

CREATE PROXY MEAN TEST BY USING MACHINE LEARNING TECHNIQUE

Bhawin Teveyanun¹

1 Faculty of Economics, Thammasat University, Thailand; plyebhawin@gmail.com

ARTICLE HISTORY

Received: 3 February 2025

Revised: 24 February 2025

Published: 4 March 2025

ABSTRACT

This study aims to develop a model for PMT using Machine Learning techniques. It starts by reviewing the weaknesses of the traditional model for measuring PMT, then develops a model for measuring PMT using Machine Learning techniques to find the most efficient model. The study also analyzes the impact of each variable on the model, which the traditional Machine Learning model cannot explain, using the SHAP value. The results of the study on the efficiency of the model in predicting whether a household is poor or not, show that the model accurately predicts 91% of the households identified as poor, while the model only covers 50% of the total poor households, and the overall model accuracy is 65% (F1-score). These results are significantly more efficient than the traditional methods, and policy recommendations can be made based on the top ten variables with the highest impact on the prediction.

Keywords: Proxy Means Test (PMT), Machine Learning, Poverty

CITATION INFORMATION: Teveyanun, B. (2025). Create Proxy Mean Test by Using Machine Learning Technique. *Procedia of Multidisciplinary Research*, 3(3), 16

การสร้างแบบจำลองการวัดรายได้ทางอ้อมด้วยวิธี Machine Learning

ภวินทร์ เตวียนันท์¹

1 คณะเศรษฐศาสตร์ สาขาวิชาเศรษฐศาสตร์ มหาวิทยาลัยธรรมศาสตร์; plyebhawin@gmail.com

บทคัดย่อ

งานศึกษาต้องการพัฒนาการสร้างแบบจำลองการวัดรายได้ทางอ้อมผ่านการใช้เทคนิค Machine Learning เริ่มจากการทบทวนจุดอ่อนของแบบจำลองการวัดรายได้ทางอ้อมโดยวิธีปกติ จากนั้นพัฒนาแบบจำลองการวัดรายได้ทางอ้อมผ่านการใช้เทคนิค Machine Learning เพื่อหาแบบจำลองการวัดรายได้ทางอ้อมที่มีประสิทธิภาพมากที่สุด นอกจากนี้ งานศึกษายังวิเคราะห์ผลกระทบของแต่ละตัวแปรต่อแบบจำลองที่ปกติแบบจำลองแบบ Machine Learning ไม่สามารถอธิบายได้ผ่านการใช้ค่า SHAP ผลลัพธ์จากงานศึกษาประสิทธิภาพของแบบจำลองในการทำนายว่าครัวเรือนนั้นเป็นครัวเรือนยากจนหรือไม่ พบว่า ผลการทำนายที่เป็นครัวเรือนยากจนทั้งหมดแบบจำลองจะทำนายว่าเป็นครัวเรือนยากจนได้อย่างถูกต้อง ร้อยละ 91 ในขณะที่ครัวเรือนที่ทำนายถูกต้องว่าเป็นครัวเรือนยากจนจะครอบคลุมครัวเรือนยากจนเพียงร้อยละ 50 ของครัวเรือนยากจนทั้งหมด และแบบจำลองมีค่าความแม่นยำของแบบจำลองโดยรวมอยู่ที่ร้อยละ 65 (F1-score) ซึ่งผลดังกล่าวมีประสิทธิภาพมากกว่าการใช้วิธีการแบบปกติสูงมาก และยังสามารถนำเสนอข้อเสนอแนะเชิงนโยบายตามตัวแปรที่ส่งผลกระทบต่อการทำนายสูงสุดสิบอันดับแรก

คำสำคัญ: แบบจำลองการวัดรายได้ทางอ้อม, การเรียนรู้ของเครื่อง, ความยากจน

ข้อมูลการอ้างอิง: ภวินทร์ เตวียนันท์. (2568). การสร้างแบบจำลองการวัดรายได้ทางอ้อมด้วยวิธี Machine Learning. *Procedia of Multidisciplinary Research*, 3(3), 16

บทนำ

ปัญหาความยากจนเป็นปัญหาหลักที่พบได้ทั่วไปในโลก โดยเฉพาะประเทศกลุ่มกำลังพัฒนาอย่างประเทศไทย ความยากจนไม่ใช่เพียงปัญหาทางด้านโครงสร้างเท่านั้น แต่ยังเป็นปัญหาทางด้านความเหลื่อมล้ำทางสังคม โดยหนึ่งในเครื่องมือที่ทางภาครัฐนิยมใช้เป็นเครื่องมือหลักในการช่วยเหลือคนยากจน คือ การอุดหนุนทางการเงินแบบมีเงื่อนไข (Conditional Cash Transfers: CCT)

การอุดหนุนทางการเงินแบบมีเงื่อนไขมีจุดสำคัญคือค้นหาว่าใครเป็นคนจน และเกณฑ์การคัดกรองควรจะเป็นเช่นไร เพื่อให้นโยบายมีประสิทธิภาพและมีการรั่วไหลออกมาให้น้อยที่สุด ดังนั้นอาจจะกล่าวได้ว่าการค้นหาตัวคนจนจึงเป็นประเด็นที่สำคัญมากที่สุดสำหรับนโยบายการช่วยเหลือคนจน

เครื่องมือที่ใช้ในการตามหาคนจนนั้นมีหลากหลายเครื่องมือ แต่วิธีที่ได้รับความนิยมและใช้กันอย่างแพร่หลาย คือ วิธีการวัดรายได้โดยอ้อม หรือที่เรียกว่า Proxy Means Test (PMT) เป็นวิธีการทางสถิติเพื่อใช้ในการกำหนดเกณฑ์ว่าครัวเรือนใดเป็นครัวเรือนยากจน ผ่านการวิเคราะห์ทางสถิติเพื่อกำหนดตัวชี้วัดอื่นๆ ที่ไม่ใช่รายได้ แต่มีความสามารถในการบ่งบอกระดับรายได้ (ในทางกลับกันตรงข้ามกับรายได้ คือ ความยากจน)

จากงานศึกษาเรื่องเพิ่มประสิทธิภาพหลักเกณฑ์คัดกรองความยากจนด้วยวิธีการวัดรายได้ทางอ้อมของ ชัยยุทธ ปัญญสวัสดิ์สุทธิ์ และคณะ (2565) โดยวิธีการสร้างแบบจำลองการวัดรายได้ทางอ้อมแบบดั้งเดิมพบว่า มีข้อจำกัดที่จำเป็นต้องพึงพาการเลือกตัวแปรที่มีคุณภาพมาก ทำให้วิธีดังกล่าวยังมีข้อจำกัดในการเลือกตัวแปรที่สนใจมาใช้แทนรายได้ของครัวเรือน เช่น ขั้นตอนการเลือกตัวแปรที่มีข้อจำกัดต้องเลือกผ่านตัวแปรที่คิดว่าสามารถชี้วัดรายได้ได้จากตัวแปรหลายร้อยตัว หรือยังไม่มีมุมมองของการพัฒนาตัวแปรที่ใช้ในการศึกษาหรือเทคนิคที่เรียกว่า feature engineering ซึ่งก็คือการปรับเปลี่ยนข้อมูลให้อยู่ในรูปแบบที่มีประโยชน์ (features) และตัวแปรที่เลือกยังต้องคำนึงถึงปัญหาภาวะร่วมเส้นตรงเชิงพหุอีกด้วย

โดยในปัจจุบันการเลือกตัวแปรให้มีประสิทธิภาพมากยิ่งขึ้นสามารถใช้เทคนิคที่มีใช้การประมวลผลทางคอมพิวเตอร์ที่มีประสิทธิภาพมากยิ่งขึ้นผ่านการใช้เทคนิคที่เรียกว่า Machine learning งานศึกษาของ Athley (2018) พบว่า การใช้เทคนิค machine learning กับการสร้างแบบจำลองในงานศึกษาทางด้านเศรษฐศาสตร์จะสามารถให้ได้แบบจำลองที่มีความ robust มากกว่าและยังสร้างแบบจำลองที่มีความเสถียรมากกว่า รวมทั้งยังสามารถแก้ไขปัญหาดังกล่าวของแบบจำลองได้มากขึ้น ซึ่งเป็นที่มาของโจทย์วิจัยชิ้นนี้

แบบจำลองที่ใช้ในสร้างแบบจำลองการวัดรายได้ทางอ้อมในงานศึกษาค้นคว้านี้จะใช้ แบบจำลองแบบการแบ่งกลุ่ม (Classification) ในลักษณะของ Binary โดยคำตอบมี 2 คำตอบ คือเป็นการทำนายสิ่งที่ใกล้เคียงกับข้อมูลที่ป้อน ว่าการจำแนกกลุ่มของครัวเรือนเป็น “ครัวเรือนยากจน” และ “ครัวเรือนรายได้สูง” ในรูปแบบของเทคนิคการทำ gradient boosting และศึกษาผลกระทบของแต่ละตัวแปรผ่านการวิเคราะห์จากการคำนวณค่า Shapley values (ตามแบบจำลองของทฤษฎีเกม) เพื่อให้ได้มิติมุมมองเรื่องการคัดกรองครัวเรือนยากจนที่หลากหลายมากขึ้น

ในงานศึกษามีวัตถุประสงค์สองประการคือ เพื่อพัฒนาแบบจำลองการวัดรายได้ทางอ้อมโดยอาศัยเทคนิค machine learning เพื่อเพิ่มประสิทธิภาพการประมาณค่าที่วัดจากค่าอัตราการพลาดโอกาส (Exclusion error) และอัตราการรั่วไหล (Inclusion error) เพื่อศึกษาผลกระทบของแต่ละตัวแปรต่อตัวแบบจำลองผ่านการวิเคราะห์ผ่านค่า Shapley values

การทบทวนวรรณกรรม

Proxy Means Test เป็นเทคนิคทางสถิติที่ใช้ในการระบุครัวเรือนยากจน โดยพิจารณาจากตัวชี้วัดอื่นๆ นอกเหนือจากรายได้ที่มีความสัมพันธ์กับระดับความยากจนของครัวเรือน แทนที่จะสอบถามรายได้โดยตรง วิธีการนี้มีข้อดีคือ ตัวชี้วัดที่นำมาใช้ในการประเมินต้องวัดและตรวจสอบได้ง่าย มีความน่าเชื่อถือ และยากต่อการปกปิดข้อมูลจริง ซึ่งเป็นปัญหาหากใช้การสอบถามรายได้โดยตรง วิธี PMT จึงเป็นทางเลือกที่มีประสิทธิภาพในการระบุครัวเรือนยากจน และได้รับ

การนำมาใช้อย่างแพร่หลายในหลายประเทศ ดังนั้น วิธี PMT นี้จะช่วยให้สามารถระบุครัวเรือนยากจนได้อย่างมีประสิทธิภาพ และเป็นเครื่องมือสำคัญในการกำหนดนโยบายและมาตรการช่วยเหลือครัวเรือนที่มีฐานะทางเศรษฐกิจและสังคมที่ยากจน

หนึ่งใน PMT ที่มีชื่อเสียง คือ ดัชนีความน่าจะเป็นที่จะยากจน (Poverty Probability Index: PPI) ของ Schreiner (2006, 2015, 2017) ซึ่งถูกสร้างขึ้นเพื่อใช้ในการคัดกรองคนยากจนในกว่า 60 ประเทศทั่วโลก โดยใบคะแนนความยากจน (Poverty Scorecard) ในแต่ละประเทศ ถูกสร้างขึ้น ผ่านการใช้วิธีการทางสถิติในการประมาณค่าแบบโลจิสติก (Logistic Regression) และมีการจำกัดตัวแปรที่ใช้เป็นตัวชี้วัดความยากจนเพียงแค่ 10 ตัวแปรเท่านั้น ซึ่งช่วยในด้านของความสะดวกและลดต้นทุนในการเก็บข้อมูล

สำหรับการคัดกรองคนยากจนด้วยวิธีการวัดรายได้ทางอ้อมในประเทศไทย งานศึกษาของ ชัยยุทธ ปัญญสวัสดิ์สุทธิ์ และคณะ (2565) ได้ทำการสร้างคะแนนความยากจนโดยอาศัยหลักการของ PMT ขึ้น เพื่อใช้ในการคัดกรองนักเรียนยากจนใน 10 จังหวัดของประเทศไทย โดยใช้วิธีการประมาณค่าแบบกำลังสองน้อยที่สุด (Least Squares Regression) เพื่อหาความสัมพันธ์ระหว่างรายได้ครัวเรือนเฉลี่ยต่อหัวต่อเดือน และทางด้านสถานะทางด้านครัวเรือน โดยมีการใช้ตัวแปรหุ่น (Dummy Variables) ของจังหวัด

ถึงแม้ว่าการใช้เทคนิค proxy mean test เป็นวิธีการที่นิยมใช้ในการวิเคราะห์และประเมินภาวะความยากจนในระดับครัวเรือน แต่อย่างไรก็ตาม วิธีการดังกล่าวก็มีจุดอ่อนอยู่หลากหลายประการ อาทิเช่นการใช้ PMT อาจทำให้เกิดความคลาดเคลื่อนในการประเมินภาวะความยากจน เนื่องจากตัวแปรที่ใช้เป็นตัวแทนอาจจะไม่สามารถสะท้อนความเป็นจริงของความยากจนได้อย่างแท้จริง ทำให้ผลการวิเคราะห์และการประเมินภาวะความยากจนอาจไม่ตรงกับความเป็นจริง (Ysobel & Ting, 2022)

งานศึกษาของ Houssou และคณะ (2007) มีการสร้างแบบจำลองการวัดรายได้ทางอ้อมผ่านเทคนิคอื่นๆ นอกเหนือจากวิธี Ordinary least square (OLS) ประกอบไปด้วย แบบจำลอง Linear Probability Mode แบบจำลอง Probit และแบบจำลอง Quantile regressions ผลการศึกษาพบว่า แบบจำลอง Quantile regressions สามารถใช้ในการสร้างแบบจำลองการวัดรายได้ทางอ้อมได้ดีที่สุด

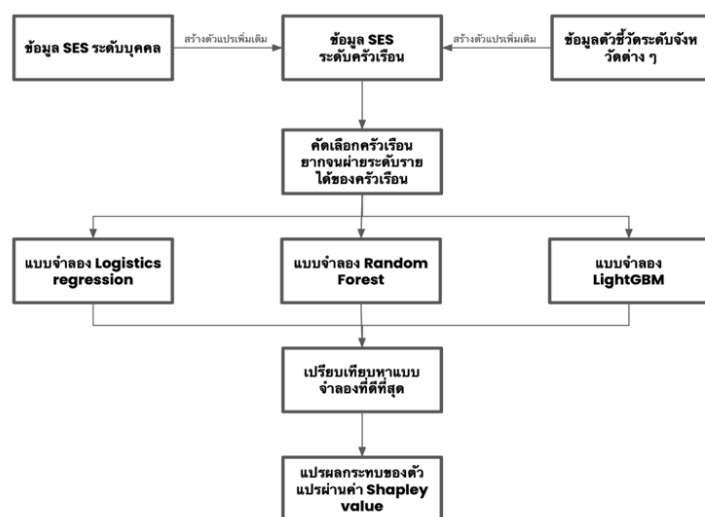
งานศึกษาของ McBride และ Nichols (2016) มีการใช้เทคนิค machine learning มาเปรียบเทียบกับแบบจำลองการวัดรายได้ทางอ้อมแบบดั้งเดิมเพื่อที่จะกำหนดเป้าหมายกลุ่มคนยากจนเพื่อได้รับนโยบาย water subsidy ในประเทศกานา โดยในงานศึกษาชิ้นนี้มีการนำเทคนิค LASSO (Least Absolute Shrinkage and Selection Operator) มาใช้ในการเลือกตัวแปรสำหรับการทำแบบจำลองการวัดรายได้ทางอ้อมพบว่าการใช้เทคนิค machine learning ดังกล่าวมีผลของความแม่นยำในการชี้วัดครัวเรือนยากจนได้มากกว่าการใช้แบบจำลองการวัดรายได้ทางอ้อมแบบดั้งเดิมอย่างมีนัยสำคัญ ทั้งนี้ในงานศึกษาชิ้นนี้ยังมีข้อเสนอแนะเพิ่มเติมอีกว่าถ้าหากจำนวนของข้อมูลมีขนาดที่ใหญ่มากยิ่งขึ้นจะทำให้ผลของการใช้เทคนิคดังกล่าวมีประสิทธิภาพมากยิ่งขึ้น

การศึกษาของ Kambuya (2017) ได้ประยุกต์ใช้แนวคิดของ Proxy Mean Test เพื่อคัดกรองครัวเรือนที่ประสบปัญหาความยากจน โดยเปรียบเทียบประสิทธิภาพของเทคนิคการคัดเลือกตัวแปรชี้วัดความยากจนด้วย 3 วิธี ได้แก่ Stepwise Regression, LASSO และ Random Forest โดยใช้ข้อมูลครัวเรือนจากการสำรวจภาวะเศรษฐกิจและสังคมของครัวเรือน (Household Socio-Economic Survey: SES) ปี พ.ศ.2559 ของสำนักงานสถิติแห่งชาติ ผลการศึกษาพบว่า ทั้งสามวิธีสามารถคัดเลือกตัวแปรชี้วัดความยากจน ที่มีความคล้ายคลึงกับตัวแปรในการศึกษาของ ชัยยุทธ ปัญญสวัสดิ์สุทธิ์ และคณะ (2560) หลังจากประมาณค่าด้วยวิธีกำลังสองน้อยที่สุด (Least Square Regression) เพื่อสร้างคะแนนความยากจนด้วยวิธี PMT พบว่า ทั้งสามเทคนิคมีประสิทธิภาพในการคัดกรองครัวเรือนยากจนค่อนข้างดี แต่ยังมีจุดอ่อนในเรื่องการแปรความหมายของตัวแปรที่หายไปเมื่อเปรียบเทียบกับแบบจำลองแบบดั้งเดิม

โดยสรุปจากการทบทวนวรรณกรรมพบว่าการใช้เทคนิค Machine learning การใช้เทคนิค machine learning จะสามารถให้ได้แบบจำลองที่มีความ robust มากกว่าและยังสร้างแบบจำลองที่มีความเสถียรมากกว่า รวมทั้งยังสามารถแก้ไขปัญหาต่างๆ ของแบบจำลองได้มากกว่าแบบจำลองแบบดั้งเดิม

วิธีการดำเนินการวิจัย

ขั้นตอนกระบวนการที่ใช้ในงานศึกษาชั้นนี้สามารถสรุปได้ดังรูปที่ 1



ภาพที่ 1 สรุปขั้นตอนกระบวนการในการศึกษา

ขั้นตอนกระบวนการที่ใช้ในงานศึกษาชั้นนี้ประกอบไปด้วย 4 ขั้นตอนหลักๆ คือ ขั้นตอนแรก การเตรียมข้อมูล ขั้นตอนคัดเลือกครัวเรือนยากจนสำหรับใช้เป็นตัวแบบ (Label) ให้แบบจำลอง machine learning เรียนรู้ ขั้นตอนการสร้างแบบจำลอง machine learning พร้อมทั้งเปรียบเทียบแบบจำลองต่างๆ และขั้นตอนสุดท้ายเป็นการแปลผลของตัวแปรต่างๆ ที่ได้จากการใช้แบบจำลอง

ข้อมูลที่ใช้ในงานศึกษาจะใช้ข้อมูลสามส่วนหลักประกอบไปด้วย ส่วนแรก ข้อมูลการสำรวจสถานะเศรษฐกิจสังคมของครัวเรือน (SES) ปี พ.ศ.2564 โดยใช้ข้อมูล ประกอบไปด้วย ข้อมูลรายได้ (ใช้สำหรับคัดเลือกครัวเรือนยากจนกับครัวเรือนที่ไม่ใช่เพื่อให้แบบจำลอง machine learning ใช้ในการเรียนรู้) ข้อมูลลักษณะของครัวเรือน ข้อมูลลักษณะของที่อยู่อาศัย ข้อมูลการถือครองทรัพย์สินต่างๆ ของครัวเรือน และข้อมูลรายจ่ายประเภทต่างๆ ส่วนที่สอง ข้อมูลการสำรวจสถานะเศรษฐกิจสังคมของครัวเรือน (SES) ปี พ.ศ.2564 รายบุคคล เพื่อใช้สำหรับสร้างตัวแปรลักษณะของครัวเรือน และส่วนที่สาม ข้อมูลคุณลักษณะของจังหวัดต่างๆ อาทิเช่น ข้อมูลผลิตภัณฑ์มวลรวมจังหวัด (Gross provincial product: GPP) ข้อมูลจำนวนประชากรรายจังหวัด ข้อมูลลักษณะอุตสาหกรรมรายจังหวัด เป็นต้น

ขั้นตอนกระบวนการศึกษาสามารถอธิบายโดยสังเขปได้ ดังนี้ การเตรียมข้อมูลจะเป็นการรวบรวมฐานข้อมูลสามส่วนเข้าด้วยกันเป็นระดับครัวเรือน หลังจากได้ข้อมูลครัวเรือนยากจนแล้วจะทำการแบ่งครัวเรือนยากจนกับครัวเรือนไม่ยากจนออกจากกันโดยการกำหนดผ่านค่ารายได้ของแต่ละครัวเรือนเพื่อใช้สำหรับเป็นตัวแบบ (Label) ให้แบบจำลอง machine learning เรียนรู้และนำไปเข้าแบบจำลองทั้งสามแบบ ประกอบไปด้วย แบบจำลองแบบ classification แบบ logistic regression แบบจำลอง random forest และแบบจำลอง LightGBM จากนั้นจะนำผลของแบบจำลองมาเปรียบเทียบหาแบบจำลองที่ดีที่สุดแล้วนำไปแปลผลค่าที่ได้เพื่อใช้ในการอธิบายผลของแบบจำลองต่อไป

การชี้วัดผลของแบบจำลองจะประกอบไปด้วย ค่าสองค่า คือ ค่า Precision และค่า Recall โดยค่าทั้งสองค่ามีความหมายดังต่อไปนี้ ค่า Precision มองในมุมว่าโมเดลทำนายได้แม่นยำแค่ไหนในค่าที่สนใจ คือ ผลการทำนายที่เป็น

ครวเรือนยากจนทั้งหมดแบบจำลองจะทำนายว่าเป็นครวเรือนยากจนได้อย่างถูกต้องมากน้อยแค่ไหน ในขณะที่ค่าของ Recall มองในมุมว่าแบบจำลองทำนายได้ถูกต้องแค่ไหน เช่น ในผลการทำนายที่ทั้งถูกและผิดแบบจำลองทำนายว่าเป็นครวเรือนยากจนได้อย่างถูกต้องมากน้อยแค่ไหน

การแปลผลของการประมาณค่าเพื่อหาค่าทิศทางของค่าตัวแปรผ่านการทำ Shapley value คือ เป็นการคำนวณภายใต้หลักทฤษฎีเกม (game theory) ในรูปแบบทฤษฎีเกมแบบร่วมมือ (coalitional game theory) โดยในหลักการดังกล่าวจะมีการมองค่าของฟีเจอร์หรือตัวแปรในแต่ละตัวเป็น “ผู้เล่น” (player) ของเกม และผลจากการทำนายของโมเดลเป็น “ผลลัพธ์” (payout) ของเกม ซึ่งค่า Shapley value เป็นค่าที่ระบุว่าควรแบ่งผลลัพธ์ของเกมให้แก่ผู้เล่นแต่ละคนอย่างไรให้ยุติธรรมโดยอ้างอิงจากสัดส่วนการมีส่วนร่วม (contribution) ของผู้เล่นแต่ละคน เมื่อนำหลักการดังกล่าวมาใช้ในการวิเคราะห์โมเดล คำนี้งสามารถใช่เพื่อบอกว่าแต่ละฟีเจอร์จะส่งผลกระทบต่อการทำนายของโมเดลมากน้อยอย่างไรได้

สรุปและอภิปรายผล

ผลของการเปรียบเทียบประสิทธิภาพของแบบจำลองทั้งสามรูปแบบ พบว่า แบบจำลอง LightGBM ได้ผลลัพธ์ของแบบจำลองที่ดีที่สุด ผ่านการเลือกตัวแบบในการเรียนรู้ผ่านเทคนิค Cross Validation คือ เครื่องมือที่ช่วยให้ตัดสินใจได้ว่าควรแบ่งข้อมูลส่วนไหนไปเป็น Training Data ซึ่ง K-Fold Cross Validation คือ การที่แบ่งข้อมูลเป็นจำนวน K ส่วน โดยการในแต่ละส่วนจะต้องมาจากสุ่มเพื่อที่จะให้ข้อมูลกระจายเท่าๆ กัน

ในแบบจำลอง LightGBM ได้ทำการสุ่มจำนวนครวเรือนทั้งหมดจาก 46,840 ครวเรือน มาใช้ในการเรียนรู้ของแบบจำลองจำนวนทั้งสิ้น 32,788 ครวเรือน และใช้ครวเรือนทั้งสิ้น 14,052 ครวเรือน เป็นตัวชี้วัดความแม่นยำของแบบจำลองซึ่งสามารถแบ่งออกได้เป็นครวเรือนไม่ยากจน 12,982 ครวเรือน และครวเรือนยากจนทั้งสิ้น 1,070 ครวเรือน ผลของประสิทธิภาพแบบจำลองสามารถได้ดังภาพที่ 2

	precision	recall	f1-score	support
0	0.91	0.50	0.65	1070
1	0.96	1.00	0.98	12982
accuracy			0.96	14052
macro avg	0.94	0.75	0.81	14052
weighted avg	0.96	0.96	0.95	14052

ภาพที่ 2 แสดงผลประสิทธิภาพของแบบจำลอง LightGBM

ตัวแปรผลของแบบจำลองสามารถแบ่งเป็นสองส่วน ส่วนแรก ค่าเลข 0 หมายถึง แบบจำลองทายว่าครวเรือนนั้นเป็นครวเรือนยากจน ในขณะที่ 1 หมายถึง ไม่ใช่ครวเรือนยากจน โดยที่แบบจำลองจะทายครวเรือนไม่ยากจนได้ครอบคลุมและมีประสิทธิภาพมากกว่า เนื่องจากจำนวนครวเรือนยากจนมีปริมาณของข้อมูลให้แบบจำลองเรียนรู้ต่ำกว่า

ในแบบจำลอง LightGBM ได้ทำการสุ่มจำนวนครวเรือนทั้งหมดจาก 46,840 ครวเรือนมาใช้ในการเรียนรู้ของแบบจำลองจำนวนทั้งสิ้น 32,788 ครวเรือน และใช้ครวเรือนทั้งสิ้น 14,052 ครวเรือนเป็นตัวชี้วัดความแม่นยำของแบบจำลองซึ่งสามารถแบ่งออกได้เป็นครวเรือนไม่ยากจน 12,982 ครวเรือน และครวเรือนยากจนทั้งสิ้น 1,070 ครวเรือน

ประสิทธิภาพของแบบจำลองในการทำนายว่าครวเรือนนั้นเป็นครวเรือนยากจนหรือไม่ พบว่า ผลการทำนายที่เป็นครวเรือนยากจนทั้งหมดแบบจำลองจะทำนายว่าเป็นครวเรือนยากจนได้อย่างถูกต้องร้อยละ 91 ในขณะที่ครวเรือนที่

ทำนายถูกต้องว่าเป็นครัวเรือนยากจนจะครอบคลุมครัวเรือนยากจนเพียงร้อยละ 50 ของครัวเรือนยากจนทั้งหมด และแบบจำลองมีความแม่นยำของแบบจำลองโดยรวมอยู่ที่ร้อยละ 65 (F1-score)

ผลการทำนายของแบบจำลอง พบว่า แบบจำลองมีตัวแปรที่สำคัญกับการทำนายครัวเรือนยากจน 10 อันดับแรก ดังนี้ จำนวนสมาชิกที่ทำงานและได้รับค่าจ้าง ค่าเช่าบ้านต่อเดือน จำนวนสมาชิกที่ได้สิทธิบัตรทอง จำนวนสมาชิกในครัวเรือนรวม ไม่รวมลูกจ้างและคนรับใช้ จำนวนสมาชิกที่ประกอบธุรกิจอุตสาหกรรมหรือวิชาชีพที่ไม่ใช่เกษตร จำนวนเครื่องปรับอากาศ อายุของหัวหน้าครัวเรือน จำนวนเตาไมโครเวฟ จำนวนเตียงนอน และจำนวนหลอดไฟนํ้าออน

การอธิบายผลการวิจัยผ่านการอธิบายค่าตัวแปรผ่านค่า SHAP ของแบบจำลอง LightGBM สามารถอธิบายผลได้ดังนี้ โดยเรียงลำดับจากความสำคัญของตัวแปรต่อแบบจำลองจากมากไปน้อย ตัวแปรที่สำคัญที่สุด คือ จำนวนสมาชิกที่ทำงานและได้รับค่าจ้าง ถ้าจำนวนสมาชิกดังกล่าวมีจำนวนมากจะยังมีโอกาสที่ครัวเรือนจะเป็นครัวเรือนไม่ยากจนในขณะที่ถ้าตัวแปรมีค่าน้อยจะยังมีโอกาสเป็นครัวเรือนยากจนเยอะ ตัวแปรที่สำคัญลำดับถัดมา คือ ค่าเช่าบ้านต่อเดือน ถ้าหากครัวเรือนมีค่าเช่าบ้านที่สูงจะยังมีแนวโน้มที่จะเป็นครัวเรือนที่ไม่ยากจนมากกว่าครัวเรือนที่มีค่าเช่าบ้านที่ต่ำกว่า โดยเปรียบเทียบ ซึ่งตัวแปรค่าเช่าบ้านดังกล่าวยังสะท้อนถึงรายจ่ายซึ่งสามารถใช้เป็นตัวแทนรายได้อีกด้วย

ตัวแปรที่สำคัญลำดับที่สาม คือ จำนวนสมาชิกที่ได้สิทธิบัตรทอง ตัวแปรนี้มีผลตรงกันข้ามกับตัวแปรอันดับแรก ถ้าจำนวนสมาชิกดังกล่าวมีจำนวนมากจะยังมีโอกาสที่ครัวเรือนจะเป็นครัวเรือนยากจนในขณะที่ถ้าตัวแปรมีค่าน้อยจะยังมีโอกาสเป็นครัวเรือนไม่ยากจน ตัวแปรสำคัญลำดับที่สี่ คือ จำนวนสมาชิกในครัวเรือนรวม ไม่รวมลูกจ้างและคนรับใช้ หากจำนวนสมาชิกมีจำนวนมากจะยังมีโอกาสที่ครัวเรือนจะเป็นครัวเรือนยากจนในขณะที่ถ้าตัวแปรมีค่าน้อยจะยังมีโอกาสเป็นครัวเรือนไม่ยากจนเช่นเดียวกับตัวแปรก่อนหน้า ตัวแปรลำดับที่ห้า และหก คือ จำนวนสมาชิกที่ประกอบธุรกิจอุตสาหกรรมหรือวิชาชีพที่ไม่ใช่เกษตร และจำนวนเครื่องปรับอากาศ โดยในตัวแปรทั้งสองถ้ามีจำนวนของตัวแปรดังกล่าวมากจะมีโอกาสไม่ใช่ครัวเรือนยากจนมากกว่าจำนวนน้อย

ตัวแปรที่มีความสำคัญอันดับที่เจ็ด คือ อายุของหัวหน้าครัวเรือน ถ้าหัวหน้าครัวเรือนมีอายุมากจะยังมีโอกาสที่จะเป็นครัวเรือนยากจนมากกว่า เนื่องจากสภาพของการเข้าถึงอาชีพและรายได้ต่างๆ มีข้อจำกัดมากขึ้นเมื่อมีอายุมากขึ้น ตัวแปรที่สำคัญลำดับถัดมา คือ จำนวนเตาไมโครเวฟที่มีใช้ในครัวเรือน ถ้ามีเตาไมโครเวฟใช้ในครัวเรือน ครัวเรือนจะยังมีโอกาสเป็นครัวเรือนยากจนน้อยกว่า ครัวเรือนที่ไม่มีเตาไมโครเวฟใช้ ตัวแปรที่สำคัญเป็นอันดับที่เก้า คือ จำนวนเตียงนอน ถ้าครัวเรือนมีเตียงใช้ในครัวเรือน ครัวเรือนจะยังมีโอกาสเป็นครัวเรือนยากจนน้อยกว่า ครัวเรือนที่ไม่มีเตียงนอนซึ่งตัวแปรส่งผลไปในลักษณะเดียวกันกับจำนวนเตาไมโครเวฟ และตัวแปรที่สำคัญเป็นลำดับที่สิบ คือ จำนวนหลอดไฟนํ้าออน ถ้าครัวเรือนมีจำนวนหลอดไฟนํ้าออนใช้ในครัวเรือนเยอะจะมีโอกาสที่ครัวเรือนจะไม่ใช้ครัวเรือนยากจนมากกว่าครัวเรือนที่มีจำนวนหลอดไฟนํ้าออนน้อยหรือไม่ใช้ในครัวเรือนเลย เนื่องจากหลอดไฟถือว่าเป็นสิ่งจำเป็นพื้นฐานในการเข้าถึงแสงสว่างซึ่งถือเป็นปัจจัยพื้นฐานในการดำรงชีพ

จะเห็นได้ว่า ตัวแปรที่แบบจำลองมองว่าสำคัญกับการทำนายครัวเรือนยากจนจะมีผลของตัวแปรไปในทิศทางที่ถูกต้องตามหลักการทางด้านเศรษฐศาสตร์และนโยบายสาธารณะทั้งหมด ดังนั้นงานวิจัยชิ้นนี้จึงได้ข้อสรุปว่านอกจากการใช้แบบจำลองเทคนิค Machine Learning ในการทำนายครัวเรือนยากจนนอกจากจะมีความแม่นยำของแบบจำลองที่มากกว่าการใช้แบบจำลองแบบปกติแล้ว การอธิบายผลของแบบจำลองยังสอดคล้องกับทฤษฎี และยังไม่ต้องคำนึงถึงปัญหาภาวะร่วมเส้นตรงเชิงพหุ หรือ Multicollinearity ทำให้ตัวแปรที่นำมาอธิบายมีความกว้างและหลากหลายไม่จำเป็นต้องเลือกตัวแปรมาวิเคราะห์เพียง 10 ตัว เช่น งานศึกษาของ Schreiner (2006, 2015, 2017) และงานศึกษาของ ชัยยุทธ ปัญญสวัสดิ์สุทธิ์ และคณะ (2560)

โดยสรุปแล้วงานศึกษาชิ้นนี้มีข้อเสนอแนะให้การคัดกรองจนควรทำเกณฑ์ตัวแปรที่ได้จากการใช้แบบจำลอง Machine Learning เป็นตัวช่วยในการคัดกรองครัวเรือนยากจนเพื่อให้ได้ประสิทธิภาพของการคัดกรองครัวเรือนยากจนที่ดีที่สุด และควรมีการประยุกต์นำเทคนิค Machine Learning ไปใช้กับงานศึกษาทางด้านเศรษฐศาสตร์ชิ้นอื่นๆ เพื่อให้เกิด

ประสิทธิภาพของแบบจำลองที่สูงที่สุด และใช้ค่า SHAP ในการแปลผลของแบบจำลองเพื่อให้แบบจำลองสูญเสียความสามารถในการอธิบายผลตัวแปรไป

เอกสารอ้างอิง

- ชัยยุทธ ปัญญสวัสดิ์สุทธิ์ และคณะ. (2560). รายงานฉบับสมบูรณ์ โครงการพัฒนาระบบสารสนเทศเพื่อหลักประกันโอกาสทางการเรียนรู้: พัฒนาระบบสารสนเทศเพื่อหลักประกันโอกาสทางการเรียนรู้ (หลักเกณฑ์และวิธีการคัดกรอง). ศูนย์บริการวิชาการเศรษฐศาสตร์ คณะเศรษฐศาสตร์ มหาวิทยาลัยธรรมศาสตร์
- ชัยยุทธ ปัญญสวัสดิ์สุทธิ์ และคณะ. (2565). โครงการวิจัยเพิ่มประสิทธิภาพหลักเกณฑ์คัดกรองความยากจนด้วยวิธีการวัดรายได้ทางอ้อม เพื่อจัดสรรเงินอุดหนุนแบบมีเงื่อนไขประจำปีการศึกษา 2565.
- Athey, S. (2018). *The Impact of Machine Learning on Economics*.
- Houssou, N., Zeller, M., Alcaraz, G., Schwarze, S., & Johannsen, J. (2007). *Proxy Means Tests for Targeting the Poorest Households Applications to Uganda Pro-poor development in low income countries: Food, agriculture, trade, and environment*.
- Kambuya, P. (2017). *Better Model Selection for Poverty Targeting through Machine Learning: A Case Study in Thailand (Vol. 38, Issue 1)*.
- Mcbride, L., & Nichols, A. (2016). *Retooling Poverty Targeting Using Out-of-Sample Validation and Machine Learning*. Retrieved from <http://econ.worldbank.org>.
- Schreiner, M. (2006). *A simple poverty scorecard for India*. report to Grameen Foundation USA. Retrieved from http://www.microfinance.com/English/Papers/Scoring_Poverty_India.pdf.
- Schreiner, M. (2015). *An Expert-Based Poverty Scorecard for Rural China*. Retrieved from microfinance.com/English/Papers/Scoring_Poverty_China_EN.pdf.
- Schreiner, M. (2017). *Simple Poverty Scorecard Poverty-Assessment Tool: Cambodia*. Retrieved from SimplePovertyScorecard.com/KHM_2011_ENG. Pdf.
- Ysobel, D., & Ting, P. (2024). *Applying an Unsupervised Machine Learning Approach to Analyze the Non-Income Poverty Indicators Used in the Listahanan 2*.

Data Availability Statement: The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

Conflicts of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.



Copyright: © 2025 by the authors. This is a fully open-access article distributed under the terms of the Attribution-NonCommercial-NoDerivatives 4.0 International (CC BY-NC-ND 4.0).