

ระบบทำนายแนวโน้มการยกเลิกใช้บริการสำหรับลูกค้าธนาคารโดยใช้
แบบจำลองการเรียนรู้ของเครื่อง

เฟาเซีย เกษตรกาลัม

สารนิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรวิศวกรรมศาสตรมหาบัณฑิต
สาขาวิชาวิศวกรรมข้อมูลขนาดใหญ่
วิทยาลัยนวัตกรรมด้านเทคโนโลยีและวิศวกรรมศาสตร์
มหาวิทยาลัยธุรกิจบัณฑิตย์
พ.ศ.2564

**BANKING-CUSTOMER CHURN PREDICTION SYSTEM
USING MACHINE LEARNING MODELS**

FAOZIA KASETKALA

The logo of Dhurakij Pundit University (DPU) is a large, stylized, light purple 'DPU' watermark. To the right of the letters is a circular emblem with a blue and white striped pattern.

**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**

2021



ใบรับรองงานสารนิพนธ์

วิทยาลัยนวัตกรรมการด้านเทคโนโลยีและวิศวกรรมศาสตร์ มหาวิทยาลัยธุรกิจบัณฑิตย์

ปริญญา วิศวกรรมศาสตรมหาบัณฑิต

หัวข้อสารนิพนธ์ ระบบทำนายแนวโน้มการยกเลิกใช้บริการสำหรับลูกค้าธนาคารโดยใช้แบบจำลองการเรียนรู้ของเครื่อง

เสนอโดย นางสาวเฟาเซีย เกษตรกาลาม์

สาขาวิชา วิศวกรรมข้อมูลขนาดใหญ่

อาจารย์ที่ปรึกษาสารนิพนธ์ ดร.ธนภัทร ช้างคะจิตร

ได้พิจารณาเห็นชอบโดยคณะกรรมการสอบสวนนิพนธ์แล้ว

.....ประธานกรรมการ
(ดร.สรรพฤทธิ์ มฤคทัต)

..... รศ.ดร. รศ.ดร. กรรมการและอาจารย์ที่ปรึกษา
(ดร.ธนภัทร ช้างคะจิตร)

.....เอกพ.....กรรมการ
(ดร.เอกสิทธิ์ พัทธวงค์ศักดิ์ดา)

วิทยาลัยนวัตกรรมการด้านเทคโนโลยีและวิศวกรรมศาสตร์รับรองแล้ว

(ดร.ชัยพร เปมะภาตะพันธ์)

คณะบดีวิทยาลัยนวัตกรรมการด้านเทคโนโลยีและวิศวกรรมศาสตร์

วันที่ 31 เดือน ธันวาคม พ.ศ. 2564

| | |
|------------------|---|
| หัวข้อสารนิพนธ์ | ระบบทำนายแนวโน้มการยกเลิกใช้บริการสำหรับลูกค้าธนาคาร โดยใช้แบบจำลองการเรียนรู้ของเครื่อง |
| ชื่อผู้เขียน | เฟาเซีย เกษตรกาลาม์ |
| อาจารย์ที่ปรึกษา | ดร. ธนภัทร มังคะจิตร |
| สาขาวิชา | วิศวกรรมข้อมูลขนาดใหญ่ |
| ปีการศึกษา | 2563 |

บทคัดย่อ

ความเสี่ยงด้านสภาพคล่องถือเป็นความเสี่ยงที่สำคัญที่สุดประเภทหนึ่งของธุรกิจภาคธนาคาร เนื่องจากธนาคารใช้การระดมทุนจากเงินฝากระยะสั้นมาเป็นเงินทุนในการให้สินเชื่อ ซึ่งมีระยะเวลาครบกำหนดยาวกว่าเงินฝาก ทำให้ไม่สามารถเปลี่ยนสินทรัพย์เป็นเงินสดได้ทันกับระยะเวลาครบกำหนดของหนี้สิน โดยเฉพาะอย่างยิ่งเมื่อเกิดการถอนเงินในอัตราที่สูงกว่าสัดส่วนปกติ

สารนิพนธ์ชิ้นนี้มีวัตถุประสงค์เพื่อพัฒนาแบบจำลองที่สามารถทำนายแนวโน้มการยกเลิกใช้บริการของลูกค้าเงินฝากประเภทเงินฝากประจำ เพื่อให้ธนาคารสามารถนำไปใช้ในการป้องกันก่อนเกิดภาวะขาดสภาพคล่อง โดยใช้การจำแนกประเภทข้อมูลด้วยเทคนิคการเรียนรู้ของเครื่องบนข้อมูลเงินฝากประจำ 3 เดือน ผลจากการทดสอบแบบจำลองพบว่า เทคนิค Random Forest ให้ความแม่นยำสูงสุดโดยมีค่า Recall โดยรวมอยู่ที่ 79% สำหรับกรณีบัญชีที่ปิดให้ค่า Recall อยู่ที่ 64% คิดเป็นจำนวนเงิน 99.67% ของจำนวนเงินที่ปิดบัญชีทั้งหมด โดยพบว่าจำนวนเดือนตั้งแต่เปิดบัญชีเป็นปัจจัยสำคัญที่สุด อย่างไรก็ตามเมื่อตัดปัจจัยนี้ออกไปพบว่ามี 3 ปัจจัยที่สำคัญได้แก่ จำนวนปีที่ลูกค้า ยอดรวมเงินฝากประจำ 3 เดือน และอายุของลูกค้า พบว่าจำนวนปีที่ลูกค้า ในช่วง 2 – 12 ปีมีผลอย่างมากต่อแนวโน้มในการยกเลิกใช้บริการ และส่งผลลดลงเมื่อจำนวนปีเพิ่มขึ้น ลูกค้าที่มีอายุในช่วง 30 – 55 ปีมีแนวโน้มสูงที่จะยกเลิกใช้บริการ ในขณะที่เมื่ออายุ 55 ปีขึ้นไปมีแนวโน้มการยกเลิกลดลง ยอดรวมเงินฝากประจำ 3 เดือน ส่งผลต่อแนวโน้มที่ลูกค้าจะยกเลิกเพิ่มขึ้นเมื่อมียอดเงินตั้งแต่ 800,000 บาทขึ้นไป ทั้งนี้ปัจจัยดังกล่าวข้างต้นสามารถนำไปใช้ในการออกนโยบายเพื่อจูงใจให้ลูกค้ากลุ่มนี้ยังคงใช้บริการและฝากเงินกับธนาคารต่อไปในอนาคต

| | |
|------------------------|---|
| A Thematic Paper Title | BANKING-CUSTOMER CHURN PREDICTION SYSTEM USING MACHINE LEARNING MODELS |
| Author | Faozia Kasetkala |
| Thesis Advisor | Dr. Thanapat Kangkachit |
| Department | Big Data Engineering |
| Academic Year | 2020 |

ABSTRACT

Liquidity risk is one of the most important risks in the banking sector since the short-term deposit is used to be the finance loan funding, which has a longer duration than the deposit. As a result, assets cannot be converted to cash in the maturity of the liabilities' time. With this business practice, liquidity shortages possibly happen, especially when withdrawals are proportionately higher than usual.

The objective of this study is to develop a model to predict the possibility of services cancellation of fixed deposit customers using machine learning techniques with data classification model on 3-month fixed deposit data. Therefore, the bank can apply it to prevent liquidity shortages. The churn model testing results show that the Random Forest technique provided the highest accuracy with an overall recall value of 79%. In the case of closed accounts, the recall is 64%, representing 99.67% of the total closing amount. We found that the most important feature is the number of months since account opening. Besides, if this feature is not considered, we found another three important features which are the number of years as a customer, the total amount of 3-month fixed deposit, and the customer's age. The results show that the range of 2 to 12 years of being a customer has a significant effect on the tendency to churn, and the chance of cancellation decrease as the number of years as a customer increase. Furthermore, the customers with age between 30 and 55 years old have a high possibility to churn, and decreases as the age increases. The total amount of 3-month fixed deposits affects the tendency of customers to churn more when the balance is 800,000 baht or higher. The aforementioned factors can be used in issuing a policy to motivate these groups of customers to further use the service and deposit money with the bank in the future.

กิตติกรรมประกาศ

สารนิพนธ์ฉบับนี้สำเร็จลุล่วงไปได้ด้วยดี จากการให้ความช่วยเหลือ ของ ดร.ชนภัทร มังคะจิตร ซึ่งเป็นอาจารย์ที่ปรึกษาสารนิพนธ์ ที่ได้กรุณาให้ความรู้ คำแนะนำ ตรวจสอบและแก้ไขข้อบกพร่องต่าง ๆ รวมทั้งคอยผลักดันและให้กำลังใจมาโดยตลอดจนกระทั่งสารนิพนธ์ฉบับนี้เสร็จสมบูรณ์ ผู้เขียนจึงขอขอบคุณอย่างสูงไว้ ณ โอกาสนี้

นอกจากนี้แล้วผู้เขียนขอขอบคุณ ผู้ช่วยศาสตราจารย์ ดร.วรพล พงษ์เพชร ดร.เอกสิทธิ์ พัทธวงษ์ศักดิ์ดา ดร.ปณิตา ฐุสรานนท์ อ.เฉลิมพล ศิริกายน และอาจารย์ทุกท่าน ที่ได้กรุณาให้ความรู้ ถ่ายทอดประสบการณ์และคำแนะนำต่างๆ ทำให้ผู้เขียนได้ใช้ความรู้ความสามารถเพื่อทำให้สารนิพนธ์ฉบับนี้สมบูรณ์ ขอขอบคุณนางสาวกุลธิดา รอดบุญ รวมถึงเจ้าหน้าที่บัณฑิตมหาวิทยาลัยธุรกิจบัณฑิต ที่ให้ความช่วยเหลือ อำนวยความสะดวกในด้านต่างๆตลอดเวลาในการศึกษาของผู้เขียน

ผู้เขียนขอขอบคุณเพื่อนนักศึกษาทุกท่านที่ช่วยเหลือ แบ่งปัน ตลอดระยะเวลาที่ได้เรียนร่วมกัน โดยเฉพาะอย่างยิ่งนายประพลเวท บุญประเสริฐ และ นางสาวนุสรา ศาสนะประดิษฐ์ ที่สนับสนุนทั้งในด้านวิชาการและด้านความบันเทิง ช่วยเป็นพลังให้ผู้เขียนสามารถจัดทำสารนิพนธ์ในครั้งนี้สำเร็จลุล่วงไปได้ด้วยดี

เฟาเซีย เกษตรกาลาม

สารบัญ

| | หน้า |
|---|------|
| บทคัดย่อภาษาไทย | ฅ |
| บทคัดย่อภาษาอังกฤษ..... | ง |
| กิตติกรรมประกาศ..... | จ |
| สารบัญ..... | ฉ |
| สารบัญตาราง..... | ช |
| สารบัญภาพ | ฌ |
| บทที่ | |
| 1. บทนำ..... | 1 |
| 1.1 ที่มาและความสำคัญของปัญหา..... | 1 |
| 1.2 วัตถุประสงค์ของงานวิจัย..... | 2 |
| 1.3 ขอบเขตงานวิจัย..... | 2 |
| 1.4 ประโยชน์ที่คาดว่าจะได้รับ..... | 2 |
| 1.5 นิยามศัพท์..... | 2 |
| 2. ทฤษฎี และผลงานวิจัยที่เกี่ยวข้อง..... | 3 |
| 2.1 Random Forest | 3 |
| 2.2 Gradient Boosting | 4 |
| 2.3 SVM..... | 5 |
| 2.4 Ensemble..... | 6 |
| 2.5 Imbalance Data..... | 6 |
| 2.6 งานวิจัยที่เกี่ยวข้อง | 7 |
| 3. วิธีวิจัย | 9 |
| 3.1 ความเข้าใจทางธุรกิจ (Business understanding)..... | 9 |
| 3.2 ความเข้าใจข้อมูลที่ใช้ในงาน (Data Understanding)..... | 10 |

สารบัญ (ต่อ)

| บทที่ | หน้า |
|---|------|
| 3.3 ตรวจสอบความถูกต้องของข้อมูลและเตรียมข้อมูลที่ใช้ในงาน (Data Preparation) | 14 |
| 3.4 การพัฒนา Model (Modeling) | 19 |
| 3.5 กระบวนการวัดประสิทธิภาพแบบจำลอง (Model Evaluation) | 21 |
| 3.6 กระบวนการ Deployment..... | 28 |
| 3.7 เครื่องมือที่ใช้ในการวิจัย..... | 28 |
| 4. ผลการศึกษา | 29 |
| 4.1 ผลการวัดประสิทธิภาพความถูกต้องของแบบจำลองด้วยข้อมูลทดสอบ..... | 29 |
| 4.2 ผลการวัดประสิทธิภาพความถูกต้องของแบบจำลองด้วยข้อมูล Unseen | 31 |
| 4.3 สรุปผลการเปรียบเทียบประสิทธิภาพ..... | 36 |
| 4.4 ผลการวัดความพึงพอใจของผู้ใช้งาน..... | 36 |
| 5. บทสรุป และข้อเสนอแนะ..... | 40 |
| 5.1 สรุปผลการศึกษา..... | 40 |
| 5.2 ข้อเสนอแนะ | 45 |
| บรรณานุกรม..... | 46 |
| ภาคผนวก..... | 48 |
| ก | 49 |
| ประวัติผู้เขียน | 51 |

| ตารางที่ | หน้า |
|---|------|
| 3.1 ข้อมูลระดับบัญชี..... | 11 |
| 3.2 ข้อมูลระดับลูกค้า..... | 11 |
| 3.3 ข้อมูลการทำรายการ..... | 13 |
| 3.4 ข้อมูลป้ายกำกับ..... | 13 |
| 3.5 ข้อมูลตัวแปรที่มีค่าว่าง..... | 15 |
| 3.6 แสดงการแปลงค่าของตัวแปร CORECUSTYPE | 18 |
| 3.7 แบบจำลอง : Base Line | 20 |
| 4.1 ผลการทดสอบประสิทธิภาพของแบบจำลองแยกตามป้ายกำกับ ข้อมูลชุดที่ 1... | 29 |
| 4.2 ผลการทดสอบประสิทธิภาพของแบบจำลอง ข้อมูลชุดที่ 1..... | 30 |
| 4.3 ผลการทดสอบประสิทธิภาพของแบบจำลองแยกตามป้ายกำกับ ข้อมูลชุดที่ 2... | 30 |
| 4.4 ผลการทดสอบประสิทธิภาพของแบบจำลอง ข้อมูลชุดที่ 2..... | 30 |
| 5.1 Confusion Matrix (Unseen ชุดที่ 1): จำนวนบัญชีและยอดเงินแบบจำลอง Random Forest..... | 41 |
| 5.2 Confusion Matrix (Unseen ชุดที่ 2) : จำนวนบัญชีและยอดเงินแบบจำลอง Random Forest..... | 41 |

| ภาพที่ | หน้า |
|---|------|
| 2.1 อธิบายหลักการของ Random Forest | 4 |
| 2.2 หลักการของ Gradient Boosting | 4 |
| 2.3 SVM 2 มิติ..... | 5 |
| 2.4 เปรียบเทียบการทำงาน Bagging และ Boosting | 6 |
| 3.1 กระบวนการ CRISP-DM | 9 |
| 3.2 สัดส่วนจำนวนบัญชี Churn : Not Churn | 13 |
| 3.3 ประเภทของข้อมูลและแสดงจำนวนข้อมูลที่มีค่า..... | 14 |
| 3.4 Boxplot สำหรับข้อมูลชุดที่ 1..... | 15 |
| 3.5 Boxplot สำหรับข้อมูลชุดที่ 2..... | 16 |
| 3.6 Boxplot ของตัวแปรที่มีค่าผิดปกติ..... | 16 |
| 3.7 Continuous Variable : P-Value < 0.05..... | 17 |
| 3.8 Categorical Variable : P-Value < 0.05..... | 18 |
| 3.9 กระบวนการแก้ปัญหาข้อมูลไม่สมดุล..... | 19 |
| 3.10 เปรียบเทียบผลจากแบบจำลอง..... | 20 |
| 3.11 ค่าพารามิเตอร์ที่เหมาะสมของ Random Forest..... | 20 |
| 3.12 ค่าพารามิเตอร์ที่เหมาะสมของ Gradient Boosting..... | 21 |
| 3.13 ค่าพารามิเตอร์ที่เหมาะสมของ SVM..... | 21 |
| 3.14 Confusion Matrix : Random Forest ข้อมูลชุดที่ 1..... | 22 |
| 3.15 Feature Importance : Random Forest ข้อมูลชุดที่ 1..... | 22 |
| 3.17 Feature Importance : Gradient Boosting ข้อมูลชุดที่ 1..... | 23 |
| 3.18 Confusion Matrix: SVM ข้อมูลชุดที่ 1..... | 24 |
| 3.19 ผลการวัดประสิทธิภาพของเทคนิค Ensemble ข้อมูลชุดที่ 1..... | 24 |
| 3.20 Confusion Matrix : Random Forest ข้อมูลชุดที่ 2..... | 25 |
| 3.21 Feature Importance : Random Forest ข้อมูลชุดที่ 2..... | 25 |
| 3.22 Confusion Matrix: Gradient Boosting ข้อมูลชุดที่ 2..... | 26 |
| 3.23 Feature Importance : Gradient Boosting ข้อมูลชุดที่ 2..... | 26 |

สารบัญภาพ (ต่อ)

| ภาพที่ | หน้า |
|--|------|
| 3.24 Confusion Matrix: SVM ข้อมูลชุดที่ 2..... | 27 |
| 3.25 ผลการวัดประสิทธิภาพของเทคนิค Ensemble ข้อมูลชุดที่ 2..... | 27 |
| 4.1 Confusion Matrix (Unseen ชุดที่ 1) : Random Forest..... | 31 |
| 4.2 Confusion Matrix (Unseen ชุดที่ 1) : Gradient Boosting..... | 32 |
| 4.3 Confusion Matrix (Unseen ชุดที่ 1) : SVM..... | 32 |
| 4.4 Confusion Matrix (Unseen ชุดที่ 1) : Voting Classifier..... | 33 |
| 4.5 Measurement Score (Unseen ชุดที่ 1) : Voting Classifier..... | 33 |
| 4.6 Confusion Matrix (Unseen ชุดที่ 2) : Random Forest..... | 34 |
| 4.7 Confusion Matrix (Unseen ชุดที่ 2) : Gradient Boosting..... | 34 |
| 4.8 Confusion Matrix (Unseen ชุดที่ 2) : SVM..... | 35 |
| 4.9 Confusion Matrix (Unseen ชุดที่ 2) : Voting Classifier..... | 35 |
| 4.10 Measurement Score (Unseen ชุดที่ 2) : Voting Classifier..... | 36 |
| 4.11 คำถามประเมินความพึงพอใจข้อที่ 1..... | 37 |
| 4.12 คำถามประเมินความพึงพอใจข้อที่ 2..... | 38 |
| 4.13 คำถามประเมินความพึงพอใจข้อที่ 3..... | 38 |
| 4.14 คำถามประเมินความพึงพอใจข้อที่ 4..... | 38 |
| 4.15 คำถามประเมินความพึงพอใจข้อที่ 5..... | 39 |
| 4.16 ความคิดเห็นเพิ่มเติม..... | 39 |
| 4.17 ข้อเสนอแนะ..... | 39 |
| 5.1 ทิศทางที่มีผลต่อแนวโน้มในการยกเลิกใช้บริการของ 3 ตัวแปรหลัก..... | 43 |
| 5.2 บัญชีที่ทำนายว่ามีแนวโน้มปิดบัญชีของแต่ละสาขา (Unseen ชุดที่ 1)..... | 44 |
| 5.3 บัญชีที่ทำนายว่ามีแนวโน้มปิดบัญชีของแต่ละสาขา (Unseen ชุดที่ 2)..... | 44 |
| 6.1 ตัวอย่างข้อมูลใน Input File..... | 49 |
| 6.2 หน้าจอการเรียกใช้ Model | 50 |
| 6.3 ตัวอย่างข้อมูลใน Output File..... | 50 |

บทที่ 1

บทนำ

1.1 ที่มาและความสำคัญของปัญหา

การดำเนินธุรกิจของธนาคาร มีความเกี่ยวข้องโดยตรงกับการบริหารความเสี่ยง ดังนั้นความสามารถในการบริหารความเสี่ยง จึงเป็นปัจจัยสำคัญต่อความสำเร็จในการดำเนินธุรกิจ ทั้งนี้ความเสี่ยงด้านสภาพคล่อง ถือเป็นความเสี่ยงที่สำคัญที่สุดประเภทหนึ่งของธุรกิจธนาคารไทย เนื่องจากโดยส่วนใหญ่แล้ว ธนาคารใช้วิธีการระดมทุนจากเงินฝากระยะสั้น เช่น เงินฝากประจำที่มีระยะเวลากำหนด 3 เดือน 6 เดือน หรือ 12 เดือน เป็นต้น ธนาคารใช้เงินทุนเหล่านี้ในการให้สินเชื่อซึ่งจะมีระยะเวลากำหนดยาวกว่าเงินฝาก การดำเนินธุรกิจลักษณะนี้ทำให้ธนาคารมีความเสี่ยงด้านสภาพคล่อง จากการที่ไม่สามารถเปลี่ยนแปลงสินทรัพย์ให้เป็นเงินสดได้ทันกับระยะเวลากำหนดของหนี้สิน

นอกจากนี้ปัญหาการยกเลิกการใช้บริการของลูกค้าถือเป็นปัญหาพื้นฐานในธุรกิจธนาคารซึ่งเป็นภาคธุรกิจที่มีการแข่งขันสูง ธนาคารเล็งเห็นถึงความสำคัญในการหาวิธีรักษาลูกค้าเดิม ซึ่งเป็นกลุ่มลูกค้าที่สามารถสร้างผลกำไรให้กับองค์กรได้ในระยะยาว อีกทั้งยังเป็นที่ทราบกันดีว่าต้นทุนที่ใช้ในการรักษาลูกค้าเดิมนั้นมีมูลค่าต่ำกว่าต้นทุนในการแสวงหาลูกค้าใหม่ ด้วยเหตุผลด้านการบริหารความเสี่ยงและการสร้างผลกำไรในทางธุรกิจตามกล่าวนมาข้างต้น ธนาคารจำเป็นต้องมีเครื่องมือที่ช่วยให้สามารถคาดการณ์ความเปลี่ยนแปลงของกระแสเงินสด โดยเฉพาะในกรณีที่เกิดจากการถอนเงินและยกเลิกการใช้บริการของลูกค้า

งานวิจัยนี้มีวัตถุประสงค์เพื่อพัฒนาแบบจำลองที่สามารถทำนายโอกาสในการยกเลิกการใช้บริการของลูกค้าเงินฝากประจำของธนาคาร เพื่อให้ธนาคารสามารถนำไปใช้ในการป้องกันก่อนเกิดเหตุการณ์ อีกทั้งเพื่อช่วยให้ธนาคารสามารถเตรียมการจัดหาแหล่งเงินในกรณีที่มีเหตุการณ์กระแสเงินสดไหลออกมากเกินไปเกินอัตราที่ยอมรับได้ ข้อมูลที่ใช้แบ่งออกเป็นกลุ่มได้ดังนี้ ข้อมูลระดับลูกค้า ข้อมูลระดับบัญชี และข้อมูลระดับ transaction โดยมีปัจจัยที่สำคัญคือ จำนวนเดือนตั้งแต่เปิดบัญชี จำนวนปีที่เงินลูกค้า ยอดรวมเงินฝากประจำ 3 เดือน และอายุของลูกค้า

1.2 วัตถุประสงค์ของงานวิจัย

1.2.1 เพื่อนำเสนอแบบจำลองสำหรับพยากรณ์แนวโน้มการยกเลิกใช้บริการของลูกค้าเงินฝากประเภทเงินฝากประจำ

1.2.2 เพื่อหาปัจจัยที่มีความสัมพันธ์ต่อแนวโน้มในการยกเลิกใช้บริการด้านเงินฝากประจำเพื่อนำไปสู่การออกนโยบายที่เหมาะสม

1.3 ขอบเขตงานวิจัย

1.3.1 เงินฝากประจำประเภท 3 เดือนธนาคารอิสลามแห่งประเทศไทย

1.3.2 พัฒนาแบบจำลองคาดการณ์โอกาสการยกเลิกใช้บริการเงินฝากประจำโดยคาดการณ์ล่วงหน้า 1 เดือน

1.4 ประโยชน์ที่คาดว่าจะได้รับ

1.4.1 เป็นเครื่องมือที่ช่วยให้ธนาคารสามารถรักษาลูกค้าที่มีแนวโน้มจะเลิกใช้บริการเงินฝากประจำกับธนาคาร

1.4.2 ช่วยให้ธนาคารสามารถบริหารจัดการความเสี่ยงด้านสภาพคล่องได้อย่างมีประสิทธิภาพ มากขึ้น

1.5 นิยามศัพท์

1.5.1 Customer churn หมายถึง ลูกค้าที่ปิดบัญชีเงินฝากประจำประเภท 3 เดือน

1.5.2 Imbalance data หมายถึง ข้อมูลคำตอบของแต่ละคลาสมีจำนวนแตกต่างกันมาก

1.5.3 Classification หมายถึง แบบจำลองที่ต้องมี ป้ายกำกับ หรือตัวแปรที่ใช้วัดเป้าหมาย เป็นตัวตั้งต้นให้เรียนรู้

1.5.4 True Positive หมายถึง สิ่ง que แบบจำลองทำนายว่า “จริง” และมีค่าเป็น “จริง”

1.5.5 True Negative หมายถึง สิ่ง que แบบจำลองทำนายว่า “ไม่จริง” และมีค่า “ไม่จริง”

1.5.6 False Positive หมายถึง สิ่ง que แบบจำลองทำนายว่า “จริง” แต่มีค่าเป็น “ไม่จริง”

1.5.7 False Negative หมายถึง สิ่ง que แบบจำลองทำนายว่า “ไม่จริง” แต่มีค่าเป็น “จริง”

1.5.8 Precision หมายถึง ค่าความแม่นยำ เป็นการเปรียบเทียบ การทำนายที่ถูกต้องว่า จริง และ que เกิดขึ้นจริง (TP) กับ การทำนายว่า จริง แต่สิ่ง que เกิดขึ้น คือ ไม่จริง (FP)

1.5.9 Recall หมายถึง ความถูกต้องของการทำนายว่าจะเป็น “จริง” เทียบกับ จำนวนครั้งของเหตุการณ์ทั้งทำนาย และ เกิดขึ้น ว่า “เป็นจริง”

1.5.10 F1-Score หมายถึง ค่าเฉลี่ยแบบ harmonic mean ระหว่าง precision และ recall

บทที่ 2

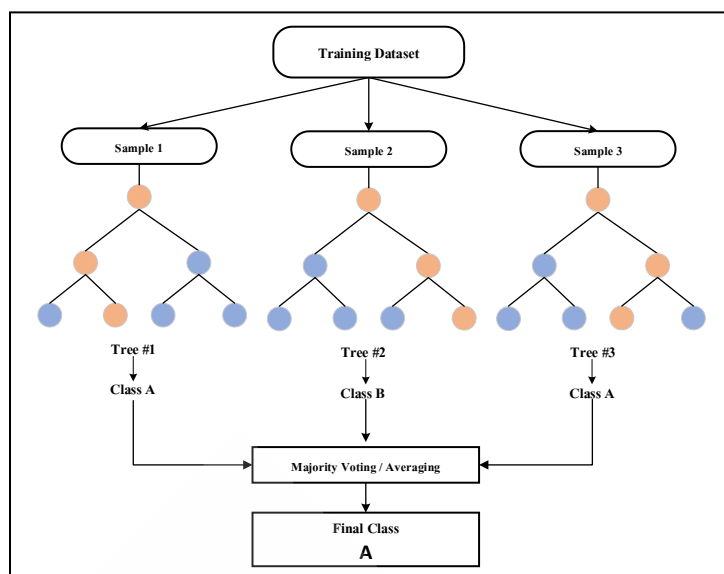
ทฤษฎีและผลงานวิจัยที่เกี่ยวข้อง

การพัฒนาแบบจำลองสำหรับพยากรณ์แนวโน้มการยกเลิกใช้บริการของลูกค้าเงินฝากประเภทเงินฝากประจำ ต้องศึกษาเอกสารและงานวิจัยที่เกี่ยวข้อง ดังรายการต่อไปนี้

- 2.1 Random Forest
- 2.2 Gradient Boosting
- 2.3 SVM
- 2.4 Ensemble
- 2.5 Imbalance Data
- 2.6 งานวิจัยที่เกี่ยวข้อง

2.1 Random Forest

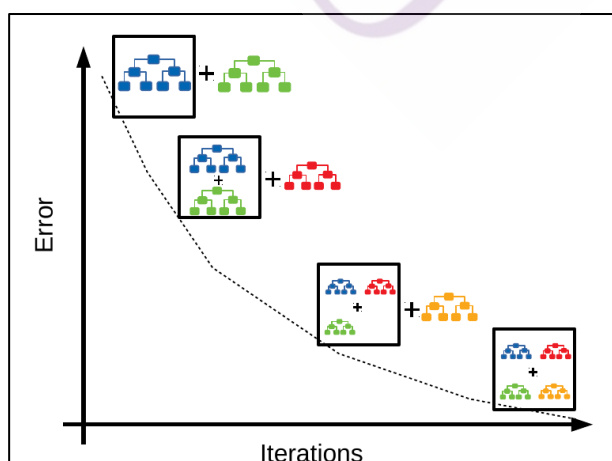
หลักการของ Random Forest คือ สร้างแบบจำลองจาก Decision Tree หลายๆ แบบจำลอง แต่ละแบบจำลองย่อยจะถูกสอนจากข้อมูลชุดเดียวกัน แต่ใช้ข้อมูลย่อยที่แตกต่างกัน โดยข้อมูลย่อยที่ถูกใช้ในการสอนแบบจำลองได้มาจากการสุ่ม ทั้งนี้คำตอบที่ได้จากการทำนายของ Random Forest จะเกิดจากการเฉลี่ยหรือการโหวตเลือกผลลัพธ์จาก Decision Tree แล้วแต่ว่าเป็นการทำนายความน่าจะเป็นหรือทำนายประเภทข้อมูล



ภาพที่ 2.1 อธิบายหลักการของ Random Forest

2.2 Gradient Boosting

หลักการของ Gradient Boosting คือ สร้างแบบจำลองจาก Decision Tree หลายๆ แบบจำลองทำงานร่วมกัน โดย อินพุตของแบบจำลองหนึ่งจะมาจากเอาต์พุตของแบบจำลองก่อนหน้า โดยแนวคิดคือ Gradient Boosting จะทำการสร้าง Tree เพื่อลดค่า Error ที่เกิดจาก Tree ก่อนหน้าด้วยวิธี Gradient Descend



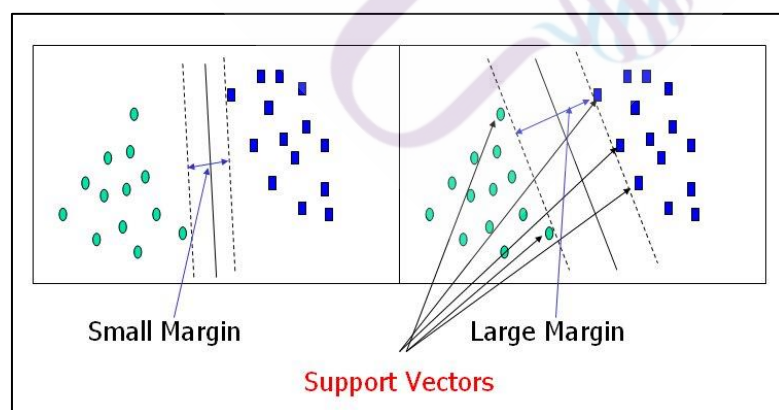
ภาพที่ 2.2 หลักการของ Gradient Boosting

ที่มา: <https://morioh.com/p/e108a4521555>

2.3 SVM

หลักการของ SVM คือการให้อินพุตที่ใช้ฝึกเป็นเวกเตอร์ในสเปซ N มิติ เช่นถ้าในกรณีของ 2 มิติ และ 3 มิติ จะเป็นจุดที่อยู่ในระนาบ xy และสเปซ xyz ตามลำดับ จากนั้นทำการสร้างไฮเปอร์เพลน(Hyperplane) ที่จะแยกกลุ่มของเวกเตอร์อินพุตออกเป็นประเภทต่างๆ ในกรณีที่ เป็น 2 มิติ และ 3 มิติ ไฮเปอร์เพลน คือเส้นตรงและระนาบตามลำดับ ข้อเด่นของ SVM คือสามารถแปลงข้อมูลที่ไม่สามารถแบ่งกลุ่มได้ด้วยสมการเชิงเส้น โดยจะทำการเก็บแมพ (Map) เวกเตอร์ในสเปซอินพุตให้เข้าสู่ Feature Space โดยใช้เคอร์เนล (kernel) ชนิดต่างๆ เช่น โพลีโนเมียล (Polynomial) เรเดียล (Radial) เป็นต้น ใน Feature Space ดังกล่าวเวกเตอร์อินพุต สามารถแยกประเภทได้โดยไฮเปอร์เพลน

กรณีที่สามารถแบ่งข้อมูลได้ด้วยเส้นตรง สามารถใช้ linear algebra ในการสร้างแบบจำลอง โดยให้นิยาม Margin เป็นผลรวมระยะห่างของเส้นตรงที่เป็นไฮเปอร์เพลน ถึงเส้นตรงที่ผ่านอินพุตที่ ใกล้ที่สุดและขนานกับไฮเปอร์เพลน ของทั้งสองกลุ่ม ระยะดังกล่าวอาจมองเป็นเวกเตอร์และมีชื่อว่า ซัพพอร์ตเวกเตอร์แมชชีน (Support Vector Machine) อัลกอริทึม SVM จะเลือกไฮเปอร์เพลนที่ให้ค่า Margin มีค่าสูงสุด ดังแสดงในภาพที่ 2.3



ภาพที่ 2.3 SVM 2 มิติ

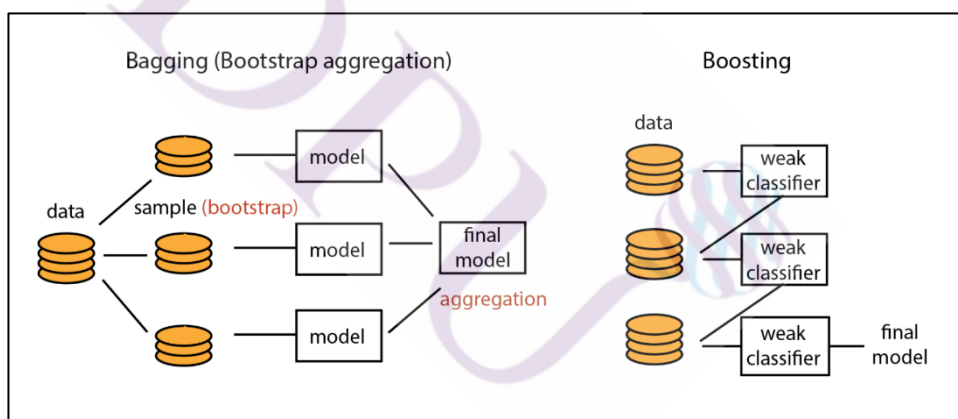
ที่มา: <http://kokzard.blogspot.com/2011/10/jfjkdshfkjsldf.html>

2.4 Ensemble

คือการที่นำแบบจำลองหลายๆแบบ ซึ่งสร้างมาจากข้อมูลชุดเดียวกัน มารวมกัน เพื่อหาคำตอบด้วยวิธีการหาค่าเฉลี่ย (averaging) หรือการเลือกจากเสียงข้างมาก (majority vote) โดยมีเทคนิคที่นิยมใช้ ดังนี้

2.4.1 Bagging (bootstrap aggregation) คือการสุ่มตัวอย่างข้อมูลออกมาแล้วสร้าง classifier ขึ้นมา โดยใช้วิธีการสุ่มแบบแทนที่ (random with replacement) นั่นคือข้อมูลที่มียังคงเหมือนเดิม ไม่ลดลงหลังจากการสุ่ม สามารถสุ่มข้อมูลหลายๆรอบเพื่อให้ได้ classifier หลายๆตัว ในการทำนายจะใช้ classifiers ทุกตัวที่ถูกสร้างขึ้นมาเพื่อทำนายชุดข้อมูลใหม่ วิธีการทำนายมีทั้งแบบหาค่าเฉลี่ยและการโหวต ขึ้นอยู่กับว่าต้องการทำนายความน่าจะเป็นหรือทำนายประเภทข้อมูล

2.4.2 Boosting คือการนำ weak classifier หรือ classifier ที่มีความแม่นยำต่ำมา ทำนายข้อมูลที่มี จากนั้นจะให้ weak classifier ตัวใหม่มาแก้ไข error โดยผลรวมของ classifier จะเกิดเป็น classifier ใหม่ขึ้นมา ทำแบบนี้ไปเรื่อยๆจนได้แบบจำลองที่ดีที่สุดจากผลรวมของ classifier



ภาพที่ 2.4 เปรียบเทียบการทำงาน Bagging และ Boosting

ที่มา: <http://analyticsth.blogspot.com/2015/09/ensemble-method.html>

2.5 Imbalance Data

ข้อมูลไม่สมดุลคือ ข้อมูลที่มีจำนวนในแต่ละกลุ่มแตกต่างกันมาก ซึ่งจะมีผลต่อการจำแนกคลาสส่วนน้อย เนื่องจากอัลกอริทึมมีความเอนเอียงไปทางกลุ่มข้อมูลส่วนใหญ่ ทำให้ทำนายคลาสส่วนน้อยผิดพลาด เทคนิคที่นำมาใช้ในการแก้ปัญหาข้อมูลไม่สมดุลมีดังนี้

2.5.1 วิธีสุ่มเกิน (Over sampling) คือ การสุ่มเลือกข้อมูลจากคลาสส่วนน้อย เพื่อให้คลาสส่วนน้อยมีจำนวนใกล้เคียงกับคลาสส่วนมาก โดยการสุ่มเกินนี้อาจเป็นการสุ่มจากข้อมูลเดิมหรือสร้างข้อมูลขึ้นมาใหม่จากตัวอย่างข้อมูลเดิม

SMOTE เป็นเทคนิคการสุ่มเกินอีกรูปแบบหนึ่ง โดยจะมีการสุ่มข้อมูลที่ได้จากการสร้างข้อมูลจากข้อมูลส่วนน้อยขึ้นมาใหม่

2.5.2 วิธีสุ่มลด (Under sampling) คือ การสุ่มลดข้อมูลจากคลาสส่วนมาก เพื่อให้คลาสส่วนมากมีจำนวนใกล้เคียงกับคลาสส่วนน้อย

2.6 งานวิจัยที่เกี่ยวข้อง

งานวิจัยที่เกี่ยวกับการพัฒนาแบบจำลองสำหรับพยากรณ์แนวโน้มการยกเลิกใช้บริการของลูกค้าเงินฝากประเภทเงินฝากประจำ และที่เกี่ยวข้องการแก้ปัญหาข้อมูลไม่สมดุล ที่ได้ศึกษามาสรุปได้ดังนี้

จิรกฤต บุญหมื่นไวย, เจษฎา ตันทนุช, เบญจวรรณ โรจนดิษฐ์. (2563) งานวิจัยนี้ศึกษาความสัมพันธ์ของตัวแปรที่มีผลต่อการยกเลิกใช้บริการโดยใช้การถดถอยโลจิสติกส์ จากนั้นนำตัวแปรที่ได้ไปสร้างแบบจำลองด้วยวิธีการ ต้นไม้ตัดสินใจ การถดถอยโลจิสติกส์ และ ตัวจำแนกประเภทแบบเบย์อย่างง่าย

Abbas Keramati, Hajar Ghaneei and Seyed Mohammad Mirmohammadi. (2016) งานวิจัยนี้ศึกษาปัจจัยที่มีผลในการยกเลิกการให้บริการธนาคารอิเล็กทรอนิกส์ ใช้วิธีการ CRISP ในการดำเนินการวิจัย เลือกใช้วิธีการ forward selection และ backward elimination ในขั้นการเลือกตัวแปรโดยผลการวิจัยพบว่า backward elimination ให้ผลลัพธ์ที่ดีกว่า แบบจำลองที่ใช้คือต้นไม้ตัดสินใจ เนื่องจากผู้วิจัยต้องการทราบคุณลักษณะของผู้ใช้บริการที่จะยกเลิก จึงเลือกใช้แบบจำลองที่สามารถอธิบายปัจจัยที่เป็นที่มีผลได้ง่าย

Nelson Rosa. (2018) งานวิจัยนี้ใช้เทคนิคโครงข่ายประสาทเทียมในการทำนายลูกค้าที่จะยกเลิกบริการของธนาคารรายย่อยในประเทศโปรตุเกส โดยใช้วิธีการ CRISP-DM ในการดำเนินการวิจัย โดยใช้ข้อมูล มกราคม-มิถุนายน 2017 เป็นตัวแปรในการสร้างแบบจำลอง เพื่อทำนายลูกค้าที่จะปิดบริการในช่วง กรกฎาคม - ธันวาคม 2017

Irfan Ullah, Basit Raza, Ahmad Kamran Malik, Muhammad Imran, Saif Ul islam, AND Sung Won Kim. (2019) งานวิจัยนี้ใช้ Random Forest ในการทำนายลูกค้าในธุรกิจโทรคมนาคม จากนั้นนำลูกค้าที่ทำนายว่าจะยกเลิกบริการไปทำ clustering เพื่อแบ่งกลุ่มลูกค้าตามพฤติกรรม เนื่องจากต้องการหาสาเหตุที่ลูกค้ายกเลิกบริการ เพื่อนำไปกำหนดกลยุทธ์ด้าน CRM

Francesco Pochetti. (2019) บทความนี้นำเสนอเทคนิคในการแก้ปัญหาข้อมูลไม่สมดุล ด้วยเทคนิค Ensembling + Oversampling มีวิธีการคือแบ่งข้อมูลคลาสส่วนใหญ่ออกเป็น N ส่วน นำข้อมูลที่แบ่งออกมาแต่ละส่วนรวมกับข้อมูลคลาสส่วนน้อย ได้ข้อมูลชุดสอนใหม่ N ชุด แต่ละชุดทำการสุ่มลด/สุ่มเกิน ด้วยการสุ่ม นำข้อมูลแต่ละชุดเข้าแบบจำลอง นำผลลัพธ์จากแต่ละแบบจำลอง มาหาค่าเฉลี่ย

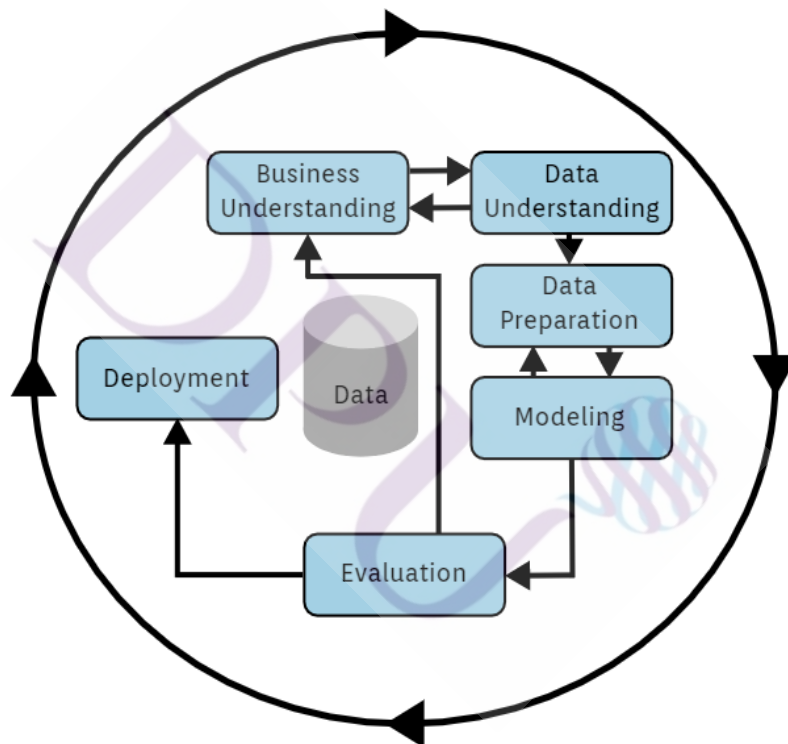
Himanshu147. (2020) บทความนี้นำเสนอโค้ดไล่น์ในการพัฒนาแบบจำลองเพื่อทำนายลูกค้าที่จะยกเลิกใช้บริการเงินฝากกระแสรายวัน โดยใช้การถดถอยโลจิสติกส์ในการสร้างแบบจำลอง ใช้ SMOTE ในการแก้ปัญหาข้อมูลไม่สมดุล



บทที่ 3

วิธีวิจัย

การศึกษาวิจัยครั้งนี้ เป็นการนำเสนอเทคนิคทำนายแนวโน้มการยกเลิกใช้บริการ สำหรับลูกค้าธนาคารโดยวิธีการเรียนรู้ของเครื่อง ใช้วิธีการ CRISP-DM ในกระบวนการทำวิจัย ซึ่งประกอบด้วย 6 กระบวนการดังภาพ



ภาพที่ 3.1 กระบวนการ CRISP-DM

3.1 ความเข้าใจทางธุรกิจ (Business Understanding)

การให้บริการด้านเงินฝากของธนาคารสามารถแบ่งประเภทเงินฝากออกเป็น 2 ประเภทใหญ่ๆ ได้ดังนี้

เงินฝากเพื่อเรียก

คือเงินฝากที่ผู้ฝากสามารถถอนได้เมื่อต้องการ โดยไม่มีเงื่อนไข แบ่งย่อยออกได้เป็นเงินฝากกระแสรายวัน และเงินฝากออมทรัพย์

เงินฝากประจำ

คือเงินฝากที่มีกำหนดเวลา เช่น ครอบกำหนด 3,9 หรือ 12 เดือนเป็นต้น เมื่อฝากครบกำหนดเวลาผู้ฝากจะได้ผลตอบแทนตามเงื่อนไขที่ตกลงกัน ในกรณีที่ถอนก่อนครบกำหนด ผู้ฝากอาจไม่ได้รับผลตอบแทนเต็มจำนวน ขึ้นอยู่กับเงื่อนไขที่ตกลงกันได้

ในการให้สินเชื่อกับลูกค้า โดยส่วนใหญ่แล้วธนาคารจะนำเงินฝากประจำมาบริหารจัดการเนื่องจากมีกำหนดเวลาครบกำหนดที่แน่นอน สามารถบริหารและควบคุมความเสี่ยงได้ ด้วยลักษณะของระยะเวลาครบกำหนดเงินฝากที่สั้นกว่าสินเชื่อมาก ทำให้การติดตามปริมาณการไหลออกของเงินฝากซึ่งสามารถเกิดได้ทั้งกรณีที่ครบกำหนดสัญญาและการถอนก่อนครบกำหนด เป็นภาระกิจสำคัญเพื่อติดตามให้สัดส่วนของเงินฝากและสินเชื่อเป็นไปตามเกณฑ์ของธนาคารแห่งประเทศไทย

เพื่อให้ธุรกิจดำเนินได้อย่างมีประสิทธิภาพและลดต้นทุนในการต้องหาลูกค้ารายใหม่หรือการกู้เงินในกรณีที่สัดส่วนเงินฝากต่อสินเชื่ออยู่ภาวะที่วิกฤติ ธนาคารจึงมีนโยบายให้สาขาดูแลติดตามลูกค้าที่คาดว่าจะมีการถอนเงินเพื่อชะลอหรือยับยั้งการเบิกถอนเงินออกไป

ดังนั้นเพื่อเพิ่มประสิทธิภาพในการทำงานให้สามารถเข้าถึงลูกค้าเป้าหมายได้อย่างถูกต้อง การมีแบบจำลองสำหรับพยากรณ์แนวโน้มการยกเลิกใช้บริการของลูกค้าเงินฝากประเภทฝากประจำจึงมีความจำเป็น

3.2 ความเข้าใจข้อมูลที่ใช้ในงาน (Data Understanding)

3.2.1 กระบวนการเก็บข้อมูลเงินฝากประจำ

ข้อมูลที่ใช้ประกอบด้วย ข้อมูลระดับบัญชีและข้อมูลระดับลูกค้า ของบัญชีเงินฝากที่ยังมีสถานะเปิดบัญชี ณ 30 พฤษภาคม 2563 สำหรับเป็นตัวแปรในการสร้างแบบจำลอง และข้อมูลสถานะบัญชี ณ 31 ธันวาคม 2563 สำหรับเป็นข้อมูลป้ายกำกับ โดยตัวแปรที่นำมาใช้สามารถแบ่งออกเป็น 4 กลุ่ม ดังนี้

- 1) ข้อมูลระดับบัญชี ได้แก่ ยอดเงินคงเหลือ, กำไรค้างรับ, อัตรากำไร, จำนวนเดือนที่จะครบสัญญา, จำนวนเดือนตั้งแต่เปิดบัญชี, ภูมิภาคของสาขาเจ้าของบัญชี
- 2) ข้อมูลระดับลูกค้า เช่น อายุ, จำนวนปีที่เป็ลูกค้า, ประเภทลูกค้า, จำนวนบัญชีเงินฝากประจำ เป็นต้น
- 3) ข้อมูลระดับ Transaction ได้แก่ ยอดรวมรายการถอนเงินของเดือนก่อนหน้า และยอดกำไรจ่ายของปีก่อนหน้า

4) ข้อมูลป้ายกำกับ ได้แก่ CHURN

0 หมายถึง Not Churn

1 หมายถึง Churn

รายละเอียดข้อมูลทั้ง 4 กลุ่มดังแสดงในตารางที่ 3.1 – 3.4

ตารางที่ 3.1 ข้อมูลระดับบัญชี

| อันดับ | ชื่อตัวแปร | ความหมาย | ตัวอย่าง |
|--------|----------------|----------------------------|--------------|
| 1 | MEBALLM | ยอดเงินคงเหลือ | 8,000 บาท |
| 2 | ACR | กำไรค่ารับ | 58.73957 บาท |
| 3 | IRN | อัตรากำไร | 0.8 |
| 4 | DEP_M_LIFETIME | จำนวนเดือนตั้งแต่เปิดบัญชี | 209 |
| 5 | MONTHTOMATURE | จำนวนเดือนที่จะครบสัญญา | 2 |
| 6 | REGION | ภูมิภาคของสาขาเจ้าของบัญชี | B360 |

ตารางที่ 3.2 ข้อมูลระดับลูกค้า

| อันดับ | ชื่อตัวแปร | ความหมาย | ตัวอย่าง |
|--------|--------------|-----------------------------|---------------------------|
| 1 | AGE | อายุลูกค้า | 8,000 บาท |
| 2 | CUSTOMER_AGE | จำนวนปีที่ เป็นลูกค้า | 58.73957 บาท |
| 3 | CORECUSTYPE | กลุ่มลูกค้าตามยอดเงินฝากรวม | Retail, Medium, Corporate |
| 4 | INVP | ประเภทลูกค้า | บุคคล, นิติบุคคล |
| 5 | GENDER | เพศ | 2 |
| 6 | INCOME | ช่วงรายได้ | 15,000 - 29,999 |
| 7 | INCOMESCR | แหล่งที่มาของรายได้ | SALARY, BUSINESS |
| 8 | EDUCATION | ระดับการศึกษา | CONDARY |
| 9 | OCCUPATON | อาชีพ | SELF-EMPLOYED |
| 10 | TRM3_CNT | จำนวนบัญชีฝากประจำ 3 เดือน | 10 , 2 เป็นต้น |

ตารางที่ 3.2 (ต่อ)

| อันดับ | ชื่อตัวแปร | ความหมาย | ตัวอย่าง |
|--------|------------------|---------------------------------------|--------------------|
| 11 | TRM6_CNT | จำนวนบัญชีฝากประจำ 6 เดือน | 10 , 2 เป็นต้น |
| 12 | TRM9_CNT | จำนวนบัญชีฝากประจำ 9 เดือน | 10 , 2 เป็นต้น |
| 13 | TRM12_CNT | จำนวนบัญชีฝากประจำ 12 เดือน | 10 , 2 , 0 เป็นต้น |
| 14 | TRM24_CNT | จำนวนบัญชีฝากประจำ 24 เดือน | 10 , 2 , 0 เป็นต้น |
| 15 | TOTAL_TIME_CNT | จำนวนบัญชีฝากประจำทั้งหมด | 10 , 2 , 0 เป็นต้น |
| 16 | TOTAL_SAVING_CNT | จำนวนบัญชีฝากออมทรัพย์ทุกประเภท | 10 , 2 , 0 เป็นต้น |
| 17 | TRM3_BAL | ยอดเงินฝากรวมของเงินฝากประจำ 3 เดือน | 12345.54 บาท |
| 18 | TRM6_BAL | ยอดเงินฝากรวมของเงินฝากประจำ 6 เดือน | 12345.54 บาท |
| 19 | TRM9_BAL | ยอดเงินฝากรวมของเงินฝากประจำ 9 เดือน | 12345.54 บาท |
| 20 | TRM12_BAL | ยอดเงินฝากรวมของเงินฝากประจำ 12 เดือน | 12345.54 บาท |
| 21 | TRM24_BAL | ยอดเงินฝากรวมของเงินฝากประจำ 24 เดือน | 12345.54 บาท |
| 22 | TOTAL_TIME_BAL | เงินฝากรวมของประจำ | 12345.54 บาท |
| 23 | TOTAL_SAVING_BAL | เงินฝากรวมของออมทรัพย์ | 12345.54 บาท |

ตารางที่ 3.3 ข้อมูลการทำรายการ

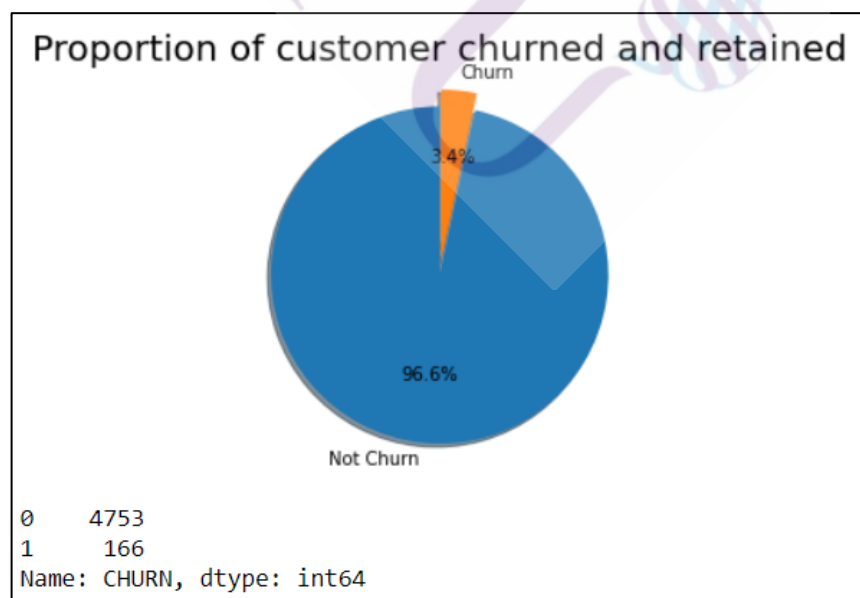
| อันดับ | ชื่อตัวแปร | ความหมาย | ตัวอย่าง |
|--------|------------|--|--------------|
| 1 | TOT_PDEBIT | ยอดรวมการถอนเงินเดือนก่อนหน้าของลูกค้า | 8,000 บาท |
| 2 | IYTD | กำไรจ่ายปีก่อนหน้า | 58.73957 บาท |

ตารางที่ 3.4 ข้อมูลป้ายกำกับ

| อันดับ | ชื่อตัวแปร | ความหมาย | ตัวอย่าง |
|--------|------------|-----------|--------------------------|
| 1 | CHURN | ป้ายกำกับ | 0 - Not Churn, 1 - Churn |

3.2.2 ทำความเข้าใจภาพรวมจำนวนบัญชี สัดส่วนของบัญชีที่มีป้ายกำกับเป็น 0 และ 1

จากภาพ 3.2 ข้อมูลมีทั้งสิ้น 4,919 ตัวอย่าง แบ่งออกเป็นบัญชีที่ Churn 166 ตัวอย่าง และบัญชีที่ Not Churn 4,753 ตัวอย่าง คิดเป็นสัดส่วนบัญชีที่ Churn ต่อบัญชีทั้งหมดอยู่ที่ 3.4% ข้อมูลมีลักษณะไม่สมดุลค่อนข้างมาก



ภาพที่ 3.2 สัดส่วนจำนวนบัญชี Churn : Not Churn

3.3 ตรวจสอบความถูกต้องของข้อมูลและเตรียมข้อมูลที่ใช้ในงาน (Data Preparation)

3.3.1 กระบวนการตรวจสอบความครบถ้วนของข้อมูล (Data Exploration)

ตรวจสอบหาข้อมูลที่มีค่าว่าง และประเภทของข้อมูล ดังภาพที่ภาพที่ 3.3 โดยพบว่าข้อมูลที่มีค่าว่างประกอบด้วยตัวแปรตามตาราง ที่ 3.5

| # | Column | Non-Null | Count | Dtype |
|----|------------------|---------------|-------|---------|
| 0 | CID | 4919 non-null | | int64 |
| 1 | MEBALLM | 4919 non-null | | float64 |
| 2 | ACR | 4919 non-null | | float64 |
| 3 | IRN | 4919 non-null | | float64 |
| 4 | IYTD | 4919 non-null | | float64 |
| 5 | DEP_M_LIFETIME | 4919 non-null | | int64 |
| 6 | MONTHTOMATURE | 4919 non-null | | int64 |
| 7 | TRM3_CNT | 4919 non-null | | int64 |
| 8 | TRM6_CNT | 4919 non-null | | int64 |
| 9 | TRM9_CNT | 4919 non-null | | int64 |
| 10 | TRM12_CNT | 4919 non-null | | int64 |
| 11 | TRM24_CNT | 4919 non-null | | int64 |
| 12 | TOTAL_TIME_CNT | 4919 non-null | | int64 |
| 13 | TOTAL_SAVING_CNT | 4919 non-null | | int64 |
| 14 | TRM3_BAL | 4919 non-null | | float64 |
| 15 | TRM6_BAL | 4919 non-null | | float64 |
| 16 | TRM9_BAL | 4919 non-null | | int64 |
| 17 | TRM12_BAL | 4919 non-null | | float64 |
| 18 | TRM24_BAL | 4919 non-null | | float64 |
| 19 | TOTAL_TIME_BAL | 4919 non-null | | float64 |
| 20 | TOTAL_SAVING_BAL | 4919 non-null | | float64 |
| 21 | TOT_PDEBIT | 1550 non-null | | float64 |
| 22 | AGE | 4919 non-null | | int64 |
| 23 | CUSTOMER_AGE | 4919 non-null | | int64 |
| 24 | CORECUSTYPE | 4919 non-null | | object |
| 25 | INVP | 4919 non-null | | object |
| 26 | GENDER | 4730 non-null | | object |
| 27 | INCOME | 4520 non-null | | object |
| 28 | INCOMESCR | 4919 non-null | | object |
| 29 | EDUCATION | 4520 non-null | | object |
| 30 | OCCUPATON | 4836 non-null | | object |
| 31 | REGION | 4919 non-null | | object |
| 32 | CHURN | 4919 non-null | | int64 |

dtypes: float64(11), int64(14), object(8)
memory usage: 1.2+ MB

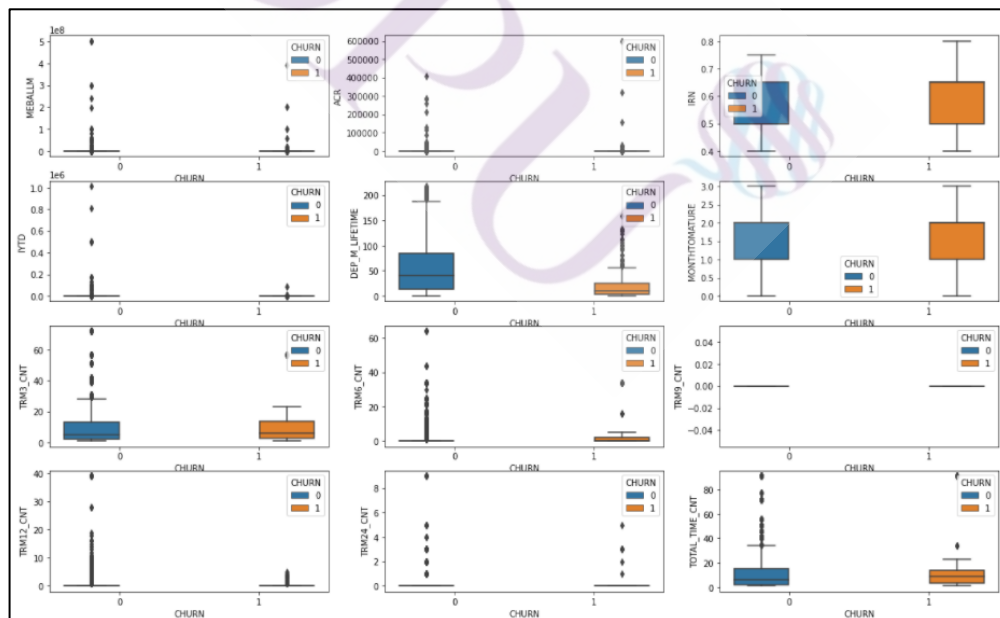
ภาพที่ 3.3 ประเภทของข้อมูลและแสดงจำนวนข้อมูลที่มีค่า

ตารางที่ 3.5 ข้อมูลตัวแปรที่มีค่าว่าง

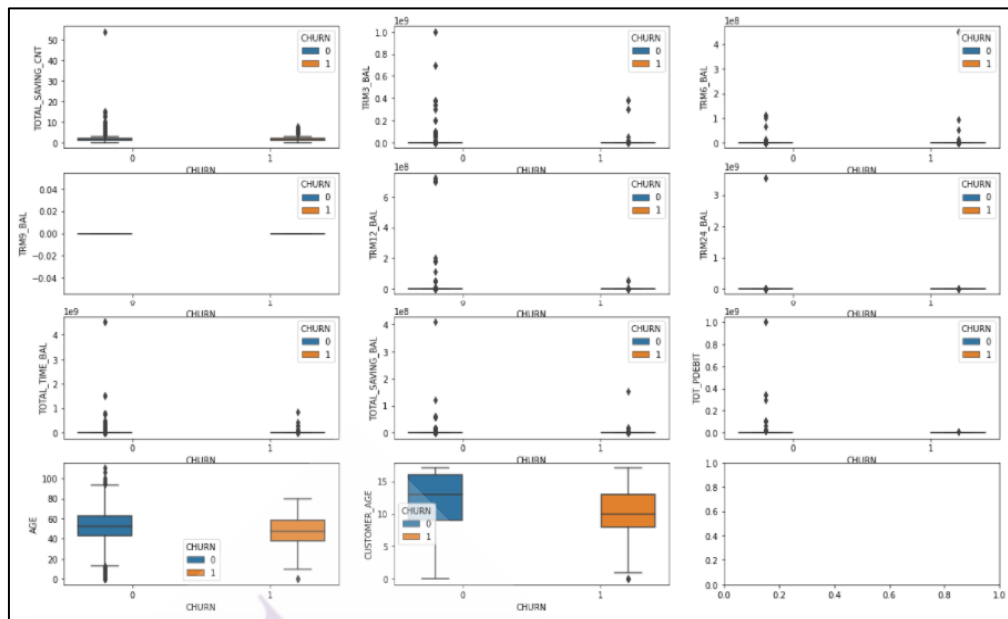
| ตัวแปร | ประเภท | จำนวนค่าว่าง |
|------------|---------|--------------|
| TOT_PDEBIT | Float64 | 3,369 |
| GENDER | Object | 189 |
| INCOME | Object | 399 |
| EDUCATION | Object | 399 |
| OCCUPATON | Object | 83 |

3.3.2 กระบวนการตรวจหาค่าผิดปกติ (Outlier)

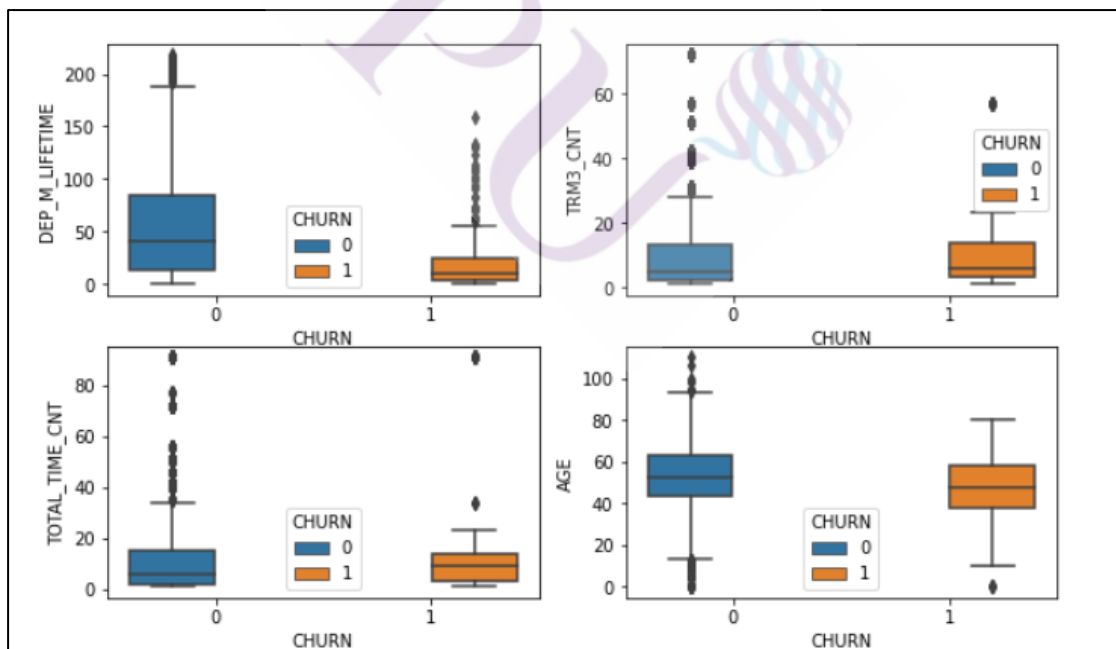
ตรวจหาค่าผิดปกติโดยการสร้างชาร์ต boxplot จากภาพที่ 3.4 และ 3.5 พบว่ามีตัวแปร 4 ค่าที่มีข้อมูลผิด ได้แก่ DEP_M_LIFETIME, TOTAL_TIME_CNT, TRM3_CNT และ AGE ดังแสดงในภาพที่ 3.6



ภาพที่ 3.4 Boxplot สำหรับข้อมูลชุดที่ 1



ภาพที่ 3.5 Boxplot สำหรับข้อมูลชุดที่ 2



ภาพที่ 3.6 Boxplot ของตัวแปรที่มีค่าผิดปกติ

3.3.3 กระบวนการทำความสะอาดข้อมูล (Data Cleansing)

3.3.3.1 เติมค่าตัวแปรที่มีค่าว่าง จากตารางที่ 3.5 ซึ่งแสดงรายละเอียดตัวแปรที่มีค่าว่าง สามารถกำหนดรูปแบบในการเติมข้อมูลที่มีค่าว่างออกเป็น 3 วิธีดังนี้

- Gender: ค่าว่างของตัวแปร Gender เกินจากลูกค้าที่ไม่ใช่บุคคล ดังนั้นใช้วิธีเติมค่าว่างด้วยค่า “C” เพื่อแยกออกจากข้อมูลเพศหญิง “F” และ เพศชาย “M”

- ข้อมูลที่มี type เป็น Object: เติมค่าว่างด้วยค่า Mode ของแต่ละตัวแปร

- ข้อมูลที่มี type ไม่เป็น Object: เติมค่าว่างด้วยค่าเฉลี่ยของแต่ละตัวแปร

3.3.3.2 กำจัดค่าผิดปกติ ด้วยวิธีการคำนวณค่าสถิติด้วยสูตร quartile ดังนี้

- หาค่าควอไทล์ที่ 1 (Q1) และ ควอไทล์ที่ 3 (Q3)

- หาค่า IQR โดย

$$IQR = Q3 - Q1$$

- หาค่า upper/ lower bound เพื่อใช้ตัดค่าผิดปกติ

$$\text{Upper bound} = Q3 + 1.5 * IQR$$

$$\text{Lower bound} = Q1 - 1.5 * IQR$$

ข้อมูลผิดปกติที่จะถูกกำจัดออกคือ ข้อมูลที่มีค่ามากกว่า Upper bound หรือ น้อยกว่า Lower bound

3.3.4 กระบวนการเลือกคุณลักษณะ (Feature Selection) ใช้การทดสอบข้อมูลทางสถิติ โดยแบ่งเป็น 2 เทคนิคดังนี้

3.3.4.1 ตัวแปรที่มีค่าต่อเนื่อง (Continuous Variable) ทดสอบโดยใช้ ANOVA ได้ตัวแปรที่มีค่า P-Value < 0.05 ดังภาพที่ 3.7

| Variable | P-Value |
|----------------|-------------|
| MEBALLM | 0.000562654 |
| ACR | 4.98166E-10 |
| DEP_M_LIFETIME | 8.87685E-17 |
| TRM3_CNT | 1.02198E-14 |
| TRM6_CNT | 1.36776E-49 |
| TOTAL_TIME_CNT | 1.97368E-29 |
| TRM3_BAL | 0.00085418 |
| TRM6_BAL | 3.11383E-10 |
| AGE | 1.24615E-08 |
| CUSTOMER_AGE | 1.61199E-17 |

ภาพที่ 3.7 Continuous Variable : P-Value < 0.05

3.3.4.2 ตัวแปรที่มีค่าไม่ต่อเนื่อง (Categorical Variable) ทดสอบโดยใช้ Chi-Square Test
ได้ตัวแปรที่มีค่า $P\text{-Value} < 0.05$ ดังภาพที่ 3.8

| Variable | P-Value |
|-------------|-------------|
| CORECUSTYPE | 1.93391E-07 |
| INVP | 0.017504555 |
| INCOME | 5.09755E-06 |
| INCOMESCR | 0.000404851 |
| EDUCATION | 6.44367E-11 |
| OCCUPATON | 0.00051516 |
| REGION | 0.046125973 |

ภาพที่ 3.8 Categorical Variable : $P\text{-Value} < 0.05$

3.3.5 กระบวนการเปลี่ยนแปลงข้อมูล (Data Transformation)

3.3.5.1 เปลี่ยนแปลงข้อมูลที่ถูกเก็บในลักษณะ Categorical โดยแบ่งออกเป็น 3 รูปแบบ ดังนี้

- ข้อมูลที่สามารถบอกลำดับขั้นได้ (Ordinal Categorical Variables) ได้แก่ ตัวแปร ซึ่งเก็บค่าการจัดกลุ่มของลูกค้าตามยอดเงินฝากรวม ทำการแปลงค่าตามลำดับขั้นดังตารางที่ 3.6

ตารางที่ 3.6 แสดงการแปลงค่าของตัวแปร CORECUSTYPE

| Categorical Values | Transform Values |
|--------------------|------------------|
| Retail | 0 |
| Medium | 1 |
| Corporate | 2 |

- ข้อมูลเชิงปริมาณที่จัดเก็บแบบมี 2 ค่า (Binary Nominal Categorical Variables) สามารถทำการเปลี่ยนค่าเป็น 0 และ 1 ตัวแปรลักษณะนี้ได้แก่ ตัว INVP ซึ่งเก็บค่าเป็น Personal และ Corporate ทำการเปลี่ยนค่าเป็น 0 และ 1 ตามลำดับ

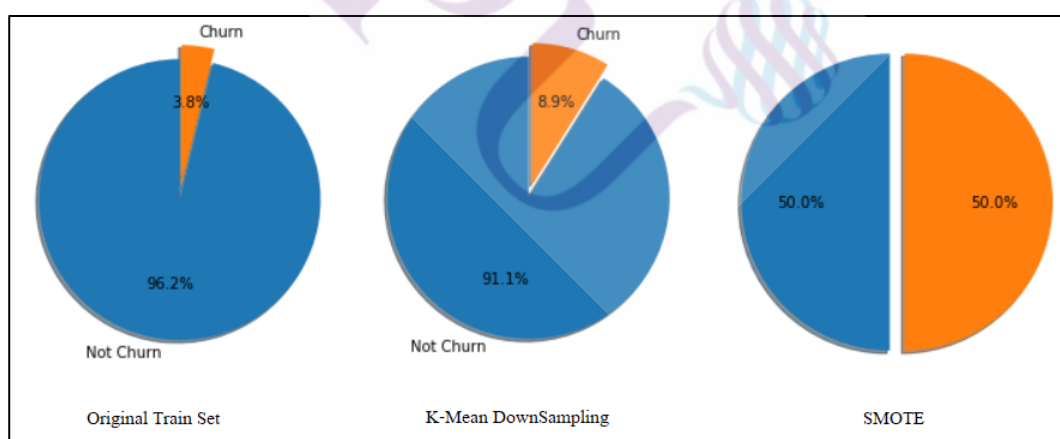
- ข้อมูลเชิงปริมาณที่ไม่มีลำดับและมามีค่ามากกว่า 2 ค่า (Nominal Categorical Variables) ใช้วิธีการ One-Hot encoding ซึ่งจะแปลงข้อมูลโดยนำค่าที่จัดเก็บในแต่ละตัวแปรแตกออกเป็นคอลัมน์ย่อย ๆ เก็บค่าในแต่ละคอลัมน์ในรูปแบบไบนารี 0/1 ตามค่าของข้อมูล

3.3.5.2 ปรับช่วงของข้อมูลให้เป็นช่วงเดียวกัน ด้วยเทคนิค Standardization เนื่องจากการที่ข้อมูลแต่ละตัวแปรมีช่วงของข้อมูลที่แตกต่างกัน เช่น ทำให้เกิด feature bias ในขั้นตอนการสร้างโมเดล

3.4 การพัฒนา Model (Modeling)

3.4.1 กระบวนการเตรียมข้อมูลสำหรับสอนแบบจำลอง

- แบ่งข้อมูลสำหรับสอนและทดสอบแบบจำลองด้วยสัดส่วน 80:20
- แก้ปัญหาข้อมูลไม่สมดุลของชุดข้อมูลสอน โดยใช้เทคนิคการจัดกลุ่มคลาสส่วนมากด้วยวิธี k-mean แล้วทำการสุ่มลดจากแต่ละคลัสเตอร์เพื่อเป็นการสร้างข้อมูลคลาสส่วนมากใหม่ให้มีขนาดใกล้เคียงกับคลาสส่วนน้อยมากขึ้น จากนั้นใช้วิธีสังเคราะห์ข้อมูลเพิ่ม (Synthetic Minority Oversampling Technique: SMOTE) เพื่อปรับให้ข้อมูลคลาสส่วนน้อยและคลาสส่วนมากมีขนาดใกล้เคียงกัน ผลที่ได้ดังภาพที่ 3.9



ภาพที่ 3.9 กระบวนการแก้ปัญหาข้อมูลไม่สมดุล

3.4.2 กระบวนการคัดเลือกแบบจำลอง (Model Selection)

ในเบื้องต้นทำการทดสอบแบบจำลองจำแนกประเภทชนิดต่างๆ ได้ผลดังตาราง 3.7 โดยมีผลการเปรียบเทียบดังภาพที่ 3.10 แบบจำลองที่มีค่า F1 มากที่สุด 3 อันดับแรกคือ Random Forest, Gradient Boosting และ Kernel SVM

ตารางที่ 3.7 แบบจำลอง : Base Line

| Model | Baseline |
|---------------------|---|
| Logistic Regression | random_state = 0 |
| SVM (Linear) | kernel = 'linear', random_state = 0 |
| K-Nearest Neighbors | n_neighbors = 22, metric = 'minkowski', p = 2 |
| Kernel SVM | kernel = 'rbf', random_state = 0 |
| Decision Tree | criterion = 'entropy', random_state = 0 |
| Random Forest | n_estimators=300, criterion = 'entropy', random_state = 0 |
| Gradient Boosting | n_estimators = 100 |

| Model | Accuracy | Precision | Recall | F1 Score | F2 Score |
|----------------------|----------|-----------|---------|----------|----------|
| Random Forest | 0.97959 | 0.74194 | 0.69697 | 0.71875 | 0.70552 |
| Gradient Boosting | 0.95918 | 0.45946 | 0.51515 | 0.48571 | 0.50296 |
| Kernel SVM | 0.94104 | 0.32727 | 0.54545 | 0.40909 | 0.48128 |
| Decision Tree | 0.91837 | 0.25926 | 0.63636 | 0.36842 | 0.49296 |
| Logistic Regression | 0.89456 | 0.19388 | 0.57576 | 0.29008 | 0.41304 |
| K-Nearest Neighbours | 0.87868 | 0.19167 | 0.69697 | 0.30065 | 0.45635 |
| SVM (Linear) | 0.87642 | 0.16667 | 0.57576 | 0.25850 | 0.38618 |

ภาพที่ 3.10 เปรียบเทียบผลจากแบบจำลอง

3.4.3 กระบวนการหาไฮเพอร์พารามิเตอร์ที่เหมาะสมที่สุด โดยใช้ฟังก์ชัน GridSearchCV ผลที่ได้สำหรับแบบจำลองทั้ง 3 แบบ ดังภาพที่ 3.11-3.13

```

1 # Fit random forest classifier
2 param_grid = {'max_depth': [3, 5, 6, 7, 8], 'max_features': [2,4,6,7,8,9],
3               'n_estimators':[50,100], 'min_samples_split': [3, 5, 6, 7]}
4 RanFor_grid = GridSearchCV(RandomForestClassifier(), param_grid, cv=5,
5                             refit=True, verbose=0)
6 RanFor_grid.fit(X_train,y_train)
7 best_model(RanFor_grid)

0.9242916449813002
{'max_depth': 8, 'max_features': 9, 'min_samples_split': 3, 'n_estimators': 100}
RandomForestClassifier(max_depth=8, max_features=9, min_samples_split=3)

```

ภาพที่ 3.11 ค่าพารามิเตอร์ที่เหมาะสมของ Random Forest

```

1 #Stochastic Gradient Boosting
2 param_grid = {'max_depth': [2,3,4,6,8], 'n_estimators': [50, 100, 300, 500]}
3 GB = GBsklearn()
4 model_GB = GridSearchCV(GB, parameters, cv = 5, n_jobs = 10, verbose = 1)
5 model_GB.fit(X_train, y_train)
6 best_model(model_GB)

```

Fitting 5 folds for each of 20 candidates, totalling 100 fits
0.9898284433604406
{'max_depth': 8, 'n_estimators': 100}
GradientBoostingClassifier(max_depth=8)

ภาพที่ 3.12 ค่าพารามิเตอร์ที่เหมาะสมของ Gradient Boosting

```

1 # Fit SVM with pol kernel
2 param_grid = {'C': [0.5,1,10,50,100], 'gamma': [0.1,0.01,0.001],
3               'probability':[True], 'kernel': ['poly'], 'degree':[2,3] }
4 SVM_grid = GridSearchCV(SVC(), param_grid, cv=3, refit=True, verbose=0)
5 SVM_grid.fit(X_train,y_train)
6 best_model(SVM_grid)

```

0.8827433628318584
{'C': 50, 'degree': 2, 'gamma': 0.1, 'kernel': 'poly', 'probability': True}
SVC(C=50, degree=2, gamma=0.1, kernel='poly', probability=True)

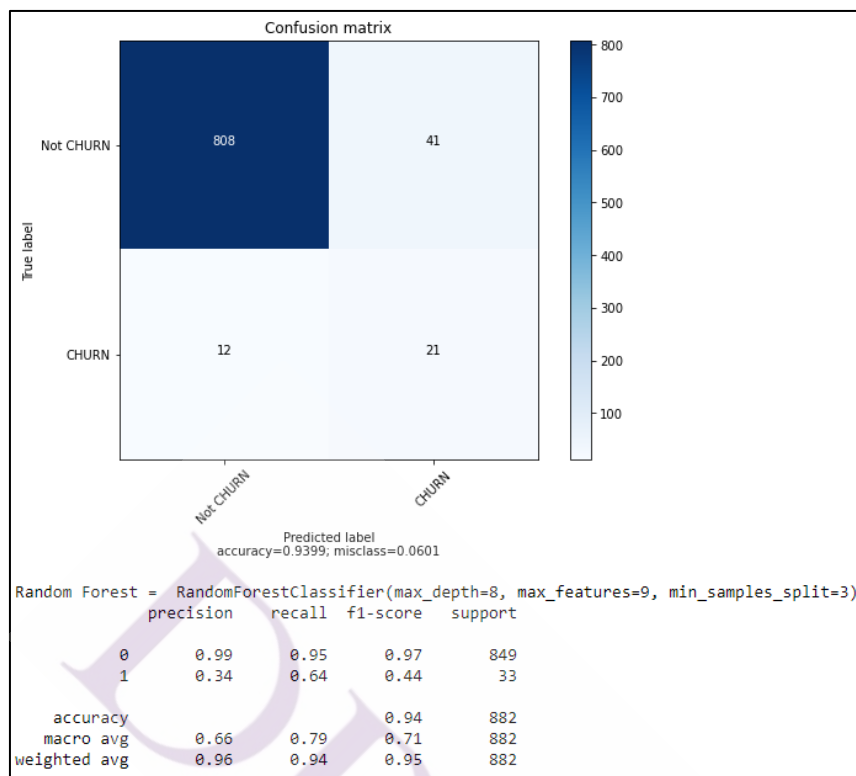
ภาพที่ 3.13 ค่าพารามิเตอร์ที่เหมาะสมของ SVM

3.5 กระบวนการวัดประสิทธิภาพแบบจำลอง (Model Evaluation)

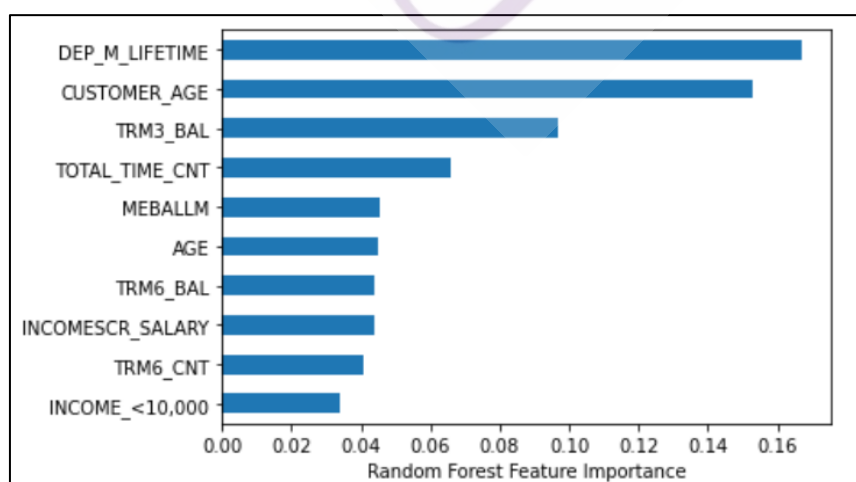
เพื่อเป็นการหาชุดข้อมูลที่เหมาะสมในการนำไปใช้ในทางธุรกิจ จึงทำการสอนและทดสอบแบบจำลองด้วยข้อมูล 2 รูปแบบคือ 1. ใช้ข้อมูลของลูกค้าทั้งบุคคลธรรมดาและนิติบุคคล และใช้ทุกคุณลักษณะที่ถูกเลือกตามหัวข้อที่ 3.3.4 และ 2. ใช้เฉพาะข้อมูลบุคคลธรรมดา และตัดคุณลักษณะ จำนวนเดือนตั้งแต่เปิดบัญชีออก เนื่องจากสมมุติฐานที่ว่า แม้ว่าในกรณีที่จำนวนเดือนตั้งแต่เปิดบัญชีจะส่งผลต่อประสิทธิภาพของแบบจำลอง ในทางธุรกิจแล้วยากที่จะหากลยุทธ์ที่มีเหมาะสมต่อคุณลักษณะนี้ได้

3.5.1 ใช้ข้อมูลของลูกค้าทั้งบุคคลธรรมดาและนิติบุคคล และใช้ทุกคุณลักษณะที่ถูกเลือกตามหัวข้อที่ 3.3.4 โดยจะเรียกข้อมูลชุดนี้ว่า ข้อมูลชุดที่ 1

วัดประสิทธิภาพของแบบจำลอง Random Forest ผลที่ได้ดังภาพที่ 3.14

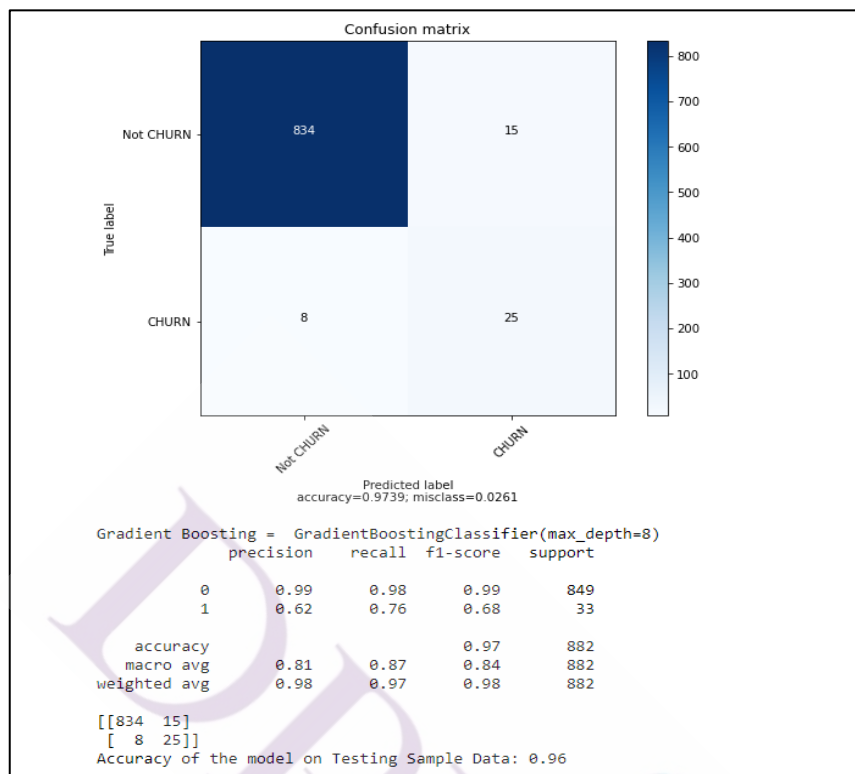


ภาพที่ 3.14 Confusion Matrix : Random Forest ข้อมูลชุดที่ 1

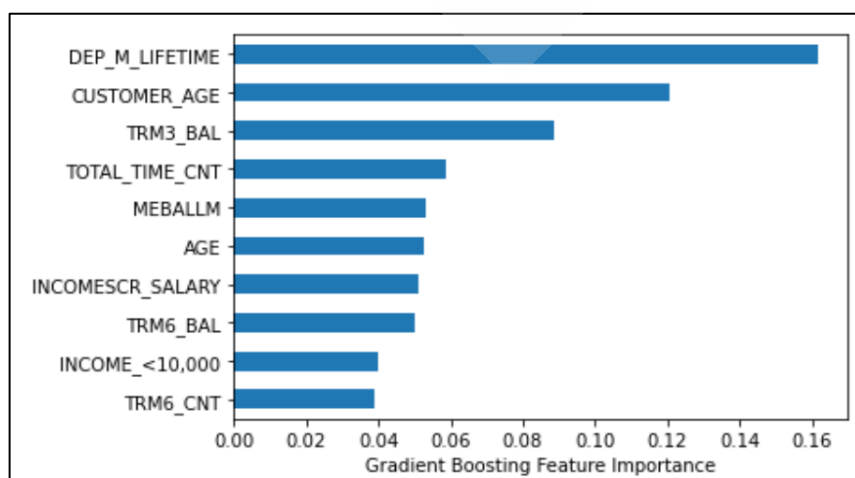


ภาพที่ 3.15 Feature Importance : Random Forest ข้อมูลชุดที่ 1

วัดประสิทธิภาพของแบบจำลอง Gradient Boosting ผลที่ได้ดังภาพที่ 3.16

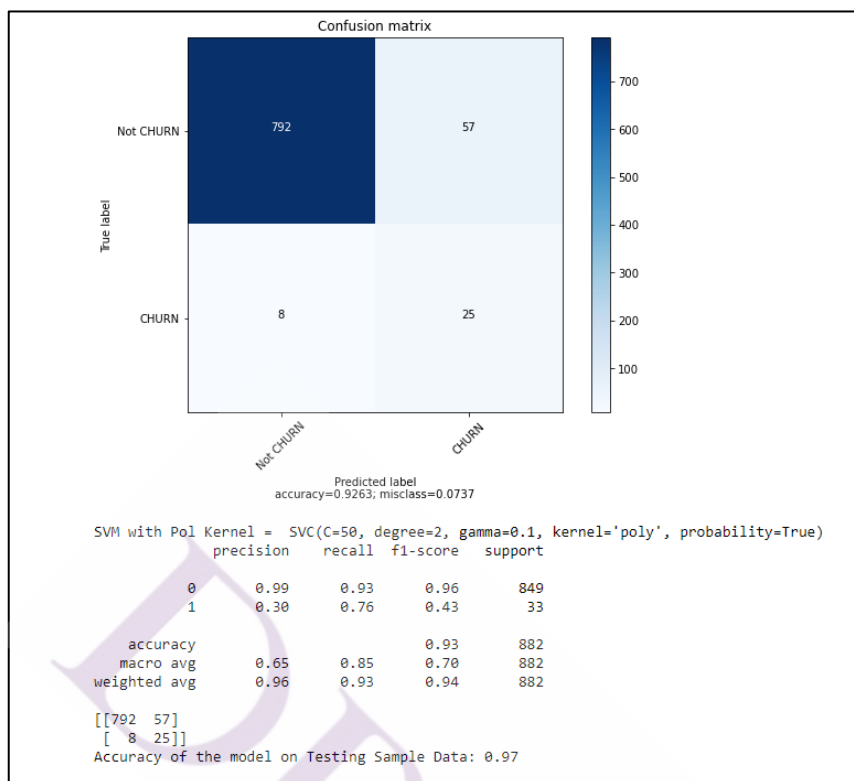


ภาพที่ 3.16 Confusion Matrix: Gradient Boosting ข้อมูลชุดที่ 1



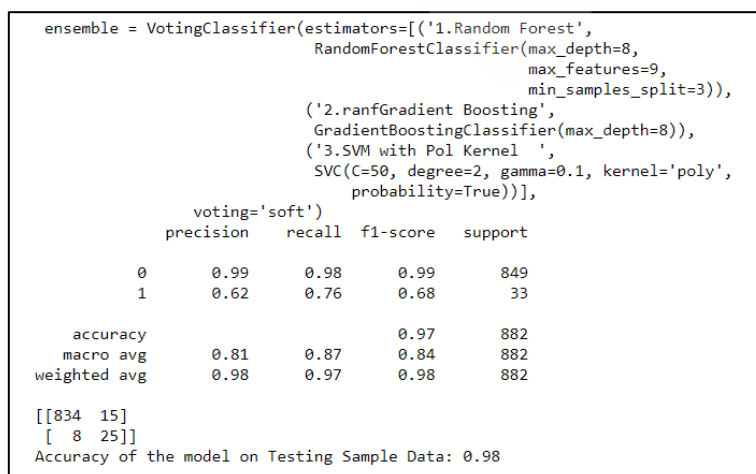
ภาพที่ 3.17 Feature Importance : Gradient Boosting ข้อมูลชุดที่ 1

วัดประสิทธิภาพของแบบจำลอง SVM โดยใช้ชุดข้อมูลสอน ผลที่ได้ดังภาพที่ 3.18



ภาพที่ 3.18 Confusion Matrix: SVM ข้อมูลชุดที่ 1

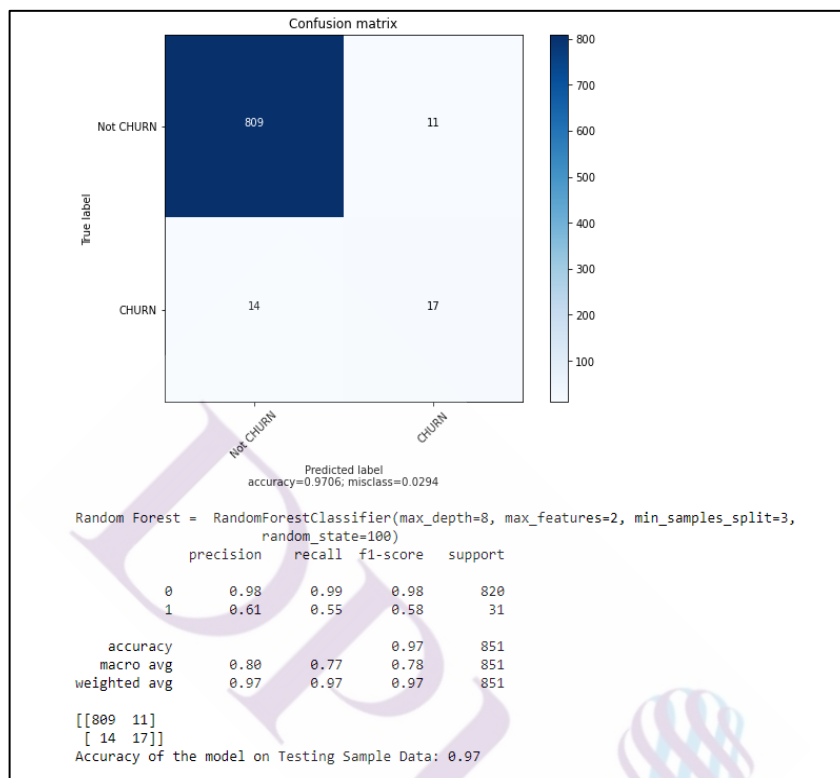
วัดประสิทธิภาพของแบบจำลองโดยวิธีการ Ensemble ด้วยแบบจำลองทั้ง 3 เทคนิคข้างต้น
ผลเป็นดังภาพที่ 3.19



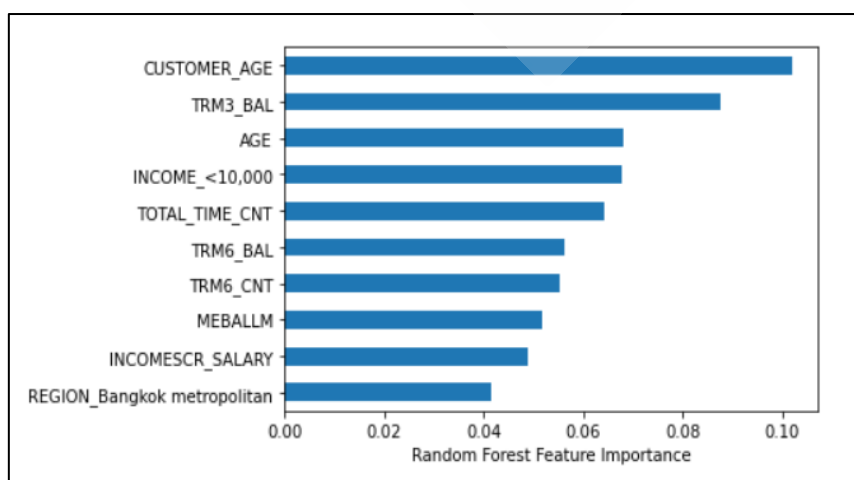
ภาพที่ 3.19 ผลการวัดประสิทธิภาพของเทคนิค Ensemble ข้อมูลชุดที่ 1

3.5.2 ใช้เฉพาะข้อมูลบุคคลธรรมดา และตัดคุณลักษณะ จำนวนเดือนตั้งแต่เปิดบัญชีออก โดยจะเรียกข้อมูลชุดนี้ว่า ข้อมูลชุดที่ 2

วัดประสิทธิภาพของแบบจำลอง Random Forest ผลที่ได้ดังภาพที่ 3.20

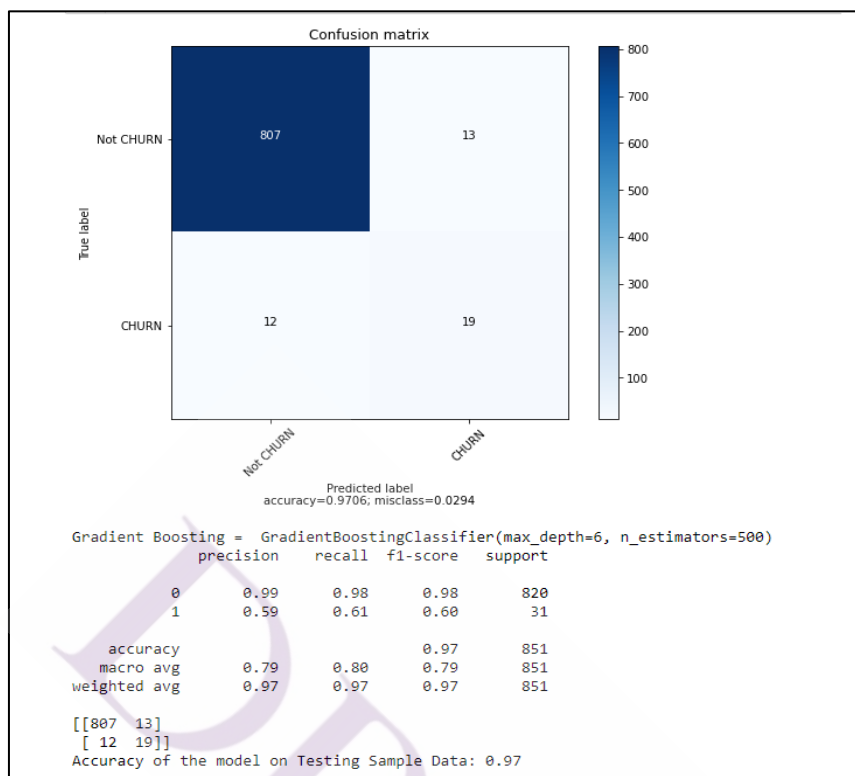


ภาพที่ 3.20 Confusion Matrix : Random Forest ข้อมูลชุดที่ 2

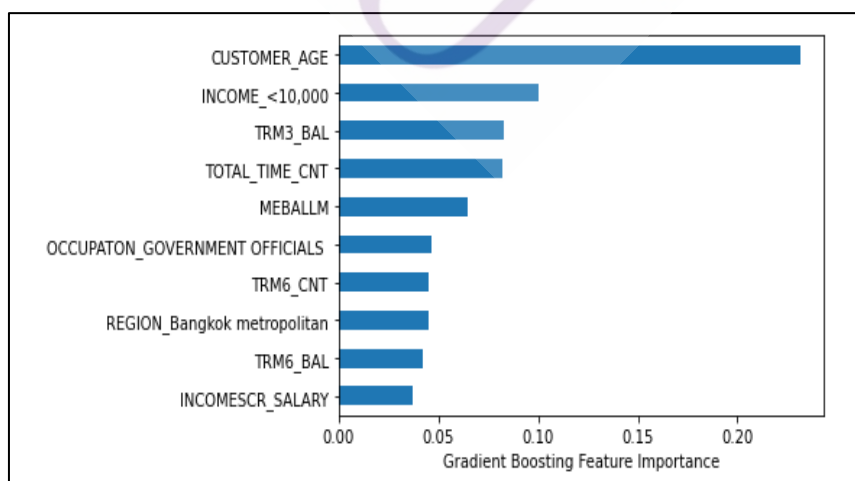


ภาพที่ 3.21 Feature Importance : Random Forest ข้อมูลชุดที่ 2

วัดประสิทธิภาพของแบบจำลอง Gradient Boosting ผลที่ได้ดังภาพที่ 3.22

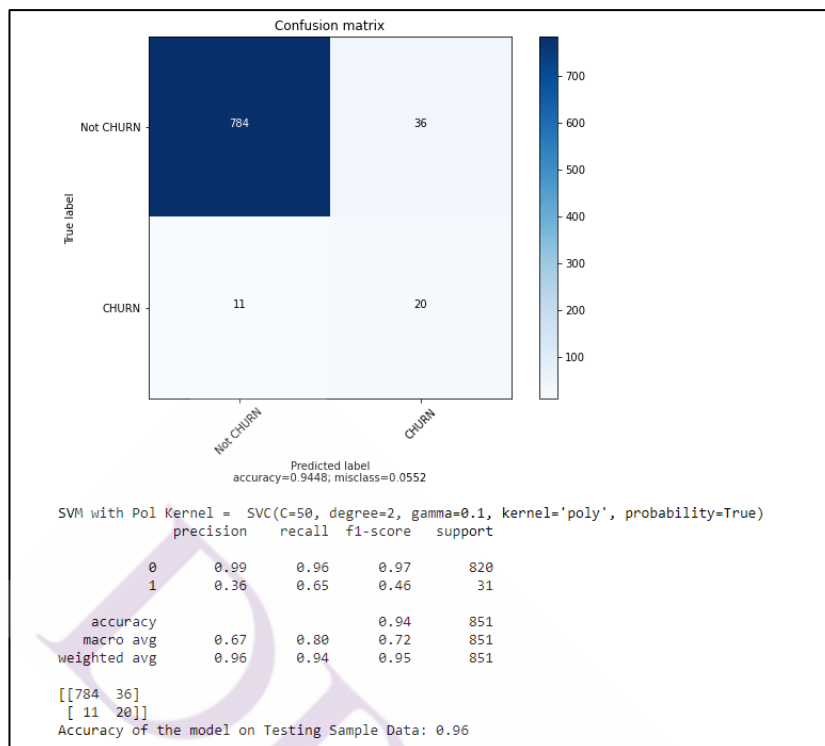


ภาพที่ 3.22 Confusion Matrix: Gradient Boosting ข้อมูลชุดที่ 2



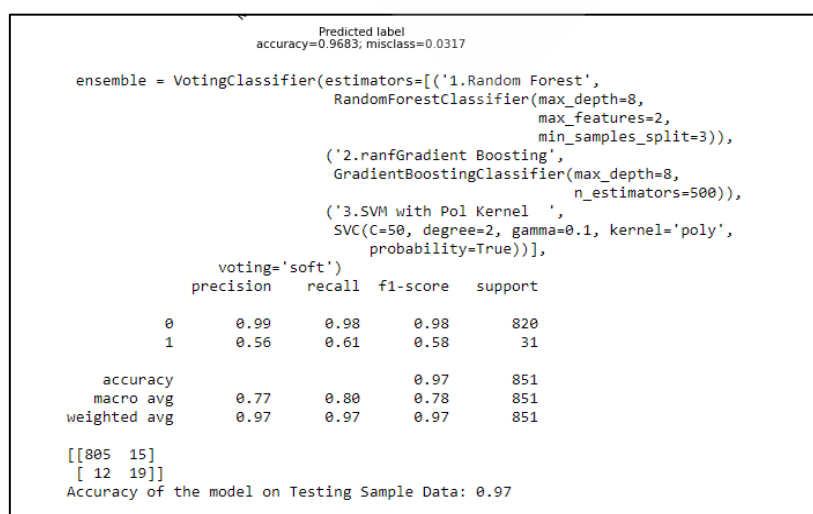
ภาพที่ 3.23 Feature Importance : Gradient Boosting ข้อมูลชุดที่ 2

วัดประสิทธิภาพของแบบจำลอง SVM โดยใช้ชุดข้อมูลสอน ผลที่ได้ดังภาพที่ 3.24



ภาพที่ 3.24 Confusion Matrix: SVM ข้อมูลชุดที่ 2

วัดประสิทธิภาพของแบบจำลองโดยวิธีการ Ensemble ด้วยแบบจำลองทั้ง 3 เทคนิคข้างต้น
ผลเป็นดังภาพที่ 3.25



ภาพที่ 3.25 ผลการวัดประสิทธิภาพของเทคนิค Ensemble ข้อมูลชุดที่ 2

3.6 กระบวนการ Deployment

แปลงแบบจำลองที่ให้มีประสิทธิภาพดีที่สุดให้อยู่ในรูปแบบ file .pkl เพื่อนำไปใช้ทำนายข้อมูลจริง

3.7 เครื่องมือที่ใช้ในการวิจัย

3.7.1 ภาษา Python ถือเป็นภาษาโปรแกรมเอนกประสงค์ที่นิยมนำมาใช้ในการงานด้าน Data Science เนื่องจากมีเครื่องมือให้เรียกใช้งานได้ง่าย

3.7.2 Jupyter Notebook เป็นเครื่องมือที่ถูกออกแบบมาให้สามารถเรียกใช้งานไลบรารี เขียนคำสั่งโปรแกรม และดูผลได้ทันที เป็นที่นิยมในงานด้าน Data Science ซึ่งจะต้องมีการตรวจสอบข้อมูล การทดสอบผลการทำงานของแบบจำลอง และมีลักษณะของการทำซ้ำเพื่อดูผลลัพธ์จนเป็นที่พอใจ

บทที่ 4

ผลการศึกษา

จากการศึกษาพัฒนาแบบจำลองเพื่อทำนายแนวโน้มการยกเลิกใช้บริการสำหรับลูกค้าธนาคาร ด้วยแบบจำลองจำแนกประเภท Random forest , Gradient Boosting , SVM และ เทคนิค Ensemble ได้ผลการทดสอบ ตามรายละเอียดดังนี้

4.1 ผลการวัดประสิทธิภาพความถูกต้องของแบบจำลองด้วยข้อมูลทดสอบ

ข้อมูลที่ใช้ในการทดสอบแบบจำลอง คือข้อมูลเงินฝากประจำ 3 เดือน ณ 30 พฤศจิกายน 2563

4.1.1 ข้อมูลชุดที่ 1 มีทั้งสิ้น 882 ตัวอย่าง โดยมีป้ายกำกับเป็น 0 (Not Churn) จำนวน 849 ตัวอย่าง และมีป้ายกำกับเป็น 1 (Churn) จำนวน 33 ตัวอย่าง ผลการทดสอบโดยใช้แบบจำลองทั้ง 4 เทคนิคตามตารางที่ 4.1 และ 4.2

ตารางที่ 4.1 ผลการทดสอบประสิทธิภาพของแบบจำลองแยกตามป้ายกำกับ ข้อมูลชุดที่ 1

| | Churn | | | Not Churn | | |
|-------------------|-----------|--------|------|-----------|--------|------|
| | precision | recall | f1 | precision | recall | f1 |
| Random Forest | 0.34 | 0.64 | 0.44 | 0.99 | 0.95 | 0.97 |
| Gradient Boosting | 0.64 | 0.76 | 0.69 | 0.99 | 0.98 | 0.99 |
| SVM | 0.30 | 0.76 | 0.43 | 0.99 | 0.93 | 0.96 |
| Voting Classifier | 0.51 | 0.70 | 0.59 | 0.99 | 0.97 | 0.98 |

ตารางที่ 4.2 ผลการทดสอบประสิทธิภาพของแบบจำลอง ข้อมูลชุดที่ 1

| | macro avg | | | weighted avg | | |
|-------------------|-----------|--------|------|--------------|--------|------|
| | precision | recall | f1 | precision | recall | f1 |
| Random Forest | 0.66 | 0.79 | 0.71 | 0.96 | 0.94 | 0.95 |
| Gradient Boosting | 0.82 | 0.87 | 0.84 | 0.98 | 0.98 | 0.98 |
| SVM | 0.65 | 0.85 | 0.70 | 0.96 | 0.93 | 0.94 |
| Voting Classifier | 0.75 | 0.84 | 0.79 | 0.97 | 0.96 | 0.97 |

4.1.2 ข้อมูลชุดที่ 2 มีทั้งสิ้น 851 ตัวอย่าง โดยมีป้ายกำกับเป็น 0 (Not Churn) จำนวน 820 ตัวอย่าง และมีป้ายกำกับเป็น 1 (Churn) จำนวน 31 ตัวอย่าง ผลการทดสอบโดยใช้แบบจำลองทั้ง 4 เทคนิคตามตารางที่ 4.3 และ 4.4

ตารางที่ 4.3 ผลการทดสอบประสิทธิภาพของแบบจำลองแยกตามป้ายกำกับ ข้อมูลชุดที่ 2

| | Churn | | | Not Churn | | |
|-------------------|-----------|--------|------|-----------|--------|------|
| | precision | recall | f1 | precision | recall | f1 |
| Random Forest | 0.61 | 0.55 | 0.58 | 0.98 | 0.99 | 0.98 |
| Gradient Boosting | 0.59 | 0.61 | 0.60 | 0.99 | 0.98 | 0.98 |
| SVM | 0.36 | 0.65 | 0.46 | 0.99 | 0.96 | 0.97 |
| Voting Classifier | 0.56 | 0.61 | 0.58 | 0.99 | 0.98 | 0.98 |

ตารางที่ 4.4 ผลการทดสอบประสิทธิภาพของแบบจำลอง ข้อมูลชุดที่ 2

| | macro avg | | | weighted avg | | |
|-------------------|-----------|--------|------|--------------|--------|------|
| | precision | recall | f1 | precision | recall | f1 |
| Random Forest | 0.80 | 0.77 | 0.78 | 0.97 | 0.97 | 0.97 |
| Gradient Boosting | 0.79 | 0.80 | 0.79 | 0.97 | 0.97 | 0.97 |
| SVM | 0.67 | 0.80 | 0.72 | 0.96 | 0.94 | 0.95 |
| Voting Classifier | 0.77 | 0.80 | 0.78 | 0.97 | 0.97 | 0.97 |

4.2 ผลการวัดประสิทธิภาพความถูกต้องของแบบจำลองด้วยข้อมูล Unseen

ใช้ข้อมูลเงินฝากประจำ 3 เดือน ณ 31 ธันวาคม 2563 2 ชุด ดังนี้

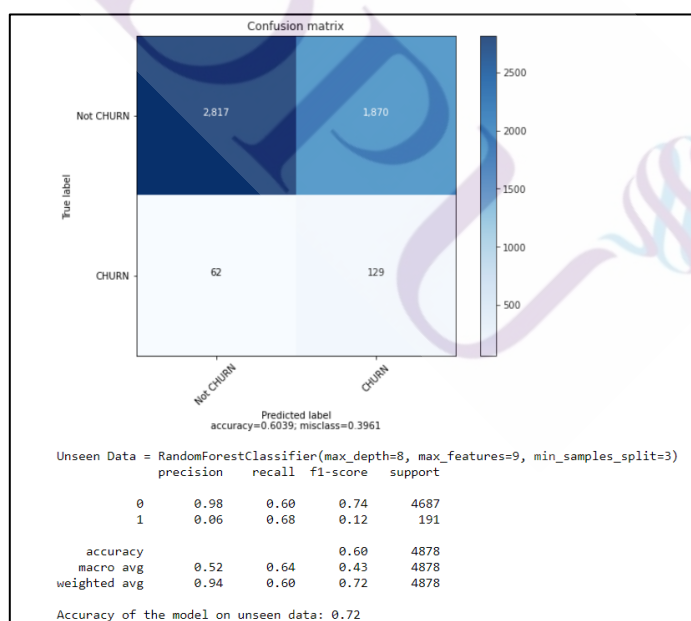
- ข้อมูล Unseen ชุดที่ 1 ข้อมูลบัญชีเงินฝากของลูกค้านิติบุคคลและบุคคลธรรมดา ทั้งสิ้น 4878 ตัวอย่าง มีป้ายกำกับเป็น 0 (Not Churn) จำนวน 4687 ตัวอย่าง และมีป้ายกำกับเป็น 1 (Churn) จำนวน 191 ตัวอย่าง

- ข้อมูล Unseen ชุดที่ 2 ข้อมูลบัญชีเงินฝากของลูกค้าบุคคลธรรมดา มีทั้งสิ้น 4686 ตัวอย่าง โดยมีป้ายกำกับเป็น 0 (Not Churn) จำนวน 4501 ตัวอย่าง และมีป้ายกำกับเป็น 1 (Churn) จำนวน 185 ตัวอย่าง

ทำนายด้วยแบบจำลองที่ได้จากการสอนด้วยข้อมูลสอนชุดที่ 1 และชุดที่ 2 ตามลำดับ ผลที่ได้ดังนี้

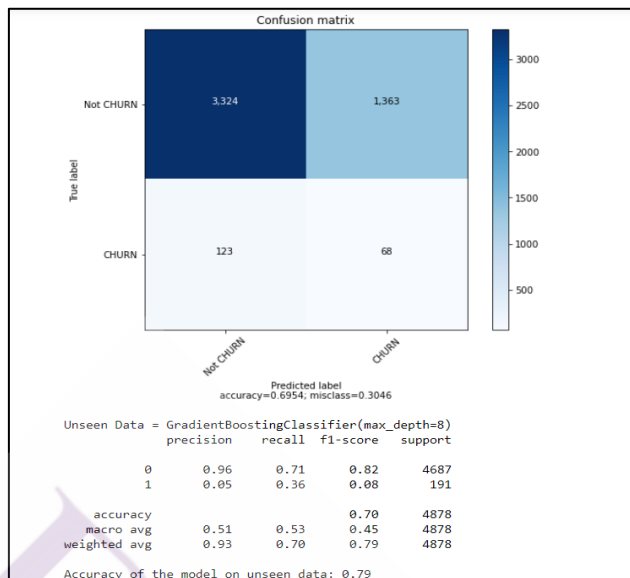
4.2.1 ประสิทธิภาพความถูกต้องของแบบจำลองโดยใช้ข้อมูล Unseen ชุดที่ 1

4.2.1.1 Random Forest



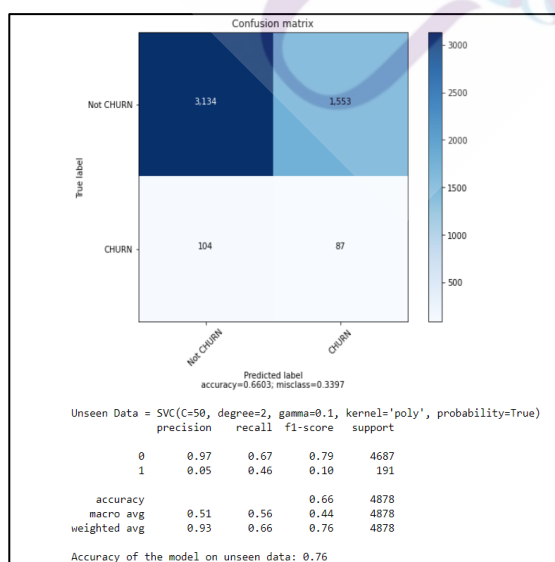
ภาพที่ 4.1 Confusion Matrix (Unseen ชุดที่ 1) : Random Forest

4.2.1.2 Gradient Boosting



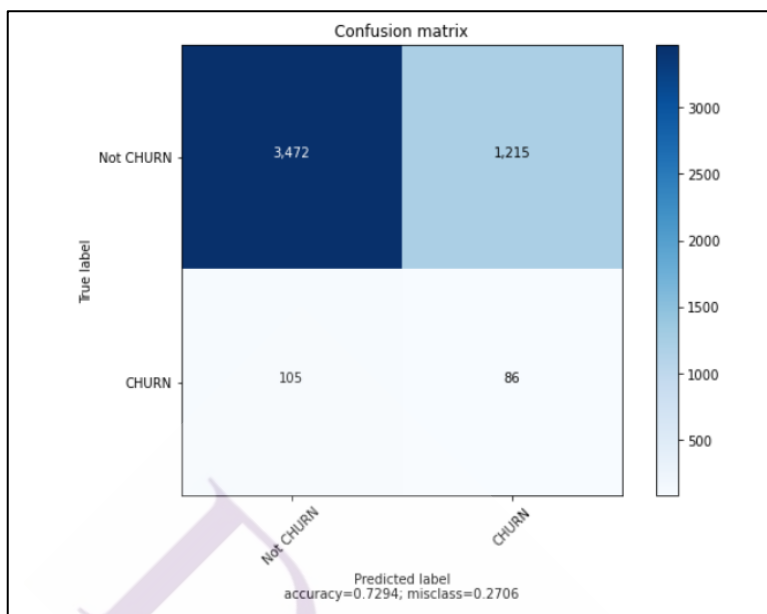
ภาพที่ 4.2 Confusion Matrix (Unseen ชุดที่ 1) : Gradient Boosting

4.2.1.3 SVM

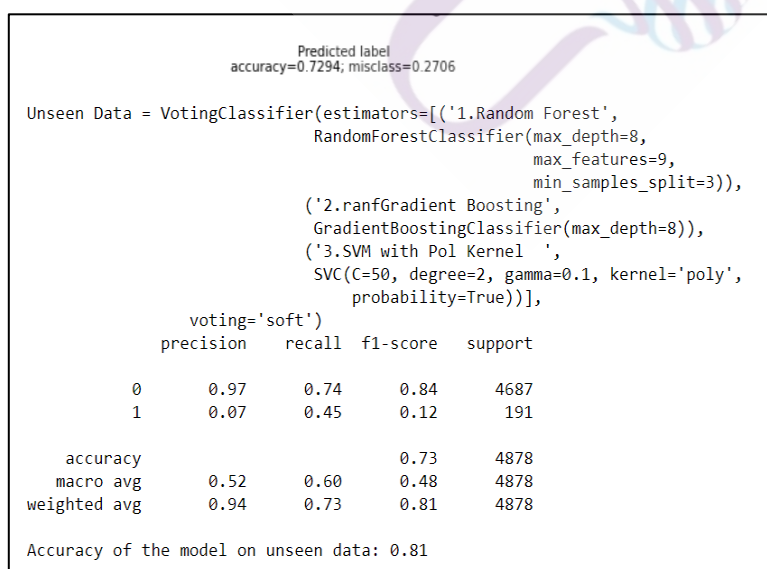


ภาพที่ 4.3 Confusion Matrix (Unseen ชุดที่ 1) : SVM

4.2.1.4 Ensemble: Voting Classifier



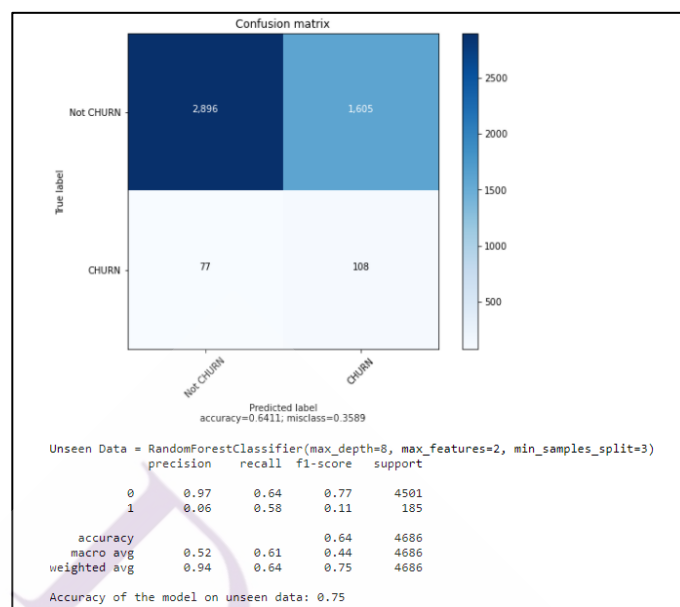
ภาพที่ 4.4 Confusion Matrix (Unseen ชุดที่ 1) : Voting Classifier



ภาพที่ 4.5 Measurement Score (Unseen ชุดที่ 1) : Voting Classifier

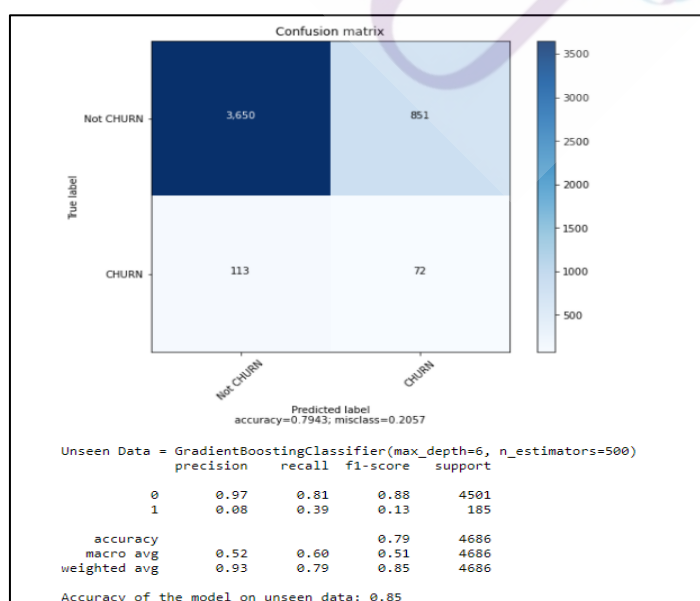
4.2.2 ประสิทธิภาพความถูกต้องของแบบจำลองโดยใช้ข้อมูล Unseen ชุดที่ 2

4.2.2.1 Random Forest



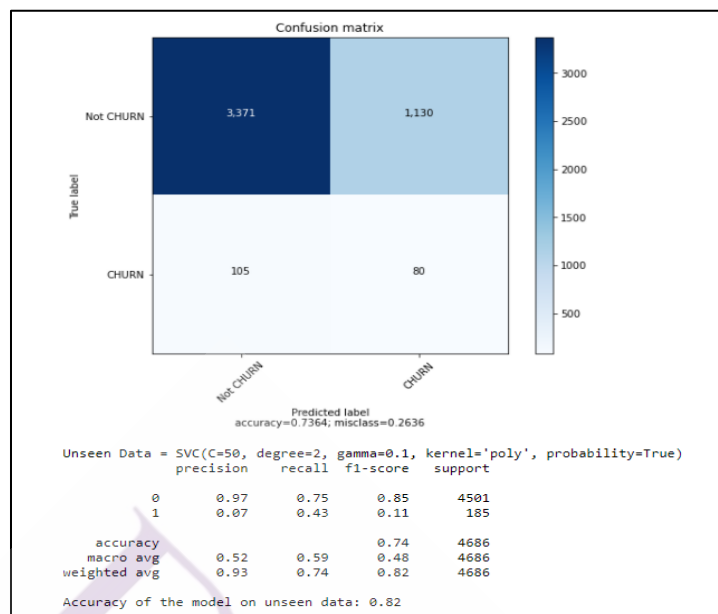
ภาพที่ 4.6 Confusion Matrix (Unseen ชุดที่ 2) : Random Forest

4.2.2.2 Gradient Boosting



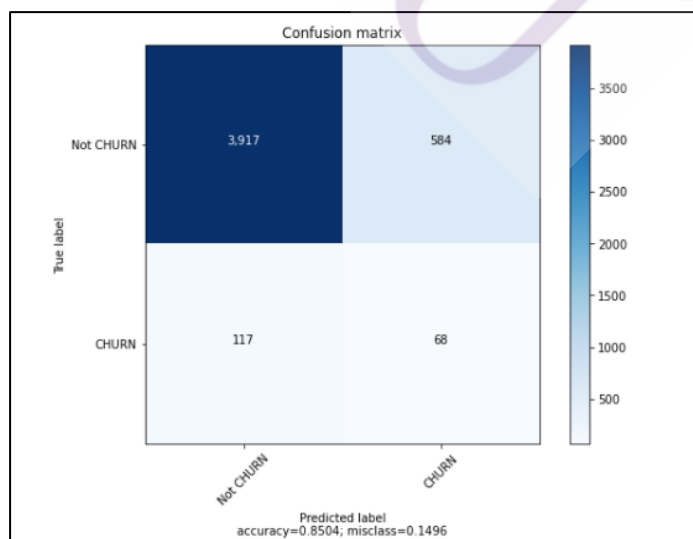
ภาพที่ 4.7 Confusion Matrix (Unseen ชุดที่ 2) : Gradient Boosting

4.2.2.3 SVM



ภาพที่ 4.8 Confusion Matrix (Unseen ชุดที่ 2) : SVM

4.2.2.4 Ensemble: Voting Classifier



ภาพที่ 4.9 Confusion Matrix (Unseen ชุดที่ 2) : Voting Classifier

| | | | | | |
|---|-----------|--------|----------|---------|--|
| Predicted label accuracy=0.8504; misclass=0.1496 | | | | | |
| Unseen Data = VotingClassifier(estimators=[('1.Random Forest', RandomForestClassifier(max_depth=8, max_features=2, min_samples_split=3)), (('2.ranfGradient Boosting', GradientBoostingClassifier(max_depth=8, n_estimators=500)), (('3.SVM with Pol Kernel ', SVC(C=50, degree=2, gamma=0.1, kernel='poly', probability=True))), voting='soft')) | | | | | |
| | precision | recall | f1-score | support | |
| 0 | 0.97 | 0.87 | 0.92 | 4501 | |
| 1 | 0.10 | 0.37 | 0.16 | 185 | |
| accuracy | | | 0.85 | 4686 | |
| macro avg | 0.54 | 0.62 | 0.54 | 4686 | |
| weighted avg | 0.94 | 0.85 | 0.89 | 4686 | |
| Accuracy of the model on unseen data: 0.89 | | | | | |

ภาพที่ 4.10 Measurement Score (Unseen ชุดที่ 2) : Voting Classifier

4.3 สรุปผลการเปรียบเทียบประสิทธิภาพ

ผลประสิทธิภาพของแต่ละแบบจำลอง จากการใช้ข้อมูลเดือน ธันวาคม 2563 พบว่า ค่า f1-score เฉลี่ยทั้ง 2 ป้ายกำกับที่ค่าใกล้เคียงกัน แต่เนื่องจากวัตถุประสงค์ที่ต้องการทำนายค่าบัญชีที่มีแนวโน้ม จะปิด จึงให้น้ำหนักกับค่า Recall และ ค่า TP จาก Confusion Matrix มากกว่าค่าอื่น ซึ่งแบบจำลอง Random Forest มีค่า recall เฉลี่ยสูงที่สุด คือ 0.64 จึงเลือกใช้ Random Forest ในการทำนาย

4.4 ผลการวัดความพึงพอใจของผู้ใช้งาน

วัดผลความพึงพอใจของผู้ใช้งาน กับกลุ่มเป้าหมาย ฝ่ายงานด้านบริหารความเสี่ยงและด้านการตลาด โดยใช้คำถามจำนวน 5 ข้อ และข้อเสนอแนะแบบจำลองอีก 1 ข้อ ผลที่ได้มีดังนี้

4.4.1 คำถาม (Questionnaire)

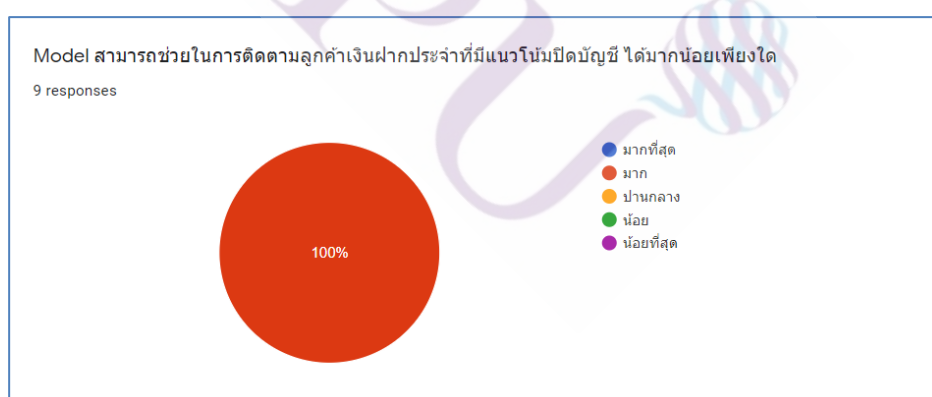
- Model สามารถช่วยในการติดตามลูกค้าเงินฝากประจำที่มีแนวโน้มปิดบัญชี ได้มากน้อยเพียงใด
- Model สามารถช่วยในการวางแผนทางและกลยุทธ์ในการรักษาลูกค้าได้มากน้อยเพียงใด
- Model จะมีประโยชน์ในการป้องกันการถอนปิดบัญชีเงินฝากประจำได้มากน้อยเพียงใด

- Model เหมาะสมที่จะนำมาใช้เป็น Early warning กระแสเงินไหลออกจากเงินฝากประจำได้มากน้อยเพียงใด
- Model เหมาะสมที่จะนำมาใช้ประกอบกับเครื่องมืออื่นเพื่อป้องกันการถอนปิดบัญชีเงินฝากประจำได้มากน้อยเพียงใด

4.4.2 กลุ่มเป้าหมาย ทั้งหมด 7 กลุ่มแบ่งเป็น

- ฝ่ายบริหารความเสี่ยง
- ฝ่ายบริหารเงิน
- ฝ่ายพัฒนาผลิตภัณฑ์
- ฝ่ายทรัพยากรบุคคล
- ฝ่ายกิจการสาขา
- ฝ่ายบริหารและพัฒนาระบบเทคโนโลยีสารสนเทศ
- ส่วนรักษาความปลอดภัยทางเทคโนโลยี

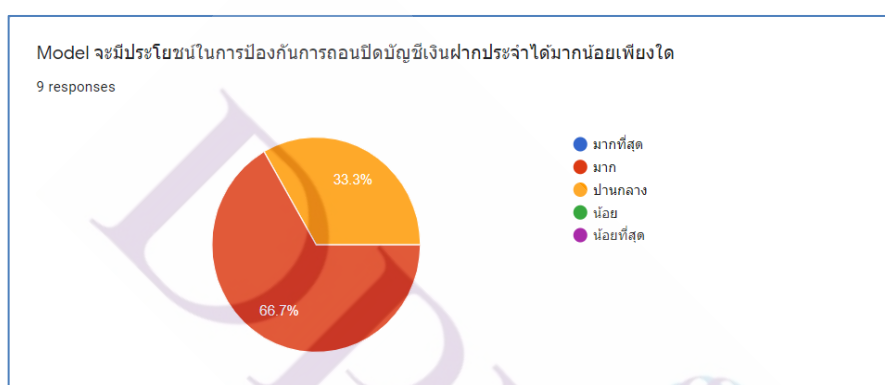
4.4.3 ผลการประเมินความพึงพอใจ



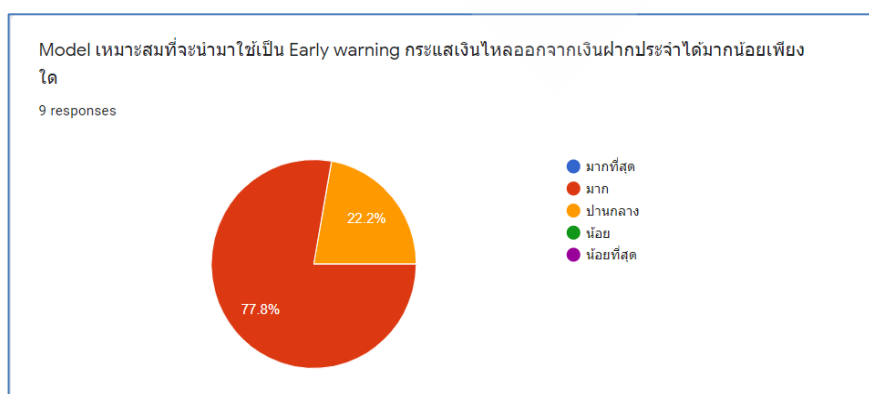
ภาพที่ 4.11 คำถามประเมินความพึงพอใจข้อที่ 1



ภาพที่ 4.12 คำถามประเมินความพึงพอใจข้อที่ 2



ภาพที่ 4.13 คำถามประเมินความพึงพอใจข้อที่ 3



ภาพที่ 4.14 คำถามประเมินความพึงพอใจข้อที่ 4



ภาพที่ 4.15 คำถามประเมินความพึงพอใจข้อที่ 5

4.4.5 สรุปความคิดเห็นและข้อเสนอแนะสำหรับผลิตภัณฑ์

ความเห็นเพิ่มเติม

9 responses

| |
|---|
| - |
| ควรแยกกลุ่มลูกค้าในการทำ model |
| เป็น Innovation ที่แมงคโมเคยมีทำมาก่อน สุดยอดเลยครับ |
| สามารถเอาไปใช้กับฝ่ายที่ดูแลลูกค้าเงินฝาก ช่วยให้สามารถ Scope ในการเลือกลูกค้าที่จะต้องเข้าไปดูแลเพื่อป้องกันลูกค้าปิดบัญชีได้ |
| ควรให้ความรู้/อบรม User มีความเข้าใจมากขึ้น เกี่ยวกับ ทฤษฎีต่างๆของการพัฒนา Model ทั้งนี้ จะส่งผลให้ User สามารถให้ความเห็นเพิ่มเติมได้มากขึ้น ในการพัฒนา Model |
| เห็นควรให้มีการพัฒนาและนำมาใช้เพื่อก่อให้เกิดประโยชน์กับองค์กร |
| เป็น Model ที่มีประโยชน์เพื่อรักษารฐานลูกค้า |

ภาพที่ 4.16 ความคิดเห็นเพิ่มเติม

ข้อเสนอแนะในการพัฒนา Model

3 responses

| |
|---|
| - |
| นำปัจจัยภายนอกมาใช้ในการทำ model เพื่อวิเคราะห์ถึงปัจจัยเพิ่มเติม |
| ควรนำปัจจัยภายนอก เข้ามาร่วมในการพัฒนา Model เช่น อัตราดอกเบี้ยเงินฝากประจำ ของธนาคารอื่น |

ภาพที่ 4.17 ข้อเสนอแนะ

บทที่ 5

บทสรุปและข้อเสนอแนะ

การศึกษานี้มีวัตถุประสงค์เพื่อพัฒนาแบบจำลองสำหรับทำนายแนวโน้มการยกเลิกใช้บริการสำหรับลูกค้าธนาคาร โดยใช้ข้อมูลเงินฝากประจำ 3 เดือน มาทำงานล่วงหน้า เพื่อคาดการณ์ว่าในอีกหนึ่งเดือนข้างหน้าจะมีบัญชีใดบ้างที่มีแนวโน้มจะปิดบัญชี ผลที่ได้จากการพัฒนาแบบจำลองโดยสรุป ดังรายละเอียดต่อไปนี้

5.1 สรุปผลการศึกษา

5.1.1 ในขั้นตอนการพัฒนาแบบจำลอง ได้แบบจำลองจาก 4 เทคนิค ที่มีประสิทธิภาพในการทำนายแนวโน้มการยกเลิกใช้บริการสำหรับลูกค้าธนาคาร ดังนี้

ขั้นตอนที่ 1 แบบจำลองจำแนกประเภท มีค่า Average Recall ดังนี้

ข้อมูลชุดที่ 1

Random Forest มีค่า Average Recall 79%

Gradient Boosting มีค่า Average Recall 87%

SVM มีค่า Average Recall 85%

ข้อมูลชุดที่ 2

Random Forest มีค่า Average Recall 77%

Gradient Boosting มีค่า Average Recall 80%

SVM มีค่า Average Recall 80%

ขั้นตอนที่ 2 นำแบบจำลองทั้ง 3 มาใช้เทคนิค Ensemble โดยวิธี Voting Classifier

ในขั้นตอนนี้แบบจำลองของ ข้อมูลชุดที่ 1 มีค่า Average Recall เท่ากับ 84% ในขณะที่ข้อมูลชุดที่ 2 มีค่า Average Recall เท่ากับ 80%

5.1.2 เมื่อใช้แบบจำลองที่ได้ทดสอบกับชุดข้อมูล Unseen ทั้ง 4 เทคนิคให้ค่า Average Recall ลดลง โดยเมื่อพิจารณาค่า แบบจำลองที่มีค่า Average Recall รวมกับ ค่า Recall ของ Class Churn แบบจำลอง Random Forest ให้ประสิทธิภาพในการทำนายดีที่สุดสำหรับทั้ง 2 ชุดข้อมูล คือ

ข้อมูล Unseen ชุดที่ 1 มีค่า Average Recall เท่ากับ 64% และมีค่า Recall เท่ากับ 68%

ข้อมูล Unseen ชุดที่ 2 มีค่า Average Recall เท่ากับ 61% และมีค่า Recall เท่ากับ 58%

5.1.3 ผลจากการทำนายข้อมูล Unseen ด้วยแบบจำลอง Random Forest เมื่อพิจารณาในรูปแบบของยอดเงินฝาก สำหรับข้อมูล Unseen ชุดที่ 1 และ ข้อมูล Unseen ชุดที่ 2 มีค่าดังตารางที่ 5.1 และ 5.2 ตามลำดับ

ตารางที่ 5.1 Confusion Matrix (Unseen ชุดที่ 1) : จำนวนบัญชีและยอดเงินแบบจำลอง Random Forest

| | | Predicted Label | |
|------------|-----------|-----------------------------------|---------------------------------|
| | | Not Churn | Churn |
| True Label | Not Churn | 2,817 บัญชี 158,268,187.19 บาท | 1,870 บัญชี 40,881,564,11.49 |
| | Churn | 62 บัญชี 3,208,730 บาท | 129 บัญชี 938,130,147.72 บาท |

ตารางที่ 5.2 Confusion Matrix (Unseen ชุดที่ 2) : จำนวนบัญชีและยอดเงินแบบจำลอง Random Forest

| | | Predicted Label | |
|------------|-----------|-----------------------------------|---------------------------------|
| | | Not Churn | Churn |
| True Label | Not Churn | 2,896 บัญชี 125,541,140.68 บาท | 1,605 บัญชี 97,522,169.5 บาท |
| | Churn | 77 บัญชี 3,679,730 บาท | 108 บัญชี 3,828,805.47 บาท |

จากตารางที่ 5.1 แบบจำลองสามารถทำนาย บัญชีที่มีแนวโน้มจะปิดได้ถูกต้อง 129 บัญชีจาก 191 บัญชี คิดเป็นยอดเงินที่สามารถนำไปสู่การป้องกันการไหลออกจากระบบได้ 938,130,147.72 บาท

จากตารางที่ 5.2 แบบจำลองสามารถทำนาย บัญชีที่มีแนวโน้มจะปิดได้ถูกต้อง 108 บัญชีจาก 185 บัญชี คิดเป็นยอดเงินที่สามารถนำไปสู่การป้องกันการไหลออกจากระบบได้ 3,828,805.47 บาท

จากผลการทำนายด้วยข้อมูล Unseen ทั้ง 2 ชุดข้อมูลคือข้อมูลชุดที่ 1 รวมลูกค้านิติบุคคลและบุคคลธรรมดา และ ข้อมูลชุดที่ 2 มีเฉพาะลูกค้าบุคคลธรรมดา พบว่าผลการทดสอบข้อมูลชุดที่ 2 จะมีค่า Recall ที่ต่ำกว่าข้อมูลชุดที่ 1 อย่างไรก็ตามเมื่อพิจารณาถึงลักษณะข้อมูลและความรับผิดชอบของหน่วยงานที่ดูแลลูกค้าที่ต่างกันแล้ว การสร้างแบบจำลองโดยใช้ข้อมูลเฉพาะกลุ่มจะช่วยให้สามารถนำแบบจำลองไปใช้งานจริงในทางธุรกิจได้ดีกว่า

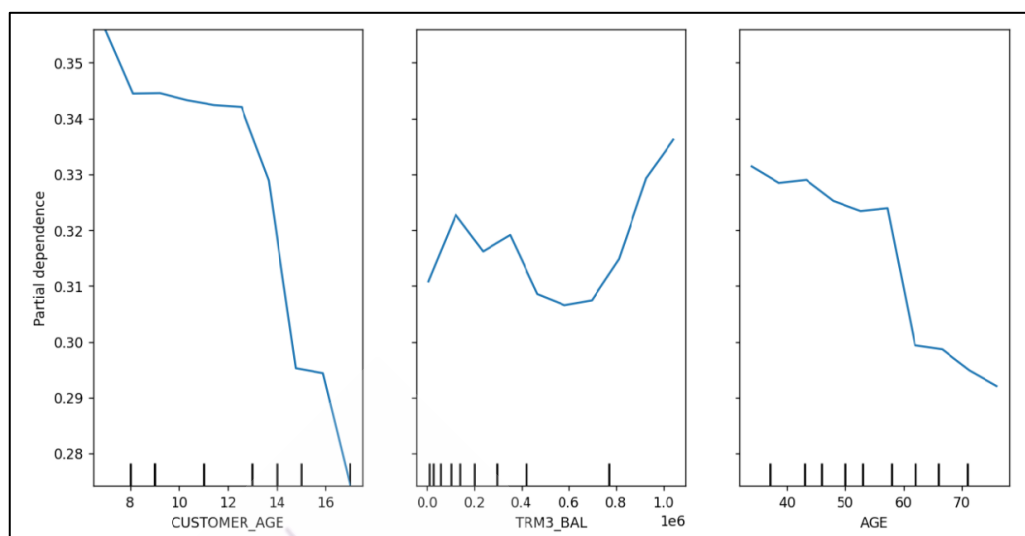
5.1.4 ปัจจัยที่มีความสัมพันธ์ต่อแนวโน้มในการยกเลิกใช้บริการ

ในขั้นตอนการวัดประสิทธิภาพของแบบจำลอง Random Forest โดยใช้ข้อมูลชุดที่ 1 พบว่าตัวแปรสำคัญ 3 อันดับแรกคือ จำนวนเดือนตั้งแต่เปิดบัญชี (DEP_M_LIFETIME), จำนวนปีที่เป็นลูกค้า (CUSTOMER_AGE), ยอดเงินฝากรวมของเงินฝากประจำ 3 เดือน (TRM3_BAL) แต่เนื่องจากในทางธุรกิจ เป็นการยากในการกำหนดกลยุทธ์ด้านการตลาดโดยการจัดหาข้อเสนอเพื่อรักษาลูกค้าด้วยการใช้ข้อมูล จำนวนเดือนตั้งแต่เปิดบัญชี เป็นหลัก จึงทำการทดสอบแบบจำลองโดยการตัดข้อมูลตัวแปรนี้ออกเพื่อหาว่ามีตัวแปรใดอีกที่มีผลต่อแนวโน้มการยกเลิกใช้บริการและสามารถนำมาใช้ในการสร้างข้อเสนอให้แก่ลูกค้าได้

ข้อมูลชุดที่ 2 เป็นข้อมูลที่ตัดตัวแปรจำนวนเดือนตั้งแต่เปิดบัญชีออก พบว่าเมื่อนำมาใช้ทดสอบแบบจำลอง Random Forest พบว่าตัวแปรสำคัญ 3 อันดับแรกคือ

- ปีที่เป็นลูกค้า(CUSTOMER_AGE)
- ยอดเงินฝากรวมของเงินฝากประจำ 3 เดือน (TRM3_BAL)
- อายุลูกค้า (AGE)

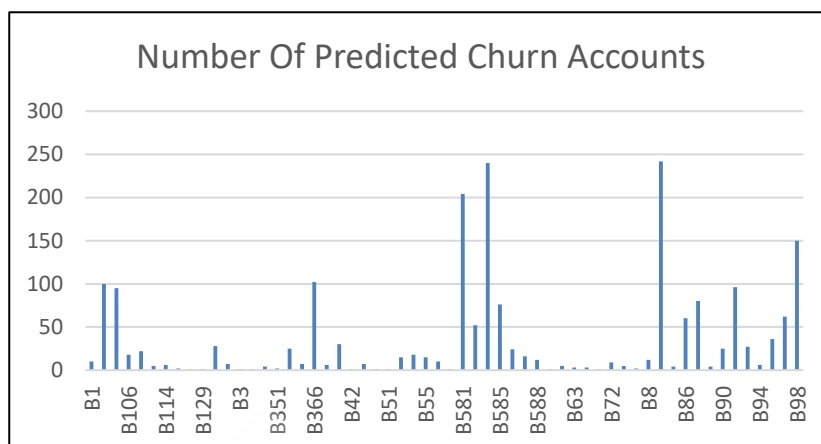
โดยตัวแปรทั้ง 3 มีผลต่อแนวโน้มในการยกเลิกใช้บริการในทิศทางดังภาพที่ 5.1



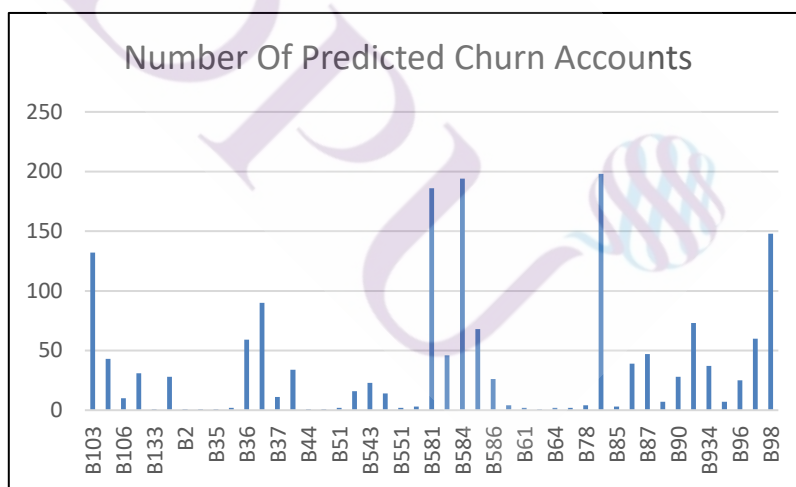
ภาพที่ 5.1 ทิศทางที่มีผลต่อแนวโน้มในการยกเลิกใช้บริการของ 3 ตัวแปรหลัก

จากภาพ จำนวนปีที่เป็ลูกค้าในช่วง 2 – 12 ปีมีผลอย่างมากต่อแนวโน้มในการยกเลิกใช้บริการ และส่งผลลดลงเมื่อจำนวนปีเพิ่มขึ้นในช่วง 13 – 17 ปี และ ลูกค้าที่มีอายุในช่วง 30 – 55 ปีมีแนวโน้มสูงที่จะยกเลิกใช้บริการ ในขณะที่อายุ 60 ปีขึ้นไปมีแนวโน้มการยกเลิกลดลง ยอดเงินฝากรวมของเงินฝากประจำ 3 เดือน ส่งผลต่อค่าความเป็นไปได้ที่ลูกค้าจะยกเลิกการให้บริการในทิศทางเดียวกันเมื่อมียอดเงินตั้งแต่ 800,000 บาทขึ้นไป

5.1.5 ปัญหาที่พบในการทำนาย คือแบบจำลองให้ค่า Fales Positive สูง ทั้ง 2 ชุดข้อมูล ซึ่งจะมีผลให้การทำงานของสาขาผู้ดูแลลูกค้าอาจจะเพิ่มขึ้นโดยไม่จำเป็น หากนำข้อมูลไปใช้โดยไม่ได้อภิปรายร่วมกับปัจจัยอื่น อย่างไรก็ตาม เมื่อทำการกระจายบัญชีที่ถูกทำนายว่าจะมีการปิดไปยังสาขาผู้ดูแลลูกค้า พบว่าผลการทำนายจากทั้ง 2 ชุดข้อมูล มีเพียง 5 สาขาเท่านั้นที่มีบัญชีมากกว่า 100 บัญชี นั่นคือโดยเฉลี่ยแล้วจำนวนบัญชีที่ต้องดูแลเป็นพิเศษตามผลการทำนายยังอยู่ในปริมาณที่สาขาสามารถดำเนินการได้ ดังภาพที่ 5.2 และ 5.3



ภาพที่ 5.2 บัญชีที่ทำนายว่ามีแนวโน้มปิดบัญชีของแต่ละสาขา (Unseen ชุดที่ 1)



ภาพที่ 5.3 บัญชีที่ทำนายว่ามีแนวโน้มปิดบัญชีของแต่ละสาขา (Unseen ชุดที่ 2)

5.2 ข้อเสนอแนะ

5.2.1 แนวทางในการเพิ่มประสิทธิภาพของแบบจำลอง คือ เพิ่มตัวแปรในมิติอื่น เช่น ข้อมูลอัตราผลตอบแทนจากภายนอก ข้อมูลผลิตภัณฑ์ด้านสินเชื่อบริษัท ซึ่งถือเป็นปัจจัยที่อาจจะมีผลในการตัดสินใจปิดบัญชี

5.2.2 ในการสร้างแบบจำลองอาจทำได้โดยแยกตามกลุ่มลูกค้า ด้วยการทำ Clustering ข้อมูลลูกค้าก่อนสร้างแบบจำลอง เพื่อให้สามารถทำนายผลและนำเสนอกลยุทธ์ในการรักษาลูกค้าได้ตรงกับความต้องการของลูกค้าได้ดีขึ้น



บรรณานุกรม



บรรณานุกรม

จิรกฤต บุญหมื่นไวย, เจษฎา ตันตานุช, เบญจวรรณ โรจนดิษฐ์. (2563).

การวิเคราะห์แนวโน้มการขอยกเลิกใช้บริการสำหรับลูกค้าธนาคารโดยวิธีการเรียนรู้ของเครื่อง. การประชุมวิชาการระดับชาติเรื่อง คุณภาพของการบริหารและนวัตกรรม ครั้งที่ 5.

Abbas Keramati, Hajar Ghaneei and Seyed Mohammad Mirmohammadi.(2016).

Developing a Prediction Model for Customer Churn from Electronic Banking Services using Data Mining.Financial Innovation, pp 2-10

Himanshu147. (2020). The Complete Guide to Checking Account Churn Prediction in

BFSI Domain. <https://www.analyticsvidhya.com/blog/2020/10/the-complete-guide-to-checking-account-churn-prediction-in-bfsi-domain/>.

Nelson Rosa. (2018). Gauging and Foreseeing Customer Churn in The Banking Industry.

NOVA Information Management School, Universidade Nova de Lisboa.

Francesco Pochetti. (2019). Extreme Label Imbalance: When You Measure the Minority Class in

Basis Points. <http://francescopochetti.com/extreme-label-imbalance-when-you-measure-the-minority-class-in-basis-points/>.

Irfan Ullah, Basit Raza, Ahmad Kamran Malik, Muhammad Imran, Saif Ul islam, AND Sung Won

Kim. (2019). A Churn Prediction Model Using Random Forest: Analysis of Machine Learning Techniques for Churn Prediction and Factor Identification in Telecom Sector. IEEE Access, pp 2169-3536.

ภาคผนวก



ภาคผนวก ก: ตัวอย่างการใช้งาน

1. Input File ประกอบด้วยข้อมูลคุณลักษณะ ซึ่งเป็นข้อมูล ณ 30 พฤศจิกายน 2020 และ ข้อมูล

ป้ายกำกับ ซึ่งเป็นข้อมูล ณ 31 ธันวาคม 2020

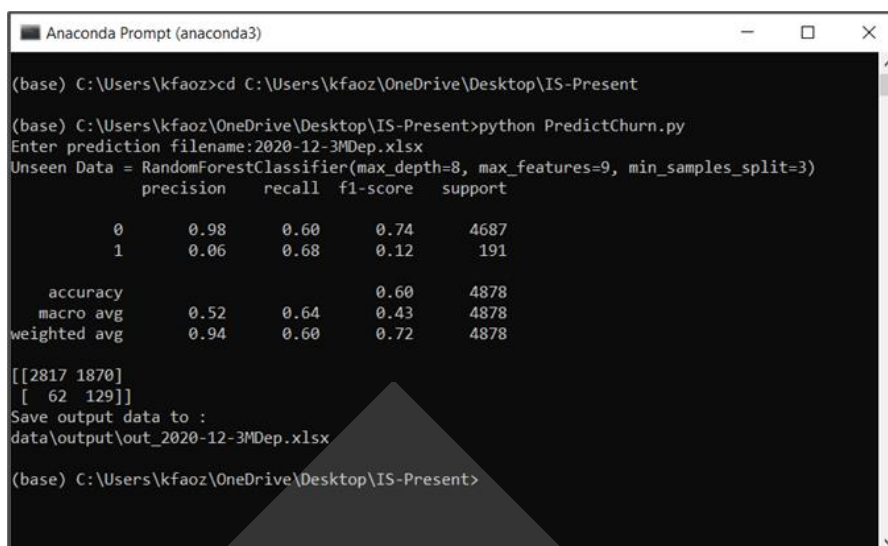
| As At 30-11-2020 | | | | | | | | | | | |
|------------------|-------------|--------|---------|------------|-----|----------|----------------|---------------|------|----------|----------|
| CID | ACN | ACTYPE | MEBALLM | ACR | IRN | IYTD | DEP_M_LIFETIME | MONTHTOMATURE | TERM | TRM3_CNT | TRM6_CNT |
| 270000000177 | 1140000087 | 3001 | 10000 | 8.49338 | 0.5 | 48.93 | 209 | 1 | 3 | 2 | 0 |
| 270000000189 | 2070000030 | 3001 | 10000 | 8.49338 | 0.5 | 48.93 | 209 | 1 | 3 | 1 | 0 |
| 270000000300 | 1590000084 | 3001 | 4000000 | 3397.26024 | 0.5 | 19572.59 | 209 | 1 | 3 | 1 | 0 |
| 270000000975 | 1860000168 | 3001 | 10000 | 8.49338 | 0.5 | 48.93 | 216 | 1 | 3 | 1 | 0 |
| 270000001485 | 1170000060 | 3001 | 10000 | 8.49338 | 0.5 | 48.93 | 216 | 1 | 3 | 3 | 0 |
| 270000002082 | 18300000342 | 3001 | 10000 | 8.49338 | 0.5 | 48.93 | 214 | 1 | 3 | 2 | 0 |
| 270000002085 | 18300000342 | 3001 | 10000 | 8.49338 | 0.5 | 48.93 | 209 | 1 | 3 | 2 | 0 |
| 270000002151 | 2040001491 | 3001 | 60000 | 50.95904 | 0.5 | 293.59 | 214 | 1 | 3 | 4 | 0 |
| 270000002154 | 2040001491 | 3001 | 20000 | 16.98614 | 0.5 | 97.87 | 211 | 1 | 3 | 4 | 0 |
| 270000002499 | 1530000141 | 3001 | 100000 | 84.93132 | 0.5 | 489.31 | 216 | 1 | 3 | 1 | 0 |
| 270000006126 | 2010002292 | 3001 | 100000 | 110.95866 | 0.5 | 501.91 | 207 | 0 | 3 | 1 | 0 |
| 270000006615 | 2190005058 | 3001 | 10000 | 8.49338 | 0.5 | 48.93 | 216 | 1 | 3 | 3 | 0 |
| 270000006618 | 2190005058 | 3001 | 10000 | 8.49338 | 0.5 | 48.93 | 215 | 1 | 3 | 3 | 0 |

| As At 30-11-2020 | | | | | | | | | | | |
|------------------|-----------|-----------|----------------|------------------|----------|----------|----------|-----------|-----------|----------------|------------------|
| TRM9_CNT | TRM12_CNT | TRM24_CNT | TOTAL_TIME_CNT | TOTAL_SAVING_CNT | TRM3_BAL | TRM6_BAL | TRM9_BAL | TRM12_BAL | TRM24_BAL | TOTAL_TIME_BAL | TOTAL_SAVING_BAL |
| 0 | 0 | 0 | 2 | 1 | 20000 | 0 | 0 | 0 | 0 | 20000 | 5302.02 |
| 0 | 0 | 0 | 1 | 2 | 10000 | 0 | 0 | 0 | 0 | 10000 | 36857.89 |
| 0 | 0 | 0 | 1 | 1 | 4000000 | 0 | 0 | 0 | 0 | 4000000 | 1282468.22 |
| 0 | 0 | 0 | 1 | 1 | 10000 | 0 | 0 | 0 | 0 | 10000 | 3285.14 |
| 0 | 4 | 0 | 7 | 3 | 30000 | 0 | 0 | 70000 | 0 | 100000 | 637674.65 |
| 0 | 0 | 0 | 2 | 3 | 20000 | 0 | 0 | 0 | 0 | 20000 | 190580.2 |
| 0 | 0 | 0 | 2 | 3 | 20000 | 0 | 0 | 0 | 0 | 20000 | 190580.2 |
| 0 | 0 | 0 | 4 | 1 | 120000 | 0 | 0 | 0 | 0 | 120000 | 55394.57 |
| 0 | 0 | 0 | 4 | 1 | 120000 | 0 | 0 | 0 | 0 | 120000 | 55394.57 |
| 0 | 0 | 0 | 1 | 2 | 100000 | 0 | 0 | 0 | 0 | 100000 | 32833.7 |
| 0 | 0 | 0 | 1 | 1 | 100000 | 0 | 0 | 0 | 0 | 100000 | 42122.41 |
| 0 | 2 | 0 | 5 | 4 | 30000 | 0 | 0 | 15000 | 0 | 45000 | 8744.37 |
| 0 | 2 | 0 | 5 | 4 | 30000 | 0 | 0 | 15000 | 0 | 45000 | 8744.37 |

| As At 30-11-2020 | | | | | | | | | | | | As At 31-12-2021 | |
|------------------|-----|--------------|-------------|------|--------|-----------------|----------|-----------------------------|-----------|-------------------------------|----------|------------------|-------|
| TOT_PDEBIT | AGE | CUSTOMER_AGE | CORECUSTYPE | INVP | GENDER | BOO | INCOME | INCOMESCR | EDUCATION | OCCUPATION | REGION | | CHURN |
| 64 | 17 | Retail | PERSONAL | M | 8360 | 30,000 - 49,999 | BUSINESS | SECONDARY | | BUSINESS OWNER (UNREGISTERED) | Southern | | 0 |
| 70 | 17 | Retail | PERSONAL | F | 8360 | 30,000 - 49,999 | BUSINESS | B.A. | | BUSINESS OWNER (UNREGISTERED) | Southern | | 0 |
| 68 | 17 | Retail | PERSONAL | F | 8360 | 15,000 - 29,999 | BUSINESS | B.A. | | BUSINESS OWNER (REGISTERED) | Southern | | 0 |
| 72 | 17 | Retail | PERSONAL | M | 8543 | <10,000 | OTHER | VOCATIONAL CERTIFICATE | | OTHER | Northern | | 0 |
| 67 | 17 | Retail | PERSONAL | F | 8551 | 15,000 - 29,999 | BUSINESS | SECONDARY | | BUSINESS OWNER (REGISTERED) | Northern | | 0 |
| 56 | 17 | Retail | PERSONAL | M | 8543 | <10,000 | SALARY | B.A. | | SPECIFIC PROFESSIONS | Northern | | 0 |
| 56 | 17 | Retail | PERSONAL | M | 8543 | <10,000 | SALARY | B.A. | | SPECIFIC PROFESSIONS | Northern | | 0 |
| 84 | 17 | Retail | PERSONAL | F | 8543 | 15,000 - 29,999 | BUSINESS | HIGH VOCATIONAL CERTIFICATE | | BUSINESS OWNER (REGISTERED) | Northern | | 0 |
| 84 | 17 | Retail | PERSONAL | F | 8543 | 15,000 - 29,999 | BUSINESS | HIGH VOCATIONAL CERTIFICATE | | BUSINESS OWNER (REGISTERED) | Northern | | 0 |
| 61 | 17 | Retail | PERSONAL | M | 8543 | <10,000 | BUSINESS | VOCATIONAL CERTIFICATE | | BUSINESS OWNER (UNREGISTERED) | Northern | | 0 |
| 50 | 17 | Retail | PERSONAL | M | 8934 | | OTHER | | | OTHER | Southern | | 0 |
| 51 | 17 | Retail | PERSONAL | F | 8366 | 15,000 - 29,999 | SALARY | B.A. | | GOVERNMENT OFFICIALS | Southern | | 0 |
| 51 | 17 | Retail | PERSONAL | F | 8366 | 15,000 - 29,999 | SALARY | B.A. | | GOVERNMENT OFFICIALS | Southern | | 0 |

ภาพที่ 6.1 ตัวอย่างข้อมูลใน Input File

2. เรียกใช้โปรแกรมเพื่อทำการทำนายด้วย Model ที่เลือก



```

Anaconda Prompt (anaconda3)

(base) C:\Users\kfaoz>cd C:\Users\kfaoz\OneDrive\Desktop\IS-Present

(base) C:\Users\kfaoz\OneDrive\Desktop\IS-Present>python PredictChurn.py
Enter prediction filename:2020-12-3MDep.xlsx
Unseen Data = RandomForestClassifier(max_depth=8, max_features=9, min_samples_split=3)
      precision    recall  f1-score   support

      0        0.98      0.60      0.74      4687
      1        0.06      0.68      0.12       191

 accuracy      0.60      4878
 macro avg      0.52      0.64      0.43      4878
weighted avg      0.94      0.60      0.72      4878

[[2817 1870]
 [ 62 129]]
Save output data to :
data\output\out_2020-12-3MDep.xlsx

(base) C:\Users\kfaoz\OneDrive\Desktop\IS-Present>
  
```

ภาพที่ 6.2 หน้าจอการเรียกใช้ Model

3. ผลการทำนายและค่าความเป็นไปได้ในการยกเลิกบริการของแต่ละบัญชี

| CID | BOO | MEBALLM | CHURN | PREDICTION | PROPENSITY_TO_CHURN(%) |
|--------------|------|---------|-------|------------|------------------------|
| 270000000177 | B360 | 10000 | 0 | 0 | 9.07 |
| 270000000189 | B360 | 10000 | 0 | 0 | 0.01 |
| 270000000300 | B360 | 4000000 | 0 | 1 | 100 |
| 270000000975 | B543 | 10000 | 0 | 0 | 0.06 |
| 270000001485 | B551 | 10000 | 0 | 0 | 1.57 |
| 270000002082 | B543 | 10000 | 0 | 0 | 1 |
| 270000002085 | B543 | 10000 | 0 | 0 | 1 |
| 270000002151 | B543 | 60000 | 0 | 0 | 3.09 |
| 270000002154 | B543 | 20000 | 0 | 0 | 5.77 |
| 270000002499 | B543 | 100000 | 0 | 0 | 12.67 |
| 270000006126 | B934 | 100000 | 0 | 1 | 88.09 |
| 270000006615 | B366 | 10000 | 0 | 1 | 54.59 |
| 270000006618 | B366 | 10000 | 0 | 1 | 54.59 |
| 270000008676 | B366 | 28000 | 0 | 0 | 14.2 |
| 270000009258 | B584 | 10000 | 0 | 1 | 60.17 |
| 270000009492 | B581 | 10000 | 0 | 0 | 13.22 |
| 270000009495 | B581 | 30000 | 0 | 0 | 15.37 |
| 270000010479 | B581 | 10000 | 0 | 1 | 86.88 |
| 270000010560 | B581 | 10000 | 0 | 0 | 10.32 |
| 270000014064 | B366 | 22000 | 0 | 0 | 19.41 |
| 270000014205 | B585 | 14500 | 0 | 0 | 0.13 |
| 270000014208 | B585 | 17900 | 0 | 1 | 79.95 |
| 270000016419 | B585 | 90150 | 0 | 0 | 0.34 |

ภาพที่ 6.3 ตัวอย่างข้อมูลใน Output File

ชื่อ-นามสกุล

ประวัติการศึกษา

ผลงานทางวิชาการ

ประวัติผู้เขียน

นางสาวฟาเซีย เกษตรกาลาม์

วิทยาศาสตร์มหาบัณฑิต

สาขาธุรกรรมอิเล็กทรอนิกส์

มหาวิทยาลัยเทคโนโลยีพระจอมเกล้าธนบุรี

ปีการศึกษา 2550

วิศวกรรมศาสตรบัณฑิต

สาขาคอมพิวเตอร์

มหาวิทยาลัยสงขลานครินทร์

ปีการศึกษา 2544

ฟาเซีย เกษตรกาลาม์, ปาณิตา รุสรานนท์

และธนภัทร นังคะจิตร. (2561). การแก้ปัญหาผู้ใช้รายใหม่ในระบบผู้แนะนำด้วยวิธีการคัดกรองร่วมแบบข้ามโดเมน. การประชุมทางวิชาการระดับชาติด้านคอมพิวเตอร์และเทคโนโลยีสารสนเทศ ครั้งที่ 14.