บทที่ 1
บทนำ

ความเป็นมาและความสำคัญของปัญหา

ในปัจจุบันได้มีการนำเทคนิคการพยากรณ์เข้ามาใช้ในงานวิจัยด้านต่างๆ หลายสาขา ซึ่งเทคนิคการพยากรณ์มีหลายวิธี การเลือกใช้ขึ้นอยู่กับลักษณะของข้อมูลที่นำมาศึกษาและจุดประสงค์ของการพยากรณ์ และการพยากรณ์ด้วยวิธีการวิเคราะห์ความถดถอยโลจิสติค (Logistic Regression Analysis) ก็เป็นวิธีหนึ่งที่นิยมนำมาใช้กันมาก ซึ่งมีจุดประสงค์เพื่อหาความสัมพันธ์ระหว่างตัวแปรตามและตัวแปรอิสระ โดยตัวแปรตามเป็นตัวแปรเชิงกลุ่ม ที่แบ่งเป็น 2 กลุ่ม หรือมากกว่า 2 กลุ่ม และตัวแปรอิสระเป็นตัวแปรเชิงปริมาณหรือตัวแปรเชิงกลุ่มหรือทั้งสองอย่าง ซึ่งจะแสดงอยู่ในรูปของสมการทางคณิตศาสตร์ เรียกว่า "ตัวแบบถดถอยโลจิสติค" เพื่อใช้พยากรณ์โอกาสหรือความน่าจะเป็นของการเกิดเหตุการณ์สนใจ สำหรับการวิเคราะห์ความถดถอยโลจิสติคสามารถทำได้ 2 แบบ คือ การวิเคราะห์ความถดถอยโลจิสติคแบบสถิตย์ (Static Logistic Regression Analysis) หรือแบบปกติ เป็นการหาตัวแบบถดถอยโลจิสติคโดยทำการประมาณค่าพารามิเตอร์จากข้อมูลเพียงใน 1 ช่วงเวลา และการวิเคราะห์ความถดถอยโลจิสติคแบบพลวัตร (Dynamic Logistic Regression Analysis) จะเป็นการสร้างตัวแบบถดถอยที่ได้ทำการแบ่งช่วงเวลาออกเป็นหลายช่วงเวลาย่อย และพิจารณาข้อมูลที่เปลี่ยนแปลงของแต่ละหน่วยตัวอย่างในช่วงเวลาย่อยเหล่านั้น เป็นเทคนิคการประมาณค่าพารามิเตอร์ในตัวแบบโดยที่มีการนำเวลา (Time) หรือตัวแปรที่เปลี่ยนตามเวลามาร่วมพิจารณาในการวิเคราะห์ด้วย ซึ่งมาจากแนวคิดที่ว่าลักษณะของหน่วยตัวอย่างมีการเปลี่ยนแปลงไปตามเวลาในลักษณะของ Hazard Model

จากงานวิจัยของ Tyler Shumway (2001) ได้ศึกษาการวิเคราะห์ความถดถอยโลจิสติคกับข้อมูลตลาดหุ้นของกลุ่มอุตสาหกรรมในนิวยอร์กและอเมริกา ปี ค.ศ.1962 ถึง ค.ศ.1992 โดยกำหนดให้แต่ละปีเป็น 1 ช่วงเวลา พบว่าตัวประมาณที่ได้จากการวิเคราะห์แบบพลวัตรเป็นตัวประมาณที่ไม่เอนเอียง มีความคงเส้นคงวา และให้เปอร์เซ็นต์ความถูกต้องของการพยากรณ์สูงกว่าตัวแบบสถิตย์ และจากงานวิจัยของ Sunti Tirapat and Seksan Kiatsupaibul (2007) ได้ศึกษาการพยากรณ์มูลค่าความเสี่ยง (Credit Value at Risk หรือ Credit VaR) ซึ่งเป็นมูลค่าความเสียหายสูงสุดที่คาดว่าจะเกิดขึ้นจากการผิดนัดชำระหนี้ของลูกค้า ภายใต้ระดับความเชื่อมั่นหนึ่งที่นักลงทุนยอมรับได้ โดยใช้ตัวแบบถดถอยโลจิสติคแบบปกติ พบว่าตัวแบบประมาณที่ได้ให้ค่า Credit VaR ของการประมาณต่ำเกินไป

โดยส่วนใหญ่นักวิจัยมักใช้การวิเคราะห์ความถดถอยโลจิสติคแบบสถิตย์เพราะเป็นวิธีที่ง่าย ทั้งที่ในความเป็นจริงแล้ว การวิเคราะห์ความถดถอยโลจิสติคแบบพลวัตรให้ความถูกต้อง

มากกว่า กระบวนการวิเคราะห์ก็ไม่ได้ซับซ้อนเกินกว่าแบบสถิตย์มากนัก ใช้วิธีการประมาณค่าพารามิเตอร์แบบสถิตย์ได้ และสามารถประยุกต์ใช้กับโปรแกรมทางสถิติได้ แต่จะมีเพียงการจัดข้อมูลสำหรับการวิเคราะห์เท่านั้นที่แตกต่างกัน และจากแนวคิดที่ว่าลักษณะของหน่วยตัวอย่างอาจมีการเปลี่ยนแปลงไปเมื่อเวลาผ่านไปและลักษณะที่เปลี่ยนไปอาจมีผลต่อการเกิดเหตุการณ์ที่สนใจได้ ด้วยเหตุนี้จึงทำให้ผู้วิจัยสนใจที่จะศึกษาและเปรียบเทียบประสิทธิภาพของตัวแบบถดถอยโลจิสติคแบบสถิตย์และแบบพลวัตร โดยศึกษากับข้อมูลระดับหน่วยย่อย (Micro unit) ที่ได้จากการจำลองขึ้นด้วยวิธีมอนติคาร์โล (Monte Carlo Method) ซึ่งข้อมูลลักษณะนี้จะมีความซับซ้อนและมีข้อมูลเป็นจำนวนมาก การจัดข้อมูลเพื่อการวิเคราะห์จึงเป็นสิ่งสำคัญที่จะต้องทำอย่างรอบคอบและถูกต้อง และงานวิจัยครั้งนี้จะแตกต่างจากงานวิจัยของ Tyler Shumway ที่ได้ทำการศึกษากับข้อมูลระดับหน่วยใหญ่ (Macro unit) ซึ่งความซับซ้อนของข้อมูลอาจมีน้อยกว่าและสัดส่วนของการเกิดเหตุการณ์ที่สนใจจะต่ำกว่า รวมทั้งการศึกษาลักษณะของข้อมูลต้องใช้ระยะเวลานานกว่าข้อมูลระดับหน่วยย่อย สำหรับการเปรียบเทียบจะพิจารณาจากพื้นที่ใต้โค้ง ROC เพื่อวัดประสิทธิภาพของการพยากรณ์ และวัดความเหมาะสมของแต่ละตัวแบบด้วยค่าสถิติ R^2 (Pseudo R^2)

วัตถุประสงค์ของการวิจัย

1. เพื่อศึกษาวิธีการวิเคราะห์ความถดถอยโลจิสติคแบบสถิตย์และการวิเคราะห์ความถดถอยโลจิสติคแบบพลวัตร
2. เพื่อเปรียบเทียบประสิทธิภาพของการพยากรณ์ระหว่างตัวแบบความถดถอยโลจิสติคแบบสถิตย์และแบบพลวัตร

สมมติฐานของการวิจัย

ตัวแบบความถดถอยโลจิสติคแบบพลวัตรน่าจะเป็นตัวแบบที่มีความเหมาะสมและมีประสิทธิภาพในการพยากรณ์สูงกว่าตัวแบบความถดถอยโลจิสติคแบบสถิตย์ เนื่องจากการประมาณค่าสัมประสิทธิ์ความถดถอยในตัวแบบได้นำสารสนเทศจากข้อมูลมาใช้มากกว่าวิธีการวิเคราะห์แบบสถิตย์

ขอบเขตของการวิจัย

ในการวิจัยครั้งนี้จะทำการศึกษาภายใต้ขอบเขตดังนี้

1. ศึกษาการวิเคราะห์ความถดถอยโลจิสติคแบบสถิตย์และแบบพลวัตร โดยตัวแปรตามมีค่าที่เป็นไปได้ 2 ค่า หรือแบ่งเป็น 2 กลุ่ม คือ กลุ่มที่เกิดเหตุการณ์ที่สนใจ และกลุ่มที่ไม่เกิดเหตุการณ์ที่สนใจ นั่นคือการวิเคราะห์ความถดถอยโลจิสติคแบบ 2 กลุ่ม (Binary Logistic Regression)

2. ศึกษากับข้อมูลที่ได้จากการจำลองขึ้นด้วยวิธีมอนติคาร์โล โดยมีจำนวนตัวแปรอิสระในแต่ละตัวแบบเท่ากับ 3 ตัวแปร ได้แก่ X_1, X_2, X_3 โดยที่ X_1 มีการแจกแจงแบบปกติด้วย $\mu = 0$ และ $\sigma^2 = 1$ X_2 มีการแจกแจงแบบเบอร์นูลีด้วยความน่าจะเป็นของความสำเร็จ $p = 0.5$ และ X_3 มีการแจกแจงแบบเอกซโพเนนเชียลด้วยพารามิเตอร์ $\lambda = 1/36$ กำหนดขนาดตัวอย่างเท่ากับ 10,000 หน่วยตัวอย่าง และกระทำซ้ำจำนวน 1,000 ครั้ง

3. สำหรับการวิเคราะห์ความถดถอยโลจิสติคแบบพลวัตรจะมีตัวแปรอิสระ 1 ตัวที่มีค่าเปลี่ยนแปลงไปในแต่ละช่วงเวลา ส่วนตัวแปรอิสระอีก 2 ตัวที่เหลือจะมีค่าคงที่ตลอดช่วงของศึกษา

4. เปรียบเทียบความความถูกต้องของการพยากรณ์ด้วยพื้นที่ใต้โค้ง ROC

5. ในภาคผนวกจะได้นำวิธีการวิเคราะห์แบบสถิตย์และแบบพลวัตรไปประยุกต์ใช้กับข้อมูลจริงเกี่ยวกับการเช่าซื้อสินค้าชนิดหนึ่ง โดยเหตุการณ์ที่สนใจ คือ การเกิดหนี้ NPL (Non Performing Loan) ของลูกค้า ใช้จำนวนตัวแปรอิสระในแต่ละตัวแบบเท่ากับ 3 ตัวแปร โดยมี 1 ตัวแปรที่มีค่าเปลี่ยนแปลงไปในแต่ละช่วงเวลา และขนาดตัวอย่างเท่ากับ 10,000 หน่วยตัวอย่าง

ข้อตกลงเบื้องต้น

1. สำหรับการวิจัยครั้งนี้จะทำการศึกษาข้อมูลในการทดลองแบบปิด คือ ไม่มีหน่วยตัวอย่างเข้าใหม่และไม่มีหน่วยตัวอย่างภายใต้การทดลองออกจากกลุ่มในระหว่างการทดลอง นอกเสียจากจะเกิดเหตุการณ์ที่สนใจหรือสิ้นสุดเวลาของหน่วยตัวอย่างนั้น

2. ลักษณะของข้อมูลที่ใช้เป็นการศึกษาแบบติดตามผล คือ สังเกตข้อมูลตั้งแต่เริ่มเข้ามาในช่วงของการสังเกตจนกระทั่งเกิดเหตุการณ์ที่สนใจ หรือสิ้นสุดเวลาของหน่วยตัวอย่างนั้น หรือสิ้นสุดช่วงเวลาของการศึกษา

3. การวิจัยครั้งนี้จะพิจารณากรณีที่ช่วงเวลาเป็นแบบไม่ต่อเนื่องและมีความยาวของแต่ละช่วงเวลาเท่ากัน

4. สำหรับข้อมูลจากการจำลองกำหนดให้เริ่มต้นเก็บข้อมูลที่เวลาเดียวกัน คือ $t = 0$ และจะเก็บข้อมูลเป็นระยะเวลา 12 และ 24 ช่วงเวลา

คำจำกัดความที่ใช้ในการวิจัย

1. R^2 (R-square)

เป็นการวัดระดับความสัมพันธ์ระหว่างตัวแปรตามกับตัวแปรอิสระในตัวแบบ หรือกล่าวได้ว่าค่า R^2 เป็นค่าที่ระบุถึงสัดส่วนของความผันแปรของตัวแปรตามที่อธิบายได้ด้วยตัวแปรอิสระ แต่สำหรับการวิเคราะห์ความถดถอยโลจิสติค ค่า R^2 ที่ได้ไม่ใช่ค่าสัดส่วนที่แท้จริงของความผันแปรเนื่องจากเราจะใช้ค่าของ Log Likelihood แทนค่า Sum Square Error ในการคำนวณค่า R^2 จึงเรียกว่า "$Pseudo\,R^2$" ซึ่งมีสูตรการคำนวณดังนี้

$$Cox\ \&\ Snell\ R^2 = 1 - \exp\left[\frac{-G}{n}\right] \qquad (1.1)$$

เมื่อ $G = [-2LL_0] - [-2LL_1]$

$-2LL_0 = -2 \times \log\ likelihood$ ของฟังก์ชันที่มีเพียงค่าคงที่

$-2LL_1 = -2 \times \log\ likelihood$ ของฟังก์ชันที่มีตัวแปรอิสระที่กำหนด

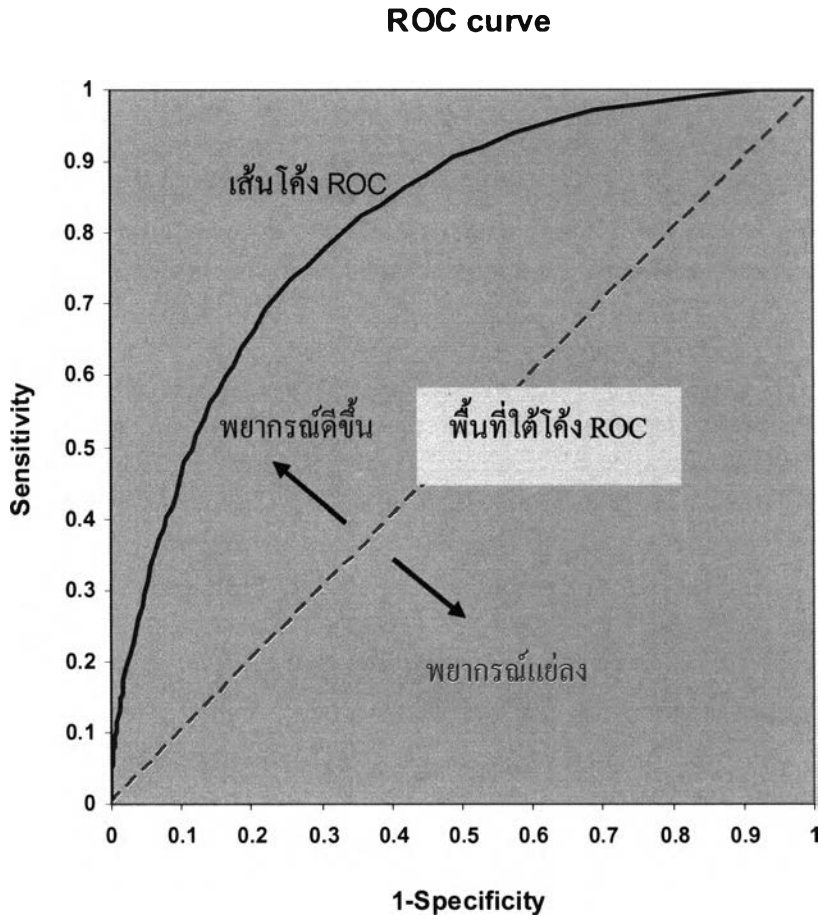
ซึ่ง $R^2_{CS} < 1$ เสมอ

2. **ROC หรือ ROC curve** (Receiver Operating Characteristic)

Receiver Operating Characteristic หรือเรียกง่ายๆ ว่า "ROC Curve" ได้ถูกนำมาใช้ครั้งแรกในช่วงสงครามโลกครั้งที่ 2 สำหรับการวิเคราะห์สัญญาณเรด้าโดยทหารของสหรัฐอเมริกา และต่อจากนั้นมาก็ได้มีงานวิจัยที่นำ ROC Curve มาใช้อย่างแพร่หลายมากขึ้น ทั้งทางด้านจิตวิทยาเกี่ยวกับมนุษย์ การวินิจฉัยโรคของทางการแพทย์ และการประเมินความถูกต้องของตัวแบบของทางด้านวิทยาศาสตร์และสังคม

ROC Curve ถูกนำมาใช้ในการประเมินความถูกต้องของการพยากรณ์เหตุการณ์ในระบบการจำแนกกลุ่มกรณีแบ่งเป็น 2 กลุ่ม ได้แก่ กลุ่มที่เกิดเหตุการณ์ที่สนใจและกลุ่มที่ไม่เกิดเหตุการณ์ที่สนใจ โดยอยู่ในรูปของกราฟที่พล็อตระหว่างค่า Sensitivity หรือ True-Positive Rate ซึ่งเป็นอัตราส่วนของการพยากรณ์ถูกต้องของการเกิดเหตุการณ์ที่สนใจ $(Y = 1)$ และค่า 1-Specificity หรือ False-Positive Rate ซึ่งเป็นอัตราส่วนของการพยากรณ์ผิดของการไม่เกิดเหตุการณ์ที่สนใจ $(Y = 0)$ โดยจะกำหนดให้ 1-Specificity อยู่บนแกน X และ Sensitivity อยู่บนแกน Y และกราฟอยู่ในช่วง $[0,1]$ ดังตัวอย่างในรูปที่ 1.1

รูปที่ 1.1 ตัวอย่างเส้นโค้งและพื้นที่ใต้โค้ง ROC



สำหรับการคำนวณค่า Sensitivity และ 1-Specificity จะทำโดยการกำหนดค่าจุดตัด (Cutoff) ที่ระดับต่างๆ ระหว่าง 0 ถึง 1 เพื่อเปรียบเทียบกับค่าความน่าจะเป็นของการเกิดเหตุการณ์ที่สนใจที่ได้จากการพยากรณ์ของแต่ละหน่วยตัวอย่างด้วยตัวแบบประมาณที่ได้จากการวิเคราะห์ $(\hat{p}_i)$ แล้วทำการจำแนกกลุ่มของตัวแปรตามซึ่งเป็นตัวแปรที่ต้องการพยากรณ์ ออกเป็น 2 กลุ่ม โดยที่

ถ้า $\hat{p}_i \le cutoff \longrightarrow \hat{Y}_i = 0$ (ตัวอย่างจะถูกพยากรณ์ให้อยู่ในกลุ่มที่ไม่เกิดเหตุการณ์)

$\hat{p}_i > cutoff \longrightarrow \hat{Y}_i = 1$ (ตัวอย่างจะถูกพยากรณ์ให้อยู่ในกลุ่มที่เกิดเหตุการณ์)

จากนั้นจึงคำนวณหาสัดส่วนของการพยากรณ์เหตุการณ์ เพื่อนำค่าที่ได้ไปทำการพล็อตโค้ง ROC และคำนวณหาพื้นที่ใต้โค้ง ซึ่งสูตรการคำนวณ Sensitivity และ 1-Specificity เป็นดังนี้

		เหตุการณ์จริง	
		เกิดเหตุการณ์ $(Y=1)$	ไม่เกิดเหตุการณ์ $(Y=0)$
พยากรณ์	เกิดเหตุการณ์ Positive $(\hat{Y}=1)$	TP	FP
	ไม่เกิดเหตุการณ์ Negative $(\hat{Y}=0)$	FN	TN
		TP+FN	FP+TN

TP (True Positive) : จำนวนตัวอย่างที่พยากรณ์ถูกต้องของการเกิดเหตุการณ์ $(\hat{Y}=1 \mid Y=1)$

FP (False Positive) : จำนวนตัวอย่างที่พยากรณ์ผิดของการไม่เกิดเหตุการณ์ $(\hat{Y}=1 \mid Y=0)$

TN (True Negative) : จำนวนตัวอย่างที่พยากรณ์ถูกต้องของการไม่เกิดเหตุการณ์ $(\hat{Y}=0 \mid Y=0)$

FN (False Negative) : จำนวนตัวอย่างที่พยากรณ์ผิดของการไม่เกิดเหตุการณ์ $(\hat{Y}=1 \mid Y=0)$

Sensitivity (True Positive Rate) เป็นความน่าจะเป็นหรืออัตราส่วนของการพยากรณ์เหตุการณ์ได้ถูกต้องของการเกิดเหตุการณ์ที่สนใจ

Specificity (True Negative Rate) เป็นความน่าจะเป็นหรืออัตราส่วนของการพยากรณ์เหตุการณ์ได้ถูกต้องของการไม่เกิดเหตุการณ์ที่สนใจ

1-Specificity (False Negative Rate) เป็นความน่าจะเป็นหรืออัตราส่วนของการพยากรณ์เหตุการณ์ผิดของการไม่เกิดเหตุการณ์ที่สนใจ

ซึ่งจะได้ว่า

$$Sensitivity = P(\hat{Y}=1 \mid Y=1) = \frac{TP}{Total\ actual\ Positive} = \frac{TP}{TP+FN}$$

$$Specificity = P(\hat{Y}=0 \mid Y=0) = \frac{TN}{Total\ actual\ Negative} = \frac{TN}{FP+TN}$$

$$1-Specificity = P(\hat{Y}=1 \mid Y=0) = \frac{FP}{Total\ actual\ Negative} = \frac{FP}{FP+TN}$$

พื้นที่ใต้โค้ง ROC เป็นค่าที่บ่งบอกถึงความสามารถในการพยากรณ์ได้ถูกต้องหรือแสดงถึงความเชื่อถือได้ของตัวแบบ มีพิสัยอยู่ระหว่าง 0 ถึง 1 ซึ่ง Hosmer, David W., and Stanley Lemeshow.(2000) ได้กำหนดเกณฑ์ทั่วไปไว้ดังนี้

ถ้าพื้นที่ใต้โค้ง $0.5 \le ROC < 0.6$	เป็นตัวแบบที่เชื่อถือไม่ได้
ถ้าพื้นที่ใต้โค้ง $0.6 \le ROC < 0.7$	เป็นตัวแบบที่เชื่อถือได้น้อย
ถ้าพื้นที่ใต้โค้ง $0.7 \le ROC < 0.8$	เป็นตัวแบบที่เชื่อถือสามารถยอมรับได้
ถ้าพื้นที่ใต้โค้ง $0.8 \le ROC < 0.9$	เป็นตัวแบบที่เชื่อถือได้ในระดับดี
ถ้าพื้นที่ใต้โค้ง $ROC \ge 0.9$	เป็นตัวแบบที่เชื่อถือได้ในระดับดีมาก

ถ้าพื้นที่ใต้โค้ง ROC เท่ากับ 0.5 แสดงว่า ตัวแบบไม่สามารถจำแนกกลุ่มได้ (เป็นเสมือนการเลือกอย่างสุ่มสำหรับการจำแนกกลุ่ม 2 กลุ่ม) เส้นโค้งจะมีลักษณะเป็นเส้นทแยงมุมจากจุด (0,0) ถึงจุด (1,1) ดังรูปที่ 1.1 และถ้าพื้นที่ใต้โค้ง ROC เท่ากับ 1 แสดงว่า ตัวแบบสามารถจำแนกกลุ่มได้อย่างถูกต้องสมบูรณ์ ในกรณีที่เส้นโค้ง ROC อยู่ใต้เส้นทแยงมุมหรือมีพื้นที่ใต้ ROC < 0.5 จำเป็นจะต้องมีการปรับเปลี่ยนให้เส้นอยู่ข้างบนเส้นทแยงมุมโดยทำการกลับด้านของการพยากรณ์ของทุกหน่วยตัวอย่าง จากการสนในการพยากรณ์ที่เป็นเกิดเหตุการณ์เปลี่ยนเป็นสนใจการพยากรณ์กลุ่มที่ไม่เกิดเหตุการณ์ นั่นคือ เปลี่ยนจาก TP เป็น FN และเปลี่ยนจาก FP เป็น TN แทน

สำหรับ 2 ตัวแบบประมาณ เมื่อทำการทดสอบการพยากรณ์กับข้อมูลชุดเดียวกันแล้ว ตัวแบบใดที่มีเส้นโค้ง ROC อยู่ข้างบนหรือมีพื้นที่ใต้โค้ง ROC มากกว่า แสดงว่าเป็นตัวแบบที่มีประสิทธิภาพหรือให้ความถูกต้องในการพยากรณ์ได้ดีกว่า

เกณฑ์การตัดสินใจ

สำหรับตัวแบบถดถอยโลจิสติคที่ได้จากการวิเคราะห์ด้วยวิธีใดมีพื้นที่ใต้โค้ง ROC ของการพยากรณ์มากกว่า จะเป็นตัวแบบที่ให้ความถูกต้องของการพยากรณ์สูงกว่าหรือมีประสิทธิภาพในการพยากรณ์ได้ดีกว่า

ประโยชน์ที่คาดว่าจะได้รับ

1. เพื่อเป็นแนวทางในการวิเคราะห์ความถดถอยโลจิสติคแบบสถิตย์และแบบพลวัตรด้วย Hazard model

2. เพื่อเป็นแนวทางในการสร้างตัวแบบถดถอยโลจิสติคที่เหมาะสมสำหรับพยากรณ์โอกาสของการเกิดเหตุการณ์ที่สนใจกับข้อมูลที่มีลักษณะเป็นหน่วยย่อย (Micro unit)

3. เพื่อเป็นแนวทางในการนำวิธีการเปรียบเทียบประสิทธิภาพของการพยากรณ์ของตัวแบบถดถอยโลจิสติคด้วยโค้ง ROC มาใช้