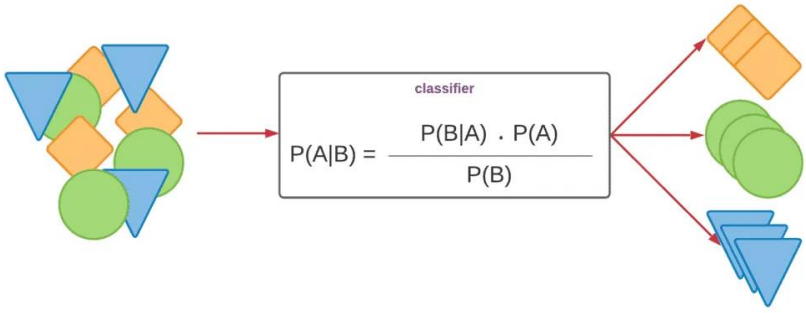


ภาพที่ 2.5 ตัวอย่างหลังจาก Run “Feature Auto Model” สำเร็จ ในส่วนเสริม “Feature Auto Model” บน RapidMiner Studio Version 10.0

2.3. เทคนิคการเรียนรู้ของเครื่อง (Machine Learning Technique)

2.3.1. การจำแนกประเภทข้อมูลแบบเบย์อย่างง่าย (Naive Bayes)

การจำแนกประเภทข้อมูลแบบเบย์อย่างง่าย เป็นเทคนิคที่อาศัย หลักความน่าจะเป็นตามทฤษฎีของเบย์ (Bayes’s theorem) มีอัลกอริทึมที่ไม่ค่อยซับซ้อน ซึ่งเป็นเทคนิคจำแนกข้อมูล จากการเรียนรู้ปัญหาที่เกิด เพื่อนำมาสร้างเงื่อนไขการจำแนกแจกข้อมูลใหม่ ในกลุ่มเดียวกัน

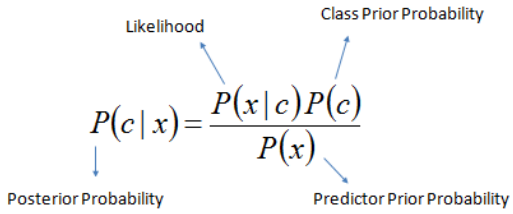


ภาพที่ 2.6 การจำแนกประเภทข้อมูลแบบเบย์อย่างง่าย

ที่มา: https://thinkinfi.com/naive-bayes-algorithm-in-machine-learning-with-python/

DPU DHURAKIJ PUNDIT UNIVERSITY มหาวิทยาลัยธุรกิจบัณฑิตย์

การจำแนกประเภทข้อมูลแบบเบย์อย่างง่ายนั้น ใช้สมมติฐานว่าปริมาณของความสนใจ จะขึ้นอยู่กับการกระจายความน่าจะเป็น ซึ่งเป็นเทคนิคของการแก้ปัญหาแบบจำแนกประเภทที่สามารถคาดการณ์ผลลัพธ์ได้ โดยวิเคราะห์ความสัมพันธ์ระหว่าง ตัวแปรในการใช้สำหรับการสร้างเงื่อนไขความน่าจะเป็น สำหรับแต่ละความสัมพันธ์ เหมาะกับกรณีของเซต ตัวอย่างที่มีจำนวนมากและ คุณลักษณะของตัวอย่างไม่ขึ้นต่อกัน โดยกำหนดให้ความน่าจะ เป็นของข้อมูลเท่ากับสมการ



$$P(c \mid \mathbf{X}) = P(x_1 \mid c) \times P(x_2 \mid c) \times \cdots \times P(x_n \mid c) \times P(c)$$

โดย P (c | x) คือ ความน่าจะเป็น ที่จะเกิดเหตุการณ์ c เมื่อเกิดเหตุการณ์ x ขึ้นก่อน

P (x | c) คือ ความน่าจะเป็น ที่จะเกิดเหตุการณ์ x เมื่อเกิดเหตุการณ์ c ขึ้นก่อน

P (x) คือ ความน่าจะเป็น ที่จะเกิดเหตุการณ์ x

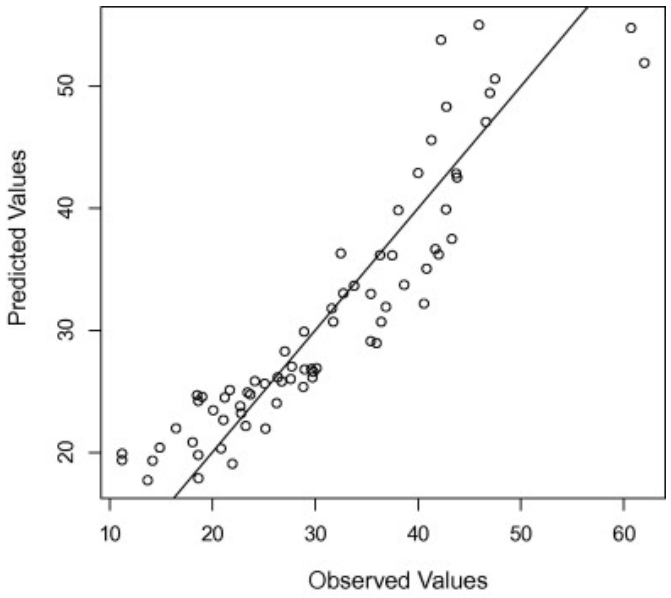
P (c) คือ ความน่าจะเป็น ที่จะเกิดเหตุการณ์ c

DPU DHURAKIJ PUNDIT UNIVERSITY มหาวิทยาลัยธุรกิจบัณฑิตย์

2.3.2. ตัวแบบเชิงเส้นนัยทั่วไป (Generalized Linear Model)

ตัวแบบเชิงเส้นนัยทั่วไป เป็นเทคนิคที่ถูกแนะนำครั้งแรกในปี ค.ศ. 1972 โดย John Nelder และ Robert Wedderburn ตัวแบบเชิงเส้นนัยทั่วไปเป็นวิธีวิเคราะห์ทางสถิติที่นิยมวิธีหนึ่ง ซึ่งสามารถแสดงผลการวิเคราะห์ที่สามารถบ่งบอกระดับความเชื่อมั่นทางสถิติ หรือ บ่งบอกผลการประเมินคุณภาพของตัวแปรได้ด้วย

แนวคิดของการวิเคราะห์ของเทคนิคนี้คือ หาความสัมพันธ์ระหว่างตัวแปรต้น (Explanatory Variables) และตัวแปรตาม (Response Variable) โดยที่ตัวแปรนำมาวิเคราะห์หาความสัมพันธ์นั้นสามารถเป็นได้ทั้งข้อมูลตัวเลข หรือ รูปแบบอื่น ๆ ที่ครอบคลุมทั้ง ตัวแปรผลลัพธ์แบบต่อเนื่องและไม่ต่อเนื่อง ซึ่งอยู่ภายใต้ฟังก์ชันการแจกแจงตระกูลเอกซ์โพเนนเชียล



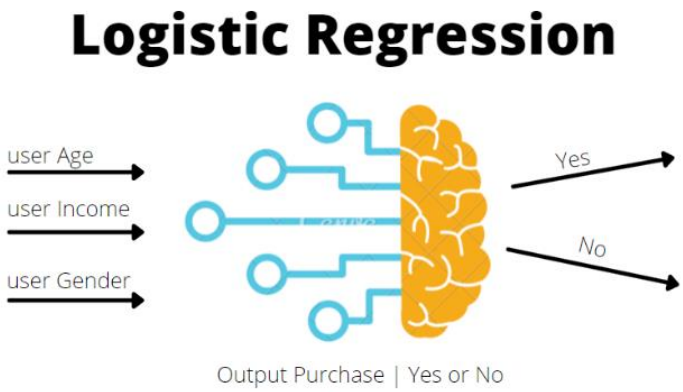
ภาพที่ 2.7 เทคนิคตัวแบบเชิงเส้นนัยทั่วไป

ที่มา: https://www.sciencedirect.com/topics/mathematics/generalized-linear-model

2.3.3. การถดถอยโลจิสติก (Logistic Regression)

การถดถอยโลจิสติก เป็นเทคนิคการวิเคราะห์ที่มีวัตถุประสงค์เพื่อ ประมาณค่า หรือ ทำนายว่าจะเกิดเหตุการณ์ขึ้น หรือ ไม่เกิดเหตุการณ์นั้นภายใต้ปัจจัยต่างๆ ซึ่งแบบจำลองโลจิสติก จะประกอบไปด้วยตัวแปรตาม หรือ ตัวแปรเกณฑ์ ที่เป็นตัวแปรแบบทวินาม (Dichotomous Variable) คือ มีได้สองค่า เช่น “เกิด” กับ “ไม่เกิด” หรือ “เสี่ยง” กับ “ไม่เสี่ยง” เป็นต้น และ ตัวแปรอิสระ หรือตัวแปรทำนาย อาจจะมีตัวเดียวหรือหลายตัว ซึ่งเป็นได้ทั้งตัวแปรเชิงกลุ่ม (Categorical Variable) หรือ ตัวแปรแบบต่อเนื่อง (Continuous Variable) แสดงดังภาพ

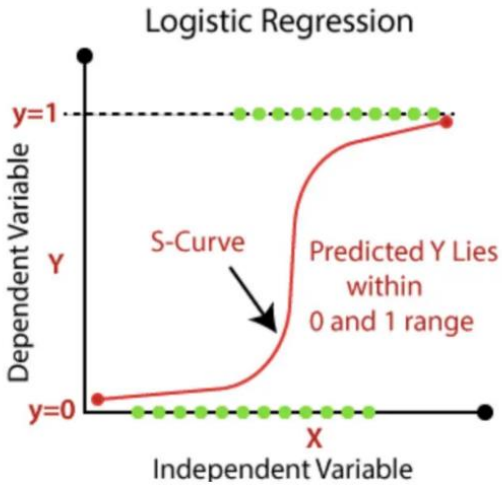
DPU DHURAKIJ PUNDIT UNIVERSITY มหาวิทยาลัยธุรกิจบัณฑิตย์



ภาพที่ 2.8 เทคนิคการถดถอยโลจิสติก

ที่มา: https://blog.devgenius.io/develop-a-logistic-regression-machine-learning-model

หลักการวิเคราะห์การถดถอยโลจีสติก เกี่ยวข้องกับทฤษฎีความน่าจะเป็นแบบทวินาม ถูกเรียกว่า Binomial Logistic Regression ถ้าตัวแปรตามเป็นพหุนามจะเรียกว่า Multinomial Logistic Regression การถดถอยโลจีสติก นั้นมีการประยุกต์ใช้ในงานวิจัยหลากหลายสาขา ทั้งสาขาทางการแพทย์วิศวกรรมศาสตร์ นิเวศวิทยา เศรษฐศาสตร์และสังคมศาสตร์ และอื่นๆอีกมากมาย



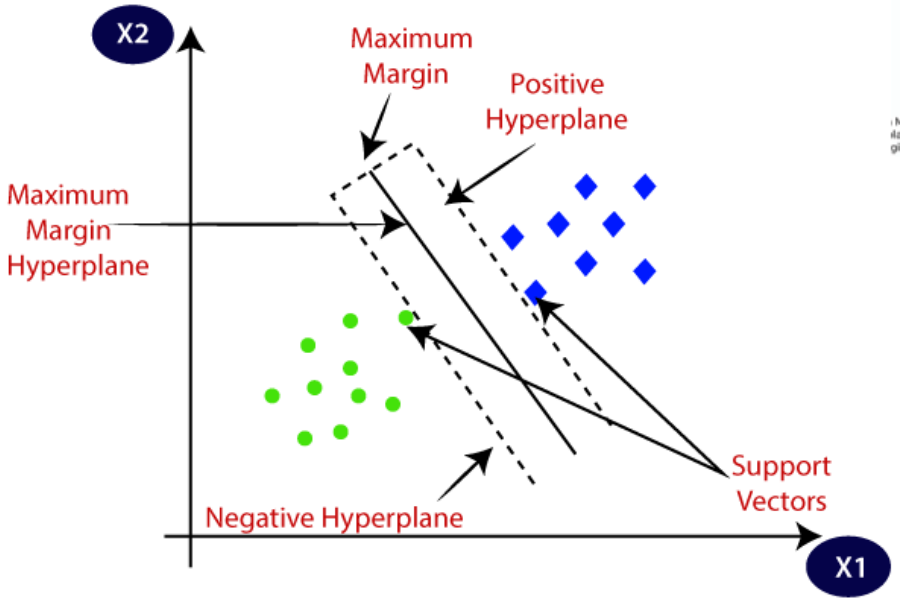
ภาพที่ 2.9 หลักการวิเคราะห์การถดถอยโลจีสติก

ที่มา: https://sparkbyexamples.com/machine-learning/logistic-regression-explained-with-examples/

DPU DHURAKIJ PUNDIT UNIVERSITY มหาวิทยาลัยธุรกิจบัณฑิตย์

2.3.4. ซัพพอร์ตเวกเตอร์แมทชีน (Support Vector Machine)

ซัพพอร์ตเวกเตอร์แมทชีน เป็นเทคนิคที่สามารถนำมาช่วยแก้ปัญหาในการจำแนกข้อมูลได้ โดยเฉพาะกับปัญหา ของข้อมูลที่มีขนาดของข้อมูลไม่ใหญ่มากนักแต่คุณลักษณะ (features) ของข้อมูลมีจำนวนมาก มันอาศัยหลักการ หาสัมประสิทธิ์ของสมการ ในการสร้างเส้นแบ่งแยกกลุ่มข้อมูล โดยจะป้อนเข้าสู่กระบวนการสอน ให้ระบบเรียนรู้ที่เน้นไปยังเส้นแบ่งแยกแยะของกลุ่มข้อมูลได้ดีที่สุดออกมา



ภาพที่ 2.10 เทคนิคซัพพอร์ตเวกเตอร์แมทชีน

ที่มา: https://www.javatpoint.com/machine-learning-support-vector-machine-algorithm

แนวความคิดของ Support Vector Machine เกิดจากการที่นำค่าของกลุ่มข้อมูล วางลงในปริภูมิคุณลักษณะ (Feature Space) จากนั้นจึงหาเส้นที่ใช้แบ่งข้อมูลทั้งสองออกจากกันโดยจะสร้างเส้นแบ่ง (Hyperplane) ที่เป็นเส้นตรงขึ้นมา และเพื่อให้ทราบว่าเส้นตรงที่แบ่งสองกลุ่มออกจากกัน เส้นตรงใดเป็นเส้นตรงที่ดีที่สุด ด้วยการนำใช้ฟังก์ชันเคอร์เนล (Kernal Function) มาใช้เพื่อแปลงข้อมูล ไปยังมิติที่สูงขึ้นในปริภูมิคุณลักษณะ (Feature Space) โดยจำแนกข้อมูลที่มีความคลุมเครือ ได้อย่างมีประสิทธิภาพ และ มีหลักการคือ การหาเส้นตรงที่มีมาร์จินที่โตที่สุด (Maximum Margin) ที่ สามารถแบ่งข้อมูลออกเป็น 2 คำตอบ

DPU DHURAKIJ PUNDIT UNIVERSITY มหาวิทยาลัยธุรกิจบัณฑิตย์

2.3.5. การจำแนกประเภทแบบเร็วโดยใช้ขอบเขตการแยกแยะที่มีขนาดกว้าง(Fast Large Margin (RapidMiner Studio Core))

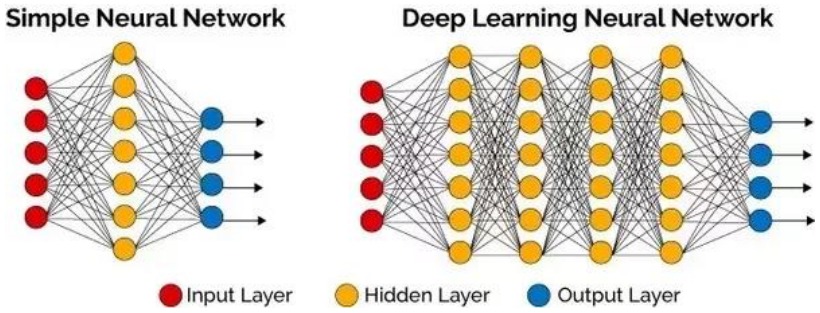
การจำแนกประเภทแบบเร็วโดยใช้ขอบเขตการแยกแยะที่มีขนาดกว้าง เป็นเทคนิคที่อ้างอิงมาจาก the linear support vector learning ของ R.E. Fan, K.W. Chang, C.J. Hsieh, X.R. Wang, และ C.J. Lin. ซึ่งเทคนิคนี้เป็นเทคนิคใหม่ที่ มีการคำนวณผลลัพธ์คล้ายกับเทคนิคของ วิธีซัพพอร์ตเวกเตอร์แมทชีน (Support Vector Machine) เป็นอย่างมาก สามารถประมวลผลข้อมูลตัวอย่างและแอททริบิวต์(Attribute) ที่มีจำนวนมากได้ เทคนิคนี้เป็นตัวดำเนินการ (Operator) ชนิดหนึ่งบน RapidMiner Studio Core ซึ่งอยู่ภายใต้ เทคนิควิธีซัพพอร์ตเวกเตอร์แมทชีน (Support Vector Machine) ที่สามารถเรียกใช้งานได้อย่างง่าย และ สามารถเลือกเทคนิคดังกล่าวเพื่อนำมาเป็นส่วนหนึ่งของการเปรียบเทียบประสิทธิภาพ แบบจำลองใน “Feature Auto Model” บน RapidMiner Studio ได้

ภาพที่ 2.11 เทคนิคการจำแนกประเภทแบบเร็วโดยใช้ขอบเขตการแยกแยะที่มีขนาดกว้าง(Fast Large Margin (RapidMiner Studio Core))

DPU DHURAKIJ PUNDIT UNIVERSITY มหาวิทยาลัยธุรกิจบัณฑิตย์

2.3.6. การเรียนรู้เชิงลึก (Deep Learning)

การเรียนรู้เชิงลึก เป็นเทคนิคการเรียนรู้แบบอัตโนมัติ ซึ่งประมวลผลข้อมูลจำนวนมาก ขนาดใหญ่ ที่เลียนแบบการทำงานของโครงข่ายประสาทของมนุษย์ (Neurons) โดยเป็นการนำโครงข่ายประสาท (Neural Network) มาซ้อนกัน หลายชั้น (Layer) จากนั้นจึงทำการเรียนรู้ข้อมูลตัวอย่าง ซึ่งข้อมูล ดังกล่าวจะถูกนำไปใช้ในการตรวจจับรูปแบบ (Pattern) หรือจัด หมวดหมู่ข้อมูล (Classify the Data) ทำให้สามารถทำการตัดสินใจด้วยการขับเคลื่อนด้วยข้อมูลได้อย่างแม่นยำ



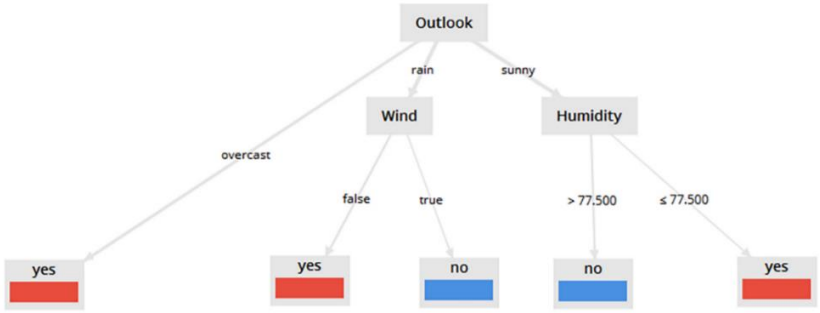
ภาพที่ 2.12 เทคนิคการเรียนรู้เชิงลึก

ที่มา: https://www.global-engage.com/life-science/deep-learning-in-digital-pathology/

2.3.7. ต้นไม้การตัดสินใจ (Decision Tree)

ต้นไม้การตัดสินใจ เป็นเทคนิคหนึ่งของ วิธีการเรียนรู้ที่ใช้ สถิติจากการเรียนรู้ของเครื่อง และการทำเหมืองข้อมูล จะสังเกตการณ์แบ่งแยกข้อมูล ซึ่งเป็นกระบวนการวิเคราะห์ข้อมูลที่นักวิทยาศาสตร์ข้อมูลทั้งมือใหม่และมือเก่านิยมใช้อย่างแพร่หลาย เนื่องจากเป็นแบบจำลองที่ง่ายต่อการเข้าใจ และ แสดงผลลัพธ์ออกมา เทคนิคนี้มีส่วนประกอบหลัก ได้แก่ ราก (Root Node) คือชุดข้อมูลและกฎเกณฑ์ตั้งต้นเมื่อทำการวิเคราะห์ไปเรื่อย ๆ ต่อมาก็คือกิ่ง (Branch) คือตัวระบุเงื่อนไขในการตัดสินใจ ถ้าผ่านเงื่อนไขในนี้ก็จะย้ายไปยังการตัดสินใจที่สำคัญรองลงมาเรื่อยจนถึงกลุ่มสุดท้ายซึ่งจะเรียกกลุ่มเหล่านั้นว่า ใบ (Leaf Node) โดยจะทำการคัดเลือกแอตทริบิวต์ ที่มีความสัมพันธ์กับคลาสมากที่สุดขึ้นมาเป็นโหนดบนสุดหลังจากนั้นก็จะหาแอตทริบิวต์ถัดไปเรื่อยๆ

DPU DHURAKIJ PUNDIT UNIVERSITY มหาวิทยาลัยธุรกิจบัณฑิตย์



ภาพที่ 2.13 ตัวอย่าง ราก (Root Node) ,กิ่ง (Branch) และ ใบ (Leaf Node) ของต้นไม้การตัดสินใจ

ที่มา: https://community.rapidminer.com/discussion/47681/visualisation-of-decision-trees

การหาความสัมพันธ์ของแอตทริบิวต์ ซึ่งจะใช้ตัววัด ที่เรียกว่า Information Gain (IG) ซึ่งเป็นสมการที่ใช้ใน การหาค่าสารสนเทศก่อนนำไปใช้ในการหาค่ามาตรฐาน อัตราส่วนเกน (Gain Ratio)

$$Gain(S,A) = Entropy(s) - \sum_{veValues(A)} \frac{|S_v|}{|S|} Entropy(S_v)$$

โดย A คือ คุณลักษณะ A

$|S_v|$ คือ สมาชิกของคุณลักษณะ A ที่มีค่า v

$|s|$ คือ จำนวนสมาชิกของกลุ่มตัวอย่าง

Entropy คือ ค่าที่ใช้วัดความแตกต่างกลุ่มข้อมูลตัวอย่าง หรือ ความไม่บริสุทธิ์ของกลุ่มข้อมูล (Entropy Measure)

$$Entropy(s) = \sum_{i=1}^{c} -P_i \log_2 P_i$$

โดย S คือ Attribute ที่นำมาวัดค่า Entropy

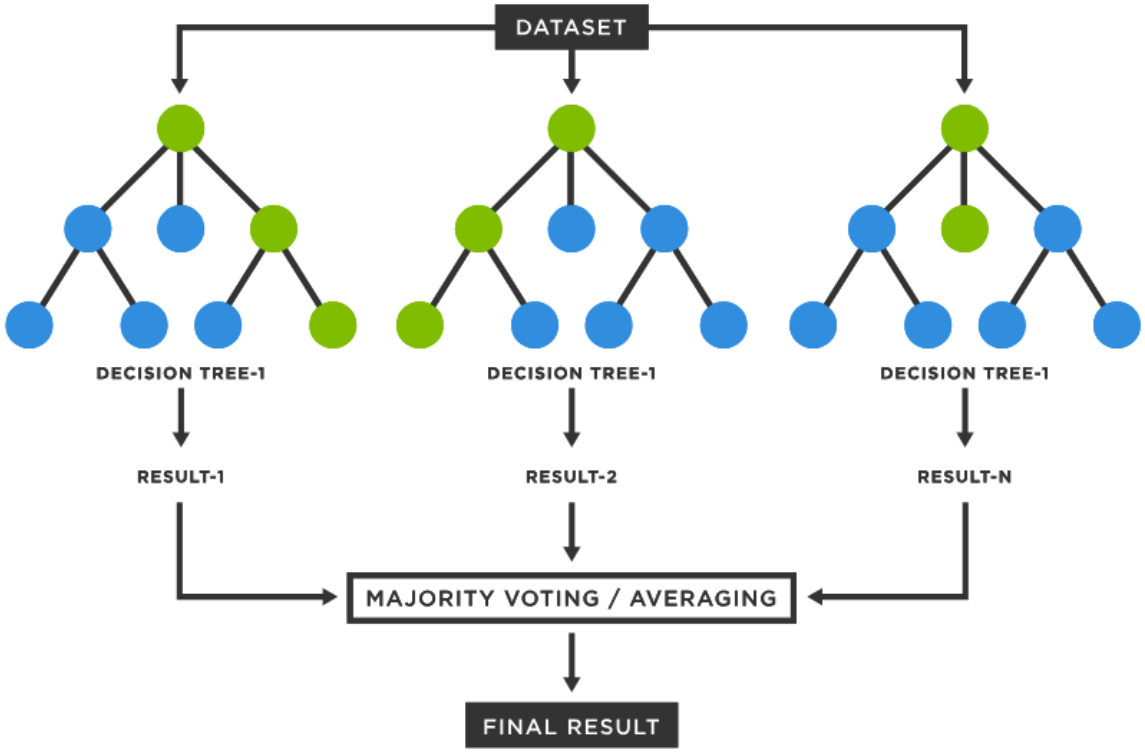
P_i คือ สัดส่วนของจำนวนสมาชิกในกลุ่ม i เท่ากับจำนวนสมาชิกทั้งหมดของกลุ่มตัวอย่าง

ดังนั้น ค่า Information Gain คือการวดัค่า Entropy ก่อนที่จะมีการแบ่ง ข้อมูลออกตามมิติของข้อมูลและหลังจากการแบ่งว่า มีประสิทธิภาพดีขึ้นหรือไม่ ถ้ามีประสิทธิภาพดีขึ้นค่า Information Gain จะมีค่าสูง

DPU DHURAKIJ PUNDIT UNIVERSITY มหาวิทยาลัยธุรกิจบัณฑิตย์

2.3.9. แบบจำลองป่าสุ่ม (Random Forest)

แบบจำลองป่าสุ่ม เป็นเทคนิควิธีหนึ่งที่นิยมเป็นอย่างมากในปัจจุบัน ซึ่งสามารถใช้กับปัญหาวิเคราะห์การถดถอย (Regression) และ ปัญหาเพื่อการจำแนก (Classification) โดย แบบจำลองป่าสุ่ม ต่อยอดมาจาก เทคนิคต้นไม้การตัดสินใจ แต่แบบจำลองป่าสุ่มนั้นมี ประสิทธิภาพการทำงาน และ พยากรณ์สูงขึ้นกว่า หลักการทำงานของเทคนิคนี้ คือจะแบ่งข้อมูลออกเป็น ต้นไม้ตัดสินใจ (Decision Tree) หลายๆ รูปแบบจำลองย่อย ๆ โดยแต่ละแบบจำลองจะได้รับคุณลักษณะ (Feature) และข้อมูล (Data) ที่ ไม่เหมือนกันทั้งหมด เพื่อทำให้ได้ต้นไม้ที่มีความหลากหลายและมีความอิสระต่อกัน



ภาพที่ 2.14 เทคนิคแบบจำลองป่าสุ่ม

ที่มา: https://www.tibco.com/reference-center/what-is-a-random-forest

โดยการทำงานของ แบบจำลองป่าสุ่ม มีลำดับดังนี้

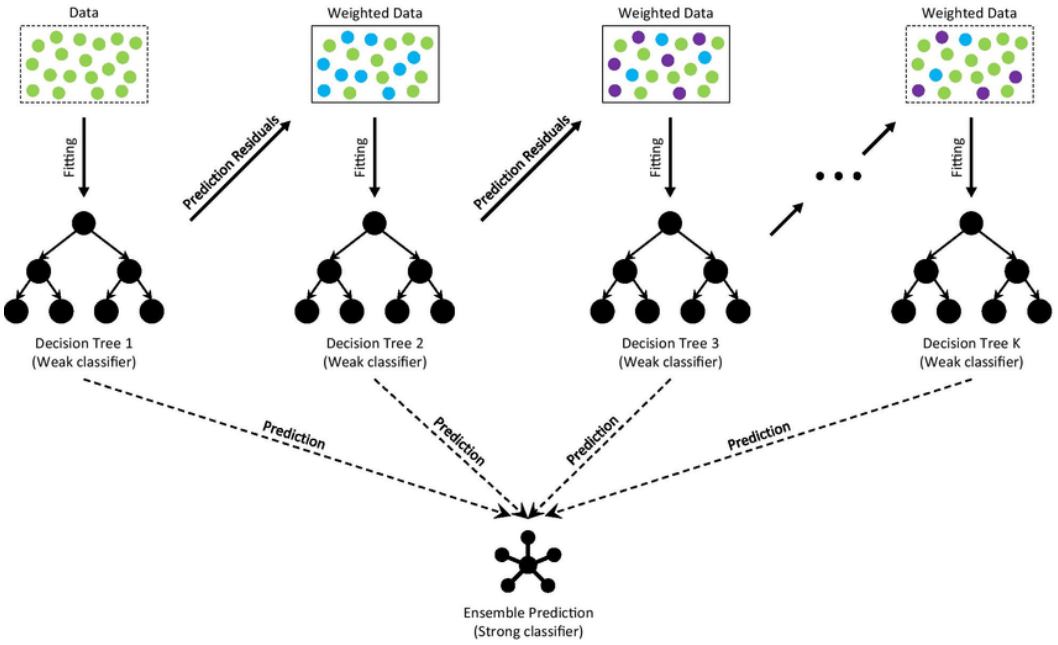
1. สุ่มเลือก คุณลักษณะ (Feature) และ ข้อมูล (Data) จากชุดข้อมูลทั้งหมดที่มี
2. สร้างต้นไม้ตัดสินใจ จากข้อมูลตัวอย่างแต่ละชุด และหาค่าพยากรณ์จากต้นไม้แต่ละต้น
3. เลือกจำนวนต้นไม้ตัดสินใจที่ต้องการจากนั้น จะทำซ้ำในขั้นตอน 1 และ 2 ในการสร้างต้นไม้

DPU DHURAKIJ PUNDIT UNIVERSITY มหาวิทยาลัยธุรกิจบัณฑิตย์

4. ค่าพยากรณ์ที่ได้จะเป็นการให้ต้นไม้ตัดสินใจแต่ละต้นหาค่าพยากรณ์ของตัวเอง จากนั้นค่าพยากรณ์ค่าสุดท้าย ในกรณีที่ปัญหาเป็นเพื่อการจำแนก (Classification) จะใช้วิธีผลโหวตมากที่สุด (Majority vote) โดยค่าพยากรณ์ของต้นไม้ตัดสินใจต้นได้รับค่าของผลโหวตมากที่สุดจะถูกเลือกให้เป็นค่าพยากรณ์ ของปัญหาซึ่งถ้าเป็นปัญหาวิเคราะห์การถดถอย (Regression) จะใช้วิธีคำนวณหาค่าเฉลี่ย (Mean) โดยจะนำเอาค่าของการพยากรณ์ของทุกต้นไม้ตัดสินใจมา หาค่าเฉลี่ยเพื่อเป็นค่าพยากรณ์ของปัญหา

2.3.9. ต้นไม้การเพิ่มไล่ระดับการตัดสินใจ (Gradient Boosted Trees)

ต้นไม้การเพิ่มไล่ระดับการตัดสินใจ เป็นชุดของแบบจำลองการถดถอย (Regression) และการจำแนกประเภท (Classification) ทั้งคู่เป็นเทคนิคการเรียนรู้แบบ Forward-learning เป็นเทคนิคที่มีพื้นฐานมาจากต้นไม้ตัดสินใจ เทคนิคนี้จะปรับปรุงประสิทธิภาพของแบบจำลองให้มีค่าสูงขึ้น โดยจะสร้างต้นไม้ตัดสินใจต่อกันหลายๆแบบ โดยแต่ละต้นไม้ตัดสินใจจะเรียนรู้จาก Error ของแบบจำลองก่อนหน้า เพื่อปรับปรุงประสิทธิภาพของแบบจำลองให้สูงยิ่งขึ้นต่อไป และ จะประเมินผลของแต่ละแบบจนกว่าจะได้ ต้นไม้ตัดสินใจ ที่สมบูรณ์ที่สุด



ภาพที่ 2.15 เทคนิคต้นไม้การเพิ่มไล่ระดับการตัดสินใจ

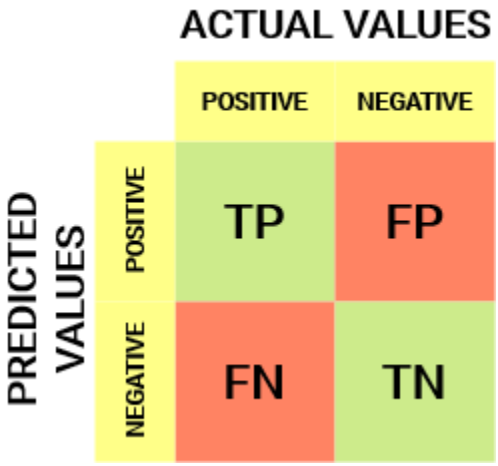
ที่มา: https://www.researchgate.net/figure/The-architecture-of-Gradient-Boosting-Decision-Tree_fig2_356698772

DPU DHURAKIJ PUNDIT UNIVERSITY มหาวิทยาลัยธุรกิจบัณฑิตย์

2.4. การวัดประสิทธิภาพของแบบจำลอง

การที่จะนำแบบจำลองไปใช้ได้จริง ๆนั้น จำเป็นจะต้องมีการวัดประสิทธิภาพของแบบจำลองดังกล่าวว่าแบบจำลองมีประสิทธิภาพเพียงพอที่จะนำมาพัฒนาต่อและสามารถนำใช้งานได้จริงหรือไม่ ตาราง Confusion Matrix เป็นหนึ่งในการวัดประสิทธิภาพประเภทหนึ่งที่เป็นที่นิยมอยู่เช่นกัน

Confusion Matrix ก็คือ การประเมินผลลัพธ์การทำนาย เปรียบเทียบกับผลลัพธ์จริง ออกมาเป็นตารางแบบจัตุรัสโดยมีจำนวนแถวเท่ากับจำนวนคอลัมน์และเท่ากับจำนวนคลาส เช่น หากคำตอบอยู่ 2 ค่า คือ “ใช่” และ “ไม่” ตาราง confusion matrix ที่เกิดขึ้นจะแสดง ตารางขนาด 2คูณ2 เป็นต้น



ภาพที่ 2.16 ตาราง Confusion Matrix

ที่มา: https://www.analyticsvidhya.com/blog/2020/04/confusion-matrix-machine-learning/

จากรูปข้างต้นตาราง Confusion Matrix ประกอบไปด้วย

ข้อมูลคอลัมน์ คือ คลาสที่อยู่ในข้อมูลเทรนนิ่ง ดาต้า (actual values)

ข้อมูลแนวแถว คือ คลาสที่แบบจำลองได้ทำนายมา (predicted values)

ตาราง Confusion Matrix นั้นสามารถนำมาใช้คำนวณ เพื่อตรวจสอบว่าการพัฒนาแบบจำลองให้มี ประสิทธิภาพเพียงพอ ที่จะนำไปใช้งานจริงได้หรือไม่ โดยการคำนวณที่ใช้ประเมินประสิทธิภาพของการทำนายจากตาราง Confusion Matrix ที่เป็นที่นิยมนำมาใช้ในงานทางด้าน Data มี 3 การคำนวณที่เป็นที่นิยม ซึ่งนำมาใช้ในงานวิจัยนี้ มีการคำนวณดังนี้

DPU DHURAKIJ PUNDIT UNIVERSITY มหาวิทยาลัยธุรกิจบัณฑิตย์

- ค่าความถูกต้อง (Accuracy) หมายถึง ค่าที่วัดได้เข้าใกลค่าจริงมากน้อยเพียงใด 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 ค่าความครบถ้วนมาพิจารณาร่วมกันเฉลี่ยแบบ ค่าเฉลี่ยฮาร์มอนิก (harmonic mean)

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

โดย

True Positive (TP) คือ ความถี่ของสิ่งที่ทำนาย ตรงกับสิ่งที่เกิดขึ้นจริงในกรณี ทำนายว่าจริง และสิ่งที่เกิดขึ้น ก็คือ จริง คำนวณได้จาก True Positive rate = ΣTrue positive / ΣCondition positive2

True Negative (TN) คือ ความถี่ของสิ่งที่ทำนายตรงกับสิ่งที่เกิดขึ้น ในกรณี ทำนายว่า ไม่จริง และสิ่งที่เกิดขึ้น ก็คือ ไม่จริง คำนวณได้จาก True Negative rate = ΣTrue negative / ΣCondition negative

False Positive (FP) คือ ความถี่ของสิ่งที่ทำนายไม่ตรงกับสิ่งที่เกิดขึ้น คือทำนายว่า จริง แต่สิ่งที่เกิดขึ้น คือ ไม่จริง คำนวณได้จาก False Positive rate = ΣFalse positive / ΣCondition negative

False Negative (FN) คือ ความถี่ของสิ่งที่ทำนายไม่ตรงกับที่เกิดขึ้นจริง คือทำนายว่าไม่จริง แต่สิ่งที่เกิดขึ้น คือ จริง คำนวณได้จาก False Negative rate= ΣFalse negative /ΣCondition positive

DPU DHURAKIJ PUNDIT UNIVERSITY มหาวิทยาลัยธุรกิจบัณฑิตย์

2.5. บทความและงานวิจัยที่เกี่ยวข้อง

Francesco Pochetti. (2019) จากบทความ Extreme Label Imbalance: When You Measure the Minority Class in Basis Points.ได้เสนอเทคนิคในการแก้ไขปัญหาที่ข้อมูลไม่สมดุลด้วยเทคนิค Essembling + Oversampling มีซึ่งมีวิธีการคือให้แบ่งข้อมูลคลาสส่วนใหญ่ออกเป็นจำนวน N ส่วน โดยนำข้อมูลแบ่งออกมาแต่ละส่วนรวมกันกับมูลในคลาสส่วนน้อย โดยจะได้มูลชุดสอนใหม่จำนวน N ชุด โดยในแต่ละชุดจะทำ การสุ่มลดหรือสุ่มเกิน จากการสุ่มจะนำข้อมูลของแต่ละชุดไปเข้าแบบ จำลองและนำผลลัพธ์ของแต่ละแบบจำลองหาค่าเฉลี่ยออกมาซึ่งสามารถได้ผลออกมาเป็นที่น่าพอใจ.

สุรวัชร ศรีเปารยะ และสายชล สินสมบูรณทอง (2560) งานวิจัยการเปรียบเทียบประสิทธิภาพวิธีการจำแนกกลุ่มการ เป็นโรคไตเรื้อรัง มีวัตถุประสงค์เพื่อเปรียบเทียบประสิทธิภาพของวิธีการจำแนกกลุ่มการเป็นโรคไตเรื้อรังของโรงพยาบาลแห่งหนึ่ง โดยเลือกใช้ 7วิธี ได้แก่ วิธีความใกล้เคียงกันมากที่สุด วิธีต้นไม้ตัดสินใจวิธีโครงข่ายประสาทเทียม วิธีซัพพอร์ตเวกเตอร์แมชชีน วิธีฐานกฎวิธีการถดถอยโลจิสติก และวิธีนาอีฟ เบย์ ซึ่งผลจากการเปรียบเทียบประสิทธิภาพของวิธีการจำแนกกลุ่มการเป็นโรคไตเรื้อรังของโรงพยาบาลแห่งหนึ่งพบว่าวิธีต้นไม้ตัดสินใจ เป็นวิธีที่มีค่าความถูกต้องและค่าความคลาดเคลื่อนกำลังสองเฉลี่ยมากที่สุด คือพบ 100%และ 0.0059 ตามลำดับ ดังนั้นจึสรุปได้ว่า วิธีต้นไม้ตัดสินใจ ในการเปรียบเทียบในงานวิจัยชิ้นนี้เป็นวิธีการจำแนกกลุ่มที่มีประสิทธิภาพการจำแนกดีที่สุดสำหรับข้อมูลดังกล่าวที่นำมาใช้กับงานวิจัยชิ้นนี้

Oleksandr Andrieiev.(2022) จากบทความ Propensity Model: Using Data to Predict Customer Behavior อธิบายว่า Propensity Model คือแบบจำลองที่ทำให้เข้าใจว่า กลุ่มเป้าหมายหรือลูกค้า ตอบสนองอย่างไรกับการส่งข้อมูลหรือดำเนินการบางประเภทออกไป การดำเนินการเหล่านี้อาจมีตั้งแต่การยอมรับข้อเสนอ หรือการซื้อ ซึ่งเป็นรูปแบบของการวิเคราะห์หรือทำนายพฤติกรรมล่วงหน้าจากพฤติกรรมในอดีตของพวกเขาเหล่านั้น เพื่อช่วยให้ประหยัดงบการตลาดเพื่อเจาะกลุ่มลูกค้าได้ตรงกับความรู้สึก และ อาจทำให้ได้ผลการตอบรับจากลูกค้ามากยิ่งขึ้นจากแบบจำลองข้างต้น และ propensity to engage จึงเป็นการทำนายพฤติกรรมการมีส่วนร่วมของลูกค้า 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 มีวัตถุประสงค์เพื่อ หาแนวทางการจัดประเภทเพื่อทำนายการตัดสินใจซื้อของลูกค้าระหว่างแบรนด์ของสินค้าตามประวัติลูกค้า และ การทำธุรกรรม จากร้านขายของชำในบันดุงประเทศอินโดนีเซีย เป็นการเปรียบเทียบ machine learning models ดังนี้ naive bayes, linear models, deep learning, and decision trees ใช้โปรแกรม RapidMiner โดยผลลัพธ์ที่ได้

DPU DHURAKIJ PUNDIT UNIVERSITY มหาวิทยาลัยธุรกิจบัณฑิตย์

ระบุว่า naive bayes สามารถทำงานได้ดีที่สุดในฐานะอัลกอริทึมการทำนาย ซึ่งมีค่า Precision เท่ากับร้อยละ 60.3 ,Recall เท่ากับร้อยละ 65.5 และ Accuracy เท่ากับร้อยละ 61.7 ซึ่งมีค่ารสูงที่สุด แต่ปัจจัยด้านประชากรศาสตร์ เช่น อายุและสถานที่ ตลอดจนปัจจัยเชิงบริบท เช่น ช่วงเวลาของสัปดาห์ มีอิทธิพลอย่างมากต่อแนวโน้มการซื้อของลูกค้า ซึ่งงานวิจัยชิ้นนี้สามารถบอกข้อมูลเชิงลึกที่เกี่ยวกับ วิธีการประยุกต์ใช้ machine learning models มาปรับปรุงการคาดการณ์พฤติกรรมผู้บริโภคและตอบสนองต่อความต้องการของตลาดที่เปลี่ยนแปลงในอนาคต

DPU DHURAKIJ PUNDIT UNIVERSITY มหาวิทยาลัยธุรกิจบัณฑิตย์

บทที่ 3
ระเบียบวิธีวิจัย

งานวิจัยนี้มุ่งเน้นเพื่อศึกษาเหตุการณ์ที่เกิดขึ้นแล้ว โดยรวบรวมข้อมูลรายงานประจำวันจากผลการสื่อสารบรอดแคสต์แบบระบุกลุ่มเป้าหมาย จากบัญชีทางการธุรกิจของแอปพลิเคชันไลน์ โดยนำเข้าสู่กระบวนการ และ เปรียบเทียบเทคนิคต่าง ๆ เพื่อหาเทคนิคแบบจำลองในการพยากรณ์ข้อมูล ที่จะนำมาใช้ในการปรับปรุงการทำนายในอนาคต โดยแนวทางการวิจัยใช้หลักการ Cross-Industry Standard Process For Data Mining (CRISPDM) เป็นกระบวนการมาตรฐานที่นิยมใช้ในการวิเคราะห์ข้อมูลสำหรับการทำเหมืองข้อมูลมาประยุกต์ใช้ ซึ่งมีรายละเอียดดังนี้

3.1. ทำความเข้าใจปัญหา (Business Understanding Phase)

3.2. ทำความเข้าใจข้อมูล (Data Understanding Phase)

3.3. การเตรียมข้อมูล (Data Preparation Phase)

3.4. การสร้างแบบจำลอง (Modeling Phase)

3.5. การประเมินผลแบบจำลอง (Evaluation Phase)

3.6. การนำไปใช้งาน (Deployment Phase)

3.1. ทำความเข้าใจปัญหา (Business Understanding Phase)

การลงโฆษณาบรอดแคส จากบัญชีทางการธุรกิจของแอปพลิเคชันไลน์ ครั้งหนึ่งหากต้องการส่งให้ตรงกลุ่มเป้าหมายมากที่สุดจะต้องส่งข้อมูลออกให้ครอบคลุมทุก ๆ กลุ่มลูกค้าเพื่อให้ลูกค้าเหล่านั้นสามารถเห็น และ ได้รับข้อมูลได้ตรงตามกลุ่มเป้าหมายที่ต้องการ แต่การส่งออกข้อมูลให้ครอบคลุมของธุรกิจใหญ่ๆนั้นจะต้องใช้ ทรัพยากรและเวลาจำนวนมากใน การจัดทำโหลดเข้าและส่งออกข้อมูล โดยการโฆษณาบรอดแคสในแต่ละครั้ง ค่าใช้จ่ายในการส่งข้อมูลแบบกระจายให้ครอบคลุมทุก ๆ กลุ่มลุ่มลูกค้าที่มีนั้นจะถูกคิดจากปริมาณในการส่งออกไปเป็นรายบุคคล หากส่งมากๆก็จะจ่ายมาก อีกทั้งรายงานประจำวันที่เก็บสถิติมาใช้เพื่อสำหรับดูสถิติการเข้าร่วมของลูกค้ารายคนจากการโฆษณาบรอดแคสออกไป เท่านั้นซึ่งใช้ประโยชน์จากมันน้อย

ดังนั้น จึงต้องการที่จะนำข้อมูลรายงานประจำวันจากผลการสื่อสารบรอดแคสต์แบบระบุกลุ่มเป้าหมาย ที่มีมาใช้ให้เกิดประโยชน์มากยิ่งขึ้น โดยนำเข้าสู่กระบวนการ และ เปรียบเทียบเทคนิคต่าง ๆ ที่ใช้แบบจำลองในการพยากรณ์ข้อมูล เพื่อนำมาใช้ในการปรับปรุงการคาดการณ์ว่าลูกค้าคนใดบ้างที่จะมีส่วนร่วมในการปล่อยแคมเปญครั้งต่อไป

DPU DHURAKIJ PUNDIT UNIVERSITY มหาวิทยาลัยธุรกิจบัณฑิตย์

3.2. ทำความเข้าใจข้อมูล (Data Understanding Phase)

จากข้อมูลรายงานประจำวันจากผลการสื่อสารบรอดแคสต์แบบระบุกลุ่มเป้าหมาย จากบัญชีทางการธุรกิจของแอปพลิเคชันไลน์ สามารถจัดกลุ่มจัดก้อนของตัวแปร ซึ่งสามารถกำหนดประชากรก็คือ ลูกค้าที่เป็นสมาชิกจากบัญชีทางการธุรกิจของแอปพลิเคชันไลน์ของธุรกิจๆหนึ่ง และ ข้อมูลที่สนใจคือ การบันทึกคำตอบจากการกดคลิกสนใจแคมเปญในครั้งแรกที่ถูกส่งไปถึงลูกค้าที่เป็นสมาชิกคนนั้น ๆ ซึ่งสามารถแบ่งกลุ่มข้อมูลได้สองประเภทดังตารางต่อไปนี้

ตารางที่ 3.1 แบ่งประเภทกลุ่มเป้าหมายออกเป็น segment 5 กลุ่มและมีความหมาย ดังนี้

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

DPU DHURAKIJ PUNDIT UNIVERSITY มหาวิทยาลัยธุรกิจบัณฑิตย์

ตารางที่ 3.2. 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

DPU DHURAKIJ PUNDIT UNIVERSITY มหาวิทยาลัยธุรกิจบัณฑิตย์

ตัวอย่างข้อมูลของรายงานที่นำมาใช้งานมีดังนี้

1. ข้อมูลรายงานประจำวันจากผลการสื่อสารบรอดแคสต์แบบระบุกลุ่มเป้าหมาย จากบัญชีทางการธุรกิจของแอปพลิเคชันไลน์ เริ่มต้น เดือนกรกฎาคม 2563 ถึง ตุลาคม 2563 คิดเป็นเดือนทั้งหมด 3 เดือน และคิดเป็นวันทั้งหมด 92 วัน โดยข้อมูลมีจำนวน 405,538 ข้อมูลตัวอย่าง โดยเก็บเป็นข้อมูลรายวันและรายแคมเปญเป็นไฟล์ Excel เดือน กรกฎาคม 2563 มี 101 ไฟล์ เดือนสิงหาคม 2563 มี 84 ไฟล์ และเดือน กันยายน 2563 มี 122 ไฟล์ รวมทั้งหมด 307 ไฟล์

user_id	[redacted]	stamp	campainge_name	answer	segment
U054df5342a0112a5f30baaa300ad72c3		2020 22:12	30June2020_SM_AF_Travel_NewNormal	#LuxuryTrip	AF
U909606ba76ccf67bc352dea4a0d948c2		2020 19:35	30June2020_SM_AF_Travel_NewNormal	#LuxuryTrip	AF
U7a4cf36be8d9a2cd3a36fa0575ced502		2020 22:25	30June2020_SM_AF_Travel_NewNormal	#NaturalTrip	AF
Uec7f002755bb435700e368853f0b477b		2020 19:26	30June2020_SM_AF_Travel_NewNormal	#LuxuryTrip	AF
U0e9c9b340015c11c306426fe77f05dc7		2020 15:31	30June2020_SM_AF_Travel_NewNormal	#LuxuryTrip	AF
U79fb0160869514268595d87572a55f17		2020 18:44	30June2020_SM_AF_Travel_NewNormal	#LuxuryTrip	AF
U64c47c9a82a7f2abae71af1f51bfd15a		2020 15:36	30June2020_SM_AF_Travel_NewNormal	#LuxuryTrip	AF
U27a8e3f351d0726ad845a184cb639c03		/2020 6:58	30June2020_SM_AF_Travel_NewNormal	#LuxuryTrip	AF
U0332559ebc68238c7552cfa954c1bc54		2020 19:26	30June2020_SM_AF_Travel_NewNormal	#NaturalTrip	AF
Ufa97bc1efbc018b83315bb66d2165ef0		2020 19:33	30June2020_SM_AF_Travel_NewNormal	#LuxuryTrip	AF
Ua4fb275035b0c1824a9e6dfcd8a648d2		2020 18:58	30June2020_SM_AF_Travel_NewNormal	#LuxuryTrip	AF
U56af70f80bea610a6ecc180a627bca60		2020 21:34	30June2020_SM_AF_Travel_NewNormal	#LuxuryTrip	AF
U12185a0008e7f3ed3d76798d1341c67d		2020 18:50	30June2020_SM_AF_Travel_NewNormal	#LuxuryTrip	AF
U89b5e3198f999f8d4fc78318a6fb32b5		2020 18:43	30June2020_SM_AF_Travel_NewNormal	#NaturalTrip	AF
Ud7c17cd51b01c45213e5f23d8480e117		2020 18:41	30June2020_SM_AF_Travel_NewNormal	#NaturalTrip	AF
U7f55483cbd900bc915610001ee9a6574		2020 11:03	30June2020_SM_AF_Travel_NewNormal	#LuxuryTrip	AF
U7b6d3fa06f7ab03baae040319912934a		2020 15:41	30June2020_SM_AF_Travel_NewNormal	#NaturalTrip	AF
U384fc275d0ef1615be075bf678337cc0		2020 20:13	30June2020_SM_AF_Travel_NewNormal	#NaturalTrip	AF
U3d022592d982c888036d19ffbf53a654		2020 16:05	30June2020_SM_AF_Travel_NewNormal	#LuxuryTrip	AF
Uea737b2729578c32f21420d8bdaa0a37		2020 20:27	30June2020_SM_AF_Travel_NewNormal	#LuxuryTrip	AF
U64d4bff0cd7e2ba6a2e256fe200c7198		2020 18:55	30June2020_SM_AF_Travel_NewNormal	#LuxuryTrip	AF
U87f7e0870cfe8e4bc51007c9028a6cc9		/2020 5:55	30June2020_SM_AF_Travel_NewNormal	#NaturalTrip	AF

ภาพที่ 3.1 ตัวอย่างชุดข้อมูล first response จากการ broadcast campaign

โดยที่คุณลักษณะที่สนใจและนำมาใช้ในการดำเนินการมีดังนี้

- User_id : Line Social ID ของลูกค้า
- Line_Name: Line Display ของลูกค้า
- Time stamp : ช่วงเวลาที่กด campaign ครั้งแรก
- Campaing_name: ชื่อแคมเปญ
- Answer : #***ข้อความคำตอบของแคมเปญ
- Segment : segment ของกลุ่มลูกค้าคนนั้น ๆ

DPU DHURAKIJ PUNDIT UNIVERSITY มหาวิทยาลัยธุรกิจบัณฑิตย์

2. ข้อมูลชุดชื่อรายการบรอดแคสต์และคำตอบ ซึ่งรวบรวมชุดรายการบรอดแคสต์และคำตอบที่จะแสดงออกมาเมื่อมีการสนใจ รายการบรอดแคสต์หรือโฆษณานั้น ๆ ทั้งหมด เป็นจำนวนทั้งหมด 400 แคมเปญ

category_ID	campainge_name_sum	answer	category
1	1Siam20%_	#Pro_OneSiamCreditCard	PRODUCTS
1	1Siam20%_	#OneSiamCard_20%	PRODUCTS
1	1Siam20_	#OneSiamCCard_20%	PRODUCTS
4	2ndHalfYearBusiness_	#ภาคธุรกิจครึ่งปีหลัง	INVEST
10	2Psy4BetterSell_	#2หลักจิตวิทยางานขาย	MARKET
10	2Psy4BetterSell_	#2จิตวิทยางานขาย	MARKET
10	2Psy4BetterSell_	#หลักจิตวิทยางานขาย	MARKET
4	3rdLanguageApp_	#3rdLanguageApp	INVEST
4	3Sart_BrandNewProgram	#3ศาสตร์เจาะหุ้นRBF	INVEST
1	3Sart_OutStanding	#OutStanding_PTTGC	PRODUCTS
4	3Sart_STGT	#3ศาสตร์_STGT	INVEST
4	5Industries_HalfYear	#5Industries_1/2Year	INVEST
10	5website_ChinaShipping	#5เว็บฯของจากจีน	MARKET
15	ActionMovieVSAnimation_	#Action	VARIETY
15	ActionMovieVSAnimation_	#Action_	VARIETY
16	Activity_for	#บริษัทจำลอง	WORKER
16	Activity_for	#สายบันเทิง_ดนตรี	WORKER
16	Activity_for	#ออกค่ายอาสา	WORKER
15	App_CreateVDOContent	#แอปทำ VDOContent	VARIETY
16	App_Teenage	#เด็กก็เก็บตังค์ได้	LEARN
15	AppFilmCamera_	#แอปกล้องฟิล์มชิคๆ	VARIETY
1	ApplyforCreditCard_	#สมัครง่าย	PRODUCTS
4	BAMandBankSector_	#BankSectorถูกพอรอได้	INVEST
4	BAMandBankSector_	#หุ้นBAM_Econขาลง	INVEST
4	BondCPALL_	#หุ้นกู้CPALL	INVEST
2	Bonus250_DolfinWallet	#ผูกจายDolfinได้โบ250	PROMOTION

ภาพที่ 3.2 ตัวอย่างชุดข้อมูล campaign category จากการ broadcast campaign

โดยที่คุณลักษณะที่สนใจและนำมาใช้ในการดำเนินการมีดังนี้

- category_ID : ID ของประเภทแคมเปญ
- campainge_name_sum : ชื่อแคมเปญ
- answer : #***ข้อความคำตอบของแคมเปญ
- category : ประเภทของแคมเปญ

DPU DHURAKIJ PUNDIT UNIVERSITY มหาวิทยาลัยธุรกิจบัณฑิตย์

3.3. การเตรียมข้อมูล (Data Preparation Phase)

การเตรียมข้อมูลเป็นขั้นตอนที่ใช้เวลามากที่สุดในขั้นตอนทั้งหมด เนื่องจากเป็นขั้นตอนที่เป็นส่วนสำคัญที่ส่งผลต่อประสิทธิภาพในการพยากรณ์ โดยขั้นตอนนี้จะทำการนำข้อมูลที่รวบรวมทั้งหมด (raw data) มาปรับให้กลายเป็นข้อมูลที่สามารถนำไปวิเคราะห์ต่อในขั้นถัดไปได้ ซึ่งอาจจะต้องมีการดำเนินการทำข้อมูลให้ถูกต้อง พร้อมที่จะใช้สำหรับวิเคราะห์ และ สร้างแบบจำลอง มีขั้นตอนการเตรียมข้อมูลประกอบด้วย ขั้นตอนต่าง ๆ ซึ่งมีหัวข้อย่อยดังนี้

3.3.1. การจัดรูปแบบข้อมูล (Data format)

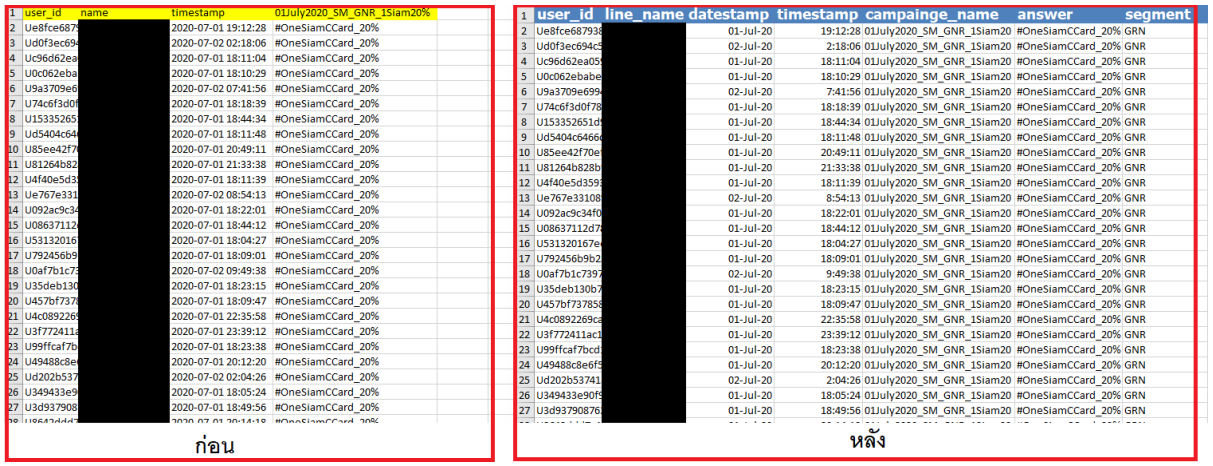
3.3.2. การเก็บรวบรวมข้อมูล (Data Collection)

3.3.3. เตรียมข้อมูลให้พร้อม (Data Preparation)

3.3.4. การแปลงรูปแบบข้อมูล (Data Transformation)

3.3.1. การจัดรูปแบบข้อมูล (Data format) เพื่อให้เป็นรูปแบบเดียวกันก่อนนำเข้าโปรแกรม

จากข้อมูลดิบ (raw data) ของ รายงานประจำวันจากผลการสื่อสารบรอดแคสต์แบบระบุกลุ่มเป้าหมาย จากบัญชีทางการธุรกิจของแอปพลิเคชันไลน์ที่ได้รวบรวมมานั้น มีการบันทึก ชื่อแคมเปญเป็นชื่อของตาราง ดังนั้นจึงจะต้องแก้ไขจัดชื่อของตารางใหม่จาก user_id ,name ,timestamp และ ชื่อของแคมเปญ มาเป็น user_id ,line_name ,datestamp ,timestamp ,campainge_name ,answer และ segment ดังภาพตัวอย่าง



	user_id	name	timestamp	01July2020_SM_GNR_1Siam20%
2	Ue8fce6875	[redacted]	2020-07-01 19:12:28	#OneSiamCCard_20%
3	Ud0f3ec694		2020-07-02 02:18:06	#OneSiamCCard_20%
4	Uc96d62ea		2020-07-01 18:11:04	#OneSiamCCard_20%
5	U0c062eba		2020-07-01 18:10:29	#OneSiamCCard_20%
6	U9a3709e6		2020-07-02 07:41:56	#OneSiamCCard_20%
7	U74c6f3d0f		2020-07-01 18:18:39	#OneSiamCCard_20%
8	U153352651		2020-07-01 18:44:34	#OneSiamCCard_20%
9	Ud5404c646		2020-07-01 18:11:48	#OneSiamCCard_20%
10	U85ee42f70		2020-07-01 20:49:11	#OneSiamCCard_20%
11	U81264b828		2020-07-01 21:33:38	#OneSiamCCard_20%
12	U4f40e5d359		2020-07-01 18:11:39	#OneSiamCCard_20%
13	Ue767e3310		2020-07-02 08:54:13	#OneSiamCCard_20%
14	U092ac9c34		2020-07-01 18:22:01	#OneSiamCCard_20%
15	U08637112		2020-07-01 18:44:12	#OneSiamCCard_20%
16	U5313201		2020-07-01 18:04:27	#OneSiamCCard_20%
17	U792456b9		2020-07-01 18:09:01	#OneSiamCCard_20%
18	U0af7b1c73		2020-07-02 09:49:38	#OneSiamCCard_20%
19	U35deb130		2020-07-01 18:23:15	#OneSiamCCard_20%
20	U457bf7378		2020-07-01 18:09:47	#OneSiamCCard_20%
21	U4c089226		2020-07-01 22:35:58	#OneSiamCCard_20%
22	U3f772411a		2020-07-01 23:39:12	#OneSiamCCard_20%
23	U99ffcaf7b		2020-07-01 18:23:38	#OneSiamCCard_20%
24	U49488c8e		2020-07-01 20:12:20	#OneSiamCCard_20%
25	Ud202b537		2020-07-02 02:04:26	#OneSiamCCard_20%
26	U349433e9		2020-07-01 18:05:24	#OneSiamCCard_20%
27	U3d937908		2020-07-01 18:49:56	#OneSiamCCard_20%

ก่อน

	user_id	line_name	datestamp	timestamp	campainge_name	answer	segment
2	Ue8fce687938	[redacted]	01-Jul-20	19:12:28	01July2020_SM_GNR_1Siam20	#OneSiamCCard_20%	GRN
3	Ud0f3ec694c9		02-Jul-20	2:18:06	01July2020_SM_GNR_1Siam20	#OneSiamCCard_20%	GNR
4	Uc96d62ea05		01-Jul-20	18:11:04	01July2020_SM_GNR_1Siam20	#OneSiamCCard_20%	GNR
5	U0c062ebabe		01-Jul-20	18:10:29	01July2020_SM_GNR_1Siam20	#OneSiamCCard_20%	GNR
6	U9a3709e699		02-Jul-20	7:41:56	01July2020_SM_GNR_1Siam20	#OneSiamCCard_20%	GNR
7	U74c6f3d0f78		01-Jul-20	18:18:39	01July2020_SM_GNR_1Siam20	#OneSiamCCard_20%	GNR
8	U153352651d		01-Jul-20	18:44:34	01July2020_SM_GNR_1Siam20	#OneSiamCCard_20%	GNR
9	Ud5404c6466		01-Jul-20	18:11:48	01July2020_SM_GNR_1Siam20	#OneSiamCCard_20%	GNR
10	U85ee42f70e		01-Jul-20	20:49:11	01July2020_SM_GNR_1Siam20	#OneSiamCCard_20%	GNR
11	U81264b828b		01-Jul-20	21:33:38	01July2020_SM_GNR_1Siam20	#OneSiamCCard_20%	GNR
12	U4f40e5d3593		01-Jul-20	18:11:39	01July2020_SM_GNR_1Siam20	#OneSiamCCard_20%	GNR
13	Ue767e33108		02-Jul-20	8:54:13	01July2020_SM_GNR_1Siam20	#OneSiamCCard_20%	GNR
14	U092ac9c34f0		01-Jul-20	18:22:01	01July2020_SM_GNR_1Siam20	#OneSiamCCard_20%	GNR
15	U08637112d7		01-Jul-20	18:44:12	01July2020_SM_GNR_1Siam20	#OneSiamCCard_20%	GNR
16	U531320167e		01-Jul-20	18:04:27	01July2020_SM_GNR_1Siam20	#OneSiamCCard_20%	GNR
17	U792456b9b2		01-Jul-20	18:09:01	01July2020_SM_GNR_1Siam20	#OneSiamCCard_20%	GNR
18	U0af7b1c7397		02-Jul-20	9:49:38	01July2020_SM_GNR_1Siam20	#OneSiamCCard_20%	GNR
19	U35deb130b7		01-Jul-20	18:23:15	01July2020_SM_GNR_1Siam20	#OneSiamCCard_20%	GNR
20	U457bf737858		01-Jul-20	18:09:47	01July2020_SM_GNR_1Siam20	#OneSiamCCard_20%	GNR
21	U4c0892269ca		01-Jul-20	22:35:58	01July2020_SM_GNR_1Siam20	#OneSiamCCard_20%	GNR
22	U3f772411ac1		01-Jul-20	23:39:12	01July2020_SM_GNR_1Siam20	#OneSiamCCard_20%	GNR
23	U99ffcaf7bcd		01-Jul-20	18:23:38	01July2020_SM_GNR_1Siam20	#OneSiamCCard_20%	GNR
24	U49488c8e6f5		01-Jul-20	20:12:20	01July2020_SM_GNR_1Siam20	#OneSiamCCard_20%	GRN
25	Ud202b53741		02-Jul-20	2:04:26	01July2020_SM_GNR_1Siam20	#OneSiamCCard_20%	GRN
26	U349433e90f9		01-Jul-20	18:05:24	01July2020_SM_GNR_1Siam20	#OneSiamCCard_20%	GNR
27	U3d9379087		01-Jul-20	18:49:56	01July2020_SM_GNR_1Siam20	#OneSiamCCard_20%	GNR

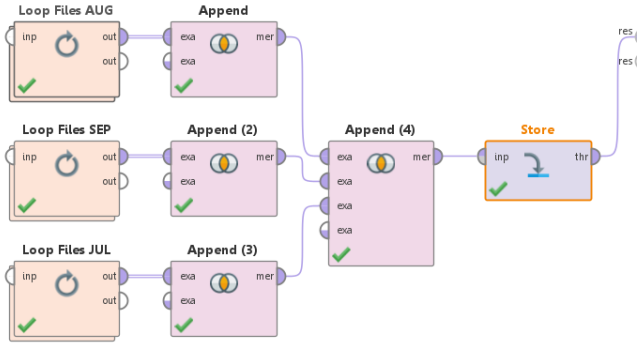
หลัง

ภาพที่ 3.3 ตัวอย่างก่อนและหลังการปรับปรุงชื่อของตารางโดยใช้ Excel Application

DPU DHURAKIJ PUNDIT UNIVERSITY มหาวิทยาลัยธุรกิจบัณฑิตย์

3.3.2. การเก็บรวบรวมข้อมูล (Data Collection) เพื่อรวบรวมไฟล์รายวันให้อยู่ในไฟล์เดียวกัน

เมื่อทำการจัดรูปแบบข้อมูลเรียบร้อยแล้ว จะต้องดำเนินการนำข้อมูลทั้งหมดเข้ากระบวนการ แต่เนื่องจากข้อมูลรายงาน 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



ภาพที่ 3.4 Operation ขั้นตอนการรวมข้อมูลในโปรแกรม RapidMiner

Open in Turbo Prep [redacted] Filter (405,538 / 405,538 examples): all

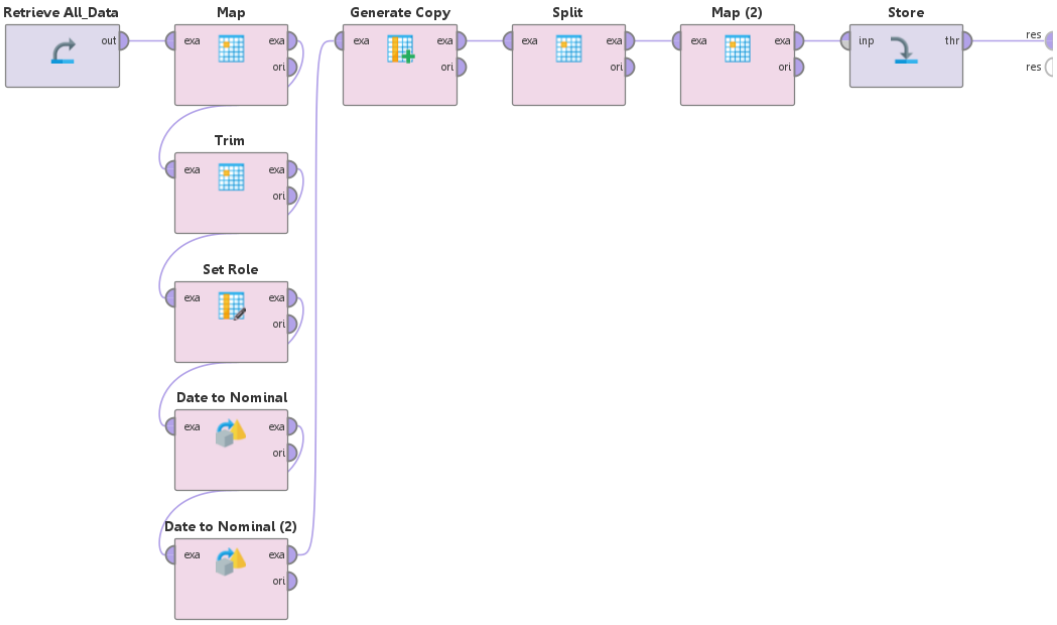
Row No.	user_id	datestamp	timestamp	campainge_...	answer	segment
1	U76d526f62b7	Aug 10, 2020, 1...	Dec 31, 1899, 7...	10August2020_...	#5Industries_1...	AF
2	U08f34e0dc19	Aug 10, 2020, 1...	Dec 31, 1899, 9...	10August2020_...	#5Industries_1...	AF
3	U515a413b54	Aug 10, 2020, 1...	Dec 31, 1899, 6...	10August2020_...	#5Industries_1...	AF
4	Ufc32e897853	Aug 10, 2020, 1...	Dec 31, 1899, 8...	10August2020_...	#5Industries_1...	AF
5	U7db56aaa08	Aug 10, 2020, 1...	Dec 31, 1899, 6...	10August2020_...	#5Industries_1...	AF
6	U47efea1e843	Aug 11, 2020, 1...	Dec 31, 1899, 8...	10August2020_...	#5Industries_1...	AF
7	U88db2d5ee0	Aug 10, 2020, 1...	Dec 31, 1899, 6...	10August2020_...	#5Industries_1...	AF
8	U62ab7111f80	Aug 11, 2020, 1...	Dec 31, 1899, 7...	10August2020_...	#5Industries_1...	AF
9	U310fe05f012	Aug 11, 2020, 1...	Dec 31, 1899, 6...	10August2020_...	#5Industries_1...	AF
10	U49e5f94ca8a	Aug 10, 2020, 1...	Dec 31, 1899, 8...	10August2020_...	#5Industries_1...	AF
11	U51e8dcd573	Aug 10, 2020, 1...	Dec 31, 1899, 6...	10August2020_...	#5Industries_1...	AF
12	U1c7946fc997	Aug 10, 2020, 1...	Dec 31, 1899, 6...	10August2020_...	#5Industries_1...	AF
13	Uf03849e7b1a	Aug 10, 2020, 1...	Dec 31, 1899, 8...	10August2020_...	#5Industries_1...	AF
14	Uf62d509608c	Aug 10, 2020, 1...	Dec 31, 1899, 8...	10August2020_...	#5Industries_1...	AF

ภาพที่ 3.5 ตัวอย่างข้อมูลหลังจากรวบรวมไฟล์ Excel 307 ไฟล์ให้อยู่ในรูปแบบเดียวกัน

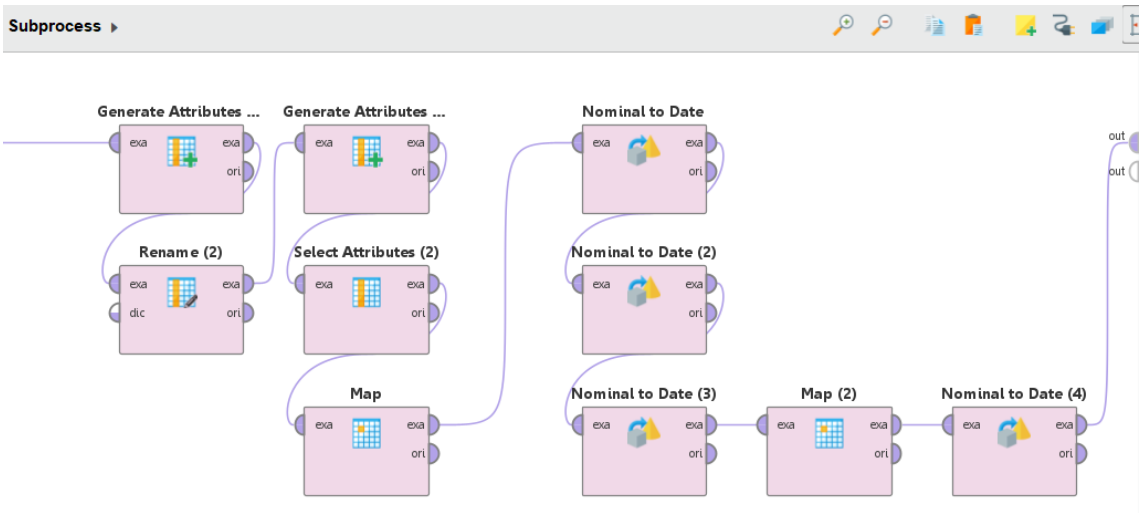
DPU DHURAKIJ PUNDIT UNIVERSITY มหาวิทยาลัยธุรกิจบัณฑิตย์

3.3.3. เตรียมข้อมูลให้พร้อม (Data Preparation) เพื่อปรับปรุงคัดเลือกให้ได้ข้อมูลที่สมบูรณ์ขึ้น

ขั้นตอนนี้จะดำเนินการคัดเลือกข้อมูล ที่ใช้ประโยชน์ได้ ให้เพียงพอและพร้อมที่จะใช้งานให้ได้ตรงกับความต้องการในทันทีในการนำมาวิเคราะห์ และ เกิดความผิดพลาดลดลง



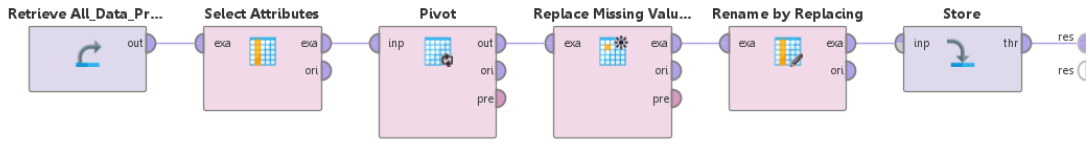
ภาพที่ 3.6 Operation การเตรียมข้อมูลในโปรแกรม RapidMiner



ภาพที่ 3.7 Operation Subprocess การเตรียมข้อมูลในโปรแกรม RapidMiner

DPU DHURAKIJ PUNDIT UNIVERSITY มหาวิทยาลัยธุรกิจบัณฑิตย์

การรวมชุดข้อมูล (Data Aggregate) วัตถุประสงค์เพื่อรวบรวมข้อมูลบางอย่างที่อยู่ใน dataset 2 dataset ให้เป็นข้อมูลเดียว โดยใช้ตารางสองตารางดำเนินดังภาพที่3.9. ตัวอย่างข้อมูลของการรวมชุดข้อมูลในโปรแกรม RapidMiner 1 และ ภาพที่ 3.11. ตัวอย่างข้อมูลของการรวมชุดข้อมูลในโปรแกรม RapidMiner 2 ดำเนินการเพื่อให้ง่ายต่อการวิเคราะห์ข้อมูล โดยรวมจำนวนสถิติการกดสนใจครั้งแรกที่ได้รับแคมเปญออกมา และ ปรับปรุงกลุ่มของข้อมูลให้ถูกต้อง



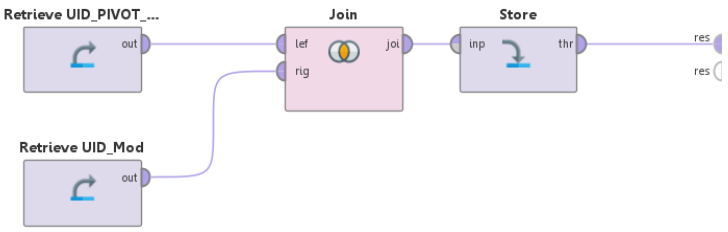
ภาพที่ 3.8 Operation การรวมชุดข้อมูลในโปรแกรม RapidMiner 1

Row No.	user_id	count_answer	mode_Segm...
1	U0000196c3fe...	1	Generic
2	U0000308c106...	1	Generic
3	U0000bcef8f1a...	2	Money loan
4	U00011fd701b...	1	Generic
5	U0001212c23e...	3	Generic
6	U000125b091b...	3	University
7	U000190a2dec...	1	Generic
8	U0001de1fa03...	1	Generic
9	U0001ffa14c92...	1	Generic
10	U0002b1ee5a1...	1	Generic
11	U0002c1fc55b...	1	Generic
12	U000328c9cec...	2	Generic
13	U00036c8cade...	1	Payroll

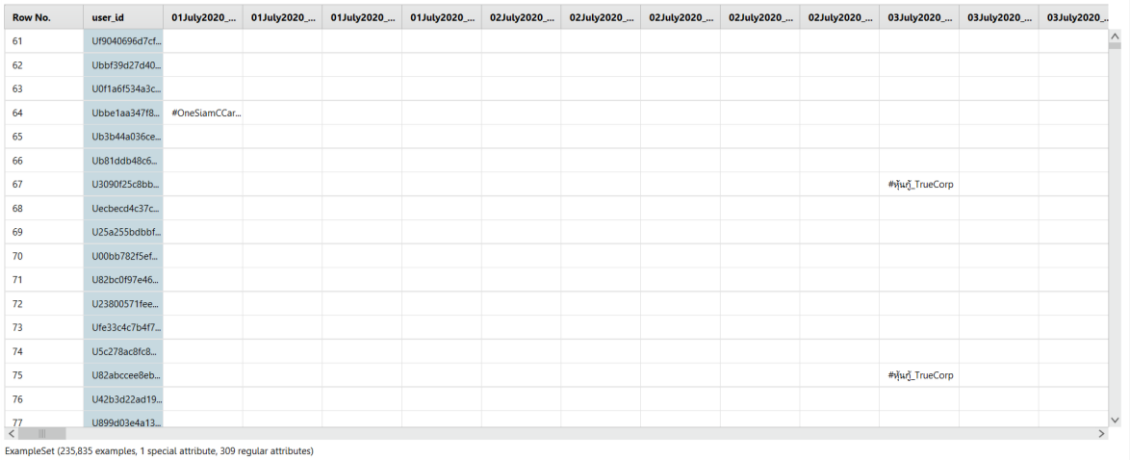
ภาพที่ 3.9 ตัวอย่างข้อมูลของการรวมชุดข้อมูลในโปรแกรม RapidMiner 1

เมื่อได้ข้อมูล Segment จากขั้นตอนข้างต้นออกมาแล้ว ดำเนินการจอยข้อมูลสองตาราง (Join Data) โดยนำตาราง 2 ตารางมารวมกันคือ ตารางข้อมูล Report Daily ที่ Pivot กับ ข้อมูล Segment โดยเงื่อนไขการรวมคือการใช้ Key ของ Segment ที่ตรงกัน

DPU DHURAKIJ PUNDIT UNIVERSITY มหาวิทยาลัยธุรกิจบัณฑิตย์



ภาพที่ 3.10 Operation การรวมชุดข้อมูลในโปรแกรม RapidMiner 2



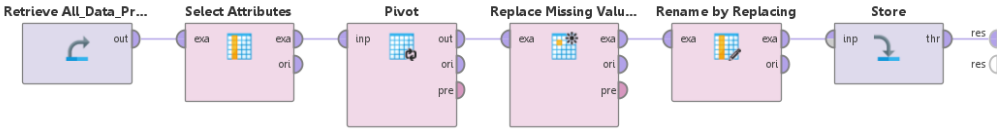
Row No.	user_id	01July2020_...	01July2020_...	01July2020_...	01July2020_...	02July2020_...	02July2020_...	02July2020_...	02July2020_...	02July2020_...	03July2020_...	03July2020_...	03July2020_...
61	Uf9040696d7cf...												
62	Ubbf39d27d40...												
63	U0f1a6f534a3c...												
64	Ubbe1aa347f8...	#OneSiamCCar...											
65	Ub3b44a036ce...												
66	Ub81ddb48c6...												
67	U3090f25c8bb...										#ทุ่นรู้_TrueCorp		
68	Uecbecd4c37c...												
69	U25a255bdbbf...												
70	U00bb782f5ef...												
71	U82bc0f97e46...												
72	U23800571fee...												
73	Ufe33c4c7b4f7...												
74	U5c278ac8fc8...												
75	U82abccee8eb...										#ทุ่นรู้_TrueCorp		
76	U42b3d22ad19...												
77	U899d03e4a13...												

ExampleSet (235,835 examples, 1 special attribute, 309 regular attributes)

ภาพที่ 3.11 ตัวอย่างข้อมูลของการรวมชุดข้อมูลในโปรแกรม RapidMiner 2

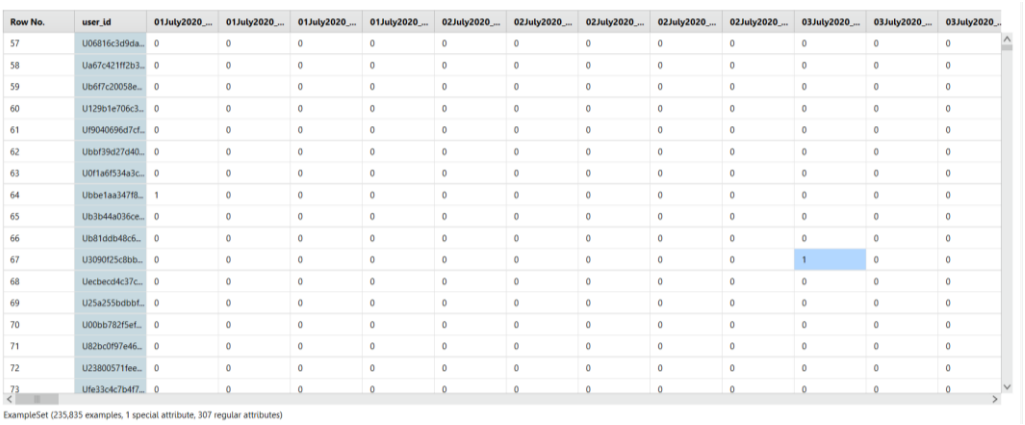
3.3.4. การแปลงรูปแบบข้อมูล (Data Transformation) เพื่อให้ได้รูปแบบข้อมูลที่นำไปใช้งานในแบบจำลองได้

เมื่อแปลงข้อมูลที่ต้องการนำไปวิเคราะห์เรียบร้อยแล้ว จึงทำการรวมและจัดกลุ่มข้อมูลให้เป็นมูลชุดเดียวกัน และนำเข้าแบบจำลองเพื่อทำการพยากรณ์ข้อมูลในขั้นตอนถัดไปโดย Operation Pivot Data แล้วแปลงเป็นข้อมูลตัวเลขหากไม่ได้กดดูแคมเปญจะมีค่าเท่ากับ 0 และ หากกดดูแคมเปญนั้น ๆจะมีค่าเท่ากับ 1 เพื่อใช้สำหรับนำเข้า “Feature Auto Model” เพื่อนำค่า 0กับ1 มารวมเป็นค่าความถี่ของการสนใจในแคมเปญประเภทนั้นๆต่อไป



ภาพที่ 3.12 Operation การแปลงรูปแบบข้อมูลในโปรแกรม RapidMiner

DPU DHURAKIJ PUNDIT UNIVERSITY มหาวิทยาลัยธุรกิจบัณฑิตย์



Row No.	user_id	01July2020_...	01July2020_...	01July2020_...	01July2020_...	02July2020_...	02July2020_...	02July2020_...	02July2020_...	02July2020_...	03July2020_...	03July2020_...	03July2020_...
57	U06816c3d9da...	0	0	0	0	0	0	0	0	0	0	0	0
58	Ua67c421ff2b3...	0	0	0	0	0	0	0	0	0	0	0	0
59	Ub6f7c20058e...	0	0	0	0	0	0	0	0	0	0	0	0
60	U129b1e706c3...	0	0	0	0	0	0	0	0	0	0	0	0
61	Uf5040696d7cf...	0	0	0	0	0	0	0	0	0	0	0	0
62	Ubbf39d27d40...	0	0	0	0	0	0	0	0	0	0	0	0
63	U0f1a6f534a3c...	0	0	0	0	0	0	0	0	0	0	0	0
64	Ubbe1aa347f8...	1	0	0	0	0	0	0	0	0	0	0	0
65	Ub3b44a036ce...	0	0	0	0	0	0	0	0	0	0	0	0
66	Ub81ddb48c6...	0	0	0	0	0	0	0	0	0	0	0	0
67	U3090f25c8bb...	0	0	0	0	0	0	0	0	0	1	0	0
68	Uecbecd4c37c...	0	0	0	0	0	0	0	0	0	0	0	0
69	U25a255bdbbf...	0	0	0	0	0	0	0	0	0	0	0	0
70	U00bb782f5ef...	0	0	0	0	0	0	0	0	0	0	0	0
71	U82bc0f97e46...	0	0	0	0	0	0	0	0	0	0	0	0
72	U23800571fee...	0	0	0	0	0	0	0	0	0	0	0	0
73	Ufe33c4c7b4f7...	0	0	0	0	0	0	0	0	0	0	0	0

ExampleSet (235,835 examples, 1 special attribute, 307 regular attributes)

ภาพที่ 3.13 ตัวอย่างข้อมูลของการแปลงรูปแบบข้อมูลในโปรแกรม RapidMiner

3.4. การสร้างแบบจำลอง (Modeling Phase)

ขั้นตอนนี้เป็นขั้นตอนการวิเคราะห์ข้อมูลด้วยวิธีการและเทคนิคต่าง ๆ โดยนำข้อมูลที่ เตรียมจากขั้นตอนก่อนหน้ามาทำการสร้างแบบจำลอง Propensity Model เพื่อการทำนายเฉพาะแคมเปญที่สำคัญที่สุด และ จากความถี่ของความสนใจของแคมเปญ ที่รวมผลจากค่า 0และ1 ก่อนหน้าออกมา จำแนกแยกประเภท พบว่าเป็นแคมเปญที่มีลูกค้ามีส่วนร่วมมากที่สุดเป็นอันดับหนึ่ง คือ PRODUCTS ซึ่งมีลูกค้าที่สนใจมากที่สุด จำนวน 112,012 คน

ขั้นตอนการสร้างแบบจำลองจะแบ่งข้อมูลออกแบบ Time Frame for Modeling และ สร้างตัวแปรใหม่เพิ่มจากข้อมูลตัวแปรเดิมที่มี (derived variable) แสดงดังตาราง

ตารางที่ 3.3 ตาราง Time Frame for Modeling

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

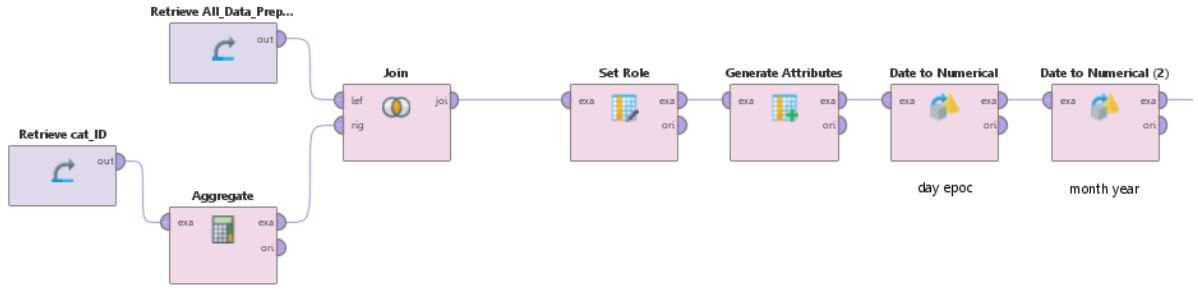
โดยที่ observation period (Observation window) คือ ช่วงเวลาที่ดูพฤติกรรมก่อนการเกิดเหตุการณ์ที่สนใจ

latency period (Buffer) คือ ช่วงเวลาที่ใช้ในการเตรียมข้อมูลหรือเตรียมการต่าง ๆ

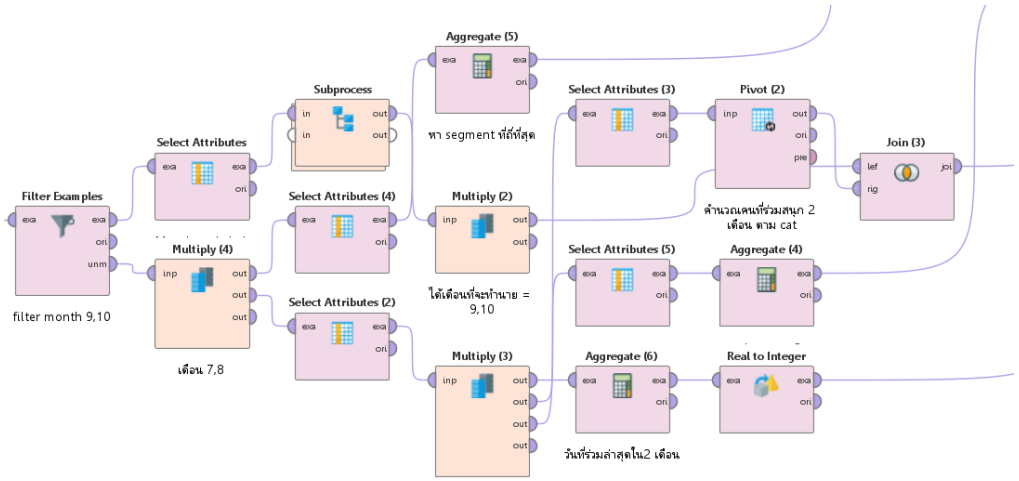
event outcome period (Target window) คือ ช่วงเวลาที่เกิดเหตุการณ์ที่สนใจ

DPU DHURAKIJ PUNDIT UNIVERSITY มหาวิทยาลัยธุรกิจบัณฑิตย์

เมื่อวางแผน Time Frame for Modeling เรียบร้อยแล้วจึงนำข้อมูลมาดำเนินการสร้างแบบจำลอง Propensity โดยใช้ Operation ดำเนินการดังนี้

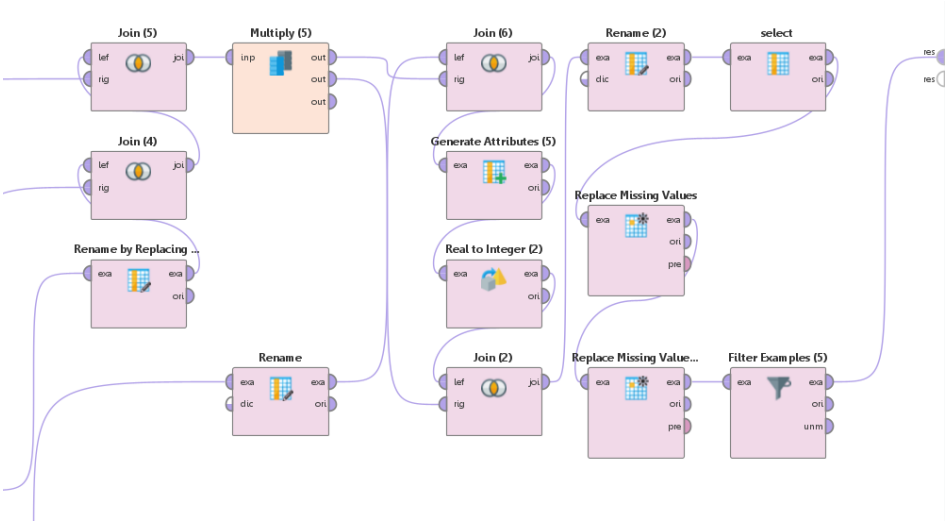


ภาพที่ 3.14 Operation ของแบบจำลอง Propensity ในโปรแกรม RapidMiner 1

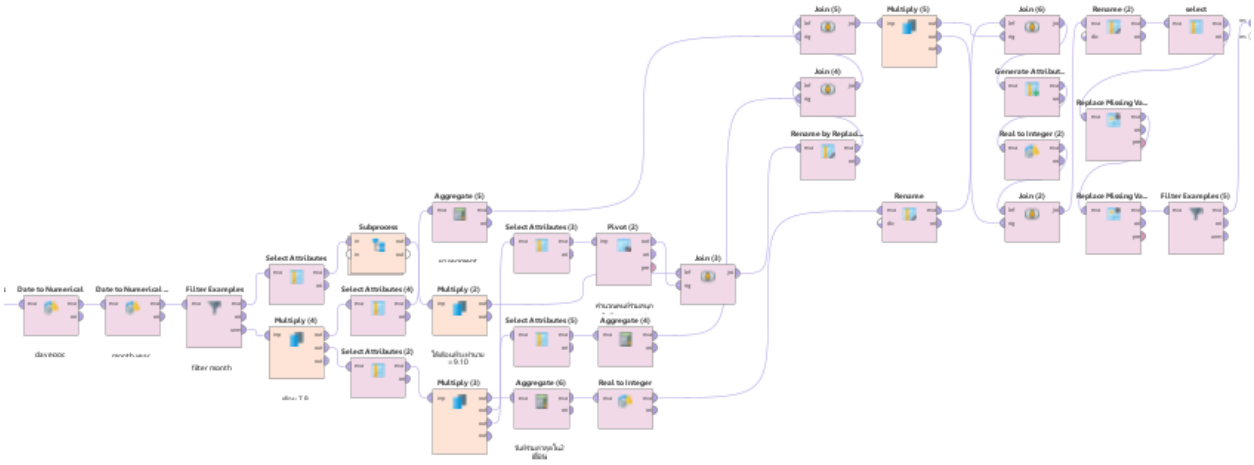


ภาพที่ 3.15 Operation ของแบบจำลอง Propensity ในโปรแกรม RapidMiner 2

DPU DHURAKIJ PUNDIT UNIVERSITY มหาวิทยาลัยธุรกิจบัณฑิตย์



ภาพที่ 3.16 Operation ของแบบจำลอง Propensity ในโปรแกรม RapidMiner 3



ภาพที่ 3.17 Operation ของแบบจำลอง Propensity ภาพรวมทั้งหมดในโปรแกรม RapidMiner

หลังจากนำข้อมูลมาดำเนินการสร้างแบบจำลอง Propensity จะดำเนินสร้างตัวแปรใหม่เพิ่มจากข้อมูลตัวแปรเดิมที่มี (derived variable) โดยจะสร้างตัวแปรใหม่ออกมาดังตารางดังนี้

DPU DHURAKIJ PUNDIT UNIVERSITY มหาวิทยาลัยธุรกิจบัณฑิตย์

ตารางที่ 3.4. ตารางตัวแปรใหม่เพิ่มจากข้อมูลตัวแปรเดิมที่มี

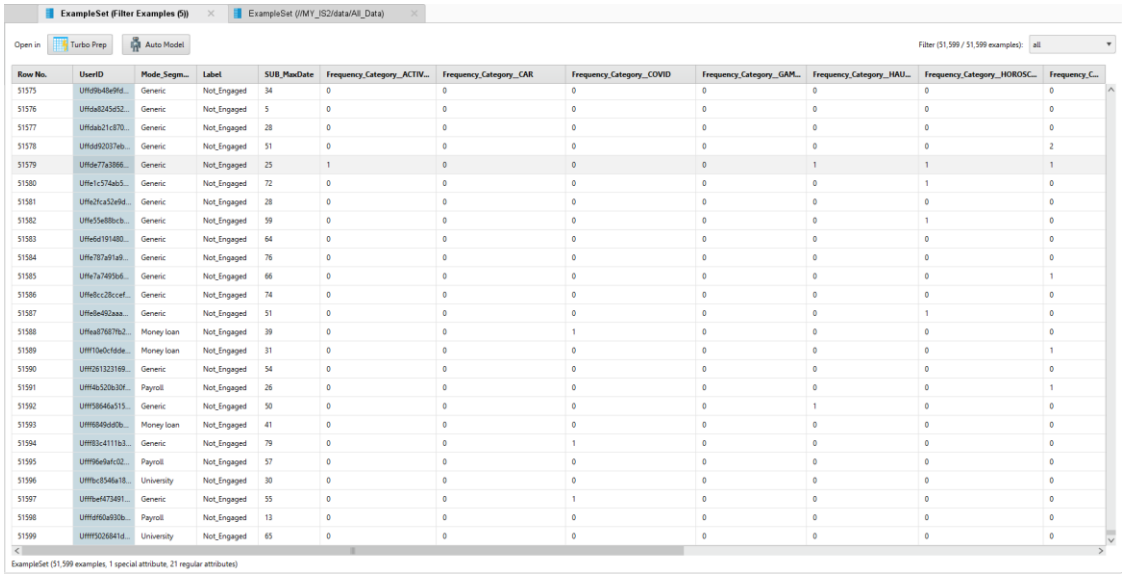
ชื่อตารางหลังเข้า Propensity Process	คำนิยาม
user_id	รหัสของลูกค้า
count(answer)	จำนวนรวมการเข้าร่วมแคมเปญทุกประเภท
Recency_Category_Product	ข้อมูลที่ลูกค้ากดสนใจล่าสุดเมื่อไหร่จากวันสุดท้ายของช่วง Training ของ Category Product เช่น คนที่1 สนใจคมเปญ Product ล่าสุดเมื่อ 60 วันที่แล้ว และ คนที่2 สนใจคมเปญ Product ล่าสุดเมื่อ 10 วันที่แล้ว เป็นต้น ซึ่งตัวแปรที่บอกได้ว่าคนที่ 2 น่าจะมีโอกาส Engage ใหม่มากกว่า(ตอนที่มีการเป็นสมาชิก ของธุรกิจแล้วได้รับ แคมเปญครั้งแรกเราจะเห่อและตื่นเต้น พร้อมที่จะกดสนใจ อะไรเข้ามาก็จะกดสนใจทั้งหมด) สูตรการคำนวณมีดังนี้ Recency_Category_Product = End of Training date - Sub_max_date Sub_max_date = ค่าที่แปลงข้อมูลวันที่ลูกค้าสนใจล่าสุดให้เป็น Integer
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

DPU DHURAKIJ PUNDIT UNIVERSITY มหาวิทยาลัยธุรกิจบัณฑิตย์

ตารางที่ 3.4 (ต่อ)

ชื่อตารางหลังเข้า Propensity Process	คำนิยาม
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

DPU DHURAKIJ PUNDIT UNIVERSITY มหาวิทยาลัยธุรกิจบัณฑิตย์

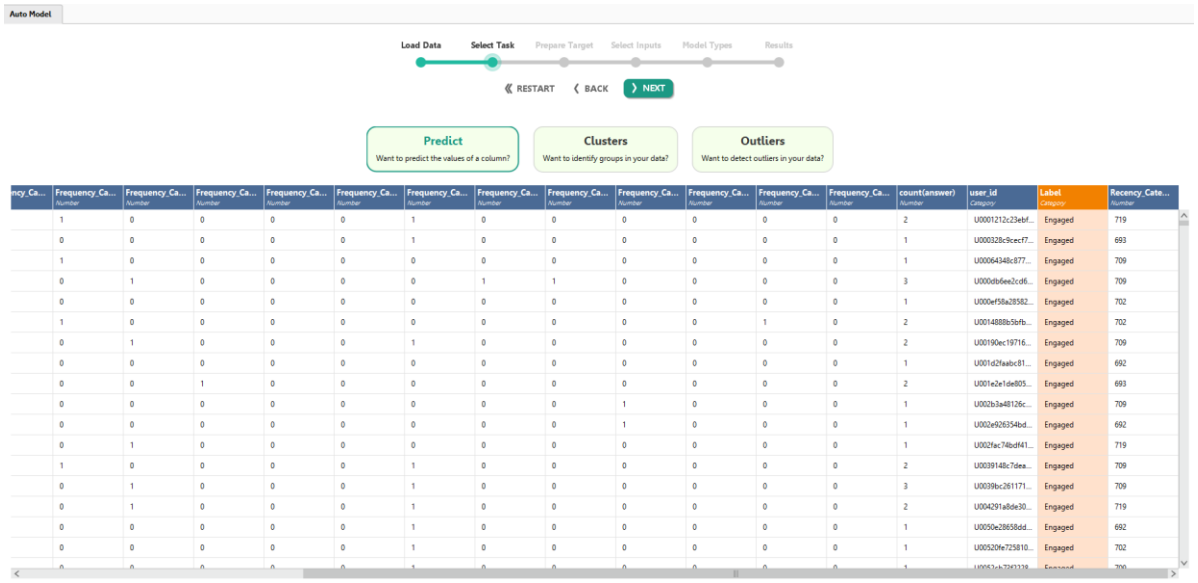


ภาพที่ 3.18 ตัวอย่างข้อมูลหลังดำเนินการสร้างแบบจำลอง Propensity

เทคนิคการเรียนรู้ของเครื่อง (Machine Learning Techniques) จากข้อมูลหลังดำเนินการสร้างแบบจำลอง Propensity โดยข้อมูลทั้งหมดจะถูกนำเข้าเพื่อสร้างแบบจำลองด้วยเทคนิคการเรียนรู้ของเครื่อง (Machine Learning) ซึ่งจะเป็นการใช้ “Feature Auto Model” บน RapidMiner เพื่อเข้ากระบวนการทั้งหมด 9 เทคนิค ซึ่งขั้นตอนต่าง ๆ ในการดำเนินการ “Feature Auto Model” มีรายละเอียดดังต่อไปนี้

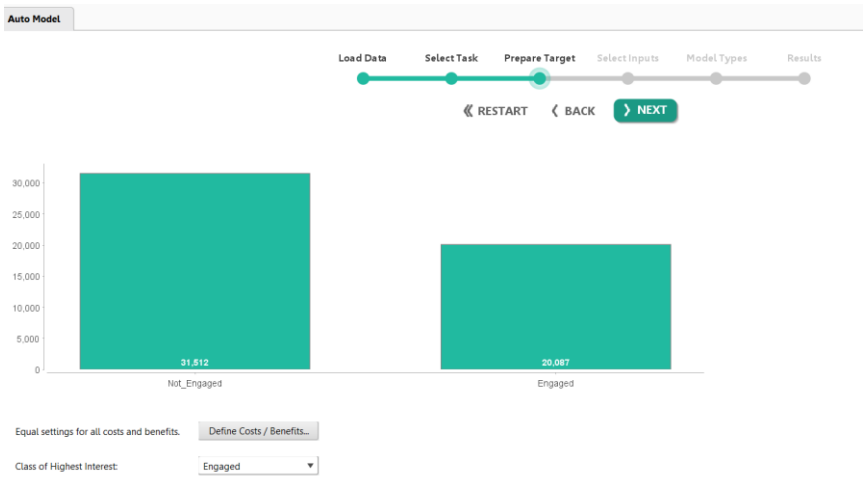
1.เมื่อเริ่มต้นเข้าสู่ “Feature Auto Model” จะต้องเลือกว่าจะให้ “Feature Auto Model” ทำงาน อะไร ซึ่งในงานวิจัยนี้เลือกใช้ “Feature Auto Model” เพื่อเปรียบเทียบประสิทธิภาพแบบจำลองของการพยากรณ์ข้อมูลของเราออกมาโดยจะต้องมี Label หรือ คำตอบที่สนใจ โดยจากภาพ จะเลือก Label ข้อมูลจะเป็น Engage(เข้าร่วมแคมเปญ) และ Not Engage(ไม่เข้าร่วมแคมเปญ)

DPU DHURAKIJ PUNDIT UNIVERSITY มหาวิทยาลัยธุรกิจบัณฑิตย์



ภาพที่ 3.19 “Feature Auto Model” ที่เลือก Label เพื่อดำเนินการพยากรณ์ข้อมูล

2.เมื่อกด Next จะเป็นการ Prepare Target ซึ่ง “Feature Auto Model” จะแสดงจำนวน Label ข้อมูลจะเป็น Engage (เข้าร่วมแคมเปญ) และ Not Engage (ไม่เข้าร่วมแคมเปญ) ที่เราเลือกออกมาว่าเป็นสัดส่วนเท่าไหร่ และต้องการ Class of Highest Interest หรือ คำตอบที่เราสนใจในการดำเนินการมากที่สุด โดยจะเลือก Class Engage(เข้าร่วมแคมเปญ)



ภาพที่ 3.20 ขั้นตอน Prepare Target

DPU DHURAKIJ PUNDIT UNIVERSITY มหาวิทยาลัยธุรกิจบัณฑิตย์

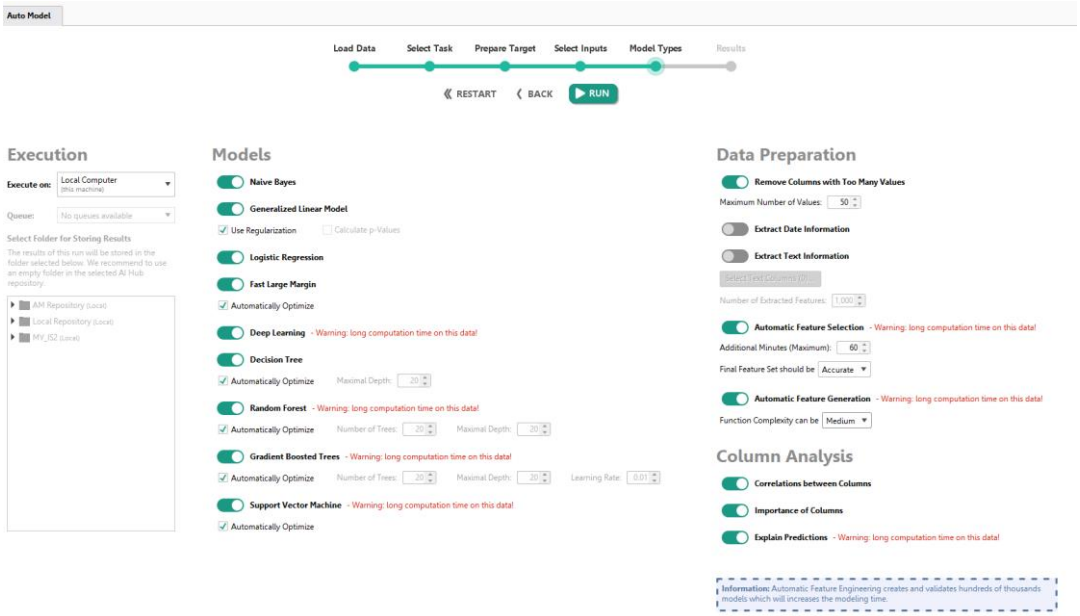
3. ถัดมาเมื่อกด Next จะเป็นการ Select Input ซึ่ง “Feature Auto Model” โดยจะแสดง ตัวแปร Input เพื่อนำเข้าแบบจำลอง ซึ่งได้แจกแจงข้อมูลของแต่ละแปร พร้อมทั้งแนะนำตัวแปรที่ควรนำเข้าแบบจำลอง ในงานวิจัยนี้ได้ใช้ตัวแปรที่ “Feature Auto Model” เลือกให้เพื่อนำเข้าแบบจำลองต่อไปดังภาพ

Selected	Status	Quality	Name	Correlation	ID-ness	Stabillity	Missing	Text-ness
☑	●		SUB_MaxDate	0.11%	0.18%	2.82%	0.00%	0.00%
☑	●		Frequency_Category__HOROSCO...	0.00%	0.01%	81.29%	0.00%	0.00%
☑	●		Frequency_Category__INSURANCE	0.00%	0.01%	85.68%	0.00%	0.00%
☑	●		Frequency_Category__INVEST	0.54%	0.03%	86.99%	0.00%	0.00%
☑	●		Frequency_Category__PRODUCTS	2.09%	0.02%	62.56%	0.00%	0.00%
☑	●		Frequency_Category__PROMOTI...	0.41%	0.02%	80.89%	0.00%	0.00%
☑	●		Frequency_Category__TECH	0.03%	0.01%	89.68%	0.00%	0.00%
☑	●		Frequency_Category__VARIETY	0.13%	0.02%	89.71%	0.00%	0.00%
☑	●		count(answer)	0.08%	0.06%	55.94%	0.00%	0.00%
☑	●		Recency_Category_Product	0.06%	0.06%	11.31%	0.00%	0.00%
☐	●		Frequency_Category__ACTIVITY	0.02%	0.01%	93.09%	0.00%	0.00%
☐	●		Frequency_Category__CAR	0.05%	0.00%	99.74%	0.00%	0.00%
☐	●		Frequency_Category__COVID	0.01%	0.01%	93.14%	0.00%	0.00%
☐	●		Frequency_Category__GAMER	0.01%	0.01%	99.78%	0.00%	0.00%
☐	●		Frequency_Category__HAUNTIN...	0.02%	0.01%	96.58%	0.00%	0.00%
☐	●		Frequency_Category__LEARN	0.13%	0.01%	92.77%	0.00%	0.00%
☐	●		Frequency_Category__MARKET	0.11%	0.01%	92.42%	0.00%	0.00%
☐	●		Frequency_Category__TAX	0.02%	0.01%	92.77%	0.00%	0.00%
☐	●		Frequency_Category__TRAVEL	0.01%	0.01%	94.83%	0.00%	0.00%
☐	●		Frequency_Category__WORKER	0.05%	0.01%	99.57%	0.00%	0.00%
☐	●		user_id	70.98%	100.00%	0.00%	0.00%	48.00%

ภาพที่ 3.21 ขั้นตอน Select Input

4. ขั้นตอนสุดท้ายก่อนจะเข้าสู่ เข้าแบบจำลอง คือการเลือกแบบจำลองที่ต้องการนำมาวัดประสิทธิภาพทั้งหมด ซึ่งจะมีการแนะนำนำแบบจำลองไว้ส่วนหนึ่งเพราะแบบจำลองเหล่านั้นจะเป็นแบบจำลองที่วิเคราะห์ข้อมูลโดยที่ใช้เวลาไม่นานมากนัก แต่งานวิจัยชิ้นนี้เลือกให้ “Feature Auto Model” นำเข้าแบบจำลองทุก ๆ แบบจำลองที่ “Feature Auto Model” มีโดยจะเลือกทั้งหมดแสดงดังภาพ

DPU DHURAKIJ PUNDIT UNIVERSITY มหาวิทยาลัยธุรกิจบัณฑิตย์



ภาพที่ 3.22 ขั้นตอน การเลือกแบบจำลองที่ต้องการนำมาวัดประสิทธิภาพทั้งหมด Model Types

3.5. การประเมินผลแบบจำลอง (Evaluation Phase)

การประเมินผลแบบจำลองจะใช้การวัดประสิทธิภาพแบบจำลองจาก Confusion Matrix ซึ่งประสิทธิภาพของผลลัพธ์ว่ามีความน่าเชื่อถือมากน้อยเพียงใด โดยคำนวณค่าทั้ง 4 แบบคำนวณดังนี้

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

จากนั้นใช้แบบจำลองที่สร้างได้จากข้อมูลที่ให้แบบจำลองเรียนรู้ทั้ง 9 แบบจำลองมาทำการพยากรณ์ข้อมูล เพื่อทำการทดสอบประสิทธิภาพ ซึ่งอยู่ในข้อมูลทดสอบ และ เปรียบเทียบการคำนวณค่าทั้ง 4 แบบคำนวณ ของเทคนิคทั้ง 9 ที่แสดงออกมาจาก “Feature Auto Model” ว่าแบบจำลองเรียนรู้ได้ดีที่สุด แล้วจึงเลือกบันทึก จัดเก็บ และเสนอ แบบจำลองที่ดีที่สุด

DPU DHURAKIJ PUNDIT UNIVERSITY มหาวิทยาลัยธุรกิจบัณฑิตย์

3.6. การนำไปใช้งาน (Deployment Phase)

นำผลลัพธ์ของการเปรียบเทียบประสิทธิภาพ แสดงผลบน Dashboard เพื่อนำแบบจำลองที่ดีที่สุดนั้นนำมาใช้ประโยชน์ โดยหลังจากที่วิจัย และ เปรียบเทียบค่าความคลาดเคลื่อนของแต่ละเทคนิคแล้ว แบบจำลองใดมีประสิทธิภาพมากที่สุด และ มีความเหมาะสมกับความต้องการของการพยากรณ์ข้อมูลในอนาคตจะ ถูกแนะนำให้นำมาประยุกต์ใช้ โดยนำเทคนิคที่ดีที่สุดที่สามารถนำข้อมูลเข้าและใช้วิเคราะห์ในการคาดการณ์ว่าลูกค้าคนใดบ้างที่จะมีส่วนร่วมในการปล่อยแคมเปญครั้งต่อไป ในอนาคต

DPU DHURAKIJ PUNDIT UNIVERSITY มหาวิทยาลัยธุรกิจบัณฑิตย์

บทที่ 4
ผลการวิจัย

การศึกษาวิจัยเพื่อเปรียบเทียบประสิทธิภาพของวิธีการพยากรณ์ของเทคนิคที่มีประสิทธิภาพที่ดีที่สุดในครั้งนี้ ได้แบ่งผลประสิทธิภาพและความถูกต้องออกเป็น 2 ส่วนคือ ผลการเปรียบเทียบประสิทธิภาพความถูกต้องของแบบจำลอง และ ผลความถูกต้องจากการนำแบบจำลองพยากรณ์ในอนาคตเทียบกับข้อมูลจริง ซึ่งแสดงผลการเปรียบเทียบประสิทธิภาพ และ ผลลัพธ์ของการทำวิจัย มีหัวข้อดังนี้

4.1. ผลการทำนาย และ ค่าน้ำหนักของแอตทริบิวต์

4.2. ผลการเปรียบเทียบประสิทธิภาพการทำนายผล

4.3. ผลความถูกต้องจากการนำแบบจำลองพยากรณ์ในอนาคตเทียบกับข้อมูลจริง

4.4. ผลของการแสดงผลเพื่อนำเสนอ

4.1 ผลการทำนาย และ ค่าน้ำหนักของแอตทริบิวต์

4.1.1. ผลการทำนาย และ ค่าน้ำหนักของแอตทริบิวต์ ด้วยวิธีการจำแนกประเภทข้อมูลแบบเบย์อย่างง่าย (Naive Bayes)

จาก “Feature Auto Model” ด้วยวิธีการจำแนกประเภทข้อมูลแบบเบย์อย่างง่าย (Naive Bayes) ผลในส่วนของตาราง Confusion Matrix และ ค่าน้ำหนักของแอตทริบิวต์ ได้ค่า Class Recall และ Class Precision ใช้ค่าพารามิเตอร์ที่กล่าวไว้แสดงได้ดังนี้

Naive Bayes - Weights

Attribute	Weight
Frequency_Category__INVEST	0.111
Frequency_Category__VARIETY	0.075
Frequency_Category__PRODUCTS	0.062
Frequency_Category__PROMOTION	0.059
count(answer)	0.044
Mode_Segment_byUID	0.029
SUB_MaxDate	0.025
Frequency_Category__TECH	0.025
Frequency_Category__INSURANCE	0.019
Frequency_Category__HOROSCOPE	0.014

ภาพที่ 4.1 ค่าน้ำหนักของแอตทริบิวต์ ของวิธีการจำแนกประเภทข้อมูลแบบเบย์อย่างง่าย (Naive Bayes)

DPU DHURAKIJ PUNDIT UNIVERSITY มหาวิทยาลัยธุรกิจบัณฑิตย์

ตารางที่ 4.1 ผลในส่วนของตาราง Confusion Matrix ของวิธีการพยากรณ์ของเทคนิคการจำแนกประเภทข้อมูลแบบเบย์อย่างง่าย (Naive Bayes)

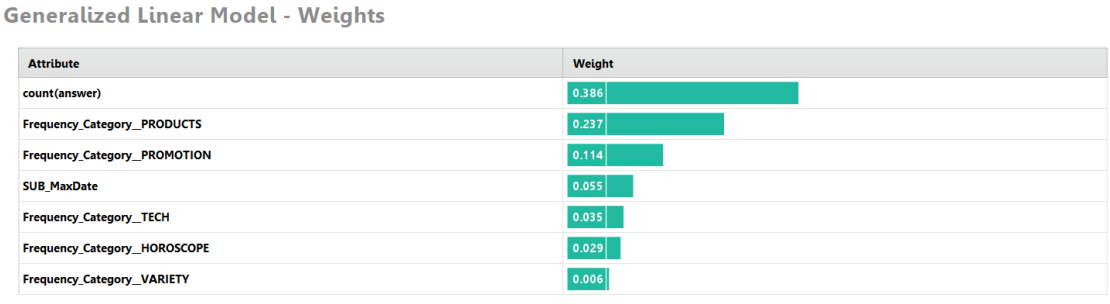
	true Not Engaged	true Engaged	class precision
pred. Not Engaged	6,784	3,309	67.21%
pred. Engaged	2,219	2,430	52.27%
class recall	75.35%	42.34%	

สรุปค่า class precision ของแบบจำลองวิธีการพยากรณ์ของเทคนิคการจำแนกประเภทข้อมูลแบบเบย์อย่างง่าย (Naive Bayes) โดย Not Engaged(ไม่เข้าร่วมแคมเปญ)จริงๆกลับมา ได้ร้อยละ 67.21 และ ตอบ Engaged(เข้าร่วมแคมเปญ) จริงๆกลับมา ได้ร้อยละ 52.27

สรุปค่า class recall ของแบบจำลองวิธีการพยากรณ์ของเทคนิคการจำแนกประเภทข้อมูลแบบเบย์อย่างง่าย (Naive Bayes) คำนวณเทียบกับคำตอบ Not Engaged(ไม่เข้าร่วมแคมเปญ)ข้างเคียง ได้ร้อยละ 75.35 และ คำนวณเทียบกับคำตอบ Engaged(เข้าร่วมแคมเปญ)ข้างเคียง ได้ร้อยละ 42.34

4.1.2. ผลการทำนาย และ ค่าน้ำหนักของแอตทริบิวต์ ด้วยวิธีตัวแบบเชิงเส้นนัยทั่วไป (Generalized Linear Model)

จาก “Feature Auto Model” ด้วยวิธีตัวแบบเชิงเส้นนัยทั่วไป (Generalized Linear Model) ผลในส่วนของตาราง Confusion Matrix และ ค่าน้ำหนักของแอตทริบิวต์ ได้ค่า Class Recall และ Class Precision ใช้ค่าพารามิเตอร์ที่กล่าวไว้แสดงได้ดังนี้



ภาพที่ 4.2 ค่าน้ำหนักของแอตทริบิวต์ ของวิธีตัวแบบเชิงเส้นนัยทั่วไป (Generalized Linear Model)

DPU DHURAKIJ PUNDIT UNIVERSITY มหาวิทยาลัยธุรกิจบัณฑิตย์

ตารางที่ 4.2 ผลในส่วนของตาราง Confusion Matrix ของวิธีการพยากรณ์ของเทคนิคตัวแบบเชิงเส้นนัยทั่วไป (Generalized Linear Model)

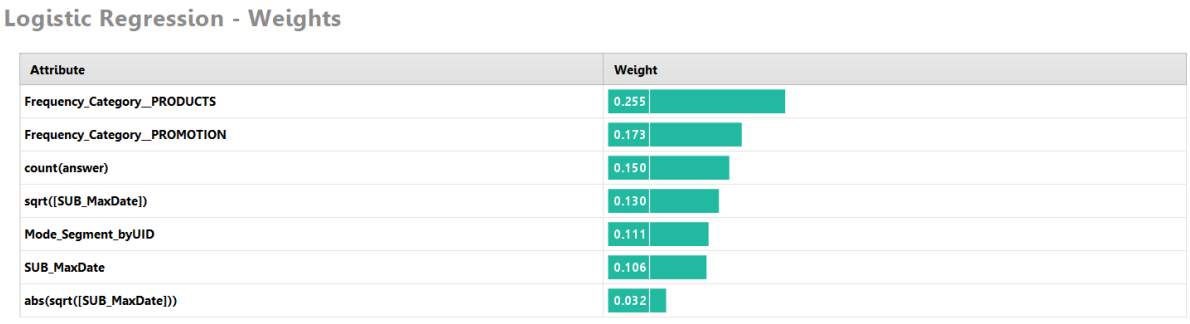
	true Not_Engaged	true Engaged	class precision
pred. Not_Engaged	7,733	3,366	69.67%
pred. Engaged	1,270	2,374	65.15%
class recall	85.89%	41.36%	

สรุปค่า class precision ของแบบจำลองวิธีการพยากรณ์ของเทคนิคตัวแบบเชิงเส้นนัยทั่วไป (Generalized Linear Model) โดย Not Engaged(ไม่เข้าร่วมแคมเปญ)จริงๆกลับมา ได้ร้อยละ 69.67 และตอบ Engaged(เข้าร่วมแคมเปญ)จริงๆกลับมา ได้ร้อยละ 65.15

สรุปค่า class recall ของแบบจำลองวิธีการพยากรณ์ของเทคนิคตัวแบบเชิงเส้นนัยทั่วไป (Generalized Linear Model) คำนวณเทียบกับคำตอบ Not Engaged(ไม่เข้าร่วมแคมเปญ)ข้างเคียง ได้ร้อยละ 85.89และ คำนวณเทียบกับคำตอบ Engaged(เข้าร่วมแคมเปญ)ข้างเคียง ได้ร้อยละ 41.36

4.1.3. ผลการทำนาย และ ค่าน้ำหนักของแอตทริบิวต์ ด้วยวิธีการถดถอยโลจีสติก (Logistic Regression)

จาก “Feature Auto Model” ด้วยวิธีการถดถอยโลจีสติก (Logistic Regression) ผลในส่วนของตาราง Confusion Matrix และ ค่าน้ำหนักของแอตทริบิวต์ ได้ค่า Class Recall และ Class Precision ใช้ค่าพารามิเตอร์ที่กล่าวไว้แสดงได้ดังนี้


Logistic Regression - Weights

Attribute	Weight
Frequency_Category_PRODUCTS	0.255
Frequency_Category_PROMOTION	0.173
count(answer)	0.150
sqrt([SUB_MaxDate])	0.130
Mode_Segment_byUID	0.111
SUB_MaxDate	0.106
abs(sqrt([SUB_MaxDate]))	0.032

ภาพที่ 4.3 ค่าน้ำหนักของแอตทริบิวต์ ของวิธีการพยากรณ์ของเทคนิคการถดถอยโลจีสติก (Logistic Regression)

DPU DHURAKIJ PUNDIT UNIVERSITY มหาวิทยาลัยธุรกิจบัณฑิตย์

ตารางที่ 4.3 ผลในส่วนของตาราง Confusion Matrix ของวิธีการพยากรณ์ของเทคนิคการถดถอยโลจีสติก (Logistic Regression)

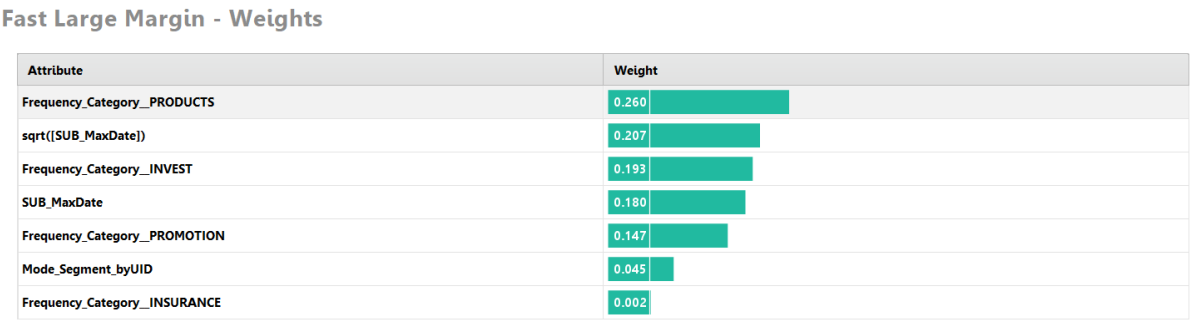
	true Not_Engaged	true Engaged	class precision
pred. Not_Engaged	7,750	3,648	67.99%
pred. Engaged	1,254	2,092	62.52%
class recall	86.07%	36.45%	

สรุปค่า class precision ของแบบจำลองวิธีการพยากรณ์ของเทคนิคการถดถอยโลจีสติก (Logistic Regression) โดย Not Engaged(ไม่เข้าร่วมแคมเปญ)จริงๆกลับมา ได้ร้อยละ 67.99 และ ตอบ Engaged(เข้าร่วมแคมเปญ)จริงๆกลับมา ได้ร้อยละ 62.52

สรุปค่า class recall ของแบบจำลองวิธีการพยากรณ์ของเทคนิคการถดถอยโลจีสติก (Logistic Regression) คำนวณเทียบกับคำตอบ Not Engaged(ไม่เข้าร่วมแคมเปญ)ข้างเคียง ได้ร้อยละ 86.07 และ คำนวณเทียบกับคำตอบ Engaged(เข้าร่วมแคมเปญ)ข้างเคียง ได้ร้อยละ 36.45

4.1.4. ผลการทำนาย และ ค่าน้ำหนักของแอตทริบิวต์ ด้วยวิธีจำแนกประเภทโดยใช้ขอบเขตที่มีขนาดกว้างที่สุดแบบเร็ว (Fast Large Margin (RapidMiner Studio Core)

จาก “Feature Auto Model” ด้วยวิธีจำแนกประเภทโดยใช้ขอบเขตที่มีขนาดกว้างที่สุดแบบเร็ว (Fast Large Margin (RapidMiner Studio Core)) ผลในส่วนของตาราง Confusion Matrix และ ค่าน้ำหนักของแอตทริบิวต์ ได้ค่า Class Recall และ Class Precision ใช้ค่าพารามิเตอร์ที่กล่าวไว้แสดงได้ดังนี้



ภาพที่ 4.4 ค่าน้ำหนักของแอตทริบิวต์ ของวิธีการพยากรณ์ของเทคนิคจำแนกประเภทโดยใช้ขอบเขตที่มีขนาดกว้างที่สุดแบบเร็ว (Fast Large Margin (RapidMiner Studio Core))

DPU DHURAKIJ PUNDIT UNIVERSITY มหาวิทยาลัยธุรกิจบัณฑิตย์

ตารางที่ 4.4 ผลในส่วนของตาราง Confusion Matrix ของวิธีการพยากรณ์ของเทคนิคจำแนกประเภทโดยใช้ขอบเขตที่มีขนาดกว้างที่สุดแบบเร็ว (Fast Large Margin (RapidMiner Studio Core))

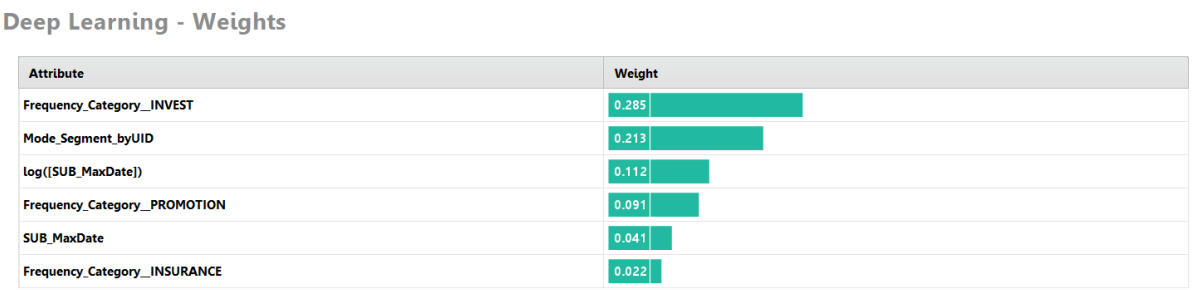
	true Not_Engaged	true Engaged	class precision
pred. Not_Engaged	5,311	2,450	68.43%
pred. Engaged	3,693	3,290	47.11%
class recall	58.98%	57.32%	

สรุปค่า class precision ของแบบจำลองวิธีการพยากรณ์ของเทคนิคจำแนกประเภทโดยใช้ขอบเขตที่มีขนาดกว้างที่สุดแบบเร็ว (Fast Large Margin (RapidMiner Studio Core)) โดย Not Engaged(ไม่เข้าร่วมแคมเปญ)จริงๆกลับมา ได้ร้อยละ 68.43 และ ตอบ Engaged(เข้าร่วมแคมเปญ)จริงๆกลับมา ได้ร้อยละ 47.11

สรุปค่า class recall ของแบบจำลองวิธีการพยากรณ์ของเทคนิคจำแนกประเภทโดยใช้ขอบเขตที่มีขนาดกว้างที่สุดแบบเร็ว (Fast Large Margin (RapidMiner Studio Core)) คำนวณเทียบกับคำตอบ Not Engaged(ไม่เข้าร่วมแคมเปญ)ข้างเคียงได้ร้อยละ 58.98 และ คำนวณเทียบกับคำตอบ Engaged(เข้าร่วมแคมเปญ)ข้างเคียง ได้ร้อยละ 57.32

4.1.5. ผลการทำนาย และ ค่าน้ำหนักของแอตทริบิวต์ ด้วยวิธีการเรียนรู้เชิงลึก (Deep Learning)

จาก “Feature Auto Model” ด้วยวิธีการเรียนรู้เชิงลึก (Deep Learning) ผลในส่วนของตาราง Confusion Matrix และ ค่าน้ำหนักของแอตทริบิวต์ ได้ค่า Class Recall และ Class Precision ใช้ค่าพารามิเตอร์ที่กล่าวไว้แสดงได้ดังนี้



ภาพที่ 4.5 ค่าน้ำหนักของแอตทริบิวต์ ของวิธีการพยากรณ์ของเทคนิคการเรียนรู้เชิงลึก (Deep Learning)

DPU DHURAKIJ PUNDIT UNIVERSITY มหาวิทยาลัยธุรกิจบัณฑิตย์

ตารางที่ 4.5 ผลในส่วนของตาราง Confusion Matrix ของวิธีการพยากรณ์ของเทคนิคการเรียนรู้เชิงลึก (Deep Learning)

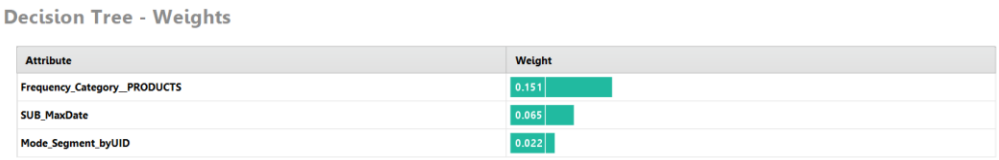
	true Not_Engaged	true Engaged	class precision
pred. Not_Engaged	8,940	5,103	63.66%
pred. Engaged	63	636	90.99%
class recall	99.30%	11.08%	

สรุปค่า class precision ของแบบจำลองวิธีการพยากรณ์ของเทคนิคการเรียนรู้เชิงลึก (Deep Learning) โดย Not Engaged(ไม่เข้าร่วมแคมเปญ)จริงๆกลับมา ได้ร้อยละ 63.66 และ ตอบ Engaged(เข้าร่วมแคมเปญ)จริงๆกลับมา ได้ร้อยละ 90.99

สรุปค่า class recall ของแบบจำลองวิธีการพยากรณ์ของเทคนิคการเรียนรู้เชิงลึก (Deep Learning)คำนวณเทียบกับคำตอบ Not Engaged(ไม่เข้าร่วมแคมเปญ)ข้างเคียง ได้ร้อยละ 99.30 และคำนวณเทียบกับคำตอบ Engaged(เข้าร่วมแคมเปญ)ข้างเคียง ได้ร้อยละ 11.08

4.1.6. ผลการทำนาย และ ค่าน้ำหนักของแอตทริบิวต์ ด้วยวิธีต้นไม้การตัดสินใจ (Decision Tree)

จาก “Feature Auto Model” ด้วยวิธีต้นไม้การตัดสินใจ (Decision Tree) ผลในส่วนของตาราง Confusion Matrix และ ค่าน้ำหนักของแอตทริบิวต์ ได้ค่า Class Recall และ Class Precision ใช้ค่าพารามิเตอร์ที่กล่าวไว้แสดงได้ดังนี้



ภาพที่ 4.6 ค่าน้ำหนักของแอตทริบิวต์ ของวิธีการพยากรณ์ของเทคนิคต้นไม้การตัดสินใจ (Decision Tree)

DPU DHURAKIJ PUNDIT UNIVERSITY มหาวิทยาลัยธุรกิจบัณฑิตย์

ตารางที่ 4.6 ผลในส่วนของตาราง Confusion Matrix ของวิธีการพยากรณ์ของเทคนิคต้นไม้การตัดสินใจ (Decision Tree)

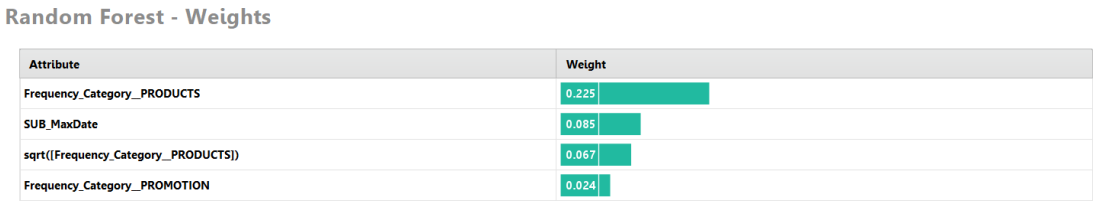
	true Not_Engaged	true Engaged	class precision
pred. Not_Engaged	8,773	3,071	74.07%
pred. Engaged	231	2,669	92.03%
class recall	97.43%	46.50%	

สรุปค่า class precision ของแบบจำลองวิธีการพยากรณ์ของเทคนิคต้นไม้การตัดสินใจ (Decision Tree) โดย Not Engaged(ไม่เข้าร่วมแคมเปญ)จริงๆกลับมา ได้ร้อยละ 74.07 และ ตอบ Engaged(เข้าร่วมแคมเปญ)จริงๆกลับมา ได้ร้อยละ 92.03

สรุปค่า class recall ของแบบจำลองวิธีการพยากรณ์ของเทคนิคต้นไม้การตัดสินใจ (Decision Tree) คำนวณเทียบกับคำตอบ Not Engaged(ไม่เข้าร่วมแคมเปญ)ข้างเคียง ได้ร้อยละ 97.43 และ คำนวณเทียบกับคำตอบ Engaged(เข้าร่วมแคมเปญ)ข้างเคียง ได้ร้อยละ 46.50

4.1.7. ผลการทำนาย และ ค่าน้ำหนักของแอตทริบิวต์ ด้วยวิธีแบบจำลองป่าสุ่ม (Random Forest)

จาก “Feature Auto Model” ด้วยวิธีแบบจำลองป่าสุ่ม (Random Forest) ผลในส่วนของตาราง Confusion Matrix และ ค่าน้ำหนักของแอตทริบิวต์ ได้ค่า Class Recall และ Class Precision ใช้ค่าพารามิเตอร์ที่กล่าวไว้แสดงได้ดังนี้



ภาพที่ 4.7 ค่าน้ำหนักของแอตทริบิวต์ ของวิธีการพยากรณ์ของเทคนิคแบบจำลองป่าสุ่ม (Random Forest)

DPU DHURAKIJ PUNDIT UNIVERSITY มหาวิทยาลัยธุรกิจบัณฑิตย์

ตารางที่ 4.7 ผลในส่วนของตาราง Confusion Matrix ของวิธีการพยากรณ์ของเทคนิคแบบจำลองป่าสุ่ม (Random Forest)

	true Not_Engaged	true Engaged	class precision
pred. Not_Engaged	8,574	1,659	83.79%
pred. Engaged	431	4,080	90.45%
class recall	95.21%	71.09%	

สรุปค่า class precision ของแบบจำลองวิธีการพยากรณ์ของเทคนิคแบบจำลองป่าสุ่ม (Random Forest) โดย Not Engaged(ไม่เข้าร่วมแคมเปญ)จริงๆกลับมา ได้ร้อยละ 83.79 และ ตอบ Engaged(เข้าร่วมแคมเปญ)จริงๆกลับมา ได้ร้อยละ 90.45

สรุปค่า class recall ของแบบจำลองวิธีการพยากรณ์ของเทคนิคแบบจำลองป่าสุ่ม (Random Forest) คำนวณเทียบกับคำตอบ Not Engaged(ไม่เข้าร่วมแคมเปญ)ข้างเคียง ได้ร้อยละ 95.21 และ คำนวณเทียบกับคำตอบ Engaged(เข้าร่วมแคมเปญ)ข้างเคียง ได้ร้อยละ 71.09

4.1.8. ผลการทำนาย และ ค่าน้ำหนักของแอตทริบิวต์ ด้วยวิธีต้นไม้การเพิ่มไล่ระดับการตัดสินใจ (Gradient Boosted Trees)

จาก “Feature Auto Model” ด้วยวิธีต้นไม้การเพิ่มไล่ระดับการตัดสินใจ (Gradient Boosted Trees) ผลในส่วนของตาราง Confusion Matrix และ ค่าน้ำหนักของแอตทริบิวต์ ได้ค่า Class Recall และ Class Precision ใช้ค่าพารามิเตอร์ที่กล่าวไว้แสดงได้ดังนี้

ภาพที่ 4.8 ค่าน้ำหนักของแอตทริบิวต์ ของวิธีการพยากรณ์ของเทคนิคต้นไม้การเพิ่มไล่ระดับการตัดสินใจ (Gradient Boosted Trees)

DPU DHURAKIJ PUNDIT UNIVERSITY มหาวิทยาลัยธุรกิจบัณฑิตย์

ตารางที่ 4.8 ผลในส่วนของตาราง Confusion Matrix ของวิธีการพยากรณ์ของเทคนิคต้นไม้การเพิ่มไล่ระดับการตัดสินใจ (Gradient Boosted Trees)

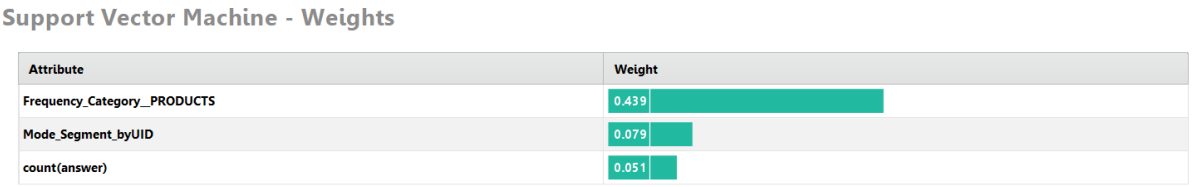
	true Not_Engaged	true Engaged	class precision
pred. Not_Engaged	8,591	1,134	88.34%
pred. Engaged	413	4,605	91.77%
class recall	95.41%	80.24%	

สรุปค่า class precision ของแบบจำลองวิธีการพยากรณ์ของเทคนิคต้นไม้การเพิ่มไล่ระดับการตัดสินใจ (Gradient Boosted Trees) โดย Not Engaged(ไม่เข้าร่วมแคมเปญ)จริงๆกลับมา ได้ร้อยละ 88.34 และ ตอบ Engaged(เข้าร่วมแคมเปญ)จริงๆกลับมา ได้ร้อยละ 91.77

สรุปค่า class recall ของแบบจำลองวิธีการพยากรณ์ของเทคนิคต้นไม้การเพิ่มไล่ระดับการตัดสินใจ (Gradient Boosted Trees) คำนวณเทียบกับคำตอบ Not Engaged(ไม่เข้าร่วมแคมเปญ)ข้างเคียง ได้ร้อยละ 95.41 และ คำนวณเทียบกับคำตอบ Engaged(เข้าร่วมแคมเปญ)ข้างเคียง ได้ร้อยละ 80.24

4.1.9. ผลการทำนาย และ ค่าน้ำหนักของแอตทริบิวต์ ด้วยวิธีซัพพอร์ตเวกเตอร์แมทชีน (Support Vector)

จาก “Feature Auto Model” ด้วยวิธีซัพพอร์ตเวกเตอร์แมทชีน (Support Vector) ผลในส่วนของตาราง Confusion Matrix และ ค่าน้ำหนักของแอตทริบิวต์ ได้ค่า Class Recall และ Class Precision ใช้ค่าพารามิเตอร์ที่กล่าวไว้แสดงได้ดังนี้



ภาพที่ 4.9 ค่าน้ำหนักของแอตทริบิวต์ ของวิธีการพยากรณ์ของเทคนิคซัพพอร์ตเวกเตอร์แมทชีน (Support Vector)

DPU DHURAKIJ PUNDIT UNIVERSITY มหาวิทยาลัยธุรกิจบัณฑิตย์

ตารางที่ 4.9 ผลในส่วนของตาราง Confusion Matrix ของวิธีการพยากรณ์ของเทคนิคซัพพอร์ตเวกเตอร์แมทชีน (Support Vector)

	true Not_Engaged	true Engaged	class precision
pred. Not_Engaged	2,562	1,475	63.46%
pred. Engaged	55	194	77.91%
class recall	97.90%	11.62%	

สรุปค่า class precision ของแบบจำลองวิธีการพยากรณ์ของเทคนิคซัพพอร์ตเวกเตอร์แมทชีน (Support Vector)โดย Not Engaged(ไม่เข้าร่วมแคมเปญ)จริงๆกลับมา ได้ร้อยละ 63.46 และ ตอบ Engaged(เข้าร่วมแคมเปญ)จริงๆกลับมา ได้ร้อยละ 77.91

สรุปค่า class recall ของแบบจำลองวิธีการพยากรณ์ของเทคนิคซัพพอร์ตเวกเตอร์แมทชีน (Support Vector) คำนวณเทียบกับคำตอบ Not Engaged(ไม่เข้าร่วมแคมเปญ)ข้างเคียง ได้ร้อยละ 97.90 และ คำนวณเทียบกับคำตอบ Engaged(เข้าร่วมแคมเปญ)ข้างเคียง ได้ร้อยละ 11.62

4.2 ผลการเปรียบเทียบประสิทธิภาพการทำนายผล

จาก “Feature Auto Model” ได้ข้อผลของประสิทธิภาพผลเปรียบเทียบค่าความถูกต้อง ของแบบจำลองโดยใช้ค่าความถูกต้องของข้อมูล (Accuracy) , ค่าสัดส่วนของข้อมูลที่ถูกต้องกับข้อมูลที่ต้องการทั้งหมด ค่าความระลึก (Recall) , ค่าสัดส่วนของข้อมูลที่ถูกต้อง ตรงตามความต้องการส่วนด้วยข้อมูลทั้งหมด ค่าความแม่นยำ (Precision) และ ค่าเฉลี่ยแบบ ค่าเฉลี่ยฮาร์มอนิก (harmonic mean) โดยจะนำ “ค่าความแม่นยำ” และ “ค่าความระลึก” มาพิจารณาร่วมกัน ค่าความถ่วงดุล (F-measure) ตามลำดับดังนี้

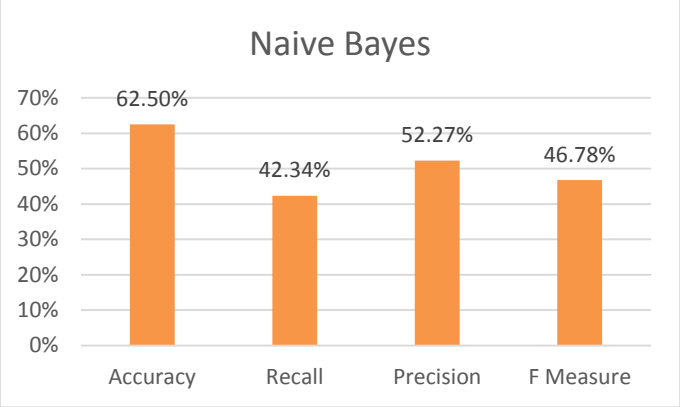
DPU DHURAKIJ PUNDIT UNIVERSITY มหาวิทยาลัยธุรกิจบัณฑิตย์

ตารางที่ 4.10 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%
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%

จากเทคนิคการเรียนรู้ของเครื่อง (Machine Learning Technique) บน “Feature Auto Model” บน RapidMiner มีประสิทธิภาพการทำนายดังนี้

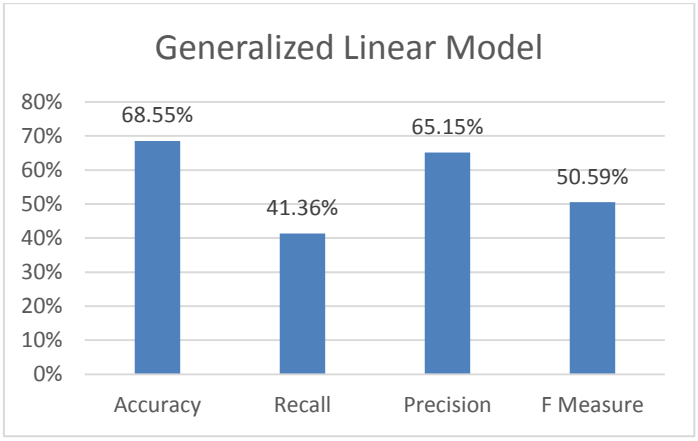
วิธีการจำแนกประเภทข้อมูลแบบเบย์อย่างง่าย (Naive Bayes) สามารถแจกแจงประสิทธิภาพได้ดังนี้ ผล Accuracy เท่ากับร้อยละ 62.50 ผล Precision เท่ากับร้อยละ 52.27 ผล Recall เท่ากับร้อยละ 42.34และ ผลของ F-measure เท่ากับร้อยละ 46.78



ภาพที่ 4.10 ผลประสิทธิภาพของแบบจำลองวิธีการพยากรณ์ของเทคนิควิธีการจำแนกประเภทข้อมูลแบบเบย์อย่างง่าย (Naive Bayes)

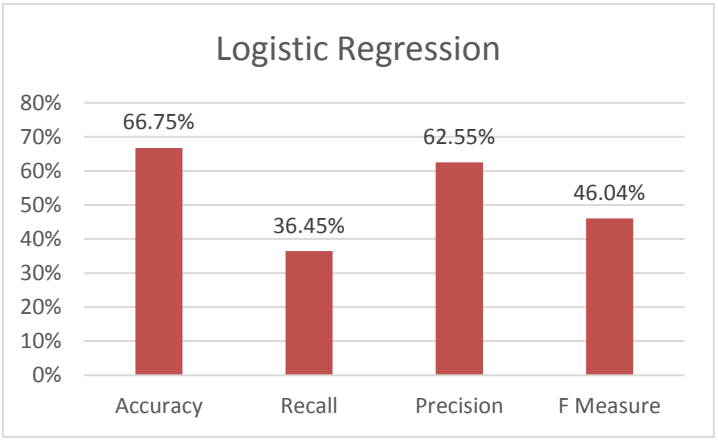
DPU DHURAKIJ PUNDIT UNIVERSITY มหาวิทยาลัยธุรกิจบัณฑิตย์

วิธีตัวแบบเชิงเส้นนัยทั่วไป (Generalized Linear Model) สามารถแจกแจงประสิทธิภาพได้ดังนี้ ผล Accuracy เท่ากับร้อยละ 68.55 ผล Precision เท่ากับร้อยละ 65.15 ผล Recall เท่ากับร้อยละ 41.36 และ ผลของ F-measure เท่ากับร้อยละ 50.59



ภาพที่ 4.11 ผลประสิทธิภาพของแบบจำลองวิธีการพยากรณ์ของเทคนิคตัวแบบเชิงเส้นนัยทั่วไป (Generalized Linear Model)

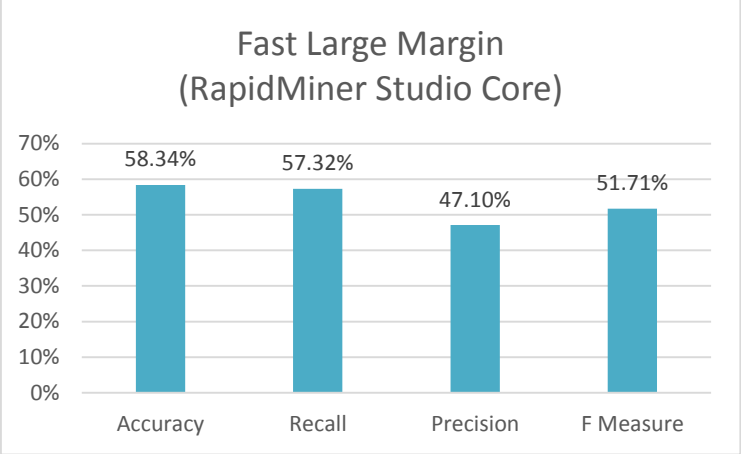
วิธีการถดถอยโลจีสติก (Logistic Regression) สามารถแจกแจงประสิทธิภาพได้ดังนี้ ผล Accuracy เท่ากับร้อยละ 66.75 ค่าPrecision มีค่าอยู่ที่ร้อยละ 62.55 ผล Recall เท่ากับร้อยละ 36.45 และ ผลของ F-measure เท่ากับร้อยละ 46.04



ภาพที่ 4.12 ผลประสิทธิภาพของแบบจำลองวิธีการพยากรณ์ของเทคนิคการถดถอยโลจีสติก (Logistic Regression)

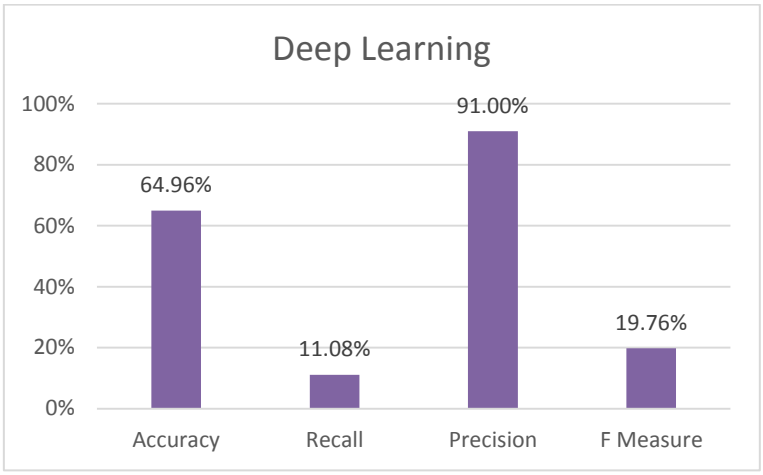
DPU DHURAKIJ PUNDIT UNIVERSITY มหาวิทยาลัยธุรกิจบัณฑิตย์

วิธีการจำแนกประเภทแบบเร็วโดยใช้ขอบเขตการแยกแยะที่มีขนาดกว้าง (Fast Large Margin (RapidMiner Studio Core)) สามารถแจกแจงประสิทธิภาพได้ดังนี้ ผล Accuracy เท่ากับร้อยละ 58.34 ผล Precision เท่ากับร้อยละ 47.10 ผล Recall เท่ากับร้อยละ 57.32 และ ผลของ F-measure เท่ากับร้อยละ 51.71



ภาพที่ 4.13 ผลประสิทธิภาพของแบบจำลองวิธีการพยากรณ์ของเทคนิคการจำแนกประเภทแบบเร็วโดยใช้ขอบเขตการแยกแยะที่มีขนาดกว้าง (Fast Large Margin (RapidMiner Studio Core))

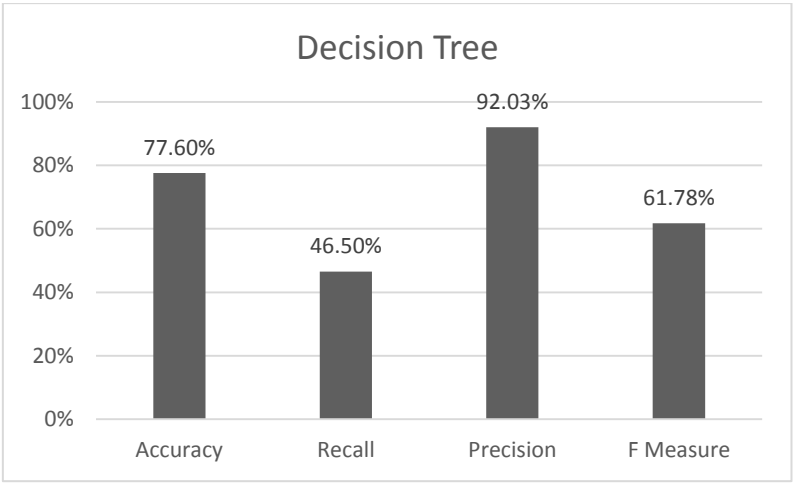
วิธีการเรียนรู้เชิงลึก (Deep Learning) สามารถแจกแจงประสิทธิภาพได้ดังนี้ ผล Accuracy เท่ากับร้อยละ 64.96 ผล Precision เท่ากับร้อยละ 91 ผล Recall เท่ากับร้อยละ 11.08 และ ผลของ F-measure เท่ากับร้อยละ 19.76



ภาพที่ 4.14 ผลประสิทธิภาพของแบบจำลองวิธีการพยากรณ์ของเทคนิคการเรียนรู้เชิงลึก (Deep Learning)

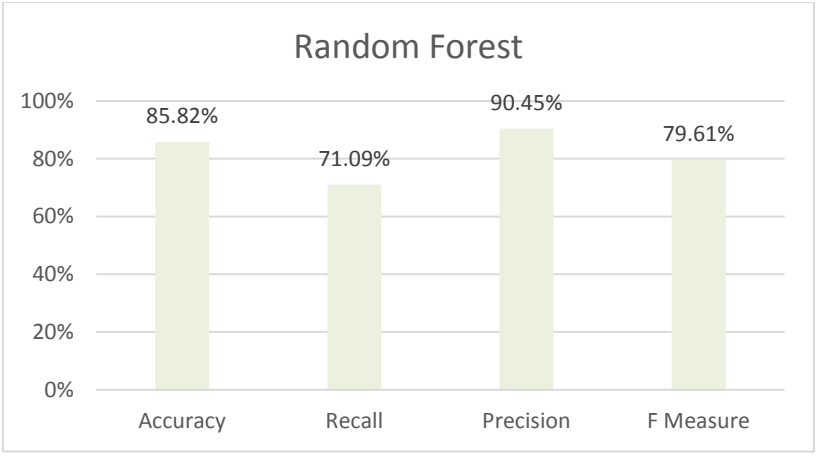
DPU DHURAKIJ PUNDIT UNIVERSITY มหาวิทยาลัยธุรกิจบัณฑิตย์

วิธีต้นไม้การตัดสินใจ (Decision Tree) สามารถแจกแจงประสิทธิภาพได้ดังนี้ ผล Accuracy เท่ากับร้อยละ 77.60 ผล Precision เท่ากับร้อยละ 92.03 ผล Recall เท่ากับร้อยละ 46.50 และ ผลของ F-measure เท่ากับร้อยละ 61.78



ภาพที่ 4.15 ผลประสิทธิภาพของแบบจำลองวิธีการพยากรณ์ของเทคนิคต้นไม้การตัดสินใจ (Decision Tree)

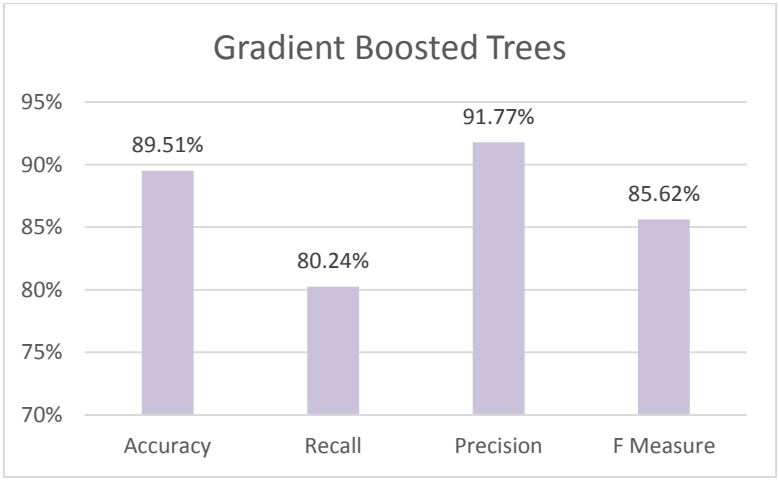
วิธีแบบจำลองป่าสุ่ม (Random Forest) สามารถแจกแจงประสิทธิภาพได้ดังนี้ ผล Accuracy เท่ากับร้อยละ 85.82 ผล Precision เท่ากับร้อยละ 90.45 ผล Recall เท่ากับร้อยละ 71.09 และ ผลของ F-measure เท่ากับร้อยละ 79.61



ภาพที่ 4.16 ผลประสิทธิภาพของแบบจำลองวิธีการพยากรณ์ของเทคนิคแบบจำลองป่าสุ่ม (Random Forest)

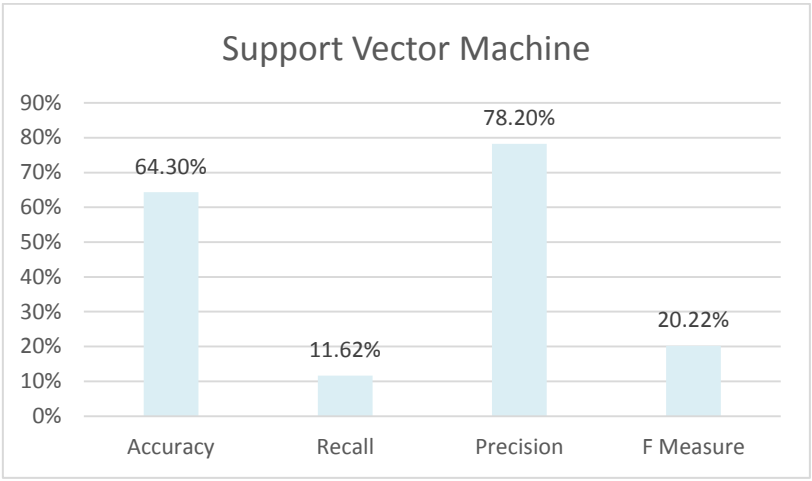
DPU DHURAKIJ PUNDIT UNIVERSITY มหาวิทยาลัยธุรกิจบัณฑิตย์

วิธีต้นไม้การเพิ่มไล่ระดับการตัดสินใจ (Gradient Boosted Trees) สามารถแจกแจงประสิทธิภาพได้ดังนี้ ผล Accuracy เท่ากับร้อยละ 89.51 ผล Precision เท่ากับร้อยละ 91.77 ผล Recall เท่ากับร้อยละ 80.24 และ ผลของ F-measure เท่ากับร้อยละ 85.62



ภาพที่ 4.17 ผลประสิทธิภาพของแบบจำลองวิธีการพยากรณ์ของเทคนิคต้นไม้การเพิ่มไล่ระดับการตัดสินใจ (Gradient Boosted Trees)

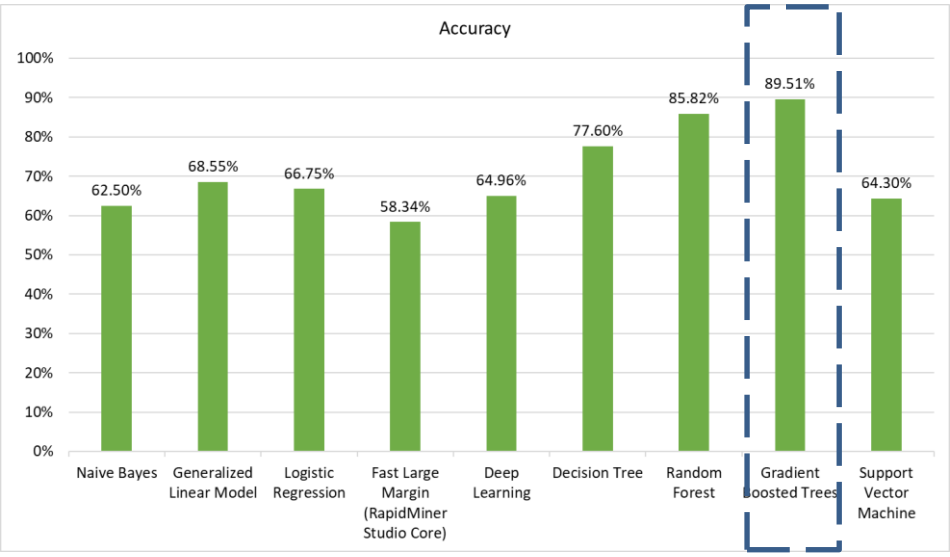
วิธีซัพพอร์ตเวกเตอร์แมทชีน (Support Vector) สามารถแจกแจงประสิทธิภาพได้ดังนี้ ค่า Accuracy มีค่าอยู่ที่ร้อยละ 64.30 ผล Precision เท่ากับร้อยละ 78.20 ผล Recall เท่ากับร้อยละ 11.62 และ ผลของ F-measure เท่ากับร้อยละ 20.22



ภาพที่ 4.18 ผลประสิทธิภาพของแบบจำลองวิธีการพยากรณ์ของเทคนิคซัพพอร์ตเวกเตอร์แมทชีน (Support Vector)

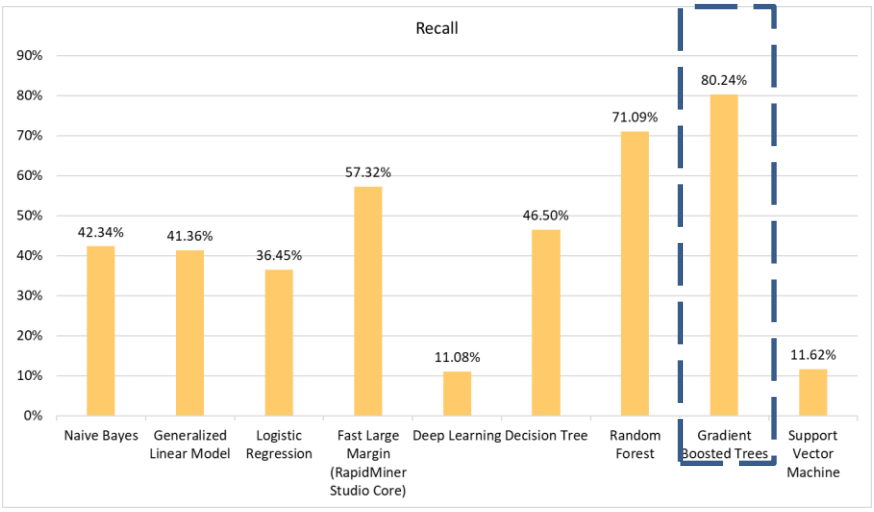
DPU DHURAKIJ PUNDIT UNIVERSITY มหาวิทยาลัยธุรกิจบัณฑิตย์

ผลความถูกต้องของแบบจำลองโดยใช้ค่าความถูกต้องของข้อมูล (Accuracy) โดยเปรียบเทียบกับ 9 แบบจำลองสามารถแจกแจงประสิทธิภาพได้ดังนี้



ภาพที่ 4.19 ผลความถูกต้องของแบบจำลองโดยใช้ค่าความถูกต้องของข้อมูล (Accuracy) โดยเปรียบเทียบกับ 9 แบบจำลอง

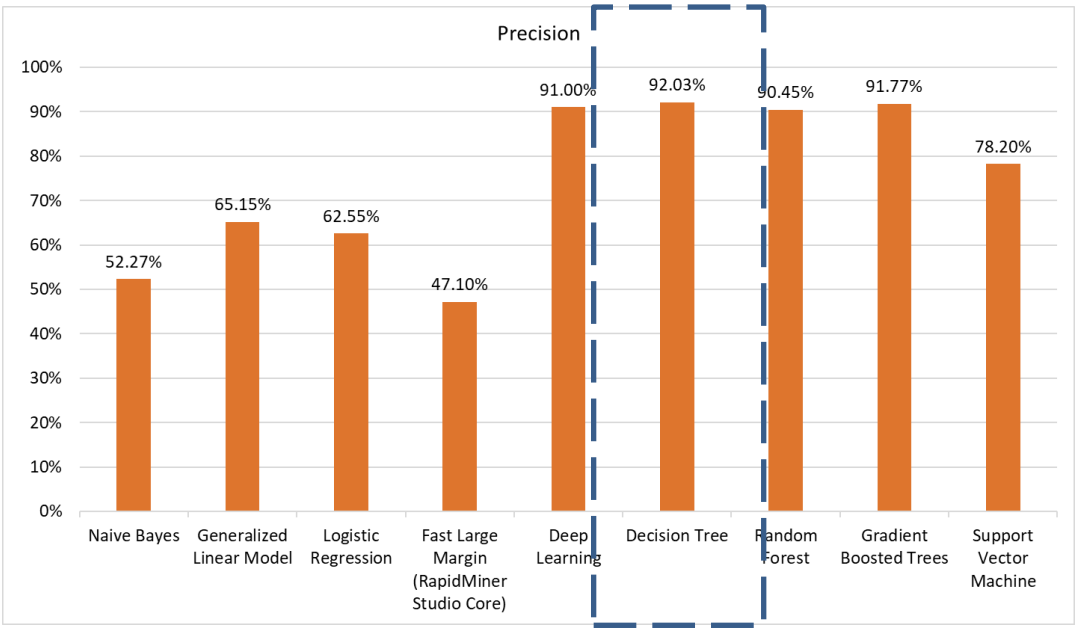
ผลค่าสัดส่วนของข้อมูลที่ถูกต้องกับข้อมูลที่ต้องการทั้งหมด ค่าความระลึก (Recall) โดยเปรียบเทียบกับ 9 แบบจำลองสามารถแจกแจงประสิทธิภาพได้ดังนี้



ภาพที่ 4.20 ผลค่าสัดส่วนของข้อมูลที่ถูกต้องกับข้อมูลที่ต้องการทั้งหมด ค่าความระลึก (Recall) โดยเปรียบเทียบกับ 9 แบบจำลอง

DPU DHURAKIJ PUNDIT UNIVERSITY มหาวิทยาลัยธุรกิจบัณฑิตย์

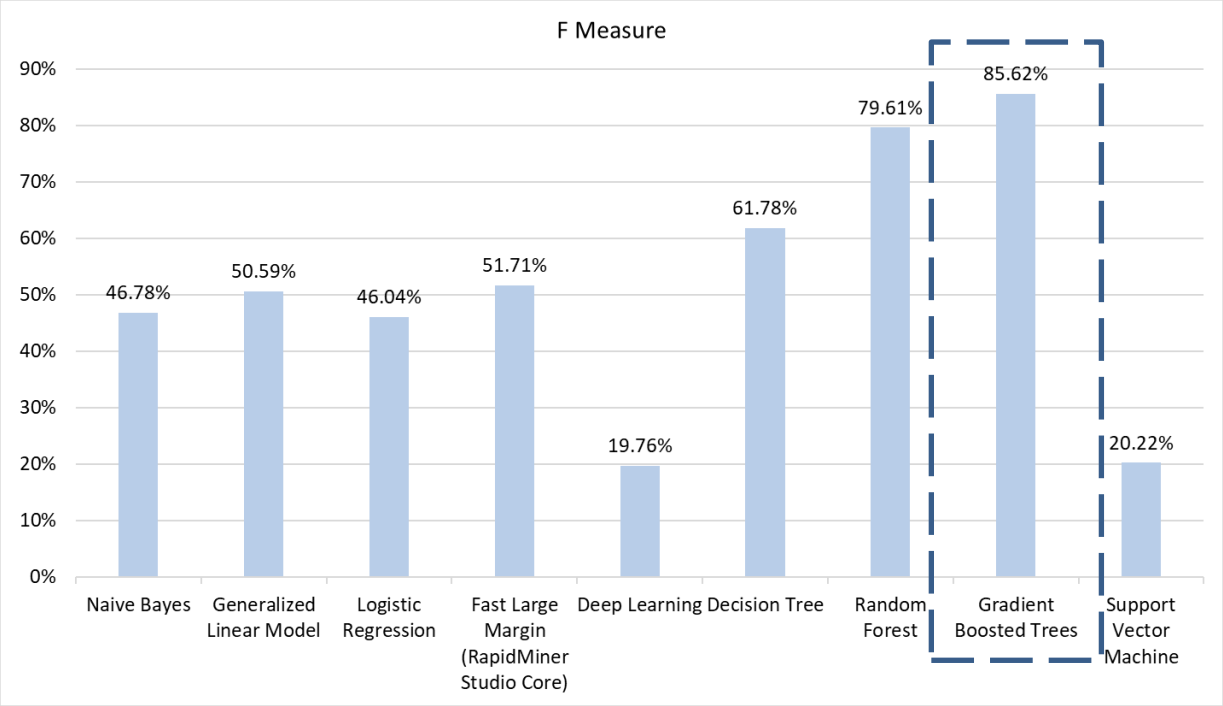
ผลค่าสัดส่วนของข้อมูลที่ถูกต้อง ตรงตามความต้องการส่วนด้วยข้อมูลทั้งหมดค่าความแม่นยำ (Precision) โดยเปรียบเทียบกับ 9 แบบจำลองสามารถแจกแจงประสิทธิภาพได้ดังนี้



ภาพที่ 4.21 ผลค่าสัดส่วนของข้อมูลที่ถูกต้อง ตรงตามความต้องการส่วนด้วยข้อมูลทั้งหมดค่าความแม่นยำ (Precision) โดยเปรียบเทียบกับ 9 แบบจำลอง

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

DPU DHURAKIJ PUNDIT UNIVERSITY มหาวิทยาลัยธุรกิจบัณฑิตย์

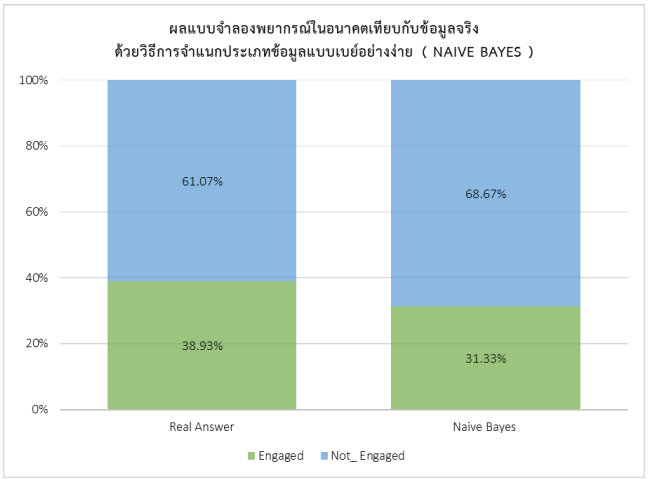


ภาพที่ 4.22 ผลค่าเฉลี่ยแบบ ค่าเฉลี่ยฮาร์มอนิก (harmonic mean) โดยจะนำ “ค่าความแม่นยำ” และ “ค่าความระลึก” มาพิจารณาร่วมกัน ค่าความถ่วงดุล (F-measure) โดยเปรียบเทียบกับ 9 แบบจำลอง

4.3 ผลความถูกต้องจากการนำแบบจำลองพยากรณ์ในอนาคตเทียบกับข้อมูลจริงเมื่อทำการวัดผลประสิทธิภาพความถูกต้องของแบบจำลองแล้วในขั้นตอนก่อนหน้า แล้วนั้นขั้นตอนนี้นั้นเป็นผลความถูกต้องจากการนำแบบจำลองพยากรณ์ในอนาคตเทียบกับ ข้อมูลจริง ซึ่งข้อมูลจริงมีค่า Engaged(เข้าร่วมแคมเปญ) ร้อยละ 38.93 และ Not_ Engaged(ไม่เข้าร่วมแคมเปญ) ร้อยละ 61.07 โดยแสดงผลได้ดังนี้

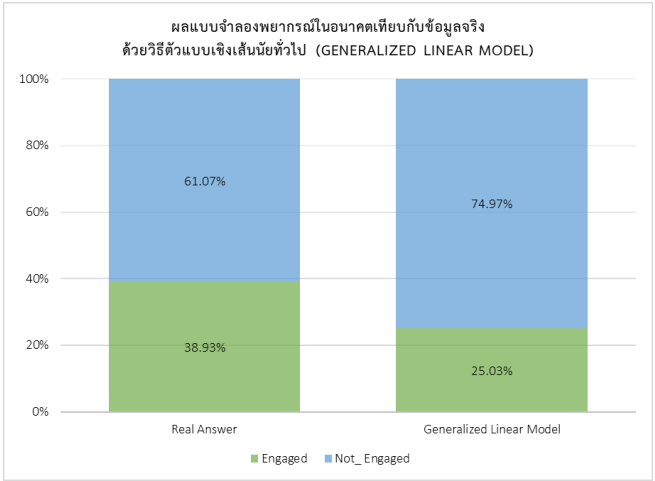
4.3.1 ผลแบบจำลองพยากรณ์ในอนาคตเทียบกับข้อมูลจริง ด้วยวิธีการจำแนกประเภทข้อมูลแบบเบย์อย่างง่าย (Naive Bayes) โดยข้อมูลพยากรณ์ Engaged(เข้าร่วมแคมเปญ) มีที่ร้อยละ 31.33 และ Not_ Engaged(ไม่เข้าร่วมแคมเปญ) มีที่ร้อยละ 68.67

DPU DHURAKIJ PUNDIT UNIVERSITY มหาวิทยาลัยธุรกิจบัณฑิตย์



ภาพที่ 4.23 ผลแบบจำลองพยากรณ์ในอนาคตเทียบกับข้อมูลจริง ด้วยวิธีการจำแนกประเภทข้อมูลแบบเบย์อย่างง่าย (Naive Bayes)

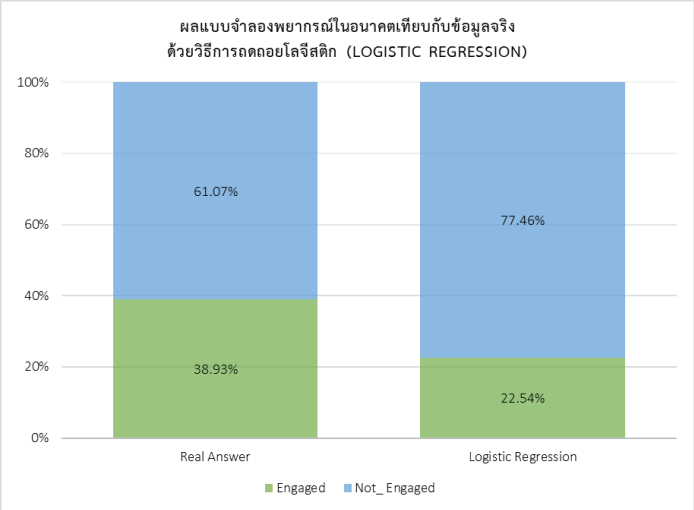
4.3.2 ผลแบบจำลองพยากรณ์ในอนาคตเทียบกับข้อมูลจริง ด้วยวิธีตัวแบบเชิงเส้นนัยทั่วไป (Generalized Linear Model) โดยข้อมูลพยากรณ์ Engaged(เข้าร่วมแคมเปญ) มีที่ร้อยละ 25.03 และ Not_ Engaged(ไม่เข้าร่วมแคมเปญ) มีที่ร้อยละ 74.97



ภาพที่ 4.24 ผลแบบจำลองพยากรณ์ในอนาคตเทียบกับข้อมูลจริง ด้วยวิธีตัวแบบเชิงเส้นนัยทั่วไป (Generalized Linear Model)

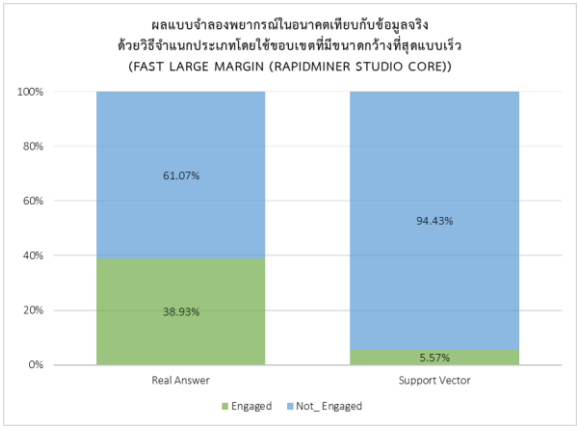
DPU DHURAKIJ PUNDIT UNIVERSITY มหาวิทยาลัยธุรกิจบัณฑิตย์

4.3.3 ผลแบบจำลองพยากรณ์ในอนาคตเทียบกับข้อมูลจริง ด้วยวิธีการถดถอยโลจีสติก (Logistic Regression) โดยข้อมูลจริง Engaged(เข้าร่วมแคมเปญ) มีที่ร้อยละ 38.93 และ Not_ Engaged(ไม่เข้าร่วมแคมเปญ) มีที่ร้อยละ 61.07



ภาพที่ 4.25 ผลแบบจำลองพยากรณ์ในอนาคตเทียบกับข้อมูลจริง ด้วยวิธีการถดถอยโลจีสติก (Logistic Regression)

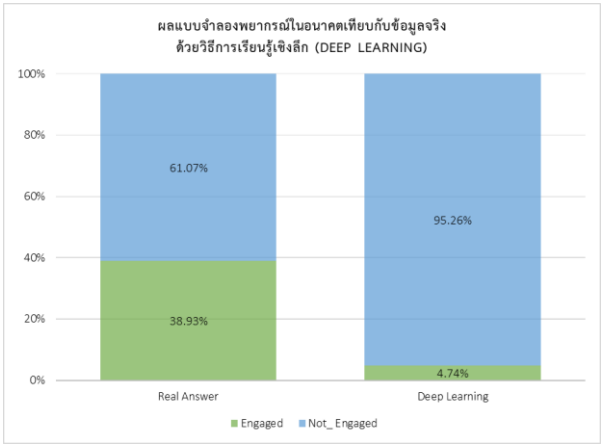
4.3.4 ผลแบบจำลองพยากรณ์ในอนาคตเทียบกับข้อมูลจริง ด้วยวิธีจำแนกประเภทโดยใช้ขอบเขตที่มีขนาดกว้างที่สุดแบบเร็ว (Fast Large Margin (RapidMiner Studio Core)) โดยข้อมูลพยากรณ์ Engaged(เข้าร่วมแคมเปญ) มีที่ร้อยละ 47.22 และ Not_ Engaged(ไม่เข้าร่วมแคมเปญ) มีที่ร้อยละ 52.78



ภาพที่ 4.26 ผลแบบจำลองพยากรณ์ในอนาคตเทียบกับข้อมูลจริง ด้วยวิธีจำแนกประเภทโดยใช้ขอบเขตที่มีขนาดกว้างที่สุดแบบเร็ว (Fast Large Margin (RapidMiner Studio Core))

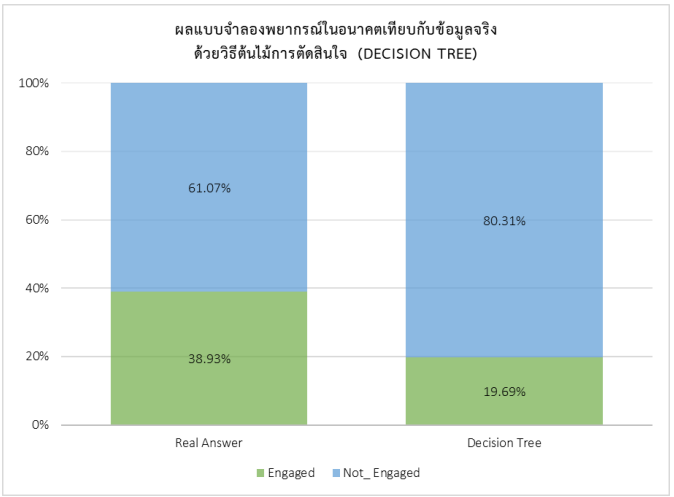
DPU DHURAKIJ PUNDIT UNIVERSITY มหาวิทยาลัยธุรกิจบัณฑิตย์

4.3.5 ผลแบบจำลองพยากรณ์ในอนาคตเทียบกับข้อมูลจริง ด้วยวิธีการเรียนรู้เชิงลึก (Deep Learning) โดยข้อมูลพยากรณ์ Engaged(เข้าร่วมแคมเปญ) มีที่ร้อยละ 4.74 และ Not_ Engaged(ไม่เข้าร่วมแคมเปญ) มีที่ร้อยละ 95.26



ภาพที่ 4.27 ผลแบบจำลองพยากรณ์ในอนาคตเทียบกับข้อมูลจริง ด้วยวิธีการเรียนรู้เชิงลึก (Deep Learning)

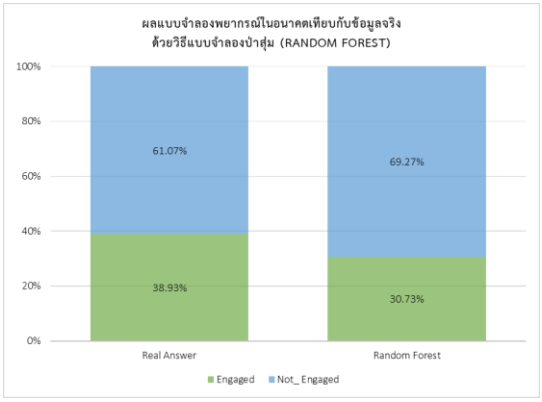
4.3.6 ผลแบบจำลองพยากรณ์ในอนาคตเทียบกับข้อมูลจริง ด้วยวิธีต้นไม้การตัดสินใจ (Decision Tree) โดยข้อมูลพยากรณ์ Engaged(เข้าร่วมแคมเปญ) มีที่ร้อยละ 19.69 และ Not_ Engaged(ไม่เข้าร่วมแคมเปญ) มีที่ร้อยละ 80.31



ภาพที่ 4.28 ผลแบบจำลองพยากรณ์ในอนาคตเทียบกับข้อมูลจริง ด้วยวิธีต้นไม้การตัดสินใจ (Decision Tree)

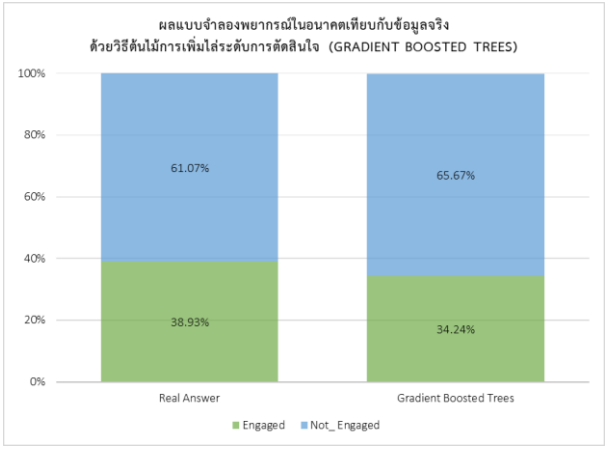
DPU DHURAKIJ PUNDIT UNIVERSITY มหาวิทยาลัยธุรกิจบัณฑิตย์

4.3.7. ผลแบบจำลองพยากรณ์ในอนาคตเทียบกับข้อมูลจริง ด้วยวิธีแบบจำลองป่าสุ่ม (Random Forest) โดยข้อมูลพยากรณ์ Engaged(เข้าร่วมแคมเปญ) มีที่ร้อยละ 30.73 และ Not_ Engaged(ไม่เข้าร่วมแคมเปญ) มีที่ร้อยละ 69.27



ภาพที่ 4.29 ผลแบบจำลองพยากรณ์ในอนาคตเทียบกับข้อมูลจริง ด้วยวิธีแบบจำลองป่าสุ่ม (Random Forest)

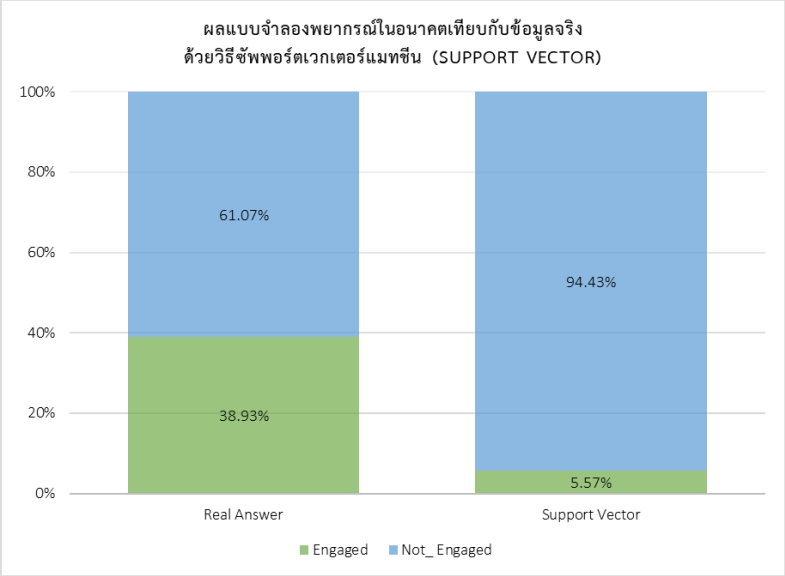
4.3.8. ผลแบบจำลองพยากรณ์ในอนาคตเทียบกับข้อมูลจริง ด้วยวิธีต้นไม้การเพิ่มไล่ระดับการตัดสินใจ (Gradient Boosted Trees) โดยข้อมูลพยากรณ์ Engaged(เข้าร่วมแคมเปญ) มีที่ร้อยละ 34.24 และ Not_ Engaged(ไม่เข้าร่วมแคมเปญ) มีที่ร้อยละ 65.67



ภาพที่ 4.30 ผลแบบจำลองพยากรณ์ในอนาคตเทียบกับข้อมูลจริง ด้วยวิธีต้นไม้การเพิ่มไล่ระดับการตัดสินใจ (Gradient Boosted Trees)

DPU DHURAKIJ PUNDIT UNIVERSITY มหาวิทยาลัยธุรกิจบัณฑิตย์

4.3.9. ผลแบบจำลองพยากรณ์ในอนาคตเทียบกับข้อมูลจริง ด้วยวิธีซัพพอร์ตเวกเตอร์แมทชีน (Support Vector) โดยข้อมูลพยากรณ์ Engaged(เข้าร่วมแคมเปญ) มีที่ร้อยละ 5.57 และ Not_ Engaged(ไม่เข้าร่วมแคมเปญ) มีที่ร้อยละ 94.43

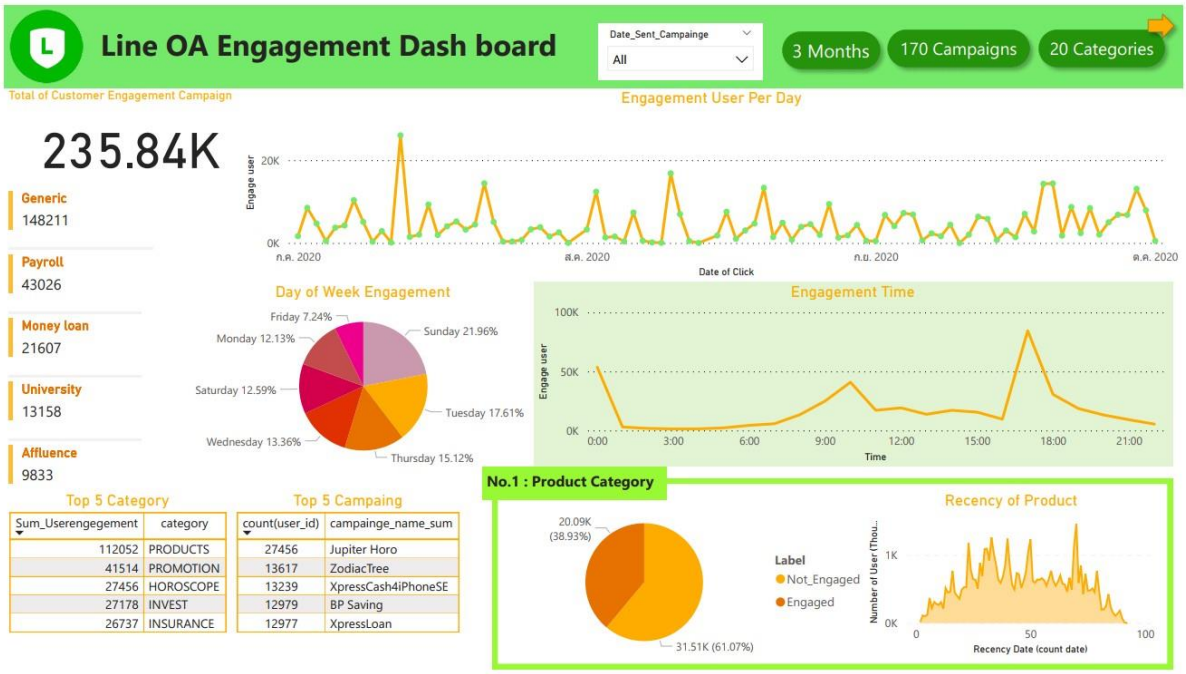


ภาพที่ 4.31 ผลแบบจำลองพยากรณ์ในอนาคตเทียบกับข้อมูลจริง ด้วยวิธีซัพพอร์ตเวกเตอร์แมทชีน (Support Vector)

DPU DHURAKIJ PUNDIT UNIVERSITY มหาวิทยาลัยธุรกิจบัณฑิตย์

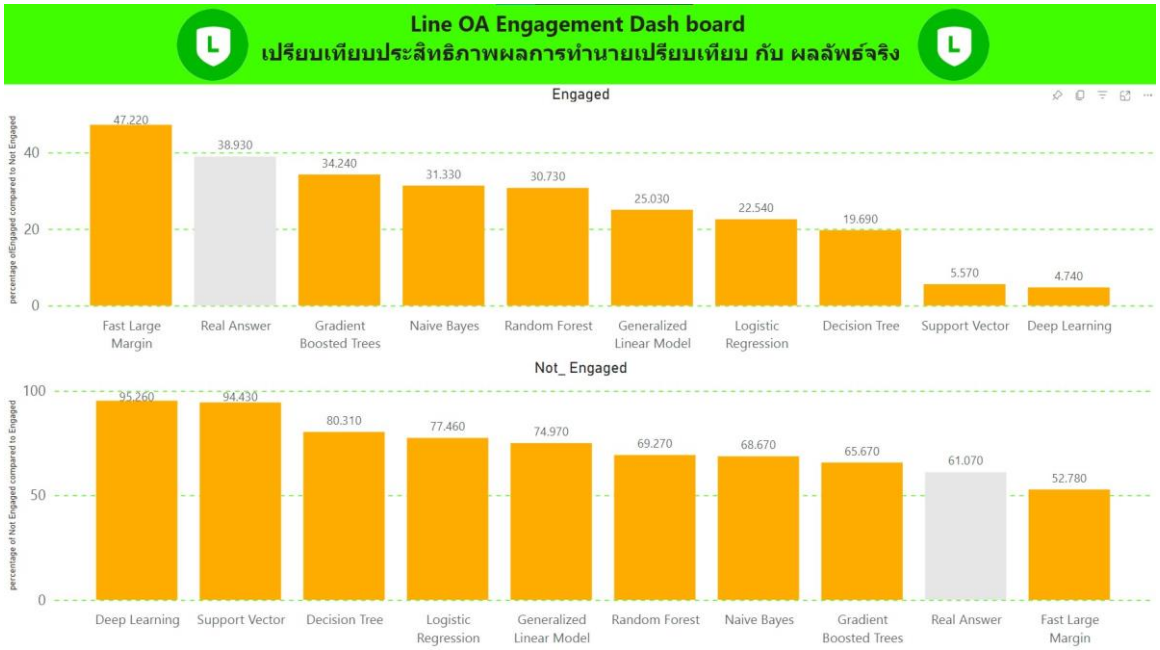
4.4. ผลของการแสดงผลเพื่อนำเสนอ

จากผลการเปรียบเทียบประสิทธิภาพการทำนายผลที่ผ่านมา จึงนำข้อมูลแสดงผลบน Dashboard เพื่อนำแบบจำลองที่ดีที่สุดนั้นนำมาใช้ให้มีประโยชน์ หลังจากที่วิจัย และ เปรียบเทียบแต่เทคนิคแล้ว หากพบว่าแบบจำลองใดมีประสิทธิภาพในการทำนายที่แม่นยำมากที่สุด และ เหมาะสมกับความต้องการของการพยากรณ์ข้อมูลในอนาคต แบบจำลองดังกล่าวจะถูกนำมาประยุกต์ใช้ในขั้นตอนนี้ ซึ่งตัวอย่างการนำแบบจำลองการพยากรณ์ข้อมูลไปใช้ประโยชน์ สามารถแสดงได้ดังภาพ



ภาพที่ 4.32 ตัวอย่างหน้าจอการแสดงผลการพยากรณ์ของข้อมูลสำหรับผู้ใช้งานหน้าที่ 1

DPU DHURAKIJ PUNDIT UNIVERSITY มหาวิทยาลัยธุรกิจบัณฑิตย์



ภาพที่ 4.33 ตัวอย่างหน้าจอการแสดงผลการพยากรณ์ของข้อมูลสำหรับผู้ใช้งานหน้าที่ 2

DPU DHURAKIJ PUNDIT UNIVERSITY มหาวิทยาลัยธุรกิจบัณฑิตย์

บทที่ 5
สรุปผลการวิจัย อภิปรายผล และข้อเสนอแนะ

งานวิจัยชิ้นนี้ได้เสนอการเปรียบเทียบการนำชุดข้อมูล เข้าแบบจำลองทั้ง 9 แบบจำลอง โดยได้ทำการวัดประสิทธิภาพของแต่ละแบบจำลอง เพื่อนำผลลัพธ์มาเปรียบเทียบและนำเสนอเทคนิคที่ดีและเหมาะสมที่สุดใช้สำหรับการพยากรณ์ข้อมูล สำหรับข้อมูลที่มีลักษณะหรือรูปแบบเดียวกันกับที่ใช้ในงานวิจัยในการใช้วิเคราะห์เพื่อคาดการณ์ว่าลูกค้าคนใดบ้างที่คาดว่าจะมีส่วนร่วมในการปล่อยแคมเปญครั้งต่อไป สามารถสรุปผลการวิจัยได้ดังนี้

5.1. สรุปผลการวิจัยและอภิปรายผล

5.2. ข้อเสนอแนะ

5.1. สรุปผลการวิจัยและอภิปรายผล

สำหรับการวิเคราะห์เปรียบเทียบ ค่าความถูกต้องของแบบจำลองโดยใช้ค่าความถูกต้องของข้อมูล (Accuracy) , ค่าสัดส่วนของข้อมูลที่ถูกต้องกับข้อมูลที่ต้องการทั้งหมด ค่าความระลึก (Recall) , ค่าสัดส่วนของข้อมูลที่ถูกต้อง ตรงตามความต้องการส่วนด้วยข้อมูลทั้งหมดค่าความแม่นยำ (Precision) และ ค่าเฉลี่ยแบบ ค่าเฉลี่ยฮาร์มอนิก (harmonic mean) โดยจะนำ “ค่าความแม่นยำ” และ “ค่าความระลึก” มาพิจารณาร่วมกัน ค่าความถ่วงดุล (F-measure) ของวิธีการสร้างแบบจำลองด้วยเทคนิคการเรียนรู้ของเครื่องโดย Feature Auto Model บนโปรแกรม RapidMiner พบว่าข้อมูลที่ใช้ สร้างแบบจำลองด้วยเทคนิคทั้งหมด 9 เทคนิคข้างต้น (การจำแนกประเภทข้อมูลแบบเบย์อย่างง่าย (Naive Bayes) , ตัวแบบเชิงเส้นนัยทั่วไป (Generalized Linear Model) , การถดถอยโลจีสติก (Logistic Regression) , การจำแนกประเภทแบบเร็วโดยใช้ขอบเขตการแยกแยะที่มีขนาดกว้าง(Fast Large Margin (RapidMiner Studio Core)) , การเรียนรู้เชิงลึก (Deep Learning) , ต้นไม้การตัดสินใจ (Decision Tree) , แบบจำลองป่าสุ่ม (Random Forest) , ต้นไม้การเพิ่มไล่ระดับการตัดสินใจ (Gradient Boosted Trees) , ซัพพอร์ตเวกเตอร์แมทชีน (Support Vector)) ด้วยชุดข้อมูลเดียวกัน

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

5.1.2. จากผลการเปรียบเทียบความถูกต้องจากการนำแบบจำลองพยากรณ์ในอนาคตเทียบกับข้อมูลจริง ซึ่งข้อมูลจริงมีค่า การเข้าร่วมแคมเปญ 38.93% และ การไม่เข้าร่วมแคมเปญ 61.07% พบว่าสองแบบจำลองมีค่าใกล้เคียง คือแบบจำลอง Naive Bayes ร้อยละของข้อมูลพยากรณ์ การเข้าร่วมแคมเปญ 31.33% และ การไม่เข้า

DPU DHURAKIJ PUNDIT UNIVERSITY มหาวิทยาลัยธุรกิจบัณฑิตย์

ร่วมแคมเปญ 68.67% และ Gradient Boosted Trees ร้อยละของข้อมูลพยากรณ์ การเข้าร่วมแคมเปญ 34.24% และ การไม่เข้าร่วมแคมเปญ 65.67%

ตารางที่ 5.1 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%

จากตารางข้างต้นสามารถสรุปได้ดังนี้

1.ค่าความแม่นยำ (Precision) ของแบบจำลองต้นไม้การเพิ่มไล่ระดับการตัดสินใจ (Gradient Boosted Trees) มีค่าที่สูงที่สุดเมื่อเทียบกับ แบบจำลองป่าสุ่ม (Random Forest), ต้นไม้การตัดสินใจ (Decision Tree) และ การเรียนรู้เชิงลึก (Deep Learning) ตามลำดับ

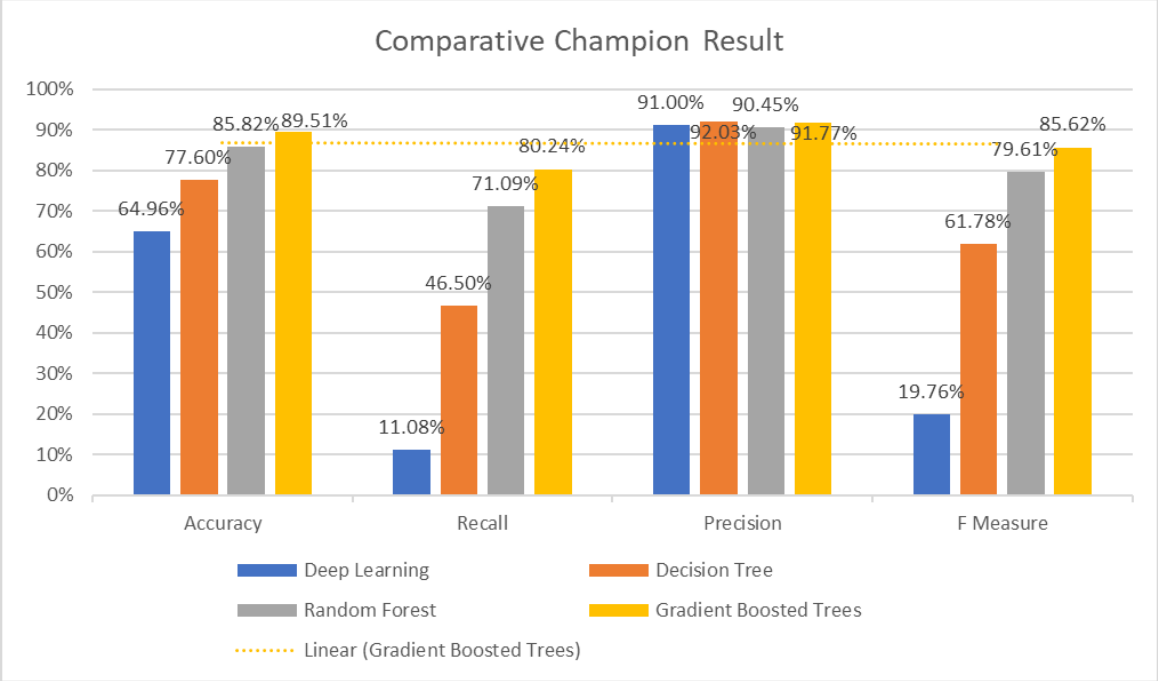
2.ค่าความถูกต้องของข้อมูล (Accuracy) ของแบบจำลองต้นไม้การเพิ่มไล่ระดับการตัดสินใจ (Gradient Boosted Trees) มีค่าที่สูงที่สุดเมื่อเทียบกับ แบบจำลองป่าสุ่ม (Random Forest), ต้นไม้การตัดสินใจ (Decision Tree) และ การเรียนรู้เชิงลึก (Deep Learning) ตามลำดับ

3. ค่าความระลึก (Recall) ของแบบจำลองต้นไม้การตัดสินใจ (Decision Tree) มีค่าที่สูงที่สุดเมื่อเทียบกับ แบบจำลองต้นไม้การเพิ่มไล่ระดับการตัดสินใจ (Gradient Boosted Trees) การเรียนรู้เชิงลึก (Deep Learning) และ ป่าสุ่ม (Random Forest) ตามลำดับ

4.ค่าความถ่วงดุล (F-measure) ของแบบจำลองต้นไม้การเพิ่มไล่ระดับการตัดสินใจ (Gradient Boosted Trees) มีค่าที่สูงที่สุดเมื่อเทียบกับ แบบจำลองป่าสุ่ม (Random Forest), ต้นไม้การตัดสินใจ (Decision Tree) และ การเรียนรู้เชิงลึก (Deep Learning) ตามลำดับ

DPU DHURAKIJ PUNDIT UNIVERSITY มหาวิทยาลัยธุรกิจบัณฑิตย์

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



ภาพที่ 5.1 กราฟการเปรียบเทียบของ4 แบบจำลองที่ดีที่สุดในการวัดประสิทธิภาพทั้ง 4 แบบคำนวณ

ดังนั้นจึงสามารถอภิปรายผลการวิจัยพบว่า แบบจำลองต้นไม้การเพิ่มไล่ระดับการตัดสินใจ (Gradient Boosted Trees) มีประสิทธิภาพในการการคาดการณ์ว่าลูกค้าคนใดบ้างที่จะมีส่วนร่วมในการปล่อยแคมเปญได้ผลลัพธ์ดีที่สุด และ เหมาะสมกับข้อมูลใช้สำหรับนำมาใช้สร้างแบบจำลองสำหรับการพยากรณ์มากที่สุดเพื่อใช้ในการคาดการณ์ว่าลูกค้าคนใดบ้างที่จะมีส่วนร่วมในการปล่อยแคมเปญต่อไปในอนาคต

DPU DHURAKIJ PUNDIT UNIVERSITY มหาวิทยาลัยธุรกิจบัณฑิตย์

5.2 ข้อเสนอแนะ

5.2.1. ตัวแปรที่นำมาใช้สำหรับงานวิจัยชิ้นนี้เป็นเพียงส่วนหนึ่งของการพิจารณา กลุ่มแคมเปญประเภทที่เกี่ยวข้องกับผลิตภัณฑ์ของธุรกิจเท่านั้น สำหรับการทำนายที่มีประสิทธิภาพมากขึ้นควรเพิ่มตัวแปรอื่น ๆที่เกี่ยวข้องอีกและ ปรับปรุงการเก็บข้อมูลเพื่อให้สามารถนำเข้าแบบจำลองและผลลัพธ์ดียิ่งขึ้น

5.2.2. ควรมีการพิจารณาตัวแปร หรือปัจจัยอื่น ๆ เพิ่มเข้าไปอีกเพราะจะส่งผลให้ผลลัพธ์กับค่าพยากรณ์ ประสิทธิภาพมากยิ่งขึ้น เช่น ข้อมูลช่วงเวลาการส่งโปรโมชั่น ช่วงที่เก็บผลการกดดูแคมเปญมากกว่า 1 วัน หรือ แผนการพัฒนาสินค้าหรือผลิตภัณฑ์ของธุรกิจ และรูปแบบแนวทางของการตลาดในช่วงนั้น ๆ เป็นต้น

5.2.3. ควรมีการประเมินผลที่ได้อย่างสม่ำเสมอ ในการที่จะพิจารณาถึง ความเหมาะสม ความถูกต้อง และ ควรปรับเพิ่มลดตัวแปรอิสระเพิ่มเติมที่อาจช่วยเพิ่มประสิทธิภาพของการสร้างแบบจำลองมากขึ้น และลองนำเทคนิคใหม่ๆที่ถูกพัฒนามาปรับใช้เพื่อให้มีประสิทธิภาพมากยิ่งขึ้น หรือเพื่อลดเวลาของการดำเนินการของ “Feature Auto Model” สามารถดูผลประสิทธิภาพของงานวิจัยชิ้นนี้แล้วเลือกแบบจำลองให้น้อยลงเพื่อลดระยะเวลา

DPU DHURAKIJ PUNDIT UNIVERSITY มหาวิทยาลัยธุรกิจบัณฑิตย์

บรรณานุกรม

DPU DHURAKIJ PUNDIT UNIVERSITY มหาวิทยาลัยธุรกิจบัณฑิตย์

บรรณานุกรม

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

[2] C. bot et al.,” Cross-industry standard process for data mining: Revision history,” Wikipedia[Online]. https://en.wikipedia.org/wiki/Cross-industry_standard_process_for_data_mining. (Accessed: May. 20, 2023).

[3] SC Spark Solution Co.Ltd., “Business Analytics กับการใช้ CRISP-DM แก้ปัญหา,” [ออยนไลน์]. https://www.sc-sparksolution.com/2021/01/25/business-analytics กับการใช้-crisp-dm-แก้ปัญหา/. (Accessed: May. 20, 2023).

[4] N. sci et al.,” การเรียนรู้ของเครื่อง,” Wikipedia[Online]. https://th.wikipedia.org/wiki/การเรียนรู้ของเครื่อง. (Accessed: May. 20, 2023).

[5] J.J. Faraway et al.,” Generalized Linear Model,” ScienceDirect[Online]. https://www.sciencedirect.com/topics/mathematics/generalized-linear-model. (Accessed: May. 20, 2023).

[6] 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).

[7] J.J. 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).

[8] RapidMiner, Inc.,” Auto Model,” RapidMiner Document[Online]. https://docs.rapidminer.com/9.3/studio/auto-model/ . (Accessed: May. 20, 2023).

[9] RapidMiner, Inc., “Operators,” RapidMiner Document[Online]. https://docs.rapidminer.com/latest/studio/operators/. (Accessed: May. 20, 2023).

[10] S. Ray, “Naive Bayes Classifier Explained: Applications and Practice Problems of Naive Bayes Classifier,” analytics Vidhya [Online]. https://www.analyticsvidhya.com/blog/2017/09/naive-bayes-explained/. (Accessed: May. 20, 2023).

DPU DHURAKIJ PUNDIT UNIVERSITY มหาวิทยาลัยธุรกิจบัณฑิตย์

บรรณานุกรม (ต่อ)

[11] B. Alam, “Naive Bayes Classifier Python Tutorial,” [Online]. https://hands-on.cloud/naive-bayes-classifier-python-tutorial/. (Accessed: May. 20, 2023).

[12] P. Sarakarn,D. Jumparway, “Overview of longitudinal data, limitations, and appropriate alternatives of statistical methods: Longitudinal data analysis in health science research,” Journal of Medicine and Health Sciences, Vol. 28 No. 1,Apr. 2021. Accessed on: May. 22, 2023 [Online]. Available: https://he01.tci-thaijo.org/index.php/jmhs/article/view/249204.

[13] H. Nawaz, “Develop a Logistic Regression Machine Learning Model,” Medium[Online]. https://blog.devgenius.io/develop-a-logistic-regression-machine-learning-model-64d2be403ba3. (Accessed: May. 22, 2023).

[14] H.J. Shalin , “Applying Decision Trees to Online Learning Data,” [Online]. https://scalar.usc.edu/works/c2c-digital-magazine-fall-2017--winter-2018/applying-decision-trees-to-online-learning-data. (Accessed: May. 22, 2023).

[15] A. Bhandari, “Understanding & Interpreting Confusion Matrices for Machine Learning (Updated 2023),” analytics Vidhya [Online]. https://www.analyticsvidhya.com/blog/2020/04/confusion-matrix-machine-learning/. (Accessed: May. 22, 2023).

[16] 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).

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

DPU DHURAKIJ PUNDIT UNIVERSITY มหาวิทยาลัยธุรกิจบัณฑิตย์

บรรณานุกรม (ต่อ)

[18] 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).

[19] Thaiall et al., “การวัดประสิทธิภาพ,” [Online]. http://www.thaiall.com/research/efficiency.htm. (Accessed: May. 23, 2023).

[20] สุรวัชร ศรีเปารยะ และ สายชล สินสมบูรณทอง, “การเปรียบเทียบประสิทธิภาพวิธีการจำแนกกลุ่มการเปนโรคไตเรื้อรัง,” ภาควิชาสถิติ คณะวิทยาศาสตร์, สถาบันเทคโนโลยีพระจอมเกลาเจ้าคุณทหารลาดกระบัง, เขตลาดกระบัง, กรุงเทพฯ, 2560

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

[22] 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

[23] 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).

[24] ชุติมณฑน์ ภักดีสิโรตม์, ”การพยากรณ์ความต้องการสินค้าด้วยเทคนิคการเรียนรู้ของเครื่องในธุรกิจค้าปลีก,” สาขาวิชาวิศวกรรมข้อมูลขนาดใหญ่ , วิทยาลัยนวัตกรรมด้านเทคโนโลยีและวิศวกรรมศาสตร์, มหาวิทยาลัยธุรกจิบัณฑิตย์, กรุงเทพฯ 2564

[25] วัชรชัย อินธิปีก,”การพยากรณ์ความต้องการสินค้าและการวางแผนการผลิต กรณีศึกษา : โรงงานแปรรูปเนื้อเป็ด,”สาขาวิชาการจัดการทางวิศวกรรม ,วิทยาลัยนวัตกรรมด้านเทคโนโลยีและวิศวกรรมศาสตร์ , มหาวิทยาลัยธุรกิจบัณฑิตย์, กรุงเทพฯ 2561

DPU DHURAKIJ PUNDIT UNIVERSITY มหาวิทยาลัยธุรกิจบัณฑิตย์

ประวัติผู้เขียน

ชื่อ – นามสกุล ธัญลักษณ์ เอียดเจริญ

ประวัติการศึกษา

พ.ศ. 2559 ปริญญาตรี วิศวกรรมศาสตร์ มหาวิทยาลัยสงขลานครินทร์

ประสบการณ์ทำงาน

พ.ศ. 2566 เจ้าหน้าที่บริหาร สังกัด ฝ่ายเทคโนโลยีสารสนเทศและการสื่อสาร สำนักงานคณะกรรมการกำกับหลักทรัพย์และตลาดหลักทรัพย์

พ.ศ. 2565 Quality Assurance Engineer SHOPEE (THAILAND) COMPANY LIMITED

พ.ศ. 2562 Application Support Engineer KASIKORN Business - Technology Group (KBTG)

พ.ศ. 2560 Senior Call Centre Application Operation Total Access Communication PLC.(Dtac)