การประยุกต์เหมืองข้อมูลในการพยากรณ์สถานะ NPLs ของลูกค้าสินเชื่อ กรณีศึกษา ธนาคารออมสิน สาขามหาวิทยาลัยเชียงใหม่

Application of Data Mining In Forecasting NPLS Status of GSB Loan Case study in Chiang Mai University branch

ปารดา ศัสตุระ[1]

[1]MBA-Online คณะบริหารธุรกิจ, มหาวิทยาลัยหอการค้าไทย,Email: Parada.cmu@gmail.com

---

บทคัดย่อ

ธนาคารออมสินมีนโยบายเร่งด่วนในการช่วยเหลือโดยการออกสินเชื่อ Covid-19 เพื่อบรรเทาภาระค่าใช้จ่ายของประชาชน แต่ส่งผลทำให้ธนาคารมีลูกหนี้ที่ไม่ก่อให้เกิดรายได้ NPLs ในการศึกษาครั้งนี้มีวัตถุประสงค์ เพื่อศึกษาปัจจัยที่ไม่ก่อให้เกิดรายได้ NPLs และการประยุกต์เหมืองข้อมูลในการพยากรณ์สถานะ NPLs ของลูกค้าสินเชื่อ กรณีศึกษา ธนาคารออมสิน สาขามหาวิทยาลัยเชียงใหม่ ใช้กระบวนการวิเคราะห์ข้อมูลมาตรฐาน เรียกว่า Cross-Industry Standard Process (CRISP) ประกอบด้วย 6 ขั้นตอน ได้แก่ เข้าใจปัญหาของธุรกิจ เข้าใจข้อมูล เตรียมข้อมูล พัฒนาแบบจำลอง การประเมิน และการนำไปใช้จริง ข้อมูลที่ใช้รวบรวมข้อมูลจากลูกค้าที่ได้รับสินเชื่อ Covid-19 ของธนาคารออมสิน สาขามหาวิทยาลัย เชียงใหม่ ตั้งแต่เดือน กรกฎาคม 2564 ถึง ธันวาคม 2564 จำนวน 7,215 ราย การเตรียมและคัดเลือกข้อมูลให้สมบูรณ์เหมาะสำหรับการทำเหมืองข้อมูลด้วยกระบวนการ Extract Transform Load (ETL) จากนั้นทำการโอนย้ายข้อมูล การลดขนาดของข้อมูล และการทำความสะอาดข้อมูล ใช้โปรแกรม WEKA เป็นเครื่องมือ วิเคราะห์หาความสัมพันธ์ และใช้อัลกอริทึมการพยากรณ์สถานะ NPLs ของลูกค้าสินเชื่อ 3 แบบ ได้แก่ Decision Tree, Naive Bayes และ Random Forest การทดสอบประสิทธิภาพอัลกอริทึมด้วยการหารากของค่าความคลาดเคลื่อนกำลังสองเฉลี่ย (Root Mean Square Error : RMSE) เพื่อหาอัลกอริทึมที่ดีที่สุด และวัดผล ผลการศึกษาพบว่า ปัจจัยที่มีผลต่อการเกิดหนี้ที่ไม่ก่อให้เกิดรายได้ ได้แก่ อาชีพ สถานะภาพ ระดับการศึกษา รายได้ เพศ และผลการพยากรณ์สถานะ NPLs ของลูกค้าสินเชื่อ จากอัลกอริทึม Decision Tree (J48) มีค่าถูกต้องสูงสุด คิดเป็นร้อยละ 96.15% เมื่อเทียบกับอัลกอริทึม Regression และ Random Forest ถูกต้องร้อยละ 94.52 % และ 96.10 % ตามลำดับ

คำสำคัญ: การเรียนรู้ของเครื่อง , NPLs

Abstract

Government Saving Bank (GSB) implemented the emergency Covid-19 relief loan policy to help individuals experiencing financial burden. The impact on the policy lead to NPLs. The purpose of this study are to study the parameters that lead to NPLs and to forecaste NPLS Status of GSB Loan using data mining. The technique in this study is the Cross-Industry Standard Process (CRISP). There are 6 processes that are business understanding, data understanding, data preparation, model development, evaluation, and deployment. The data collected from

COVID-19 relief loan from 7,215 customer GSB Loan in Chiang Mai University branch during July-December 2020. To do the data preparation, feature selection for data mining i.e., data transfer, data reduction, data cleaning using WEKA as tool. There are 3 algorithms for forecasting NPLS that are Decision Trees, Naive Bayes and XXXXX. Algorithm performance testing by using the root of the mean square error (Root Mean Square Error : RMSE) for the best algorithm. The results can demonstrate that there are 5 parameters (occupation, status, education, income and gender) that lead to NPLs and the most accuracy algorithm is Gradient Boosted Tree with accuracy 96.15 % comparing with Simple Regression and Random Forest with accuracy 94.52 % and 96.10 %, respectively.

Keywords: Machine learning , NPLs

1. บทนำ / วัตถุประสงค์การวิจัย

ธนาคารออมสินเป็นธนาคารเพื่อสังคม โดยจะดูแลลูกค้าและประชาชน 3 กลุ่มหลัก ได้แก่ ผู้มีรายได้น้อย ผู้ประกอบการรายย่อย และองค์การชุมชน จากสถานการณ์โควิดส่งผลกระทบกับลูกค้าเป็นจำนวนมาก ทำให้กลุ่มธนาคารต้องเร่งเข้าช่วยเหลือลูกค้าที่ประสบปัญหาดังกล่าว โดยธนาคารออมสินได้ตอบรับนโยบายเร่งด่วน เพื่อบรรเทาความเดือนร้อนจากการขาดรายได้อันเนื่องมาจากมาตรการป้องกันและควบคุมการแพร่ระบาดที่ทางรัฐบาลประกาศขึ้นตามมติ ศบค. โดยการปล่อยสินเชื่อผู้มีอาชีพอิสระ Covid-19 ซึ่งในระยะแรกของโครงการจะเริ่มจากผู้ที่ได้รับผลกระทบที่อยู่ในพื้นที่สีแดงเข้ม และต่อมาก็ได้ขยายโครงการไปยังทุกพื้นที่ทั่วประเทศ ผลกระทบจากสถานการณ์ ส่งผลทำให้กำไรสุทธิปรับตัวลดลงส่งผลต่อปัญหาคุณภาพหนี้ จากรายงานผลการดำเนินงานตามแผนปฏิบัติการและงบประมาณ ณ ไตรมาส 1 ปี 2564 ธนาคารออมสินมีอัตราหนี้ค้างชำระเกิน 3 เดือน (NPLs) ของสินเชื่อรวม ร้อยละ 1.94 (ณ ก.พ. 2564) ทำให้ธนาคารต้องตั้งสำรองฯ ในระดับที่สูงขึ้นกว่าช่วงเวลาปกติ ซึ่งแนวโน้มในระยะต่อไป ธนาคารต้องเตรียมสภาพคล่องเพื่อสนับสนุนในการฟื้นกิจการและดำเนินการควบคู่ไปกับการเร่งปรับโครงสร้างหนี้และให้ความช่วยเหลือลูกหนี้ในกลุ่มที่ได้รับผลกระทบ เพื่อเป็นแนวทางในการรักษาคุณภาพหนี้ของธนาคารไว้ให้อยู่ในระดับที่ยอมรับได้ ดังนั้น ธนาคารออมสิน ซึ่งตอบรับนโยบายเร่งด่วนนี้ จึงทำการศึกษาเพื่อพยากรณ์โอกาสเกิด NPLs ของลูกค้าสินเชื่อผู้มีอาชีพอิสระ Covid–19 เพื่อวางแผนในอนาคตต่อไป

2. การดำเนินการวิจัย

1. แนวคิดในการศึกษาในครั้งนี้ จากการช่วยเหลือเพื่อบรรเทาภาระค่าใช้จ่ายของประชาชนที่ธนาคารออมสินได้ตอบรับนโยบายการออกสินเชื่อนั้น ส่งผลทำให้ธนาคารมีลูกหนี้ที่ไม่ก่อให้เกิดรายได้ NPLs เนื่องจากกลุ่มลูกค้าที่ได้รับสินเชื่อและไม่สามารถชำระหนี้ได้ตามปกติ ว่าเกิดจากกปัจจัยใด และทำนายปัจจัยการเกิดหนี้ที่ไม่ก่อให้เกิดรายได้ (NPLs)

2. การพยากรณ์ (Forecasting) คือ การประมาณ หรือการคาดคะเนเหตุการณ์ที่จะเกิดขึ้นในอนาคต เช่น การพยากรณ์รายได้-รายจ่าย การพยากรณ์ยอดขาย หรือการพยากรณ์จำนวนลูกค้าที่เข้ามาใช้บริการ เป็นต้น ซึ่งการพยากรณ์จะมีประโยชน์ในการวางแผนและการตัดสินใจขององค์กร การพยากรณ์จำแนกได้เป็น 2 ประเภท คือ การพยากรณ์เชิงปริมาณ จะเป็นการพยากรณ์จากการวิเคราะห์ข้อมูลตัวเลข ส่วนการพยากรณ์เชิงคุณภาพ จะเป็นการพยากรณ์จากความคิดเห็นและประสบการณ์ ในกระบวนการพยากรณ์ (Forecasting Process) มีทั้งหมด 5 ขั้นตอน โดย

เริ่มจากการกำหนดวัตถุประสงค์ กำหนดช่วงเวลาที่ต้องการพยากรณ์ เลือกเทคนิคการพยากรณ์ เก็บข้อมูลที่ต้องใช้ในการพยากรณ์ และขั้นตอนสุดท้าย คือการทำการพยากรณ์

3. NPLs (Non-Performing Loan) หรือหนี้เสีย คือ สินเชื่อที่ไม่ก่อให้เกิดรายได้ หรือ หนี้เสีย ซึ่งหมายถึงหนี้ที่ไม่ได้รับการชำระคืนตามข้อมกำหนดตกลงกันตามสัญญา ซึ่งจะเกิดขึ้นเมื่อได้ทำการขอสินเชื่อเอาไว้ แล้วไม่ได้ชำระหนี้นานเกินจำนวนวันที่กำหนด โดยปกติแล้วกำหนดไว้ที่ 90 วัน โดยสาเหตุการเกิดมาจากปัจจัยหลายอย่าง ไม่ว่าจะเป็นรายได้ สภาพรวมของตลาดในประเทศ วินัยทางการเงินในการชำระหนี้ของลูกหนี้ เมื่อเกิด NPL จะส่งผลระยะยาวในการขอยื่นกู้สินเชื่อในอนาคต เหมือนกับประวัติที่ทำให้สถาบันการเงินไม่แน่ใจในความสามารถในการชำระหนี้ และทำให้สถาบันการเงินต้องทำการกันเงินสำรองตามเกณฑ์ของธนาคารแห่งประเทศไทยไว้เผื่อเกิดกรณี NPL

4. การทำเหมืองข้อมูล (Data Mining) เป็นกระบวนการที่กระทำกับข้อมูลจำนวนมากเพื่อค้นหารูปแบบและความสัมพันธ์ที่ซ่อนอยู่ในชุดข้อมูลนั้น การทำเหมืองข้อมูลได้ถูกนำไปประยุกต์ใช้กับงานหลากหลายประเภท ซึ่งแบ่งประเภทของการทำเหมืองข้อมูลเป็น 2 ประเภท ได้แก่ 1.การทำเหมืองข้อมูลแบบบรรยาย (Descriptive data mining) เป็นการนำข้อมูลที่มีอยู่มาใช้เพื่อการอธิบายเพื่อเป็นแนวทางในการตัดสินใจ โดยไม่ได้มีวัตถุประสงค์เพื่อการทำนาย เช่น การหาความสัมพันธ์ (Association) และ 2.การทำเหมืองข้อมูลแบบทำนาย (Predictive data mining) เป็นการนำข้อมูลที่เกิดขึ้นในอดีตมาทำนายผลข้อมูลที่จะเกิดขึ้นในอนาคต ซึ่งในขั้นตอนของการทำเหมืองข้อมูล จะมีเทคนิคแบบ Supervised classification และ Unsupervised classification โดยเทคนิคทั้ง 2 รูปแบบ จะมีการใช้มาตรฐานที่เรียกว่า CRISP-DM (Cross Industry Standard Process for Data Mining) ซึ่งเป็นกระบวนการในการทำเหมืองข้อมูลที่ทำให้ข้อมูลที่ได้รับมีประสิทธิภาพและมีผลลัพธ์ที่มีความถูกต้อง โดยมีขั้นตอน ดังนี้

ขั้นตอนที่ 1 ทำความเข้าใจทางธุรกิจ (Business Understanding)
ขั้นตอนที่ 2 ทำความเข้าใจข้อมูล (Data Understanding)
ขั้นตอนที่ 3 การเตรียมข้อมูล (Data Preparation)
ขั้นตอนที่ 4 การพัฒนาแบบจำลอง (Modeling)
ขั้นตอนที่ 5 การทดสอบผลแบบจำลอง (Evaluation)
ขั้นตอนที่ 6 การนำแบบจำลองไปใช้งาน (Deployment)

ซึ่งการศึกษาครั้งนี้ ผู้ศึกษาได้ใช้โปรแกรม Weka Vision 3.8.5 เป็นโปรแกรมที่ใช้ในการวิเคราะห์ข้อมูลด้วยเทคนิคเหมืองข้อมูล คือ การวิเคราะห์ข้อมูลเพื่อหาความสัมพันธ์ รูปแบบและแยกประเภทข้อมูล นอกจากนั้นยังรวบรวมเทคนิคในการวิเคราะห์ข้อมูลที่หลากหลายไว้ เช่น เทคนิคแบบ Tree J48 , Decision Rule: Part และ Bayes: naive Bayes เป็นต้น

งานวิจัยที่เกี่ยวข้อง

1. การประยุกต์ใช้เทคนิคการเรียนรู้ของคอมพิวเตอร์เพื่อพัฒนาตัวแบบทางเหมืองข้อมูลสำหรับพยากรณ์ระดับฮีโมโกลบินของผู้บริจาคโลหิต นำเสนอเกี่ยวกับปริมาณโลหิตและส่วนประกอบที่เพียงพอในธนาคารเลือดเพื่อใช้ในการรักษาผู้ป่วย ดังนั้นจึงต้องมีการส่งเสริมให้ผู้บริจาคโลหิตมีสุขภาพดีและมีคุณสมบัติที่เหมาะสมในการบริจาคเลือดอย่างสม่ำเสมอ โดยวัตถุประสงค์เพื่อพัฒนาและเปรียบเทียบประสิทธิภาพการเรียนรู้ในการพยากรณ์

จำแนกกลุ่มผลตรวจ Hb ของผู้บริจาคโลหิต จากการศึกษาเก็บข้อมูล 44 ตัวแปรของผู้บริจาคโลหิตจำนวน 2,180 ราย จากภาคบริจาคโลหิตแห่งชาติ 12 แห่งและสถานีกาชาดหัวหินเฉลิมพระเกียรติ และพัฒนาตัวแบบทางเหมืองข้อมูล ได้แก่ ต้นไม้ตัดสินใจ (Decision Tree) ซัพพอร์ตเวกเตอร์แมชชีน (SVM) การจำแนกแบบเบส์อย่างง่าย (Naive Bayesain Classifier) และโครงข่ายประสาทเทียม (Artificial Neural Networks) จากนั้นเปรียบเทียบประสิทธิภาพการจำแนกกลุ่มของตัวแบบพยากรณ์พบว่าต้นไม้ตัดสินใจเป็นตัวแบบการจำแนกกลุ่มที่เหมาะสมที่สุดโดยให้ค่าความถูกต้อง (accuracy) ค่าความไว (sensitivity) ค่าความจำเพาะ (specificity) ค่าการพยากรณ์ผลบวก (positive predictive value) และ ค่าพยากรณ์ผลลบ (negative predictive value) สูงสุด ซึ่งผู้วิจัยจะได้นำไปพัฒนาเป็นระบบประเมินผลออนไลน์ในการบริจาคโลหิตได้ (สาธิต เทศสมบูรณ์, 2562)

2. การใช้เหมืองข้อมูลช่วยในการตัดสินใจการให้สินเชื่อ กรณีศึกษา: บริษัท กรุงไทยคาร์เร้นท์ แอนด์ ลีส จำกัด (มหาชน) โดยมีวัตถุประสงค์ในการศึกษาเพื่อสนับสนุนการอนุมัติสินเชื่อของบริษัทได้อย่างมีประสิทธิภาพและลดปริมาณหนี้สูญ จากการเก็บรวบรวมตัวอย่างจำนวน 441 รายและใช้เทคนิคการจำแนกกลุ่มเพื่อหากฎที่ใช้ในการจำแนกลูกค้ากลุ่มดี และลูกค้ากลุ่มไม่ดี ซึ่งได้ทำการทดลองจำแนกกลุ่มข้อมูลด้วยเทคนิคการจำแนก 3 เทคนิค ได้แก่ โมเดลการจำแนกกลุ่มแบบ Decision Tree: C4.5 หรือเรียกว่าโมเดล Classifiers Tree J48 โมเดลการจำแนกกลุ่มแบบ Decision Rule: Part หรือเรียกว่าโมเดล Classifiers Part และ โมเดลการจำแนกกลุ่มแบบ Bayes: Naïve Bayes หรือเรียกว่า โมเดล Classifiers Bayes ซึ่งผลลัพธ์ที่ได้คือ การจำแนกกลุ่มลูกค้าด้วยเทคนิค J48 เป็นเทคนิคที่เหมาะสมกับแนวทางงานวิจัย และมีความถูกต้องมากที่สุด สามารถแบ่งกลุ่มได้ตามเงื่อนไขที่ชัดเจนมากที่สุด ซึ่งผู้ศึกษาจะได้นำไปใช้ในการเขียนโปรแกรมระบบการตัดสินใจอนุมัติสินเชื่อออนไลน์ต่อไป (ทิทย์ธิดา วงศ์พิพันธ์,2556)

3. การพยากรณ์ความเสี่ยงการเกิดโรคเบาหวานโดยเทคนิคการทำเหมืองข้อมูล:กรณีศึกษาโรงพยาบาลมหาสารคาม โดยใช้ปัจจัยเฉพาะด้านจากข้อมูลผู้ป่วยของโรงพยาบาล จำนวน 50,000 ราย ประกอบด้วยข้อมูลผู้ป่วยที่เป็นโรคเบาหวานและไม่ได้เป็น และนำเสนอประสิทธิภาพของตัวแบบที่พัฒนาด้วยเทคนิค Decision Tree ID3 โดยดำเนินการตามกระบวนการตามมาตรฐานสำหรับการทำเหมืองข้อมูล CRISP-DM พบว่า ปัจจัยที่มีผลต่อการเป็นโรคเบาหวานมีทั้งหมด 8 ตัวแปร ได้แก่ อายุ เพศ BMI ประวัติเบาหวานในพ่อ แม่ พี่น้อง ไตรกลีเซอไรด์ ระดับไขมันประวัติการสูบบุหรี่ และประวัติการดื่มสุรา ซึ่งการพยากรณ์ด้วยอัลกอรึทึม Decision Tree พบว่ามีความถูกต้องและมีประสิทธิภาพในการพยากรณ์ที่สูง (รักถิ่น เหลาหา, 2560)

4. การพัฒนาตัวแบบการพยากรณ์ผลผลิตมันสำปะหลังด้วยเทคนิคการทำเหมืองข้อมูล โดยใช้เทคนิคการทำเหมืองข้อมูลในการจำแนกประเภท (Classification) โดยใช้อัลกอริทึม J48, Random Tree, Simple Cart, Naïve Bayes, LAD Tree เพื่อสร้างต้นไม้ตัดสินใจ (Decision Tree) จากนั้นทำการทดสอบความแม่นจำในการพยากรณ์ด้วยเทคนิค Cress-validation Test โดยทำการแบ่งข้อมูลออกเป็น 2 ช่วง ได้แก่ 5-fold cress-validation และ 10-fold cress-validation ซึ่งผลการทดสอบพบว่าให้ค่าความแม่นยำสูงสุดเพียง 70.96% ดังนั้นผู้วิจัยจึงทำการปรับปรุงวิธีการทดสอบให้มีความแม่นยำ โดยทำการแบ่งข้อมูลออกเป็น 2 ส่วน ได้แก่ ข้อมูลเรียนรู้ (Training Set) และข้อมูลทดสอบ (Data Set) จำนวน 5 ชุด ผลการทดสอบพบว่ามีความแม่นยำเพิ่มขึ้นสูงสุดถึง 80.12% โดยสรุปได้ดังนี้ อัลกอริทึมของชุดข้อมูล Test Set ที่ได้ความแม่นยำสูงสุดจำนวน 3 อัลกอริทึม คือ ชุดข้อมูลที่ 5 อัลกอริทึม J48 ให้ค่าความแม่นยำที่ 75.64% ชุดข้อมูลที่ 5 อัลกอริทึม Simple Cart ให้ค่าความแม่นยำที่ 80.12% และชุดข้อมูลที่ 4 อัลกอริทึม LAD Tree ให้ค่าความแม่นยำที่ 78.68% ส่วนอัลกอริทึมที่เหลือได้แก่ Random Tree และ Naïve Bayes ให้ค่าความแม่นยำ 66.85%

และ 67.91% ตามลำดับ ซึ่งค่าที่ได้มีความแม่นยำที่ต่ำ จึงไม่นำมาใช้ในการสร้างตัวแบบพยากรณ์ จากนั้นจึงดำเนินการนำอัลกอริทึมที่ได้ค่าความแม่นยำมากที่สุดไปใช้ในการออกแบบและพัฒนาระบบสารสนเทศการพยากรณ์ผลผลิตมันสำปะหลัง การสืบค้นข้อมูล และการประเมินความพึงพอใจผ่านเว็บไซต์ ต่อไป (ปรีชา ลิ้มตระกูล, 2558)

5. เทคนิคการทำเหมืองข้อมูลเพื่อการบริหารจัดการหนี้ที่ไม่ก่อให้เกิดรายได้ (NPLs) จากการศึกษาตามกระบวนการวิเคราะห์ข้อมูลมาตรฐาน Cross-Industry Standard Process for Data Mining (CRISP-DM) จากประวัติลูกค้าที่ผ่านการพิจารณาการให้สินเชื่อของธนาคารจำนวน 5,000 คน ด้วยการใช้เทคนิคการจำแนกกลุ่ม 2 เทคนิค ได้แก่ Decision tree และ Naïve Bayes ผลจากการศึกษา พบว่าการคัดเลือกปัจจัย ที่จำเป็นในการศึกษาหนี้ที่ไม่ก่อให้เกิดรายได้ 6 แอทริบิวต์ ได้แก่จำนวนชั้นหนี้ อายุ ประเภทอาชีพ วัตถุประสงค์ของการกู้ เงินเดือน และรายได้สุทธิ ด้วยการจำแนกกลุ่มข้อมูลด้วยเทคนิค ID3 Algorithm และ Naïve Bayes ให้ความถูกต้องร้อยละ 91.8% และ 54.7% ตามลำดับ และอาชีพที่ก่อให้เกิดหนี้ที่ไม่ก่อให้เกิดรายได้สูงสุด 3 อันดับแรก ได้แก่ คนขับแท็กซี่ ครู และตำรวจ กล่าวโดยสรุป เทคนิค ID3 Algorithm ในการทำเหมืองข้อมูล ให้ความถูกต้องมากกว่า Naïve Bayes (อานนท์ บุพพวงษ, 2563)

3. ขั้นตอนการศึกษา

1. การรวบรวมข้อมูล เป็นแบบทุติยภูมิ คือ รวบรวมข้อมูลที่ต้องการศึกษา จากแหล่งข้อมูล ที่ได้มีการเก็บรวบรวมไว้แล้ว โดยการศึกษานี้ได้อ้างอิงข้อมูลจากสินเชื่อ Covid-19 ของธนาคารออมสิน สาขามหาวิทยาลัยเชียงใหม่ เวลาทั้งหมด 6 เดือน ตั้งแต่เดือน กรกฎาคม 2564 ถึง ธันวาคม 2564 จำนวนลูกค้า 7,215 ราย

2. การคัดเลือกข้อมูล ผู้ศึกษาได้คัดเลือกปัจจัยที่สำคัญในการทำเหมืองข้อมูล ด้วยวิธีการคัดเลือกข้อมูลจากกลุ่มลูกค้าที่เริ่มมีการค้างชำระสินเชื่อ จากกลุ่มอาชีพของลูกค้า และจากรายได้

3. การเตรียมข้อมูล ประกอบด้วยขั้นตอนดังนี้

3.1 การทำความสะอาดข้อมูล (Data Cleaning) ทำการเพิ่มคุณภาพของข้อมูลเพื่อ ให้ข้อมูลมีความถูกต้อง โดยการจัดกลุ่มข้อมูล การกำจัดข้อมูลที่ผิดปกติ หรือข้อมูลที่เป็นค่าว่าง (Null)

3.2 การผสานข้อมูล (Data Integration) ทำการรวมข้อมูลที่ได้มาจากหน่วยงานที่มีการจัดเก็บที่แตกต่างกัน ให้มีการจัดเก็บข้อมูลในรูปแบบที่เป็นมาตรฐานเดียวกัน

3.3 การแปลงข้อมูล (Data Transformation) เป็นการจัดการข้อมูลที่ได้ให้มีความเหมาะสมต่อการนำไปใช้ในพยากรณ์ ดังตารางที่ 1.

ตารางที่ 1: แสดงการแปลงข้อมูลของกลุ่มลูกค้า

Variables	Variable value	Description
Education (การศึกษา)	1	ต่ำกว่าปริญญาตรี
	2	ปริญญาตรี
	3	สูงกว่าปริญญาตรี
Gender (เพศ)	Female	หญิง
	Male	ชาย

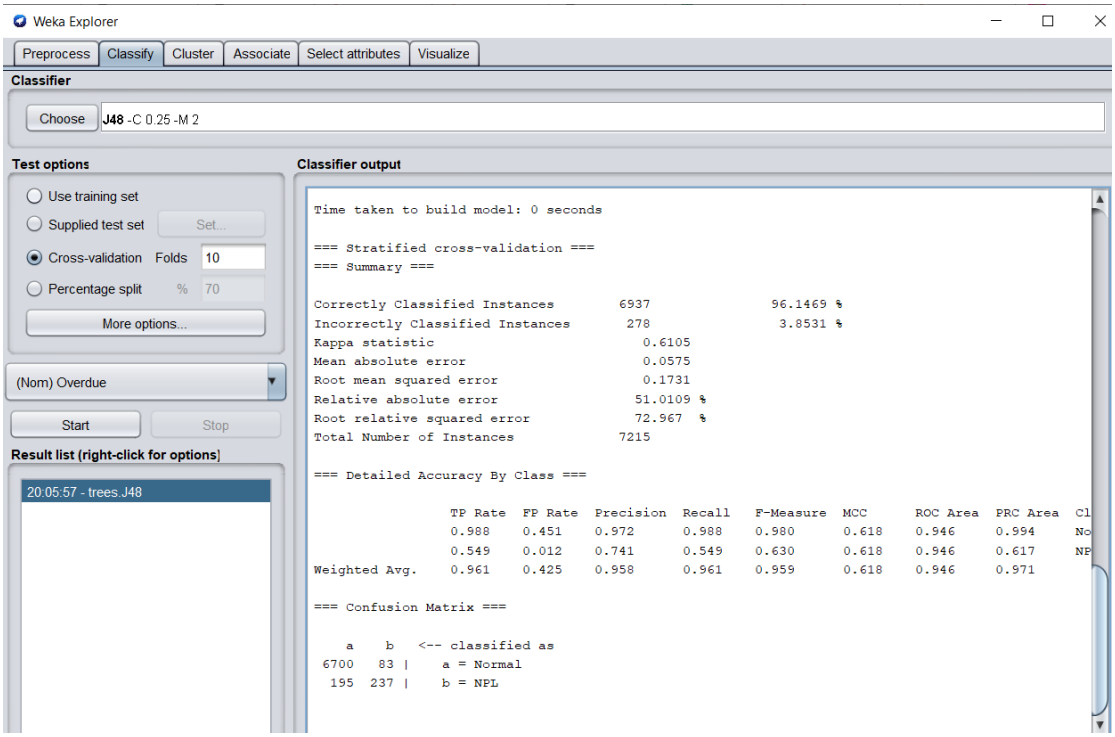
3.4 การลดรูปข้อมูล(Data Reduction) เนื่องจากบางข้อมูลมีการจัดเก็บข้อมูลที่มีความละเอียดมาก ทำให้ผู้ศึกษาต้องทำการลดรูปข้อมูล เพื่อที่จะได้นำข้อมูลไปวิเคราะห์ได้ง่ายขึ้น

4. การวิเคราะห์โดยการทำเหมืองข้อมูล ผู้ศึกษาได้ใช้โปรแกรม Weka ซึ่งใช้เป็นเครื่องมือในการวิเคราะห์ข้อมูลและสร้างตัวแบบทดสอบ ซึ่งข้อมูลจะถูกจัดให้อยู่ในรูปแบบของ .csv เพื่อที่จะได้นำข้อมูลเข้าโปรแกรมเพื่อวิเคราะห์ โดยจะใช้วิธีการจัดหมวดหมู่ (Classification) และวิธีการเรียนรู้ (Learning Methods) จากการศึกษาครั้งนี้ ผู้ศึกษาทำการสร้างแบบทดสอบจำนวน 3 Algorithm เพื่อใช้ในการเรียนรู้และสร้างแบบจำลอง โดยการใช้วิธี K-Fold Cross-Validation Test ซึ่งมีการกำหนดค่า K=5 และ K=10 ตามลำดับ โดยให้แต่ละส่วนมีจำนวนข้อมูลที่เท่ากัน โดยข้อมูลหนึ่งส่วนจะใช้เป็นตัวทดสอบประสิทธิภาพของโมเดล ทำงานวนไปจนครบ และข้อมูลที่เหลือจะใช้สำหรับการทดสอบ โดยข้อมูลที่ได้จากการวิเคราะห์ด้วยโปรแกรม Weka แสดงผลได้ดังต่อไปนี้

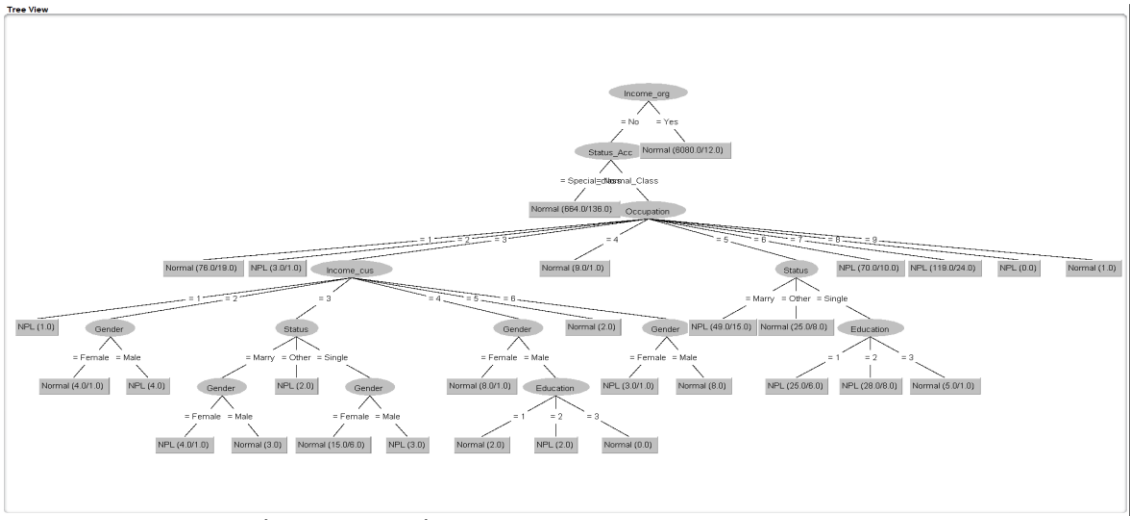
ตารางที่ 2: ค่าแสดงประสิทธิภาพของตัวแบบทดสอบ

Classification Model	(Training Set)	(5-fold cross-validation)	(10-fold cross-validation)
Trees.J48	96.52 %	96.20 %	96.15 %
Rules: Part	96.94 %	96.06 %	96.10 %
Bayes: NaiveBayes	94.57 %	94.44 %	94.52 %

จากข้อมูลดังตารางข้างต้น การวัดประสิทธิภาพของแบบจำลองแบบ Classification: Trees.J48 หรือวิธีต้นไม้ตัดสินใจ โดยวิธี 10-Fold Cross-Validation เป็นการแบ่งข้อมูลแบบสุ่มออกเป็นจำนวน 10 กลุ่ม และมีการทดสอบจำนวน 10 รอบ โดยรอบที่ 1 เป็นการเอาข้อมูลในกลุ่มที่ 1 ออกมาเพื่อใช้ในกลุ่มที่ 2 ถึง 10 เมื่อทำจนครบการเรียนรู้ตามจำนวนที่กำหนด ผู้ศึกษาจึงได้ข้อสรุปจากการเรียนรู้แบบจำลองแบบ Classification: Trees.J48 นั้น มีค่าความถูกต้องมากที่สุดได้ 96.15% และมีค่าความไม่ถูกต้อง 3.9% และสามารถสร้างกฎการเรียนรู้ได้ 42 กฎ ซึ่งในการวัดค่าประสิทธิภาพของตัวแบบจำลองที่ได้นั้น สามารถแสดงได้ถึงค่าความถูกต้อง (Correct) ค่าความแม่นยำ (Precision) ค่าความระลึก (Recall) และค่าความถ่วงดุล (F-Measure) จะแสดงในภาพ ดังนี้



รูปที่ 1: ผลที่ได้จากแบบจำลองแบบ Classification: Trees.J48



รูปที่ 2: Tree View ที่ได้จากแบบจำลองแบบ Classification: Trees.J48

จากภาพที่ 2 สามารถจำแนกสถานะลูกค้าที่มีสถานะหนี้ที่ไม่ก่อให้เกิดรายได้ (NPLs) โดยใช้กฎที่ได้จากแบบจำลองแบบต้นไม้ตัดสินใจ โดยมีโหนดราก (Root Node) คือ การรับรู้รายได้องค์กร ประกอบด้วย ค่า Yes และ ค่า No เมื่อการรับรู้รายได้ขององค์กรเท่ากับค่า No ดังจะอธิบายรายละเอียดดังนี้ ยกตัวอย่าง เช่น

Rule	คำอธิบาย
Income_org=No -> Status_Acc= Normal Class -> Occupation=5 -> Status=Marry -> NPLs	องค์กรไม่มีการรับรู้รายได้ มีสถานะทางบัญชีเป็นปกติ มีอาชีพผู้ประกอบกิจการ/เจ้าของกิจการ มีสถานะภาพ สมรส จะเป็นลูกค้าที่มีสถานะ “NPLs”
Income_org=(No -> Status_Acc= Normal Class -> Occupation=7 -> NPLs	องค์กรไม่มีการรับรู้รายได้ มีสถานะทางบัญชีเป็นปกติ มีอาชีพรับจ้างทั่วไป/ผู้ใช้แรงงาน จะเป็นลูกค้าที่มีสถานะ “NPLs”
Income_org =No -> Status_Acc =Normal Class -> Occupation =3 -> Income_cus =1 -> NPLs	องค์กรไม่มีการรับรู้รายได้ มีสถานะทางบัญชีเป็นปกติ มีอาชีพพนักงานและลูกจ้างในบริษัท มีรายได้ไม่เกิน 10,000 บาท จะเป็นลูกค้าที่มีสถานะ “NPLs”
Income_org=No) -> Status_Acc =Normal Class -> Occupation=5 -> Status=Single -> Education= 1 -> NPLs	องค์กรไม่มีการรับรู้รายได้ มีสถานะทางบัญชีเป็นปกติ มีอาชีพผู้ประกอบกิจการ/เจ้าของกิจการ มีสถานะภาพ โสด มีระดับการศึกษาต่ำกว่าปริญญาตรี จะเป็นลูกค้าที่มีสถานะ “NPLs”

4. ผลการวิจัย

จากการศึกษาข้อมูล ลูกค้าสินเชื่อ สาขามหาวิทยาลัยเชียงใหม่ จากกลุ่มตัวอย่างของลูกค้าจำนวนทั้งสิ้น 7,215 ราย ระยะเวลาในการศึกษาข้อมูลตั้งแต่เดือนกรกฎาคม 2564 ถึงเดือน ธันวาคม 2564 โดยใช้การสร้างตัวแบบทดสอบแบบ Trees.J48 ตัว และดำเนินการศึกษาตามกระบวนการมาตรฐานสำหรับการทำเหมืองข้อมูล CRISP ด้วยโปรแกรม Weka พบว่ากลุ่มลูกค้าที่มีปัญหาการค้างชำระ ดังนี้

1. อาชีพ: กลุ่มลูกค้าที่เป็นผู้ประกอบการธุรกิจ/เจ้าของธุรกิจจะมีอัตราส่วนในการเกิดหนี้ที่ไม่ก่อให้เกิดรายได้ (NPLs) มากกว่ากลุ่มลูกค้าผู้รับจ้างทั่วไป/ผู้ใช้แรงงาน(5) และกลุ่มอาชีพเกษตรกร(8) มีอัตราส่วนในการค้างชำระน้อยที่สุด
2. สถานะภาพ: กลุ่มลูกค้าที่มีสถานะภาพโสด(Single)จะมีอัตราส่วนการเกิดหนี้ที่ไม่ก่อให้เกิดรายได้ (NPLs) มากกว่ากลุ่มลูกค้าสถานะภาพสมรส(Marry) และสถานภาพอื่นๆ(Other) ตามลำดับ
3. ระดับการศึกษา: กลุ่มลูกค้าที่มีการศึกษาระดับต่ำกว่าปริญญาตรี(1)จะมีอัตราส่วนการเกิดหนี้ที่ไม่ก่อให้เกิดรายได้ (NPLs) มากกว่าระดับปริญญาตรี(2) และระดับสูงกว่าปริญญาตรี(3) ตามลำดับ ผลจากการศึกษาทำนายข้อมูลลูกค้าจากกลุ่มตัวอย่างหนี้ที่ไม่ก่อให้เกิดรายได้ ดังนี้

OCCUPATION	กลุ่มข้อมูล	ร้อยละ NPLs
1	2581	2.79%
2	113	0.23%
3	1093	2.13%
4	53	0.08%
5	1644	4.57%
6	608	2.48%
7	956	4.22%
8	15	0.00%
9	152	0.23%

Status	กลุ่มข้อมูล	ร้อยละ NPLs
Marry	3183	6.16%
Single	3394	9.07%
Other	638	1.51%

Education	กลุ่มข้อมูล	ร้อยละ NPLs
1	2817	8.80%
2	3775	7.17%
3	623	0.77%

และผลจากการทดสอบตัวแบบทดสอบ Classification: Trees.J48 ที่ได้จาก Confusion Matrix ซึ่งมีค่าความถูกต้องอยู่ที่ 96.15% ดังนี้

a	b	classified as
6,700	83	a = Normal
195	237	b = NPL

Positive	Negative	
TP	FP	Positive
FN	TN	Negative

อธิบายผลลัพธ์จากข้อมูล Confusion Matrix ดังนี้

1. แบบทดสอบ จำนวนลูกค้า 6,700 ราย มีสถานะบัญชีปกติ (TP)
2. แบบทดสอบ จำนวนลูกค้า 237 ราย มีสถานะบัญชีหนี้ที่ไม่ก่อให้เกิดรายได้ NPLs (TN)
3. แบบทดสอบ จำนวนลูกค้า 83 ราย ระบบประมวลผลว่ามีสถานะทางบัญชีหนี้ที่ไม่ก่อให้เกิดรายได้ NPLs แต่จากข้อมูลมีสถานะทางบัญชีปกติ
4. แบบทดสอบ จำนวนลูกค้า 195 ราย ระบบประมวลผลว่ามีสถานะทางบัญชีปกติ แต่จากข้อมูลมีสถานะทางบัญชีหนี้ที่ไม่ก่อให้เกิดรายได้ NPLs

5. อภิปรายผล / บทสรุป

การพยากรณ์สถานะทางบัญชีหนี้ที่ไม่ก่อให้เกิดรายได้ NPLs ของลูกค้าสินเชื่อ Covid-19 โดยการสร้างแบบทดสอบจำนวน 3 Algorithm ได้แก่ แบบจำลอง Trees.J48 แบบจำลอง Rules: Part และแบบจำลอง Bayes: NaiveBayes ซึ่งจากการเปรียบเทียบประสิทธิภาพของแบบจำลองนั้นมีความแตกต่างกันไม่มาก ผู้ศึกษาจึงทำการเลือกศึกษาแบบจำลอง Trees.J48 เนื่องจากเมื่อเปรียบเทียบประสิทธิภาพแล้วได้ ค่าความถูกต้องมากที่สุดอยู่ที่ 96.15% ค่าความไม่ถูกต้อง 3.86% และได้ทำการสร้างแบบจำลองด้วยวิธี N-Fold Cross-Validation เป็นการแบ่งข้อมูลแบบสุ่มออกเป็นจำนวน N กลุ่ม จากวิธีดังกล่าวสามารถสร้างกฎการจำแนกได้ 42 กฎ เมื่อวัดค่าประสิทธิภาพของแบบทดสอบแล้วมีค่ามากกว่าร้อยละ 70 ขึ้นไป แสดงว่าวิธีการแบ่งข้อมูลชุดการเรียนรู้และทดสอบสามารถนำไปใช้ในการพัฒนาแบบการพยากรณ์สถานะทางบัญชีหนี้ที่ไม่ก่อให้เกิดรายได้ NPLs ได้ ดังนั้นผลที่ได้จากการสร้างตัวแบบทดสอบ Trees:J48 สามารถนำไปประยุกต์ใช้ในหลักเกณฑ์การพิจารณาสินเชื่อของธนาคารหรือช่วยเหลือกลุ่มลูกค้าที่มีสถานะทางบัญชีหนี้ที่ไม่ก่อให้เกิดรายได้ NPLs ได้

6. ข้อเสนอแนะ

1. จากการศึกษาในครั้งนี้ปัจจัยที่ทำให้เกิดหนี้ที่ไม่ก่อให้เกิดรายได้ NPLs นั้นยังคงเป็นปัจจัยในเรื่องของรายได้ที่ได้รับ และจากสถานการณ์ Covid-19 ทำให้เห็นถึงกลุ่มลูกค้าที่ประสบปัญหาในเรื่องรายได้อย่างแท้จริง
2. สำหรับการศึกษาในครั้งต่อไปจะทำการการเก็บรวบรวมข้อมูลของสาขาใกล้เคียงหรือสาขาที่อยู่ในสังกัดธนาคารออมสินภาค 8 เพื่อนำข้อมูลมาวิเคราะห์ความแตกต่างและทำการเปรียบเทียบถึงปัจจัยการเกิดหนี้ที่ไม่ก่อให้เกิดรายได้ NPLs เมื่อทำการวิเคราะห์ได้จะทำให้เห็นแนวโน้มข้อมูลลูกค้าในภาพรวมของระดับภาค

เอกสารอ้างอิง

วีรพงษ์ แสงลู่ทอง.(2564). *การวิเคราะห์พฤติกรรมลูกค้าจาก E-Marketplace ด้วย Apriori Algorithm.* สืบค้น 1 พฤษภาคม 2565,จาก https://proceeding_academicday.pdf (utcc.ac.th)

สาธิต เทศสมบูรณ์.(2562). *การประยุกต์ใช้เทคนิคการเรียนรู้ของคอมพิวเตอร์เพื่อพัฒนาตัวแบบทางเหมืองข้อมูลสำหรับพยากรณ์ระดับฮีโมโกลบินของผู้บริจาคโลหิต.* (วิทยานิพนธ์ วิทยาศาสตร์มหาบัณฑิต). กรุงเทพฯ: มหาวิทยาลัยจุฬาลงกรณ์มหาวิทยาลัย

ทิพย์ธิดา วงศ์พิพันธ์. (2556). *การใช้เหมืองข้อมูลช่วยในการตัดสินใจการให้สินเชื่อกรณีศึกษา:บริษัทกรุงไทยคาร์เร้นท์ แอนด์ ลีส จำกัด (มหาชน).* สืบค้น 9 พฤษภาคม 2565,จาก http://libdoc.dpu.ac.th/thesis/147907.pdf

รักถิ่น เหลาหา.(2560). *การพยากรณ์ความเสี่ยงการเกิดโรคเบาหวานโดยเทคนิคการทำเหมืองข้อมูล: กรณีศึกษาโรงพยาบาลมหาสารคาม.* สืบค้น 9 พฤษภาคม 2565,จาก http://research.rmu.ac.th/rdi-mis//upload/fullreport/1615219126.pdf

ปรีชา ลิ้มตระกูล.(2558). *การพัฒนาตัวแบบการพยากรณ์ผลผลิตมันสำปะหลังด้วยเทคนิค การทำเหมืองข้อมูล.* สืบค้น 9 พฤษภาคม 2565,จาก https://doi.nrct.go.th/ListDoi /listDetail?Resolve_DOI=10.14457/STOU.the.2015.22

นิเวศ จิระวิชิตชัย.(2555). *การประยุกต์ใช้ซอฟต์แวร์เวก้าในการจำแนกประเภท.* สืบค้น 11 พฤษภาคม 2565,https://e-library.siam.edu/e-journal/wp-content/uploads/2012/12/EJSU-vol13-no25-jul-dec2012-5.pdf

อานนท์ บุพพวงษ.(2563). *เทคนิคการทำเหมืองข้อมูลเพื่อการบริหารจัดการหนี้ที่ไม่ก่อให้เกิดรายได้ (NPLs).*สืบค้น 11 พฤษภาคม 2565,จาก https://proceeding_academicday.pdf (utcc.ac.th)

ธนาวัฒน์ แก้วแฟง และ ชุติพนธ์ ศรีสวัสด.(2563). *การคัดเลือกคุณลักษณะที่เหมาะสม และการเปรียบเทียบประสิทธิภาพอัลกอริทึมเพื่อพัฒนาต้นแบบ สารสนเทศเพื่อการจัดการคัดกรองผู้ป่วยที่มีความเสี่ยง COVID-19.* สืบค้น 25 พฤษภาคม 2565,จาก การประชุมวิชาการระดับชาติวิทยาศาสตร์เทคโนโลยีและนวัตกรรม (มหาวิทยาลัยแม่โจ้) ครั้งที่ 2

โกเมศ อัมพวัน.(2559). *การจำแนกประเภทและการทำนายข้อมูล.* สืบค้น 25 พฤษภาคม 2565,จาก https://staff.informatics.buu.ac.th/~komate/886464/%5B6%5D-lassification.pdf

คลลักษณ์ บางศุข. (2559). *การประยุกต์ใช้เทคนิคการทำเหมืองข้อมูลเพื่อจัดการข้อมูลสารสนเทศของเครือข่ายการท่องเที่ยวโดยชุมชนในจังหวัดสุราษฎร์ธานี.* (รายงานผลการวิจัย). สุราษฎร์ธานี: สถาบันพระปกเกล้า