

การทำนายแนวโน้มการยกเลิกใช้บริการของลูกค้าเงินฝากประจำ

โดยใช้แบบจำลองการเรียนรู้ของเครื่อง

Fixed-Deposit Customers Churn Prediction Using Machine Learning Models

เฟาเซีย เกษตรกาลัม*

ธนภัทร มังคะจิตร**

บทคัดย่อ

ความเสี่ยงด้านสภาพคล่องถือเป็นความเสี่ยงที่สำคัญที่สุดประเภทหนึ่งของธุรกิจภาคธนาคาร เนื่องจากธนาคารใช้การระดมทุนจากเงินฝากระยะสั้นมาเป็นเงินทุนในการให้สินเชื่อ ซึ่งมีระยะเวลาคงกำหนดยาวกว่าเงินฝาก ทำให้มีโอกาสเกิดภาวะขาดสภาพคล่องได้โดยเฉพาะอย่างยิ่งเมื่อเกิดการถอนเงินในอัตราที่สูงกว่าสัดส่วนปกติ ธนาคารจำเป็นต้องมีเครื่องมือที่ช่วยให้สามารถคาดการณ์ความเปลี่ยนแปลงของกระแสเงินสดได้ล่วงหน้าเพื่อช่วยให้สามารถบริหารความเสี่ยงได้อย่างมีประสิทธิภาพ

สารนิพนธ์ชิ้นนี้มีวัตถุประสงค์เพื่อพัฒนาแบบจำลองที่สามารถทำนายแนวโน้มการถอนปิดบัญชีเงินฝากประจำด้วยเทคนิคการเรียนรู้ของเครื่อง เพื่อให้ธนาคารสามารถนำไปใช้ในการป้องกันก่อนเกิดภาวะขาดสภาพคล่อง ผลการศึกษาโดยใช้ข้อมูลเงินฝากประจำ 3 เดือน จำนวน 4,919 ตัวอย่าง พบว่าเทคนิค Random Forest ให้ความแม่นยำสูงสุดโดยมีค่า Recall โดยรวมอยู่ที่ 79% ในขณะที่ค่า Recall ลดลงเป็น 64%

ในการทำนายบัญชีที่ปิดจำนวน 191 ตัวอย่าง จากข้อมูลชุดทดสอบทั้งสิ้นจำนวน 4,878 ตัวอย่าง อย่างไรก็ตามเมื่อคิดเป็นจำนวนเงินของบัญชีที่ถูกทำนายว่าปิดจากบัญชีที่ปิดจริงอยู่ที่ 99.67% ซึ่งกล่าวได้ว่าแบบจำลองที่ได้สามารถนำมาใช้เป็นเครื่องมือในการป้องกันวิกฤตด้านสภาพคล่องได้

คำสำคัญ: เงินฝากประจำ การเรียนรู้ของเครื่อง ทำนายแนวโน้มการยกเลิกใช้บริการ

ABSTRACT

Liquidity risk is one of the most important risks in the banking sector. Generally, financial loans offer a long period of liabilities' time. Meanwhile, the source of loans funding is

*นักศึกษาหลักสูตรวิศวกรรมศาสตรมหาบัณฑิต สาขาวิศวกรรมข้อมูลขนาดใหญ่

**ที่ปรึกษาสารนิพนธ์หลัก

from short-term deposits. As a result, liquidity shortages possibly happen especially when withdrawals are proportionately higher than usual. Therefore, banks need tools to anticipate changes in cash flow in advance to manage their risks effectively.

The objective of this study is to mitigate the effectiveness of liquidity shortages by the use of predicting the churn of fixed-deposit accounts using various machine learning techniques. In this study, a Random Forest model offers the best recall of 79% on the validation data contained 4,919 three-month fixed deposit accounts. Meanwhile, an average recall reduces to 64% when applies the model to the tested data having 191 closed accounts from 4,878 accounts in total. Contrastingly, the total balance of predicted closed accounts is 99.67% of the actual closed accounts. It can imply that our model is a useful tool to prevent the further liquidity crisis.

Keyword: Fixed-Deposit, Machine Learning, Churn Prediction

บทนำ

การดำเนินธุรกิจของธนาคาร มีความเกี่ยวข้องโดยตรงกับการบริหารความเสี่ยง ดังนั้นความสามารถในการบริหารความเสี่ยง จึงเป็นปัจจัยสำคัญต่อความสำเร็จในการดำเนินธุรกิจ ทั้งนี้ ความเสี่ยงด้านสภาพคล่อง ถือเป็นความเสี่ยงที่สำคัญที่สุดประเภทหนึ่งของธุรกิจธนาคารไทย เนื่องจากโดยส่วนใหญ่แล้ว ธนาคารใช้วิธีการระดมทุนจากเงินฝากระยะสั้น เช่น เงินฝากประจำที่มีระยะเวลาครบกำหนด 3 เดือน 6 เดือน หรือ 12 เดือน เป็นต้น ธนาคารใช้เงินทุนเหล่านี้ในการให้สินเชื่อซึ่งจะมีระยะเวลาครบกำหนดยาวกว่าเงินฝาก การดำเนินธุรกิจลักษณะนี้ทำให้ธนาคารมีความเสี่ยงด้านสภาพคล่อง จากการที่ไม่สามารถเปลี่ยนแปลงสินทรัพย์ให้เป็นเงินสดได้ทันกับระยะเวลาครบกำหนดของหนี้สิน

นอกจากนี้ปัญหาการยกเลิกการใช้บริการของลูกค้าถือเป็นปัญหาพื้นฐานในธุรกิจธนาคารซึ่งมีการแข่งขันสูง ธนาคารเล็งเห็นถึงความสำคัญในการหาวิธีรักษาลูกค้าเดิม ซึ่งเป็นกลุ่มลูกค้าที่สามารถสร้างผลกำไรให้กับองค์กรได้ในระยะยาว ประกอบกับจากการประเมินความเสี่ยงของธนาคารในไตรมาสที่สองของปี 2563 พบแนวโน้มการเกิดสัดส่วนกระแสเงินสดจ่ายมากกว่ากระแสเงินสดรับในช่วงหนึ่งปีข้างหน้า ด้วยเหตุผลด้านการบริหารความเสี่ยงและการสร้างผลกำไรในทางธุรกิจตามกล่าวมาข้างต้น ธนาคารจำเป็นต้องมีเครื่องมือที่ช่วยให้สามารถคาดการณ์

ความเปลี่ยนแปลงของกระแสเงินสด โดยเฉพาะในกรณีที่เกิดจากการถอนเงินและยกเลิกการใช้บริการของลูกค้า

งานวิจัยนี้มีวัตถุประสงค์เพื่อพัฒนาแบบจำลองที่สามารถทำนายแนวโน้มการถอนปิดบัญชีเงินฝากประจำ เพื่อให้ธนาคารสามารถนำไปใช้ในการป้องกันก่อนเกิดเหตุการณ์ อีกทั้งเพื่อช่วยให้ธนาคารสามารถเตรียมการจัดหาแหล่งเงินในกรณีที่มีเหตุการณ์กระแสเงินสดไหลออกมากเกินไปเกินอัตราที่ยอมรับได้ ข้อมูลที่ใช้แบ่งออกเป็นกลุ่มได้ดังนี้ ข้อมูลระดับลูกค้า ข้อมูลระดับบัญชี และข้อมูลระดับ transaction โดยมีปัจจัยที่สำคัญคือ จำนวนเดือนตั้งแต่เปิดบัญชี จำนวนปีที่为客户 ยอดรวมเงินฝากประจำ 3 เดือน และอายุของลูกค้า

วัตถุประสงค์

เพื่อนำเสนอแบบจำลองสำหรับพยากรณ์แนวโน้มการยกเลิกใช้บริการของลูกค้าเงินฝากประเภทเงินฝากประจำ

ขอบเขตของการวิจัย

1. ขอบเขตของแผนงาน

การศึกษานี้เป็นการศึกษาเพื่อพัฒนาแบบจำลองที่สามารถทำนายแนวโน้มการยกเลิกใช้บริการของลูกค้าเงินฝากประเภทเงินฝากประจำล่วงหน้า 1 เดือน ด้วยเทคนิคการเรียนรู้ของเครื่อง โดยข้อมูลที่ใช้ในการสอนแบบจำลองและทดสอบ ใช้เฉพาะบัญชีที่มีสถานะเปิดบัญชี ณ สิ้นเดือนเท่านั้น ไม่รวมบัญชีที่มีการปิดบัญชีระหว่างเดือน

2. ขอบเขตของข้อมูล

ข้อมูลที่น่ามาใช้เป็นข้อมูลบัญชีเงินฝากประจำประเภท 3 เดือนที่มีการกำหนดคุณลักษณะการต่ออายุเมื่อครบอายุสัญญาเป็นแบบต่ออัตโนมัติที่บัญชีเดิมเท่านั้น โดยเก็บข้อมูลในปี 2563 เพื่อวิเคราะห์อัตราการปิดบัญชี และใช้ข้อมูล ณ สิ้นเดือน พฤศจิกายน และ ธันวาคม 2563 ในการทดสอบแบบจำลอง ทั้งนี้ข้อมูลที่ใช้ทำนายจะมีป้ายกำกับ 2 ค่าเท่านั้น คือ 1 หมายถึง Churn และ 0 หมายถึง Not Churn

ประโยชน์ที่คาดว่าจะได้รับ

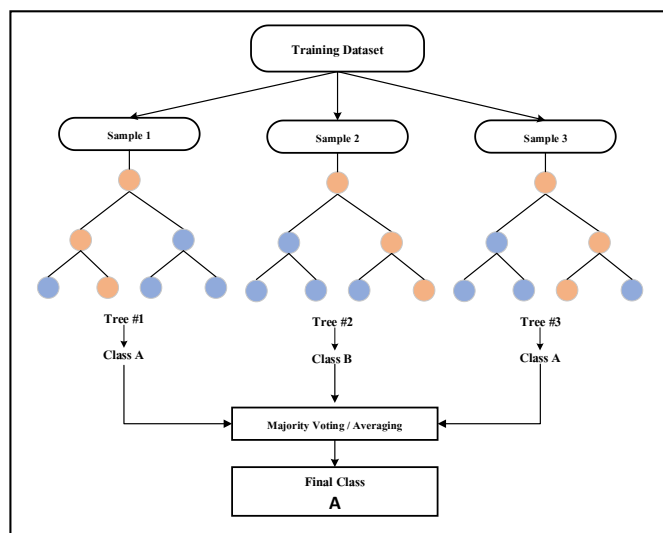
1. เป็นเครื่องมือที่ช่วยให้ธนาคารสามารถรักษาลูกค้าที่มีแนวโน้มจะเลิกใช้บริการเงินฝากประจำกับธนาคาร
2. ช่วยให้ธนาคารสามารถบริหารความเสี่ยงด้านสภาพคล่องได้อย่างมีประสิทธิภาพมากขึ้น

แนวคิด ทฤษฎี และงานวิจัยที่เกี่ยวข้อง

การพัฒนาแบบจำลองสำหรับพยากรณ์แนวโน้มการยกเลิกใช้บริการของลูกค้าเงินฝากประจำต้องศึกษาทฤษฎีต่อไปนี้

1. Random Forest

Tony Yiu (2019) ได้อธิบายหลักการของ Random Forest ไว้ดังนี้ Random Forest สร้างแบบจำลองจาก Decision Tree หลายๆแบบจำลอง แต่ละแบบจำลองย่อยจะถูกสอนจากข้อมูลชุดเดียวกัน แต่ใช้ข้อมูลย่อยที่แตกต่างกัน โดยข้อมูลย่อยที่ถูกใช้ในการสอนแบบจำลองได้มาจากการสุ่ม ทั้งนี้คำตอบที่ได้จากการทำนายของ Random Forest จะเกิดจากการเฉลี่ยหรือการโหวตเลือกผลลัพธ์จาก Decision Tree แล้วแต่ว่าเป็นการทำนายความน่าจะเป็นหรือทำนายประเภทข้อมูล



ภาพที่ 1 อธิบายหลักการของ Random Forest

2. Imbalance Data

ข้อมูลไม่สมดุลคือ ข้อมูลที่มีจำนวนในแต่ละกลุ่มแตกต่างกันมาก ซึ่งจะมีผลต่อการจำแนกคลาสส่วนน้อย เนื่องจากอัลกอริทึมมีความเอนเอียงไปทางกลุ่มข้อมูลส่วนใหญ่ ทำให้ทำนายคลาสส่วนน้อยผิดพลาด เทคนิคที่นำมาใช้ในการแก้ปัญหาข้อมูลไม่สมดุลมีดังนี้ (Francesco Pochetti, 2019)

2.1 วิธีสุ่มเกิน (Over sampling) คือ การสุ่มเลือกข้อมูลจากคลาสส่วนน้อย เพื่อให้คลาสส่วนน้อยมีจำนวนใกล้เคียงกับคลาสส่วนมาก โดยการสุ่มเกินนี้อาจเป็นการสุ่มจากข้อมูลเดิมหรือสร้างข้อมูลขึ้นมาใหม่จากตัวอย่างข้อมูลเดิม

SMOTE เป็นเทคนิคการสุ่มเกินอีกรูปแบบหนึ่ง โดยจะมีการสุ่มข้อมูลที่ได้จากการสร้างข้อมูลจากข้อมูลส่วนน้อยขึ้นมาใหม่ (Cornellius Yudha Wijaya, 2020)

2.2 วิธีสุ่มลด (Under sampling) คือ การสุ่มลดข้อมูลจากคลาสส่วนมาก เพื่อให้คลาสส่วนมากมีจำนวนใกล้เคียงกับคลาสส่วนน้อย

งานวิจัยที่เกี่ยวกับการพัฒนาแบบจำลองสำหรับพยากรณ์แนวโน้มการยกเลิกใช้บริการของลูกค้าเงินฝากประจำ และที่เกี่ยวข้องการแก้ปัญหาข้อมูลไม่สมดุล ที่ได้ศึกษามาสรุปได้ดังนี้

จิรกฤต บุญหมื่นไวย, เจษฎา ตันตานุช, เบญจวรรณ โรจนดิษฐ์ (2563) ได้นำเสนองานวิจัยที่ศึกษาความสัมพันธ์ของตัวแปรที่มีผลต่อการขอยกเลิกใช้บริการ โดยใช้การถดถอยโลจิสติกส์ จากนั้นนำตัวแปรที่ได้ไปสร้างแบบจำลองด้วยวิธีการต้นไม้ตัดสินใจ การถดถอยโลจิสติกส์ และ ตัวจำแนกประเภทแบบเบย์อย่างง่าย จากการศึกษาพบว่าปัจจัยที่มีผลต่อการคงสถานะของลูกค้าได้แก่ คะแนนเครดิต ที่อยู่อาศัย เพศ อายุ ยอดเงินคงเหลือในบัญชี จำนวนผลิตภัณฑ์ และการเป็นสมาชิกของธนาคาร

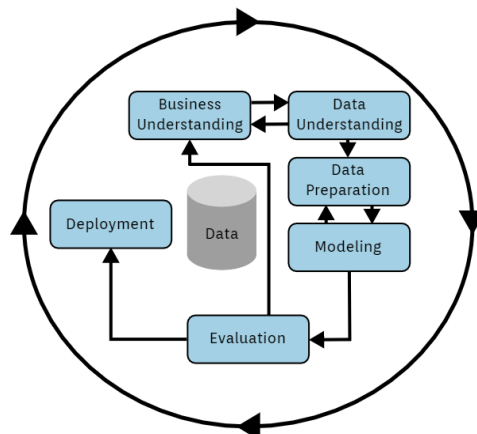
Abbas Keramati, Hajar Ghaneei and Seyed Mohammad Mirmohammadi (2016) นำเสนองานวิจัยที่ศึกษาปัจจัยที่มีผลในการยกเลิกการใช้บริการธนาคารอิเล็กทรอนิกส์ โดยใช้วิธีการ CRISP ในการดำเนินการวิจัย เลือกใช้วิธีการ forward selection และ backward elimination ในขั้นการเลือกตัวแปร โดยผลการวิจัยพบว่า backward elimination ให้ผลลัพธ์ที่ดีกว่า แบบจำลองที่ใช้คือต้นไม้ตัดสินใจ เนื่องจากผู้วิจัยต้องการทราบคุณลักษณะของผู้ใช้บริการที่จะยกเลิกเพื่อให้ผู้บริหารของธนาคารสามารถกำหนดกลยุทธ์ในการป้องกันการยกเลิกการใช้บริการ จึงเลือกใช้แบบจำลองที่สามารถอธิบายปัจจัยที่เป็นที่มีผลได้ง่าย

Nelson Rosa (2018) ได้นำเสนองานวิจัยโดยใช้เทคนิคโครงข่ายประสาทเทียมในการทำนายลูกค้าที่จะยกเลิกบริการของธนาคารรายย่อยในประเทศโปรตุเกส โดยใช้วิธีการ CRISP-DM ในการดำเนินการวิจัย ใช้ข้อมูล มกราคม-มิถุนายน 2017 เป็นข้อมูลสำหรับสร้างแบบจำลอง เพื่อทำนายลูกค้าที่จะปิดบริการในช่วง กรกฎาคม - ธันวาคม 2017 ผลจากการสอนแบบจำลองพบว่า ปัจจัยที่มีผลคือ อายุ และ ผลิตภัณฑ์ด้านเงินฝาก ผู้วิจัยทดสอบโดยใช้เทคนิคโครงข่ายประสาทเทียม 4 โมเดล ที่ต่างกันในระดับ Hidden layers คือ มี 1 ชั้น, 2 ชั้น, 3 ชั้น และ 4 ชั้น โดยโมเดลที่มีให้ค่าความถูกต้องมากที่สุดคือ โมเดลที่มี Hidden layers 3 ชั้น

Himanshu147 (2020) ได้นำเสนอแนวทางในการพัฒนาแบบจำลองเพื่อทำนายลูกค้าที่จะยกเลิกใช้บริการเงินฝากกระแสรายวัน โดยใช้การถดถอยโลจิสติกส์ในการสร้างแบบจำลอง และใช้ SMOTE ในการแก้ปัญหาข้อมูลไม่สมดุล ออกแบบการทดลองโดยการแบ่งข้อมูลออกเป็น 3 ชุด คือ ข้อมูลตั้งเดือนธันวาคม 2016 ถึง พฤศจิกายน 2017 ใช้เป็น Observation window ข้อมูลเดือนธันวาคม 2017 ใช้เป็น Static pool month ซึ่งคือจุดอ้างอิงข้อมูลซึ่งใช้สำหรับเลือกลูกค้าที่มีบัญชีกระแสรายวันในเดือนนั้นๆ เพื่อใช้ทำนาย ข้อมูลในช่วง มกราคม 2018 ถึง มิถุนายน 2018 ใช้เป็น Performance Window ซึ่งคือช่วงเวลาใช้ทดสอบผลการทำนาย จากการทำนายพบว่าปัจจัยที่สำคัญคือ ยอดเงินคงเหลือเฉลี่ยใน 3 เดือน จำนวนการทำรายการ ยอดรวมค่าธรรมเนียมของ 3 เดือนล่าสุด ประเภทของลูกค้า และ ผลิตภัณฑ์สินเชื่อที่มีหลักประกัน

ระเบียบวิธีวิจัย

การศึกษาครั้งนี้ เป็นการนำเสนอเทคนิคทำนายแนวโน้มการยกเลิกใช้บริการสำหรับลูกค้าธนาคารโดยวิธีการเรียนรู้ของเครื่อง ใช้วิธีการ CRISP-DM ในกระบวนการทำวิจัย ซึ่งประกอบด้วย 6 กระบวนการดังภาพที่ 2



ภาพที่ 2 กระบวนการ CRISP-DM

1. ความเข้าใจทางธุรกิจ (Business Understanding)

การให้บริการด้านเงินฝากของธนาคารสามารถแบ่งประเภทเงินฝากออกเป็น 2 ประเภท ใหญ่ๆ ได้แก่ เงินฝากเพื่อเรียก คือเงินฝากที่ผู้ฝากสามารถถอนได้เมื่อต้องการโดยไม่มีเงื่อนไข และ เงินฝากประจำ คือเงินฝากที่มีกำหนดเวลา เมื่อฝากครบกำหนดเวลาผู้ฝากจะได้ผลตอบแทนตาม เงื่อนไขที่ตกลงกัน

ในการให้สินเชื่อกับลูกค้า โดยส่วนใหญ่แล้วธนาคารจะนำเงินฝากประจำมาบริหารจัดการเนื่องจากมีกำหนดเวลาครบกำหนดที่แน่นอน สามารถบริหารและควบคุมความเสี่ยงได้ ด้วย ลักษณะของระยะเวลาครบกำหนดเงินฝากที่สั้นกว่าสินเชื่อมาก ทำให้การติดตามปริมาณการไหล ออกของเงินฝากซึ่งสามารถเกิดได้ทั้งกรณีที่ครบกำหนดสัญญาและการถอนก่อนครบกำหนด เป็น การกิจสำคัญเพื่อติดตามให้สัดส่วนของเงินฝากและสินเชื่อเป็นไปตามเกณฑ์ของธนาคารแห่ง ประเทศไทย

2. ความเข้าใจข้อมูลที่ใช้ในงาน (Data Understanding)

2.1 กระบวนการเก็บข้อมูลเงินฝากประจำ ข้อมูลที่ใช้ประกอบด้วย ข้อมูลระดับบัญชี และข้อมูลระดับลูกค้า ของบัญชีเงินฝากที่ยังมีสถานะเปิดบัญชี ณ 30 พฤษภาคม 2563 สำหรับเป็น ตัวแปรในการสร้างแบบจำลอง และข้อมูลสถานะบัญชี ณ 31 ธันวาคม 2563 สำหรับเป็นข้อมูลป้ายกำกับ โดยตัวแปรที่นำมาใช้สามารถแบ่งออกเป็น 4 กลุ่ม ดังนี้

1) ข้อมูลระดับบัญชี ได้แก่ ยอดเงินคงเหลือ, กำไรค้างรับ, อัตรากำไร, จำนวน เดือนที่จะครบสัญญา, จำนวนเดือนตั้งแต่เปิดบัญชี, ภูมิภาคของสาขาเจ้าของบัญชี

2) ข้อมูลระดับลูกค้า ได้แก่ อายุ, จำนวนปีที่เป็นลูกค้า, กลุ่มลูกค้าตามยอดเงินฝากรวม, ประเภทลูกค้า, เพศ, ช่วงรายได้, แหล่งที่มาของรายได้, ระดับการศึกษา, อาชีพ, จำนวนบัญชีฝากประจำ 3 เดือน, จำนวนบัญชีฝากประจำ 6 เดือน, จำนวนบัญชีฝากประจำ 9 เดือน, จำนวนบัญชีฝากประจำ 12 เดือน, จำนวนบัญชีฝากประจำ 24 เดือน, จำนวนบัญชีฝากประจำทุกประเภท, จำนวนบัญชีเงินฝากออมทรัพย์, เงินฝากรวมของฝากประจำ 3 เดือน, เงินฝากรวมของฝากประจำ 6 เดือน, เงินฝากรวมของฝากประจำ 9 เดือน, เงินฝากรวมของฝากประจำ 12 เดือน, เงินฝากรวมของฝากประจำ 24 เดือน, เงินฝากรวมของฝากประจำ, เงินฝากรวมของออมทรัพย์

3) ข้อมูลระดับ Transaction ได้แก่ ยอดรวมรายการถอนเงินของเดือนก่อนหน้า และ ยอดกำไรจ่ายของปีก่อนหน้า

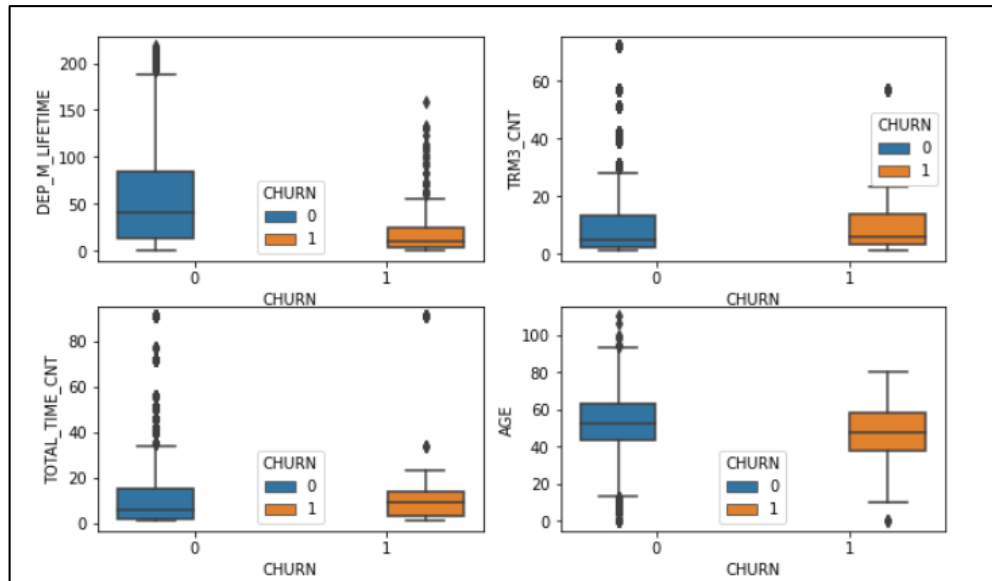
4) ข้อมูลป้ายกำกับ ได้แก่ 1 หมายถึง Churn และ 0 หมายถึง Not Churn

2.2 ทำความเข้าใจภาพรวมจำนวนบัญชี โดยสัดส่วนของบัญชีที่มีป้ายกำกับเป็น 0 และ 1 ข้อมูลมีทั้งสิ้น 4,919 ตัวอย่าง แบ่งออกเป็นบัญชีที่มีป้ายกำกับเป็น 1 (Churn) 166 ตัวอย่าง และบัญชีที่มีป้ายกำกับเป็น 0 (Not Churn) 4,753 ตัวอย่าง คิดเป็นสัดส่วนบัญชีที่ Churn ต่อบัญชีทั้งหมดอยู่ที่ 3.4% ลักษณะข้อมูลมีความไม่สมดุลค่อนข้างมาก

3. ตรวจสอบความถูกต้องของข้อมูลและเตรียมข้อมูลที่ใช้ในงาน (Data Preparation)

3.1 กระบวนการตรวจสอบความครบถ้วนของข้อมูล (Data Exploration) พบว่ามีตัวแปรที่มีค่าว่าง ได้แก่ ยอดรวมการถอนเงินเดือนก่อนหน้าของลูกค้า, เพศ, รายได้, การศึกษา และ อาชีพ

3.2 กระบวนการตรวจหาค่าผิดปกติ (Outlier) โดยการสร้างชาร์ต boxplot พบว่ามีตัวแปร 4 ค่าที่มีข้อมูลผิด ได้แก่ จำนวนเดือนตั้งแต่เปิดบัญชี (DEP_M_LIFETIME), จำนวนบัญชีฝากประจำทั้งหมด (TOTAL_TIME_CNT), จำนวนบัญชีฝากประจำ 3 เดือน (TRM3_CNT) และ อายุ (AGE) ดังภาพที่ 3



ภาพที่ 3 Boxplot ของตัวแปรที่มีค่าผิดปกติ

3.3 กระบวนการทำความสะอาดข้อมูล (Data Cleansing)

3.3.1 เติมค่าตัวแปรที่มีค่าว่าง โดยมีรูปแบบในการเติมข้อมูลที่มีค่าว่างออกเป็น 3 วิธีดังนี้

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

- ข้อมูลที่มี type เป็น Object: เติมค่าว่างด้วยค่า Mode ของแต่ละตัวแปร

- ข้อมูลที่มี type ไม่เป็น Object: เติมค่าว่างด้วยค่าเฉลี่ยของแต่ละตัวแปร

3.3.2 กำจัดค่าผิดปกติ ด้วยวิธีการคำนวณค่าสถิติด้วยสูตร Quartile ดังนี้

$$\text{Upper bound} = Q3 + 1.5 * \text{IQR} \text{ และ } \text{Lower bound} = Q1 - 1.5 * \text{IQR}$$

โดยข้อมูลผิดปกติที่จะถูกกำจัดออกคือ ข้อมูลที่มีค่ามากกว่า Upper bound หรือ น้อยกว่า Lower bound

3.4 กระบวนการเลือกคุณลักษณะ (Feature Selection) ใช้การทดสอบข้อมูลทางสถิติ โดยแบ่งเป็น 2 เทคนิคได้แก่

ตัวแปรที่มีค่าต่อเนื่อง (Continuous Variable) ทดสอบโดยใช้ ANOVA ได้ตัวแปรที่มีค่า P-Value < 0.05 ดังนี้ ยอดเงินคงเหลือ,กำไรค้างรับ,จำนวนเดือนตั้งแต่เปิดบัญชี,จำนวนบัญชีฝาก

ประจำ 3 เดือน และ 6 เดือน,จำนวนบัญชีฝากประจำทั้งหมด,เงินฝากรวมของฝากประจำ 3 เดือน และ 6 เดือน,อายุ และ จำนวนปีที่เป็นลูกค้า

ตัวแปรที่มีค่าไม่ต่อเนื่อง (Categorical Variable) ทดสอบโดยใช้ Chi-Square Test ได้ตัวแปรที่มีค่า $P\text{-Value} < 0.05$ ดังนี้ ลูกค้าตามยอดเงินฝากรวม, ประเภทลูกค้า,ช่วงรายได้,แหล่งที่มาของรายได้,ระดับการศึกษา,อาชีพ และ ภูมิภาคของสาขาเจ้าของบัญชี

3.5 กระบวนการเปลี่ยนแปลงข้อมูล (Data Transformation)

3.5.1 แปลงข้อมูลที่ถูกเก็บในลักษณะ Categorical โดยแบ่งออกเป็น 3 รูปแบบ ดังนี้

- ข้อมูลที่สามารถบอกลำดับขั้นได้ (Ordinal Categorical Variables) ได้แก่ ตัวแปร ซึ่งเก็บค่าการจัดกลุ่มของลูกค้าตามยอดเงินฝากรวม ทำการแปลงค่าตามลำดับขั้นคือ Retail แทนด้วย 0 , Medium แทนด้วย 1 และ Corporate แทนด้วย 2

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

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

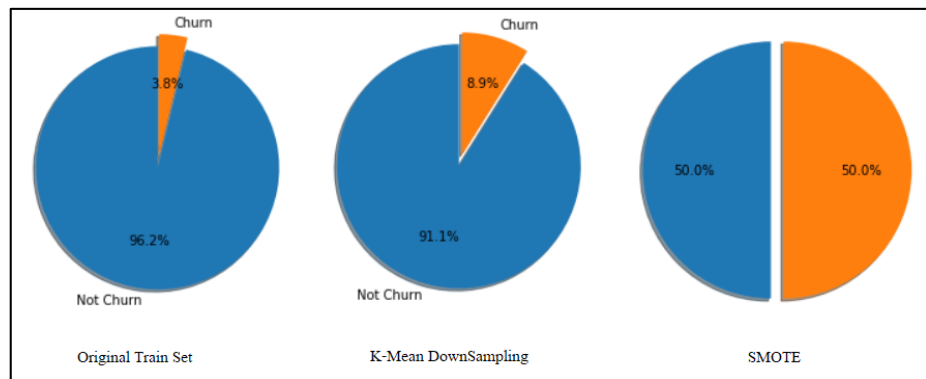
3.5.2 ปรับช่วงของข้อมูลให้เป็นช่วงเดียวกัน ด้วยเทคนิค Standardization เนื่องจากการที่ข้อมูลแต่ละตัวแปรมีช่วงของข้อมูลที่แตกต่างกัน เช่น ทำให้เกิด feature bias ในขั้นตอนการสร้างโมเดล

4. การพัฒนา Model) Modeling)

4.1 กระบวนการเตรียมข้อมูลสำหรับสอนแบบจำลอง

- แบ่งข้อมูลสำหรับสอนและทดสอบแบบจำลองด้วยสัดส่วน 80:20
- แก้ปัญหาข้อมูลไม่สมดุลของชุดข้อมูลสอน โดยใช้เทคนิคการจัดกลุ่มคลาสส่วนมากด้วยวิธี k-mean แล้วทำการสุ่มลดจากแต่ละคลาสเตอร์เพื่อเป็นการสร้างข้อมูลคลาสส่วนมากใหม่ให้มีขนาดใกล้เคียงกับคลาสส่วนน้อยมากขึ้น จากนั้นใช้วิธีสังเคราะห์ข้อมูลเพิ่ม

(Synthetic Minority Oversampling Technique: SMOTE) เพื่อปรับให้ข้อมูลคลาสส่วนน้อยและคลาสส่วนมากมีขนาดใกล้เคียงกัน ผลที่ได้ดังภาพที่ 4



ภาพที่ 4 กระบวนการแก้ปัญหาข้อมูลไม่สมดุล

4.2 กระบวนการหาไฮเพอร์พารามิเตอร์ที่เหมาะสมที่สุดสำหรับอัลกอริทึม Random Forest โดยใช้ฟังก์ชัน GridSearchCV ผลที่ได้ ดังภาพที่ 5

```

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)

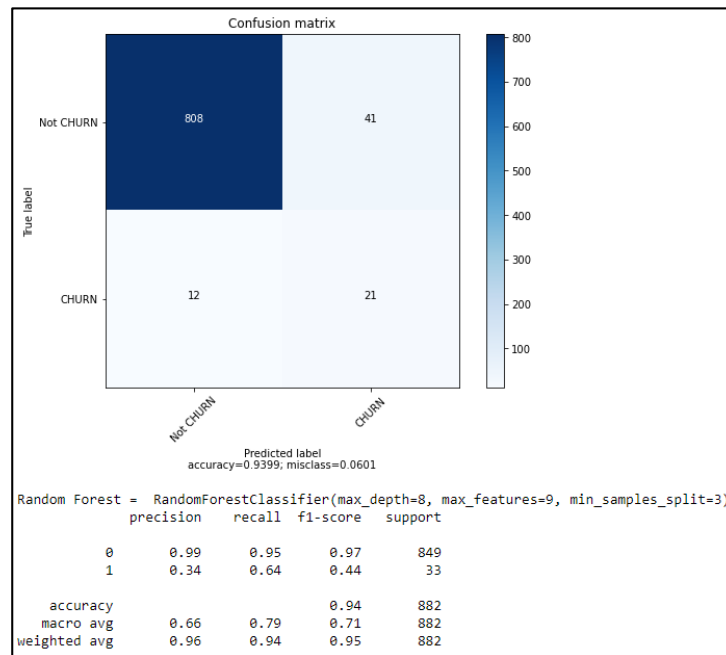
0.9242916449813002
{'max_depth': 8, 'max_features': 9, 'min_samples_split': 3, 'n_estimators': 100}
RandomForestClassifier(max_depth=8, max_features=9, min_samples_split=3)

```

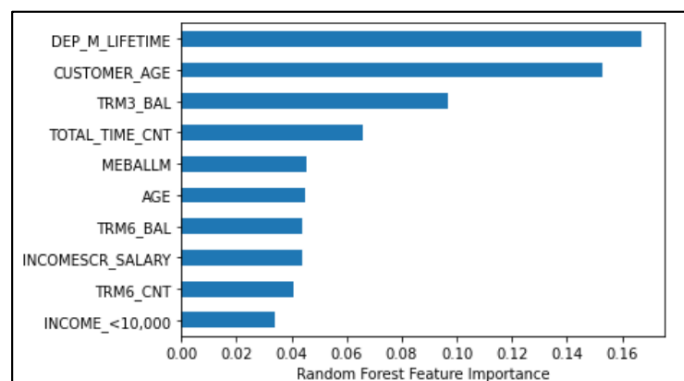
ภาพที่ 5 ค่าพารามิเตอร์ที่เหมาะสมของ Random Forest

5. กระบวนการวัดประสิทธิภาพแบบจำลอง (Model Evaluation)

วัดประสิทธิภาพของแบบจำลอง Random Forest โดยใช้ชุดข้อมูลสอน ผลที่ได้ดังภาพที่ 6 โดยมีตัวแปรที่เป็นปัจจัยสำคัญดังภาพที่ 7



ภาพที่ 6 Confusion Matrix : Random Forest



ภาพที่ 7 Feature Importance : Random Forest

6. การนำมาใช้งาน

แปลงแบบจำลอง Random Forest ให้อยู่ในรูปแบบ file .pkl เพื่อนำไปใช้ทำนายข้อมูลจริง

6.1 เตรียมข้อมูลเพื่อนำเข้าในรูปแบบ Excel หรือ CVS File

6.2 เรียกใช้โปรแกรมเพื่อการทำนายด้วย Model ที่เลือก ดังภาพที่ 8

6.3 นำผลการทำนายซึ่งระบุความเป็นไปได้ในการยกเลิกบริการของแต่ละบัญชีไป

ใช้งาน

```
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

   0.00      0.98      0.60      0.74      4687
   1.00      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 8 การเรียกใช้ Model

ผลการวิจัย

1. ผลการวัดประสิทธิภาพความถูกต้องของแบบจำลองด้วยข้อมูลทดสอบ

ข้อมูลที่ใช้ในการทดสอบแบบจำลอง คือข้อมูลเงินฝากประจำ 3 เดือน ณ 30

พฤศจิกายน 2020

มีทั้งสิ้น 882 ตัวอย่าง โดยมีป้ายกำกับเป็น 0 (Not Churn) จำนวน 849 ตัวอย่าง และมีป้ายกำกับเป็น 1 (Churn) จำนวน 33 ตัวอย่าง ผลการทดสอบดังตามตารางที่ 1 และ ตารางที่ 2

ตารางที่ 1 ผลการทดสอบประสิทธิภาพของแบบจำลอง Random Forest แยกตามป้ายกำกับ

Churn			Not Churn		
precision	recall	f1	precision	recall	f1
0.34	0.64	0.44	0.99	0.95	0.97

ตารางที่ 2 ผลการทดสอบประสิทธิภาพของแบบจำลอง Random Forest แบบค่าเฉลี่ย

macro avg			weighted avg		
precision	recall	f1	precision	recall	f1
0.66	0.79	0.71	0.96	0.94	0.95

2. ผลการวัดประสิทธิภาพความถูกต้องของแบบจำลองด้วยข้อมูล Unseen

ใช้ชุดข้อมูลเงินฝากประจำ 3 เดือน ณ 31 ธันวาคม 2020 ในการทดสอบ ผลดังตารางที่ 3 และ 4

ตารางที่ 3 ผลการทดสอบประสิทธิภาพของแบบจำลองด้วยข้อมูล Unseen แยกตามป้ายกำกับ

Churn			Not Churn		
precision	recall	f1	precision	recall	f1
0.06	0.68	0.12	0.98	0.60	0.74

ตารางที่ 4 ผลการทดสอบประสิทธิภาพของแบบจำลองด้วยข้อมูล Unseen แบบค่าเฉลี่ย

macro avg			weighted avg		
precision	recall	f1	precision	recall	f1
0.52	0.64	0.43	0.94	0.60	0.72

สรุปและอภิปรายผลการวิจัย

การศึกษานี้มีวัตถุประสงค์เพื่อพัฒนาแบบจำลองสำหรับทำนายแนวโน้มการยกเลิกใช้บริการสำหรับลูกค้าธนาคาร โดยใช้ข้อมูลเงินฝากประจำ 3 เดือน มาทำงานล่วงหน้า เพื่อคาดการณ์ว่าในอีกหนึ่งเดือนข้างหน้าจะมีบัญชีใดบ้างที่มีแนวโน้มจะปิดบัญชี ผลที่ได้โดยสรุป ดังรายละเอียดต่อไปนี้

1. สรุปผลการศึกษา

1.1 ในขั้นตอนการพัฒนาแบบจำลอง แบบจำลองที่ได้มีประสิทธิภาพในการทำนายแนวโน้มการยกเลิกใช้บริการของลูกค้า มีค่า Average Recall เท่ากับ 79% และมีค่า Recall ของป้ายกำกับ 1 เท่ากับ 64% ทั้งนี้เมื่อใช้ทดสอบกับชุดข้อมูล Unseen ให้ค่า Average Recall ลดลง คือ ค่า Average Recall เท่ากับ 64% แต่มีค่า Recall ของป้ายกำกับ 1 เพิ่มขึ้นเป็น 68%

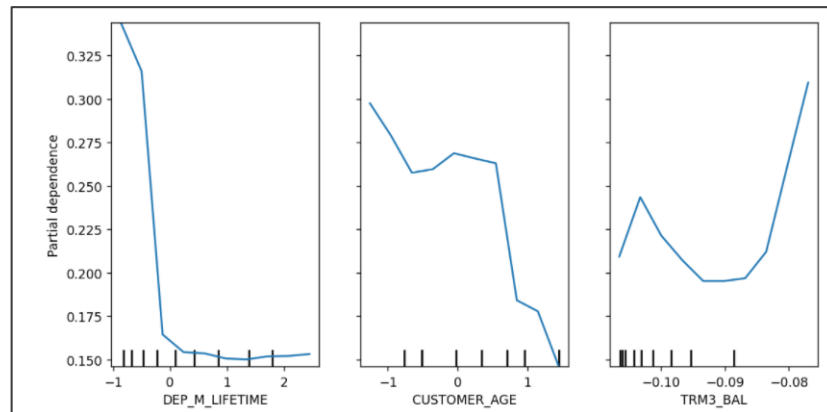
1.2 ผลจากการทำนายข้อมูล Unseen เมื่อพิจารณาในรูปแบบของยอดเงินฝาก ผลดังตารางที่ 5

ตารางที่ 5 Confusion Matrix : จำนวนบัญชีและยอดเงินจากการทำนายข้อมูล Unseen

		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 บาท
	Churn	62 บัญชี 3,208,730 บาท	129 บัญชี 938,130,147.72 บาท

จากตารางที่ 5 แบบจำลองสามารถทำนาย บัญชีที่มีแนวโน้มจะปิดได้ถูกต้อง 129 บัญชี จาก 191 บัญชี เป็นยอดเงินที่สามารถนำไปสู่การป้องกันการไหลออกจากระบบได้ 938,130,147.72 บาท ซึ่งคิดเป็น 99.67% ของจำนวนเงินที่ปิดบัญชีทั้งหมด

1.3 ปัจจัยที่มีความสัมพันธ์ต่อแนวโน้มในการยกเลิกใช้บริการ ในขั้นตอนการวัดประสิทธิภาพของแบบจำลองพบว่าตัวแปรสำคัญ 3 อันดับแรกคือ จำนวนเดือนตั้งแต่เปิดบัญชี (DEP_M_LIFETIME), จำนวนปีที่เป็ลูกค้า (CUSTOMER_AGE), ยอดเงินฝากรวมของเงินฝากประจำ 3 เดือน (TRM3_BAL) โดยตัวแปรทั้ง 3 มีผลต่อแนวโน้มในการยกเลิกใช้บริการในทิศทางดังภาพที่ 9



ภาพที่ 9 ทิศทางที่มีผลต่อแนวโน้มในการยกเลิกใช้บริการของ 3 ตัวแปรหลัก

จากภาพ จำนวนเดือนตั้งแต่เปิดบัญชีที่มีค่าน้อยมีความเป็นไปได้สูงที่ลูกค้าจะยกเลิกการใช้บริการ โดยจะส่งผลน้อยลงอย่างมากเมื่อเปิดบัญชีนานขึ้น อายุของลูกค้ามีผลในทิศทางผกผันกับค่าความเป็นไปได้ที่ลูกค้าจะยกเลิกการใช้บริการ ในขณะที่ยอดเงินฝากรวมของเงินฝากประจำ 3 เดือน จะส่งผลต่อค่าความเป็นไปได้ที่ลูกค้าจะยกเลิกการใช้บริการในทิศทางเดียวกันเมื่อยอดเงินฝากมีจำนวนมากขึ้น

ข้อเสนอแนะ

1. แนวทางในการเพิ่มประสิทธิภาพของแบบจำลอง คือ เพิ่มตัวแปรในมิติอื่น เช่น ข้อมูลอัตราผลตอบแทนจากภายนอก ข้อมูลผลิตภัณฑ์ด้านสินเชื่อของลูกค้า ซึ่งถือเป็นปัจจัยที่อาจจะมีผลในการตัดสินใจปิดบัญชี
2. ในการสร้างแบบจำลองอาจทำได้โดยแยกกลุ่มลูกค้า เนื่องจากลูกค้ารายและรายย่อยอาจมีปัจจัยในการตัดสินใจปิดบัญชีที่แตกต่างกัน

บรรณานุกรม

จิรกฤต บุญหมื่นไวย, เจษฎา ตัณฑนุช, 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

Cornellius Yudha Wijaya. (2020). 5 SMOTE Techniques for Oversampling your Imbalance Data.

Retrieved from <https://towardsdatascience.com/5-smote-techniques-for-oversampling-your-imbalance-data-b8155bde2b5>

Francesco Pochetti. (2019). Extreme Label Imbalance: When You Measure the Minority Class in Basis Points. Retrieved from <http://francescopochetti.com/extreme-label-imbalance-when-you-measure-the-minority-class-in-basis-points/>.

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

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

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.

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

NOVA Information Management School, Universidade Nova de Lisboa.

Tony Yiu. (2019). Understanding Random Forest How the Algorithm Works and Why it Is So Effective.

Retrieved from <https://towardsdatascience.com/understanding-random-forest-58381e0602d2>.