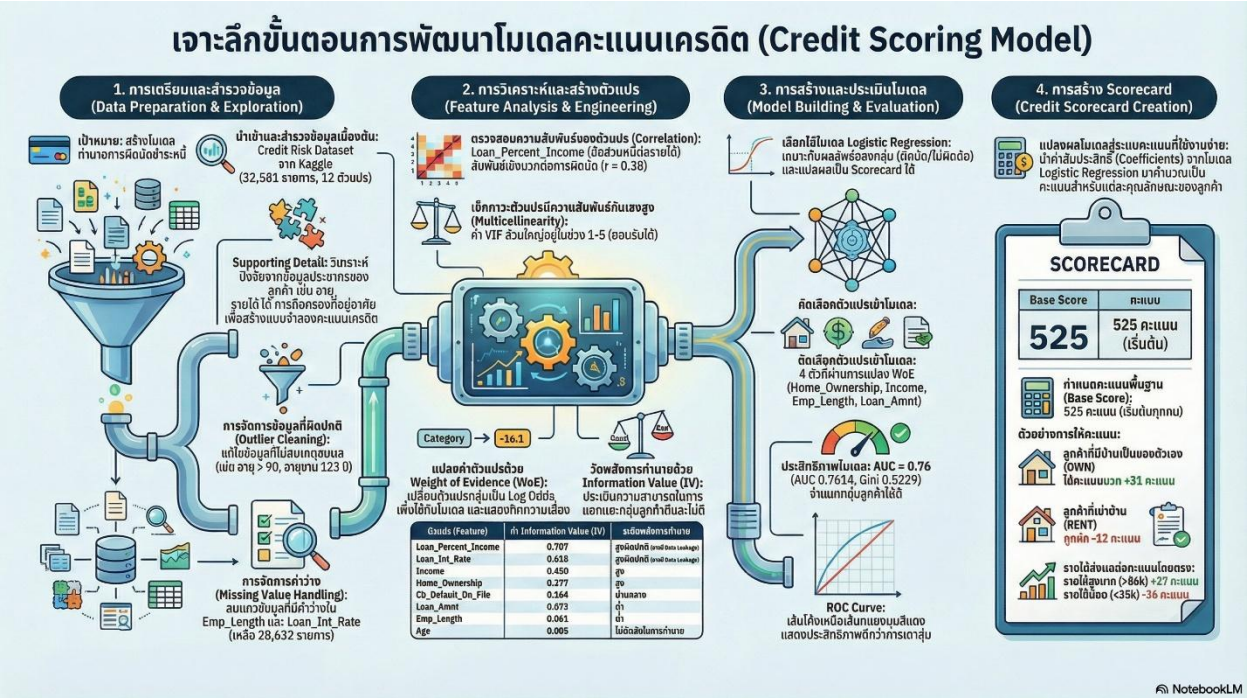
รายงานทางเทคนิค: การวิเคราะห์ความเสี่ยงด้านสินเชื่อและการพัฒนาแบบจำลอง Credit Scorecard



1.2. สรุปข้อมูลที่ใช้ในการวิเคราะห์

ชุดข้อมูลที่ใช้ในการวิเคราะห์คือ credit_risk_dataset.csv ซึ่งเป็นข้อมูลจำลองจาก Credit Bureau ที่เผยแพร่บนแพลตฟอร์ม Kaggle โดยมีรายละเอียดของตัวแปรต่างๆ ดังนี้

Feature Name	Description
person_age	Age
person_income	Annual Income
person_home_ownership	Home ownership
person_emp_length	Employment length (in years)
loan_intent	Loan intent
loan_grade	Loan grade

Feature Name	Description
loan_amnt	Loan amount
loan_int_rate	Interest rate
loan_status	Loan status (0 is non default, 1 is default)
loan_percent_income	Percent income
cb_person_default_on_file	Historical default
cb_person_cred_hist_length	Credit history length

1.3. ภาพรวมข้อมูลเบื้องต้น

- ข้อมูลเริ่มต้นประกอบด้วย 32,581 ระเบียน (Records) และ 12 ตัวแปร (Variables) โดยมีประเภทข้อมูลที่หลากหลาย ดังนี้
- **int64:** 5 ตัวแปร
- **float64:** 3 ตัวแปร
- **object:** 4 ตัวแปร (ข้อมูลเชิงกลุ่ม)

1.4. บทสรุปเบื้องต้น

ก่อนที่จะดำเนินการวิเคราะห์เชิงลึกในขั้นตอนต่อไป จำเป็นต้องมีการตรวจสอบคุณภาพของข้อมูล จัดการค่าที่ผิดปกติ (Outliers) และค่าว่าง (Missing Values)
เพื่อเตรียมข้อมูลให้มีความสมบูรณ์และพร้อมสำหรับการสร้างแบบจำลองที่มีความน่าเชื่อถือ

2.0 การเตรียมข้อมูลและการจัดการข้อมูลที่ผิดปกติ (Data Preparation & Cleaning)

2.1. ความสำคัญของการเตรียมข้อมูล

การเตรียมข้อมูล (Data Preparation) เป็นขั้นตอนพื้นฐานที่สำคัญที่สุดในกระบวนการสร้างแบบจำลอง
เพื่อให้แน่ใจว่าข้อมูลที่นำมาใช้มีความสะอาด ถูกต้อง และมีคุณภาพสูง
ซึ่งจะส่งผลโดยตรงต่อความแม่นยำและเสถียรภาพของแบบจำลองที่พัฒนาขึ้น

2.2. การจัดการค่าผิดปกติ (Outlier Handling)

2.2.1. การวิเคราะห์ค่าผิดปกติ

จากการตรวจสอบข้อมูลเบื้องต้นผ่าน Box Plot พบว่ามีค่าผิดปกติ (Outliers) ที่ไม่สมเหตุสมผลในตัวแปร person_age และ person_emp_length ซึ่งจำเป็นต้องได้รับการจัดการก่อนนำไปวิเคราะห์

2.2.2. ขั้นตอนการจัดการ

- **person_age (อายุ):** พบค่าอายุสูงสุดที่ 144 ปี ซึ่งเป็นไปไม่ได้ในความเป็นจริง
  จึงได้ทำการกรองข้อมูลให้อยู่ในช่วงอายุที่สมเหตุสมผลคือ 20 ถึง 90 ปี
  ทำให้ค่าสูงสุดใหม่ของอายุลดลงมาอยู่ที่ 84 ปี
- **person_emp_length (อายุงาน):** พบค่าอายุงานสูงสุดที่ 123 ปี ซึ่งเป็นไปไม่ได้เช่นกัน จึงได้ทำการลบระเบียนข้อมูล ที่มีค่าดังกล่าวออก

2.3. การจัดการค่าว่าง (Missing Value Handling)

2.3.1. จำนวนค่าว่างที่พบ

จากการตรวจสอบพบว่ามีค่าว่างใน 2 ตัวแปรหลัก ดังตาราง

ตัวแปร	จำนวนค่าว่าง
person_emp_length	895
loan_int_rate	3,114

2.3.2. วิธีการจัดการ

เพื่อให้ข้อมูลที่ใช้ในการสร้างแบบจำลองมีความสมบูรณ์ครบถ้วนทุกตัวแปร โครงการนี้จึงเลือกใช้วิธีการลบระเบียนข้อมูลที่มีค่าว่างออกทั้งหมด (dropna) เพื่อรักษาความสมบูรณ์ของข้อมูลในทุกมิติ และหลีกเลี่ยงการสร้างอคติที่อาจเกิดขึ้นจากการเติมค่า (imputation)
ในตัวแปรที่สำคัญต่อการประเมินความเสี่ยง

2.3.3. ผลลัพธ์หลังการจัดการ

หลังจากการลบค่าว่าง มีข้อมูลที่ถูกลบออกไปทั้งสิ้น 3,941 ระเบียน
ทำให้ข้อมูลสุดท้ายที่พร้อมสำหรับนำไปวิเคราะห์มีจำนวนทั้งสิ้น 28,632 ระเบียน

2.4. บทสรุป

หลังจากที่ข้อมูลได้รับการทำความสะอาด จัดการค่าผิดปกติ และค่าว่างเรียบร้อยแล้ว
ขั้นตอนต่อไปคือการสำรวจความสัมพันธ์ระหว่างตัวแปรต่างๆ
เพื่อทำความเข้าใจโครงสร้างและคุณลักษณะของข้อมูลในเชิงลึก

3.0 การวิเคราะห์ข้อมูลเชิงสำรวจ (Exploratory Data Analysis)

3.1. วัตถุประสงค์

ในส่วนนี้เป็นการตรวจสอบความสัมพันธ์ระหว่างตัวแปรเชิงตัวเลขต่างๆ ผ่านการวิเคราะห์ค่าสหสัมพันธ์ (Correlation) และภาวะร่วมเส้นตรง (Multicollinearity)
เพื่อประเมินความเหมาะสมของตัวแปรแต่ละตัวก่อนที่จะนำไปใช้สร้างแบบจำลอง

3.2. การวิเคราะห์สหสัมพันธ์ (Correlation Analysis)

จากการวิเคราะห์ Correlation Heatmap พบความสัมพันธ์ที่สำคัญหลายประการ:

- ความสัมพันธ์เชิงบวกที่แข็งแกร่งที่สุด: Age และ Cb_Cred_Hist_Length (r = 0.88)
  มีความสัมพันธ์กันสูงมาก ซึ่งสมเหตุสมผลเนื่องจากผู้ที่มีอายุมากมักจะมีประวัติสินเชื่อที่ยาวนานกว่า
- ปัจจัยที่ส่งผลต่อสถานะสินเชื่อ (Loan_Status):
  - Loan_Percent_Income (r = 0.38): มีความสัมพันธ์เชิงบวกปานกลาง
    หมายความว่ายิ่งสัดส่วนหนี้ต่อรายได้สูงขึ้น
    ความเสี่ยงในการผิดนัดชำระหนี้ก็จะสูงขึ้นตามไปด้วย
  - Loan_Int_Rate (r = 0.34): อัตราดอกเบี้ยมีความสัมพันธ์เชิงบวกกับสถานะการผิดนัดชำระหนี้
    ซึ่งอาจตีความได้ว่ากลุ่มลูกค้าที่มีความเสี่ยงสูงมักจะได้รับอัตราดอกเบี้ยที่สูงกว่า
  - Income (r = -0.17): รายได้มีความสัมพันธ์เชิงลบกับความเสี่ยง
    กล่าวคือผู้ที่มีรายได้สูงมีแนวโน้มที่จะผิดนัดชำระหนี้น้อยกว่า
- โครงสร้างหนี้และรายได้: Loan_Amnt มีความสัมพันธ์เชิงบวกกับ Loan_Percent_Income (r = 0.57)
  ซึ่งบ่งชี้ว่าเมื่อวงเงินกู้สูงขึ้น สัดส่วนภาระหนี้ต่อรายได้ก็จะสูงขึ้นอย่างชัดเจน

3.3. การวิเคราะห์ภาวะร่วมเส้นตรง (Variance Inflation Factor - VIF)

VIF เป็นตัวชี้วัดที่ใช้ประเมินว่าตัวแปรอิสระตัวหนึ่งมีความสัมพันธ์กับตัวแปรอิสระตัวอื่นๆ มากน้อยเพียงใด

3.3.1. ผลการคำนวณค่า VIF

Feature	VIF
Age	4
Income	2
Emp_Length	1
Loan_Amnt	3
Loan_Int_Rate	1
Loan_Status	1
Loan_Percent_Income	3
Cb_Cred_Hist_Length	4

3.3.2. การประเมินผลลัพธ์

ตามเกณฑ์ทั่วไป ค่า VIF ที่อยู่ระหว่าง 1-5 ถือว่ามีความสัมพันธ์กันในระดับปานกลางและยอมรับได้
จากตารางข้างต้นจะเห็นว่าตัวแปรทุกตัวมีค่า VIF ไม่เกิน 4 ซึ่งหมายความว่า
ชุดข้อมูลนี้ไม่มีปัญหาภาวะร่วมเส้นตรง (Multicollinearity) ที่รุนแรง

3.4. บทสรุป

เมื่อเข้าใจความสัมพันธ์เบื้องต้นและตรวจสอบความเหมาะสมของตัวแปรแล้ว
ขั้นตอนต่อไปคือการวิเคราะห์ตัวแปรแต่ละตัวอย่างละเอียดโดยใช้เทคนิค Weight of Evidence (WoE)
เพื่อวัดความสามารถในการจำแนกกลุ่มลูกค้าดีและไม่ดีออกจากกัน

4.0 การวิเคราะห์และคัดเลือกตัวแปรด้วย Weight of Evidence (WoE)

4.1. แนวคิดเบื้องต้น

Weight of Evidence (WoE) เป็นเทคนิคที่ใช้ในการแปลงค่าของตัวแปร (โดยเฉพาะตัวแปรกลุ่ม) ให้อยู่ในรูปของค่าตัวเลขที่สะท้อนถึงระดับความเสี่ยงที่เกี่ยวข้องกับการผิดนัดชำระหนี้ ส่วน Information Value (IV) เป็นตัวชี้วัดที่คำนวณจาก WoE เพื่อประเมินพลังในการพยากรณ์ (Predictive Power) ของตัวแปรนั้นๆ

4.2. เกณฑ์การประเมิน

เกณฑ์ที่ใช้ในการแปลความหมายของค่า IV และ KS Statistic มีดังนี้

เกณฑ์การแปลความหมาย Information Value (IV)

ค่า IV	การแปลผล
< 0.02	ไม่มีอำนาจในการพยากรณ์
0.02 - 0.1	อำนาจการพยากรณ์ต่ำ
0.1 - 0.3	อำนาจการพยากรณ์ปานกลาง
0.3 - 0.5	อำนาจการพยากรณ์สูง
> 0.5	สูงผิดปกติ (อาจเกิด Data Leakage)

เกณฑ์การแปลความหมาย KS Statistic

ค่า KS Statistic	การแปลผล
< 20%	ความสามารถในการแยกแยะต่ำมาก
20% - 40%	ความสามารถในการแยกแยะปานกลาง
40% - 50%	ดีมาก (Good)
50% - 70%	ยอดเยี่ยม (Excellent)
> 70%	สูงผิดปกติ (อาจเกิด Overfitting)

4.3. ผลการวิเคราะห์ WoE และ IV ของแต่ละตัวแปร

ตัวแปร (Feature)	ค่า Information Value (IV)	ค่า KS Statistic	การแปลผลพลังการพยากรณ์
Loan_Percent_Income_bin	0.707	0.3516	สูงผิดปกติ (อาจเกิด Data Leakage)
Loan_Int_Rate	0.618	0.3040	สูงผิดปกติ (อาจเกิด Data Leakage)
income_bin	0.450	0.2559	อำนาจการพยากรณ์สูง
Home_Ownership	0.377	0.2904	อำนาจการพยากรณ์สูง
Cb_Default_On_File	0.164	0.1652	อำนาจการพยากรณ์ปานกลาง
Loan_Intent	0.096	0.1447	อำนาจการพยากรณ์ต่ำ
Loan_Amnt_bin	0.073	0.1063	อำนาจการพยากรณ์ต่ำ
Emp_Length_bin	0.061	0.1076	อำนาจการพยากรณ์ต่ำ
age_bin	0.005	0.0338	ไม่มีอำนาจในการพยากรณ์

4.4. การคัดเลือกตัวแปรสำหรับสร้างแบบจำลอง

จากการวิเคราะห์ค่า IV
ได้มีการคัดเลือกตัวแปรที่มีพลังในการพยากรณ์ตั้งแต่ระดับปานกลางถึงสูงมาใช้ในการสร้างแบบจำลอง อย่างไรก็ตาม, ตัวแปร Loan_Int_Rate และ Loan_Percent_Income แม้จะมีค่า IV สูงมาก แต่ถูก ตัดออก เนื่องจากอาจก่อให้เกิดปัญหา Data Leakage เพราะในทางปฏิบัติ ค่าเหล่านี้มักจะถูกกำหนด หลังจากการประเมินความเสี่ยงของลูกค้าแล้ว

ดังนั้น, ตัวแปรที่ถูกคัดเลือกเพื่อนำไปสร้างแบบจำลอง ได้แก่:

- Home_Ownership_woe
- Income_woe
- Emp_Length_woe
- Loan_Amnt_woe

4.5. บทสรุป

เมื่อได้ทำการแปลงค่าและคัดเลือกตัวแปรที่เหมาะสมที่สุดผ่านการวิเคราะห์ด้วย WoE แล้ว
ขั้นตอนถัดไปคือการนำตัวแปรเหล่านี้ไปสร้างแบบจำลอง Logistic Regression
เพื่อทำนายความน่าจะเป็นของการผิดนัดชำระหนี้

5.0 การพัฒนาแบบจำลอง Logistic Regression

5.1. กระบวนการสร้างแบบจำลอง

ในส่วนนี้เป็นการสร้างแบบจำลอง Logistic Regression โดยนำตัวแปรที่ผ่านการแปลงค่าเป็น WoE มาใช้เป็นตัวแปรอิสระ (Independent Variables) เพื่อทำนายตัวแปรตาม (Dependent Variable) ซึ่งก็คือ Loan_Status (สถานะการผิดนัดชำระหนี้)

5.2. ขั้นตอนการสร้างแบบจำลอง

- **การแบ่งข้อมูล:** ข้อมูลถูกแบ่งออกเป็น 2 ส่วน คือ ชุดข้อมูลสำหรับฝึก (Train Set) 80% และ ชุดข้อมูลสำหรับทดสอบ (Test Set) 20% โดยกำหนด random_state=42 เพื่อให้สามารถทำซ้ำผลลัพธ์ได้
- **การสร้างแบบจำลอง:** ใช้ไลบรารี statsmodels ในการสร้างและฝึกแบบจำลอง Logit ซึ่งเป็นรูปแบบหนึ่งของ Logistic Regression
- **ผลลัพธ์ของแบบจำลอง:** สรุปผลลัพธ์ทางสถิติของแบบจำลองได้ดังตารางต่อไปนี้

ตัวแปร	coef	std err	z	P>|z|
const	-1.3031	0.017	-75.582	0.000
Home_Ownership_woe	-0.8540	0.028	-30.414	0.000
Income_woe	-1.2511	0.028	-45.191	0.000
Emp_Length_woe	-0.2201	0.070	-3.161	0.002
Loan_Amnt_woe	-2.5574	0.069	-37.241	0.000

5.3. การประเมินนัยสำคัญทางสถิติ

จากตารางผลลัพธ์ จะเห็นว่าค่า P>|z| ของตัวแปรทุกตัว (รวมถึงค่าคงที่ const) มีค่าน้อยกว่า 0.05 ซึ่งหมายความว่าตัวแปรทั้งหมดที่ถูกคัดเลือกมามีความสัมพันธ์กับสถานะสินเชื่อ (Loan_Status) อย่างมีนัยสำคัญทางสถิติ และสามารถนำไปใช้ในการทำนายได้

5.4. บทสรุป

หลังจากสร้างแบบจำลองและตรวจสอบนัยสำคัญทางสถิติของตัวแปรแล้ว
ขั้นตอนต่อไปคือการประเมินประสิทธิภาพของแบบจำลองโดยใช้ชุดข้อมูลทดสอบ (Test Set) เพื่อวัดความสามารถในการทำนายกับข้อมูลที่ไม่เคยเห็นมาก่อน

6.0 การประเมินประสิทธิภาพและตรวจสอบความถูกต้องของแบบจำลอง

6.1. ความสำคัญของการประเมินประสิทธิภาพ

การประเมินประสิทธิภาพแบบจำลอง (Model Validation)
เป็นขั้นตอนที่จำเป็นเพื่อวัดความสามารถในการทำนายของแบบจำลองกับข้อมูลใหม่ที่ไม่เคยใช้ในการฝึก (Test Set) เพื่อให้มั่นใจว่าแบบจำลองไม่ได้เกิดภาวะ Overfitting และสามารถนำไปใช้งานได้จริง

6.2. ตัวชี้วัดประสิทธิภาพหลัก

- **AUC Score:** 0.7614
- **Gini Coefficient:** 0.5229

ค่า AUC (Area Under Curve) ที่ 0.7614 แสดงให้เห็นว่าแบบจำลองมีความสามารถในการจำแนกกลุ่มลูกหนี้ที่ดี (Good) และไม่ดี (Bad) ได้อย่างถูกต้องในระดับที่ดี (Good Predictive Power)

6.3. การตีความสัมประสิทธิ์ของแบบจำลอง (Model Coefficients)

ค่าสัมประสิทธิ์ (Coefficient) ของตัวแปร WoE แต่ละตัวบ่งบอกถึงน้ำหนักและความสำคัญของปัจจัยนั้นๆ ที่มีต่อความเสี่ยง

ปัจจัย	ค่าสัมประสิทธิ์ (Coef)	การตีความ
Loan_Amnt_woe	-2.5574	ปัจจัยด้านวงเงินกู้มีอิทธิพลต่อความเสี่ยงมากที่สุด
Income_woe	-1.2511	ปัจจัยด้านรายได้มีอิทธิพลเป็นอันดับสอง
Home_Ownership_woe	-0.8540	สถานะการถือครองที่อยู่อาศัยเป็นปัจจัยสำคัญลำดับถัดมา
Emp_Length_woe	-0.2201	อายุงานเป็นปัจจัยที่มีนัยสำคัญทางสถิติ แต่มีอิทธิพลน้อยที่สุดในกลุ่มนี้

จากการเปรียบเทียบขนาดของค่าสัมประสิทธิ์ (โดยไม่คำนึงถึงเครื่องหมาย)
สามารถเรียงลำดับความสำคัญของปัจจัยที่ส่งผลต่อความเสี่ยงจากมากไปน้อยได้ดังนี้: วงเงินกู้ > รายได้ > การถือครองที่อยู่อาศัย > อายุงาน

6.4. บทสรุป

เมื่อแบบจำลองได้รับการพัฒนาและตรวจสอบประสิทธิภาพจนเป็นที่น่าพอใจแล้ว
ขั้นตอนสุดท้ายคือการนำผลลัพธ์ทางสถิติมาแปลงเป็นเครื่องมือที่ใช้งานได้จริงในทางธุรกิจ นั่นคือ Credit Scorecard

7.0 การนำไปใช้: การสร้าง Credit Scorecard

7.1. หลักการ

ในส่วนนี้จะเป็นการแปลงผลลัพธ์จากแบบจำลอง Logistic Regression
ที่มีความซับซ้อนให้กลายเป็นระบบการให้คะแนน (Scorecard)
ที่บุคลากรทั่วไปสามารถเข้าใจและนำไปใช้ประกอบการตัดสินใจอนุมัติสินเชื่อได้อย่างง่ายดายและรวดเร็ว

7.2. หลักการและค่าที่ใช้ในการคำนวณ

คะแนนจะถูกคำนวณโดยใช้ค่า Offset และ Factor ซึ่งได้มาจากการตั้งค่าเป้าหมายทางธุรกิจ
เพื่อปรับสเกลของคะแนนให้เหมาะสมกับการใช้งาน

- **TARGET_SCORE (คะแนนเป้าหมาย):** 600
- **TARGET_ODDS (อัตราต่อรองเป้าหมาย):** 50
- **PDO (Points to Double the Odds):** 20
- **Factor (คำนวณได้):** 28.85
- **Offset (คำนวณได้):** 487.12

7.3. Credit Scorecard ฉบับสมบูรณ์

จากการคำนวณตามหลักการข้างต้น ได้คะแนนพื้นฐานและตารางคะแนนสำหรับแต่ละคุณลักษณะดังนี้

คะแนนพื้นฐาน (Base Score): 525 คะแนน

ลูกค้าทุกคนจะเริ่มต้นด้วยคะแนนนี้ และจะถูกบวกหรือลบด้วยคะแนนจากคุณลักษณะต่างๆ ดังตารางด้านล่าง

Feature	Bin	WoE	Points
Home_Ownership	RENT	-0.50	-12
	OTHER	-0.47	-12
	MORTGAGE	0.66	16
	OWN	1.24	31
Income	< 35,000	-1.01	-36
	35,000 - 49,000	-0.09	-3
	49,000 - 63,000	0.23	8
	63,000 - 86,000	0.49	18
	> 86,000	1.02	37
Emp_Length	< 2	-0.32	-2
	2 - 4	-0.07	0
	4 - 6	0.12	1
	6 - 8	0.19	1
	8 - 10	0.30	2
	10 - 40	0.35	2
	> 40	0.00	0
	Missing	-0.50	-3
Loan_Amnt	> 14,500	-0.47	-35
	10,000 - 14,500	-0.00	0
	< 4,400	0.05	4

Feature	Bin	WoE	Points
	6,750 - 10,000	0.18	13
	4,400 - 6,750	0.29	21

ข้อสังเกต: คะแนนสำหรับ 'อายุงาน' (Emp_Length) มีค่าน้อย
ซึ่งสอดคล้องกับค่าสัมประสิทธิ์ในแบบจำลองที่ต่ำที่สุด แม้จะมีนัยสำคัญทางสถิติ
แต่ก็เป็นปัจจัยที่มีอิทธิพลน้อยที่สุดเมื่อเทียบกับตัวแปรอื่นๆ

7.4. สรุปรายงาน

รายงานฉบับนี้ได้นำเสนอขั้นตอนการพัฒนาแบบจำลองความเสี่ยงด้านสินเชื่ออย่างเป็นระบบ
ตั้งแต่การเตรียมข้อมูล การวิเคราะห์เชิงสำรวจ การคัดเลือกตัวแปรด้วยเทคนิค Weight of Evidence
การสร้างแบบจำลอง Logistic Regression ไปจนถึงการประเมินประสิทธิภาพและการสร้าง Credit Scorecard
ที่พร้อมใช้งาน
ซึ่งสามารถใช้เป็นเอกสารอ้างอิงสำหรับการตรวจสอบและทบทวนกระบวนการทางเทคนิคได้อย่างสมบูรณ์