



ธนาคารแห่งประเทศไทย
BANK OF THAILAND



สถาบันวิจัยเศรษฐกิจ
ป๋วย อึ๊งภากรณ์

Credit Risk Database

Credit Scoring Models for Thai SMEs

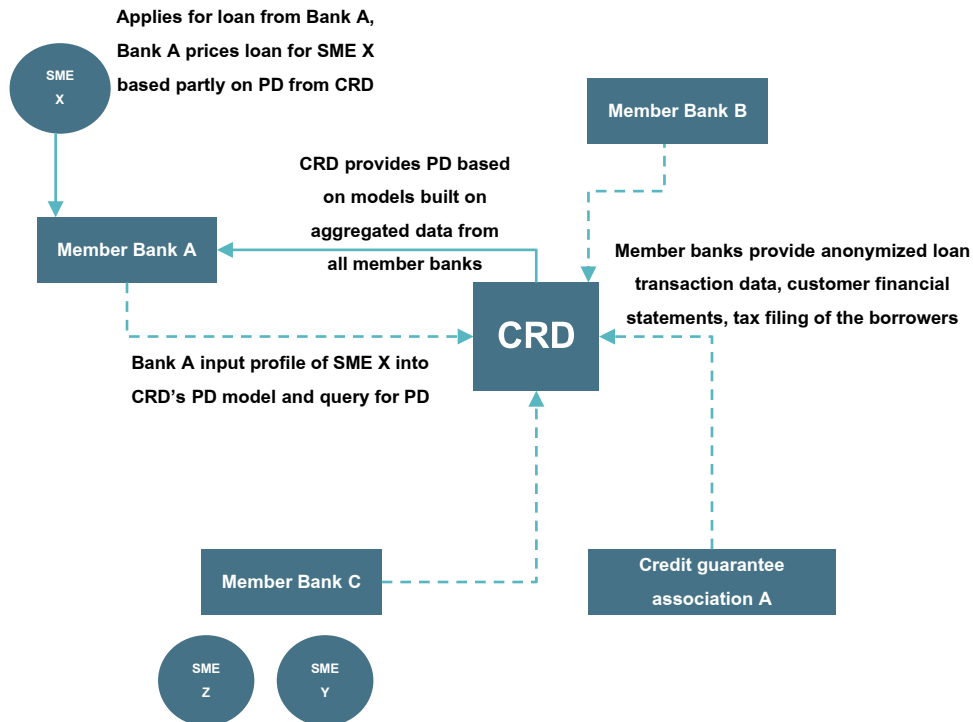
28 พฤษภาคม 2564

ธนาคารแห่งประเทศไทย





CRD in Japan and the Phillipines: A conceptual representation



Objectives:

- Help FIs and credit guarantee corps better assess risk of lending to SMEs
- Price the loans more accurately
- Reach deeper tiers of SMEs

Input into CRD from FIs:

- Anonymized financial, non-financial, and loan data

Output from CRD to FIs:

- PD models or PD based on aggregated data from all member FIs

A stylized PD model

Probability of Default

$$= \beta_1 \frac{\text{Total Equity}}{\text{Total Asset}} + \beta_2 \frac{\text{Gross Profit}}{\text{Sales}} + \beta_3 \text{Total no. of employees} + \beta_4 \text{Size of Loan} + \beta_5 \text{Business Sector} + \beta_6 \text{Location}..$$

วัตถุประสงค์ ของ CRD

1. ลดอุปสรรคการเข้าถึงแหล่งเงินทุนของ SMEs (**Information Asymmetry**)
2. สนับสนุนการให้สินเชื่อแบบ **Info-based Lending**
3. ส่งเสริมการค้าประกันสินเชื่อแบบ **Risk-based Pricing**

วัตถุประสงค์ของ CRD

	สถาบันการเงิน	บสย.	SMEs
Pain points	■ ข้อมูล: มีฐานข้อมูลจำกัด และ bias (แต่ละ สง. มีเฉพาะข้อมูลลูกค้าตนเอง)	■ รูปแบบธุรกิจ: ปัญหา Adverse Selection ของโครงการค้ำประกันแบบ Indirect Guarantee	■ ไม่สามารถเข้าถึงสินเชื่อในระบบได้ด้วยราคาที่เป็นธรรม
การมี CRD	✓ มี Credit scoring model ที่สร้างจากฐานข้อมูลที่ครอบคลุมทั้ง SMEs ที่มีและไม่มีประวัติใน NCB เพื่อใช้ประกอบกับ credit scoring model ที่มีอยู่แล้ว	✓ มี Credit scoring model ที่สร้างจากฐานข้อมูลที่ครอบคลุมทั้ง SMEs ที่มีและไม่มีประวัติใน NCB เพื่อใช้ประกอบกับ credit scoring model ที่มีอยู่แล้ว ✓ สนับสนุนการทำ direct guarantee	✓ สามารถเข้าถึงสินเชื่อได้มากขึ้น โดยเฉพาะสินเชื่อที่ไม่มีหลักประกัน ✓ อัตราดอกเบี้ยเหมาะสมกับความเสี่ยงของ SME นั้นๆเอง

Dataset





ธนาคารแห่งประเทศไทย
BANK OF THAILAND

Overview & mapping of sandbox data (data from beginning to Dec 2018)

Loan Data (Ordinary & Juristic)

Loan Data (Juristic-
only)

LAR+SMD & CPFS
(Juristic)

Financial Statement Data
(CPFS)

Data Mapping Definition

ข้อมูลนิติบุคคลที่มีสินเชื่อ รวมถึง
นำส่งงบการเงินให้กรมพัฒนาธุรกิจ
การค้าอย่างครบถ้วน ณ สิ้นปี

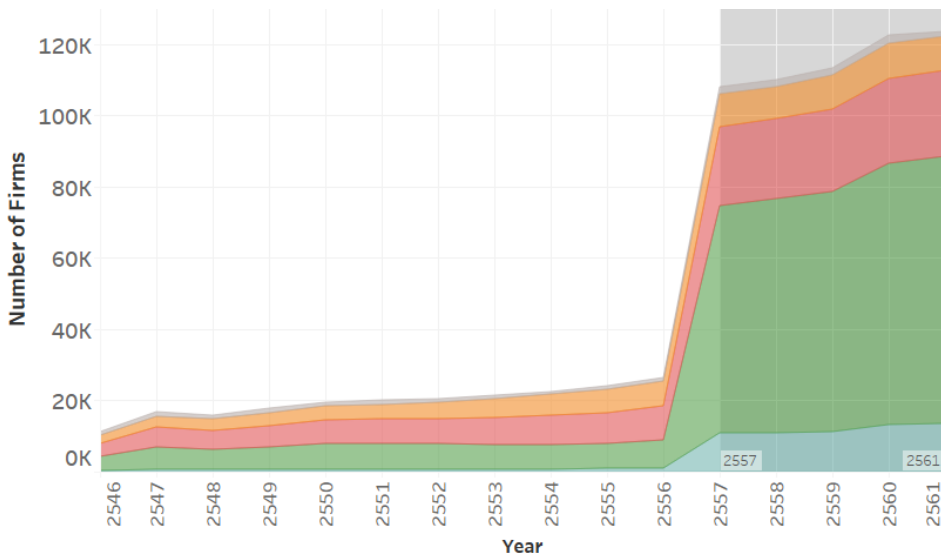
Data Sources	# Observations				
	2014	2015	2016	2017	2018
Loan Data Ordinary & Juristic (LAR+SMD)	473K	473K	489K	1.25M	1.47M
Loan Data (Juristic-only)	127K	130K	132K	146K	150K
Financial Statement Data (CPFS)	447K	458K	486K	505K	536K
Juristic with both datasets	108K	109K	112K	122K	123K

mapped results



Number of Firms with Loans and Financial Reports by Firm Sizes

จำนวนนิติบุคคลขนาดต่างๆ ที่มีสินเชื่อและนำส่งงบการเงินครบถ้วน



SMEs Size

Large

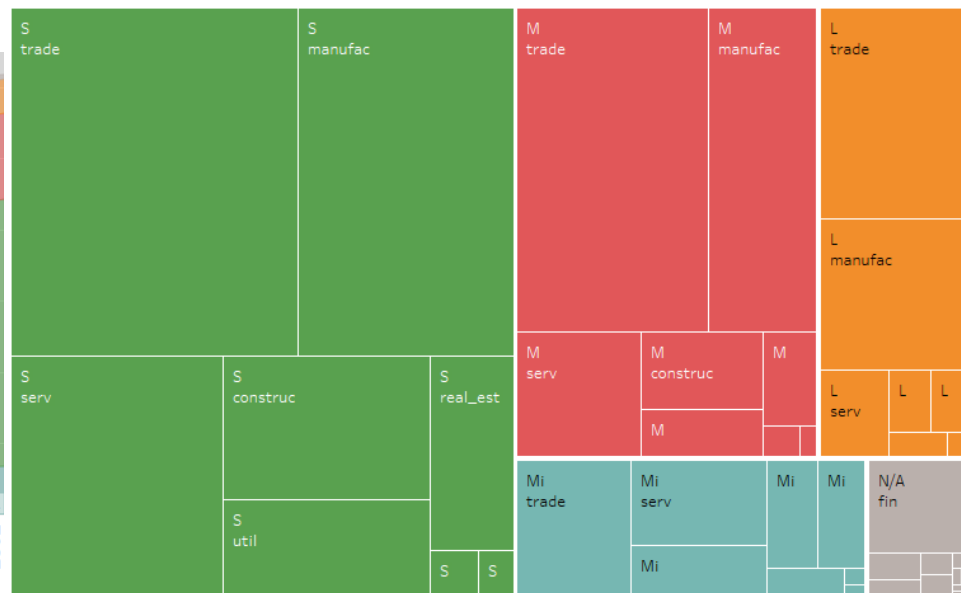
Medium

Small

Micro

N/A

จำนวนนิติบุคคลขนาดต่างๆ ที่มีสินเชื่อและนำส่งงบการเงินครบถ้วน ณ สิ้นปี 2561





Step 1

กำหนด “ภาคธุรกิจ”
จากหมวด (หลักแรก)
ของ ISIC BOT Rev 4

ISIC เามาจากหลายแหล่ง เรียงตาม priority คือ

- ข้อมูลงบการเงิน (CPFS): ISIC จากกิจกรรมทางเศรษฐกิจที่ก่อให้เกิดรายได้คิดเป็นสัดส่วนที่มากที่สุดของปีงบการเงิน
- ข้อมูล LAR: ISIC ของกิจกรรมที่มียอดวงเงินสินเชื่อสูงสุด

- หมวด 'A' ถึง 'E' เป็นภาคการผลิต
- หมวด 'G' ถึง 'S' เป็นภาคการค้าและบริการ
- หมวดนอกเหนือจากนี้เป็น NULL

Step 2

ใช้ “รายได้” เป็นตัวแบ่งหลัก
และใช้ “การจ้างงาน” เป็น
ตัวแบ่งต่อไป (หากไม่มี
ข้อมูลรายได้)

ใช้จากข้อมูลงบการเงิน (CPFS)

กรณีไม่มีข้อมูล
รายได้ใน CPFS

ใช้ข้อมูลรายได้จากฐานข้อมูล
SMP (SME Profile) แทน

กรณีไม่มีข้อมูลรายได้ใน
CPFS และ SMP

ใช้ข้อมูลการจ้างงานจากฐานข้อมูล
SMP (SME Profile) แทน

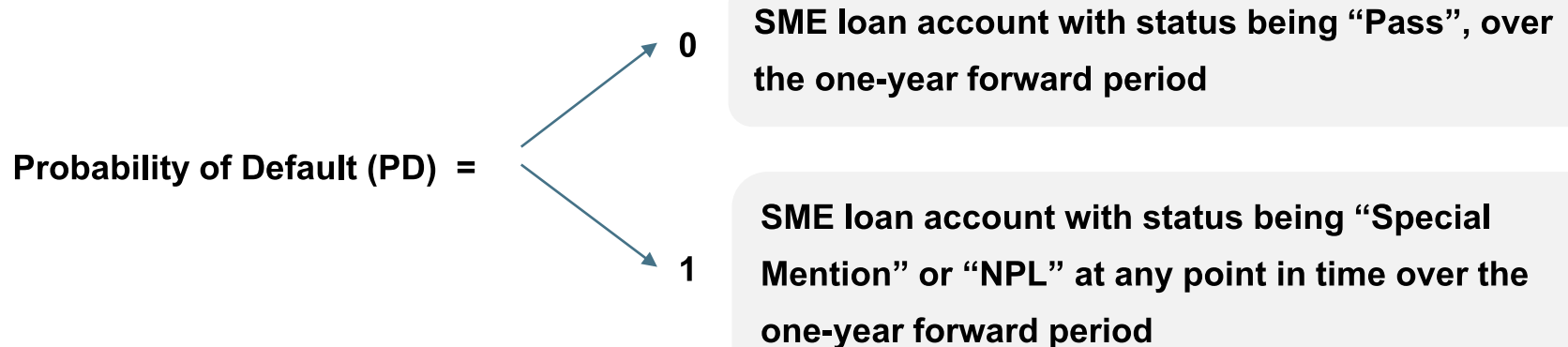
ภาคธุรกิจ	วิสาหกิจขนาดย่อม (Small)				วิสาหกิจขนาดกลาง (Medium)	
	วิสาหกิจขนาดย่อย (Micro)		วิสาหกิจขนาดย่อม			
	การจ้างงาน	รายได้	การจ้างงาน	รายได้	การจ้างงาน	รายได้
ภาคการผลิต	ไม่เกิน 5 คน	ไม่เกิน 1.8 ลบ	ไม่เกิน 50 คน	ไม่เกิน 100 ลบ.	ไม่เกิน 200 คน	ไม่เกิน 500 ลบ
ภาคการค้าและบริการ	ไม่เกิน 5 คน	ไม่เกิน 1.8 ลบ	ไม่เกิน 30 คน	ไม่เกิน 50 ลบ.	ไม่เกิน 100 คน	ไม่เกิน 300 ลบ

CRD Concept





One-year forward Probability of Default



With this definition, the number of total default observations are total of 42,958 obs (10.71%)



รูปแบบในการจัดทำ CRD ด้าน Technical



Dependent Variable: Loan Status

- ณ วันปิดรอบบัญชีทางการเงิน (CRD Japan)
- ภายในช่วง 12 เดือนข้างหน้า (BOT's credit score model)
- ณ 12 เดือนข้างหน้า (BIS the 9th IFC Conference paper)

Explanatory Variables

Repayment Behaviors
(Default Related data)



Financial Statement



Loan Usage Behaviors



Profiles
(Non-Financial Data)



Selection Criteria



Selection Methods



Training Dataset



Testing Dataset



Models

Estimating Parameters



Model Prediction Result (Credit Score และ PD)

Model Evaluation



CPFS Data

1. CPFS Data – เช็คความสมบูรณ์ของข้อมูล

- มีข้อมูลครบทุก Item
- Cross check ความถูกต้องของตัวเลข เช่น
 - $\text{Asset} = \text{Liability} + \text{Equity}$
 - $\text{Total Current Asset} = \text{Cash} + \text{Short-term Investment} + \text{Account Receivable} + \text{short-term Loans} + \text{Inventory} + \text{Other Current Asset}$
 - $\text{Total debt} = \text{Short-term debt} + \text{Long-term debt}$

Financial Ratio

2. Financial Ratio

Step 1: เช็คความสมเหตุสมผลของ financial ratio เช่น

- ROA และ ROE ควรมีค่าระหว่าง -100 ถึง 100
- Account Receivable Turnover Days ควรมีค่าระหว่าง 0 ถึง 365 วัน
- Years of Debt Redemption ควรมีค่าระหว่าง 0 ถึง 30 ปี

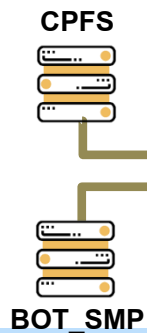
Step 2: แทนค่า Outlier ด้วยค่าที่เหมาะสม



Example of explanatory variables from sandbox data



ประเภทข้อมูล	Financial Statement Data	Default Related Data	Non-Financial Data
ตัวอย่าง	- Balance sheet statement - Income statement - Financial ratio เช่น Profitability ratio Solvency ratio , Asset utilization ratio	- วงเงิน/ยอดคงค้างสินเชื่อในแต่ละงวด - ประวัติการผิดนัดชำระหนี้: SM,NPL	- ประเภทธุรกิจ - ที่ตั้งสถานประกอบการ - จำนวนพนักงาน - ปีที่จดทะเบียนนิติบุคคล
ที่มาของข้อมูล	CPFS	BOT_LAR, BOT_SMD	BOT_SMP, CPFS



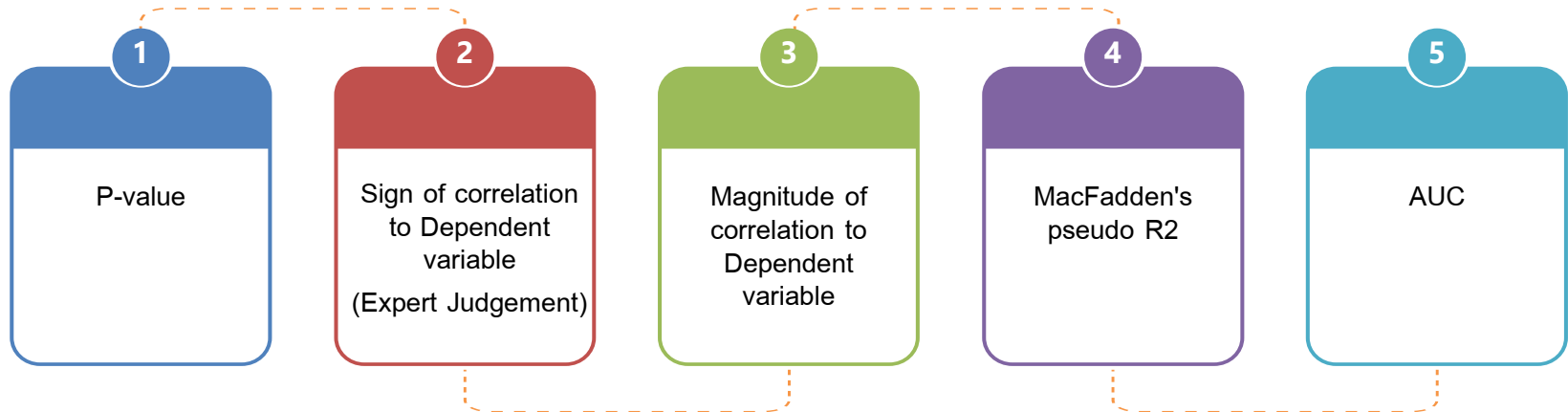
Up to 34 Financial Ratios





Transformation selection criteria

- เลือกรูปแบบของ Independent variable (Level vs Log form)
- Run Single logit regression และ ใช้ 5 criteria ในการเลือก ตามลำดับ



variables	P-value		Sign (Corr to Y)		Magnitude (Corr to Y)		MacFadden's pseudo R2		AUC		Winner
	level	ln	level	ln	level	ln	level	ln	level	ln	
ROA	0.00	0.00	-	0.06	-	0.09	0.00	0.02	0.61	0.61	ln
ROE	0.00	0.00	-	0.02	-	0.05	0.00	0.01	0.56	0.56	ln
Operating margin	0.00	0.00	-	0.07	-	0.07	0.00	0.01	0.50	0.54	level



- **Application score (A-score)** มักประกอบไปด้วยตัวแปรจาก Loan application หากเป็นนิติบุคคล จะรวมข้อมูลจากงบการเงินด้วย และข้อมูล repayment behavior ซึ่งอาจจะเป็นข้อมูลภายในหรือข้อมูลจาก NCB
- สำหรับ **Credit Scoring Model** แรกของ CRD project จัดทำขึ้นเพื่อรองรับกลุ่มลูกค้า SMES นิติบุคคล ที่ไม่มี financial footprint ซึ่งประกอบไปด้วยตัวแปรที่มีส่วนคล้ายคลึงกับ A-score เช่น ข้อมูลบริษัทและงบการเงิน (แต่จะไม่ใช้ข้อมูล repayment behavior ในการพิจารณา)
- ทั้งนี้ ข้อจำกัดของ Credit Scoring Model ที่ใช้ในแต่ละสถาบันการเงิน มีข้อจำกัด คือ มีข้อมูลเฉพาะลูกค้าของตนเท่านั้น แต่ CRD project จะมีข้อมูลของธนาคารพาณิชย์ทุกแห่ง รวมถึงสถาบันการเงินเฉพาะกิจบางแห่ง



ธนาคารแห่งประเทศไทย

Traditional NCB Data

ของบริษัท A

“พฤติกรรมการใช้สินเชื่อของบริษัท A”

A's Credit History

(สำหรับบริษัทที่มีสินเชื่อในระบบ)

Loan 1

- ประวัติการชำระเงิน

Loan 2

- ประวัติการชำระเงิน

NCB Score

ของบริษัท A

“พฤติกรรมการใช้สินเชื่อของบริษัท A
สามารถแปลงเป็น score ได้ xxx คะแนน”

A's Credit History จาก

NCB

(สำหรับบริษัทที่มีสินเชื่อในระบบ)

Loan 1

- ประวัติการชำระเงิน

Loan 2

- ประวัติการชำระเงิน

Credit History ของบริษัท

ที่มีสินเชื่อในระบบ

(บริษัท A,B,C,...)

จากถึงข้อมูล NCB

Score ของบริษัท A

PD Score vs NCB score

CRD (PD Score)

ของบริษัท ก.

“PD ของบริษัท ก. จะเท่ากับ PD โดยเฉลี่ยของ
ลูกหนี้ในฐานข้อมูลที่มี Profile แบบเดียวกัน”

ข้อมูล ของ บริษัท ก.

(สง. ต้องหาข้อมูลเอง)

Input

Model

Credit History ของบริษัท

มีสินเชื่อในระบบ

(บริษัท A,B,C,...)

จากถึงข้อมูล ธปท.

ข้อมูลอื่น ๆ เช่น

งบการเงิน Sector

(บริษัท A,B,C,...)

PD คำนวณจาก
ข้อมูลที่ สง. ใส่เข้ามา

PD (หรือ Score)

ของบริษัท ก.)

หมายเหตุ:

บริษัท A มีประวัติสินเชื่อในระบบ

บริษัท ก. ไม่เคยมีประวัติสินเชื่อในระบบ

Analytical Results

I. Logistic Regression & Scorecard





Logistic Regression Model (con vars)

Variable	Coefficient (significant)	Variable	Coefficient (significant)
(Intercept)	-0.679***	ln(Equity ratio)	-0.042***
ln(ROA)	-0.214***	ln(Asset turnover)	-0.517***
ln(ROE)	-0.036***	Cash Conversion Cycle	0.010***
ln(Ordinary Income Margin)	-0.029***	ln(Interest expense to sales)	0.042***
ln(Operating margin)	0.126***	ln(Cash to sales)	-0.351***
Net margin	-0.005***	ln(Working capital to debt)	-0.092***
ln(Sales growth)	-0.054***	Firm Age	-0.029***
ln(Asset growth)	-0.021**	ln(Authorized Share Capital)	-0.030***

**No
over & under
sampling**

AUC
74.69 %

(PD's cutoff
threshold = 0.11)



Introduction of Credit Scorecard from Logit Model

Binning & WOE Calculation

- นำแต่ละ variable ที่มีมาแบ่งกลุ่ม (Bin) ตามความเหมาะสม
- แยกจำนวน Default (Bad) และ Non-default (Good) data ในแต่ละ Bin
- Weight of Evidence (WOE) ในแต่ละ bin

$$WOE_i = \ln\left(\frac{\% Distn Good}{\% Distn Bad_i}\right)$$

Stepwise Logit Regression

- Linear Transform of Logit Model by WOE

$$\ln(Odds)_i = \beta_0 + \beta_1 WOE_{i1} + \dots + \beta_n WOE_{in} + \epsilon$$

- Run Stepwise Logit Regression เพื่อเลือก variables ที่ถูกแปลงเป็น WOE แล้ว

Credit Score Result

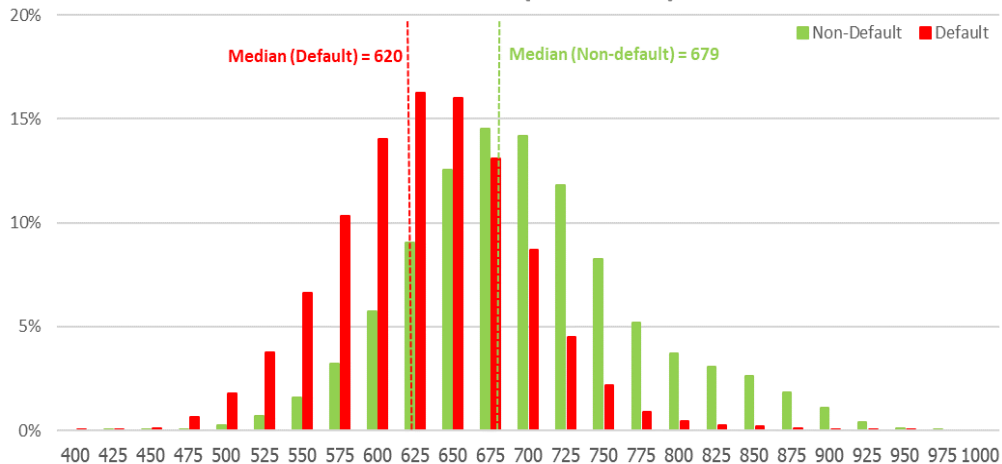
- นำผลการเลือก variables จาก stepwise regression มาคำนวณ Credit Score ตาม range ที่ต้องการ

$$\text{Total Score} = \text{Base Score} + \text{Contributions by Predictors}$$

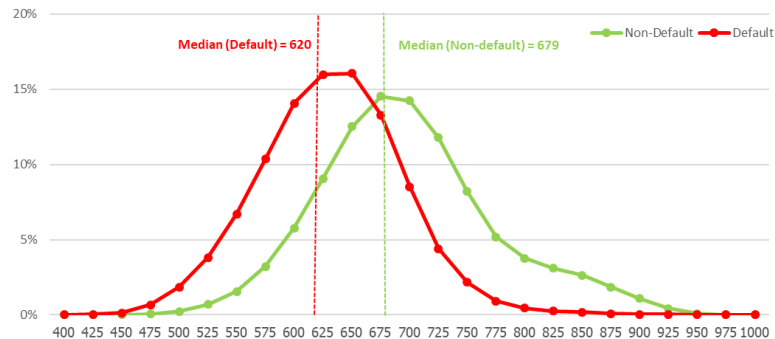


Credit Score Distribution (Good vs Bad)

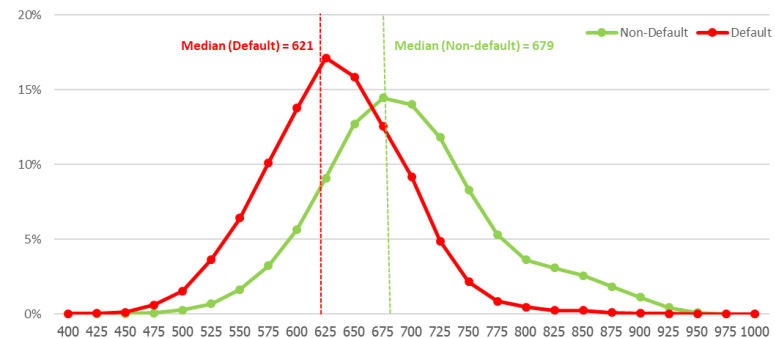
Credit Score (Full Data)



Credit Score (Trained Data)



Credit Score (Test Data)





Scorecard Threshold Score

Accuracy: Ability to differentiate Defaults & Non-defaults

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN}$$

Sensitivity: Ability to determine Defaulters

$$\text{Sensitivity} = \frac{TP}{TP+FN}$$

Specificity: Ability to determine Non-Defaulters

$$\text{Specificity} = \frac{TN}{TN+FP}$$

True positive (TP)

No. of Correctly identified Defaulters

False positive (FP)

No. of Incorrectly identified Defaulters

True negative (TN)

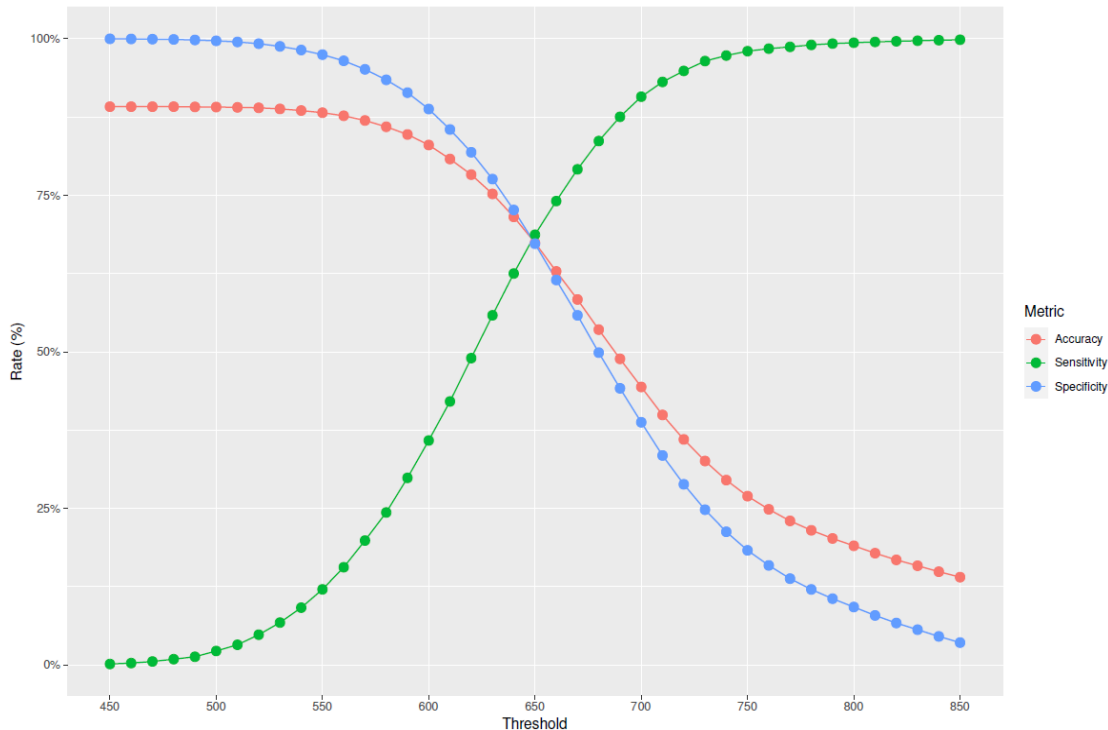
No. of Correctly identified Non-defaulters

False negative (FN)

No. of Incorrectly identified Non-defaulters

Variation of Logistic Classifier's Metrics by Threshold of Scorecard

Data Used: 100,249 Firm Loans in Test Set



Analytical Results

II. Tree-based model





Good

นิติบุคคลที่ไม่มีสินเชื่อสัญญาใดมี
สถานะเป็น SM หรือ NPL ตลอด
ระยะเวลา 1 ปีข้างหน้า

Bad

vice versa

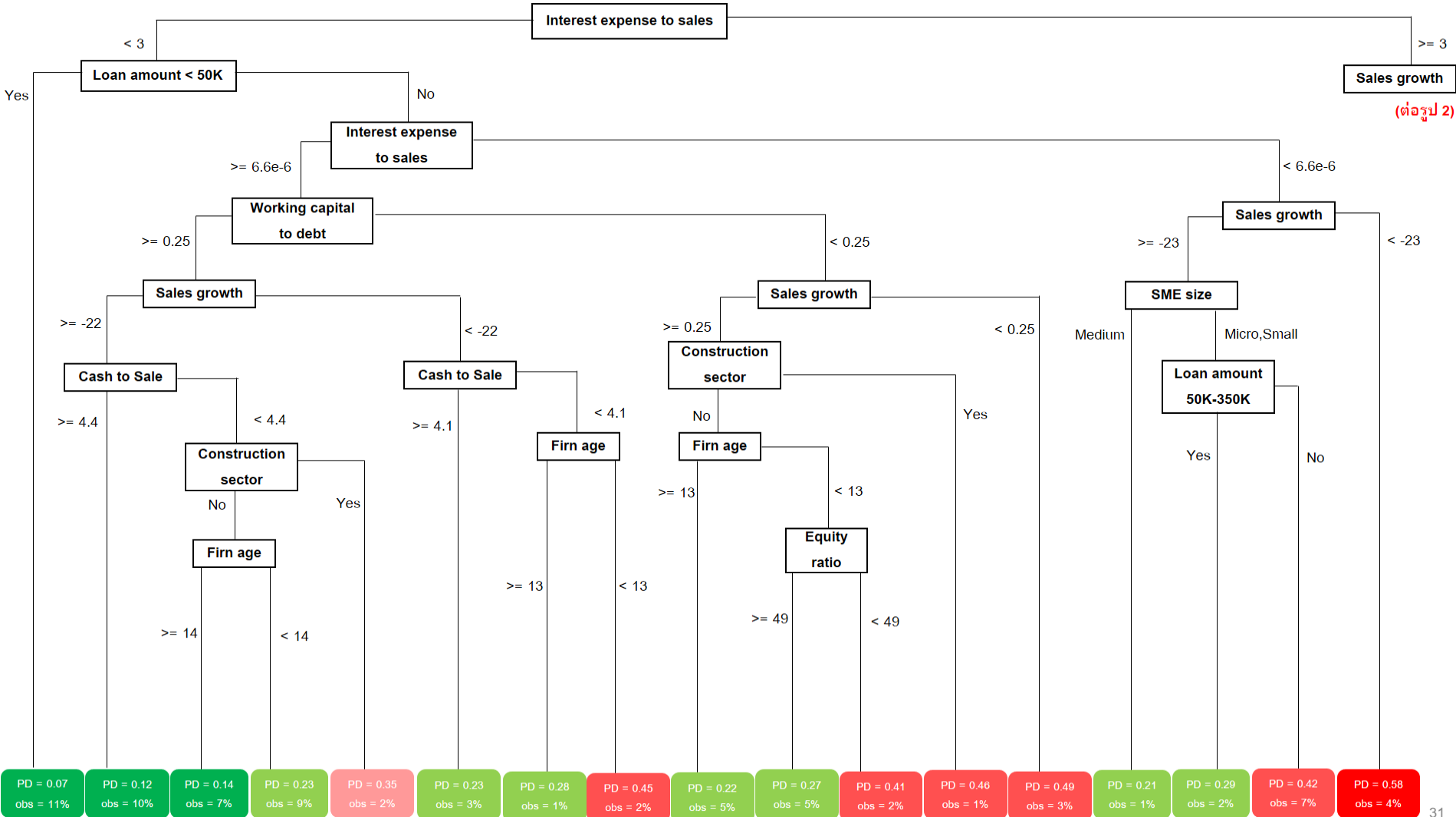
Default & Non-default Proportion



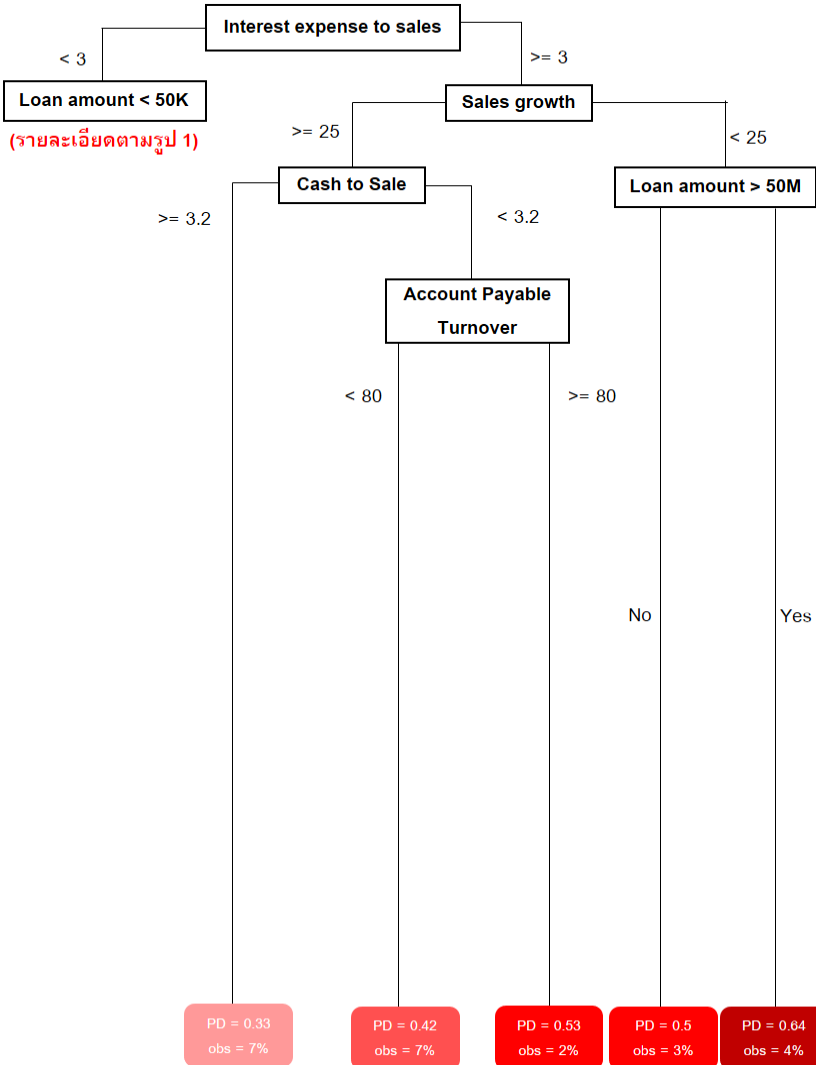
Imbalanced Data

Imbalanced Data

- เกิด imbalanced data ขึ้นระหว่างกลุ่ม Default และ Non-default เนื่องจากสัดส่วนระหว่าง minority และ majority แตกต่างกันมาก
- อาจก่อให้เกิดปัญหาการประมาณค่า PD ที่ต่ำกว่าความเป็นจริงขึ้นมา รวมถึง pattern ของกลุ่ม non-default จะถูกกลืนหายไป
- แก้ไข undersampling ข้อมูลจากกลุ่ม non-default ลงและ oversampling ข้อมูลจากกลุ่ม default ขึ้น



Decision Tree Model



Over & Under sampling

- Before: Default / Non-default obs = 1:9
- After: 3:7

Hyperparameter settings

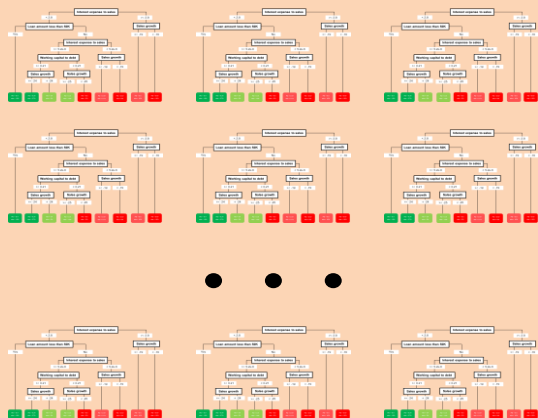
- min split = 18
- min bucket = 7
- cp = 0.001
- max depth = 8

AUC = 72.43 %

(PD's cutoff threshold = 0.3)



Concept



Ensemble Method

- Majority Voting
- Average

Final Result

Random Forest Model

Variable	Importance (node purity)
Interest expense to sales	3087.14
Sales growth	2971.31
Firm age	2136.88
Cash to sales	2068.83
Reserves ratio	1856.36
Total assets turnover Rate	1771.39
Working capital to debt	1753.08
Asset growth	1584.89
Account payable turnover	1532.72
Fixed asset turnover	1417.06



Random Forest Model

Variable	Importance (node purity)
Account receivable turnover	1337.75
Fixed assets to fixed liabilities and total equity	1295.51
SGA ratio	1284.38
Ordinary income Margin	1247.77
Net margin	1194.37
Quick ratio	1190.10
Equity ratio	1168.57
Current ratio	1165.05
Fixed assets to total equity	1158.13
Operating profit to total asset	1146.12

Over & Under sampling

- Before: Default / Non-default obs = 1:9
- After: 3:7

Hyperparameter settings

- Number of tree = 500
- mtry = 36 (num of vars = 147)
- node size = 2

AUC = 80.45 %

(PD's cutoff threshold = 0.21)

Summary





Comparison of models performance

Econometric Models

Model	AUC (%)
Logistic regression (continuous vars)	74.69
Logistic regression (weight of evidence)	74.57

นำไปทำ scorecard

Tree-based Models

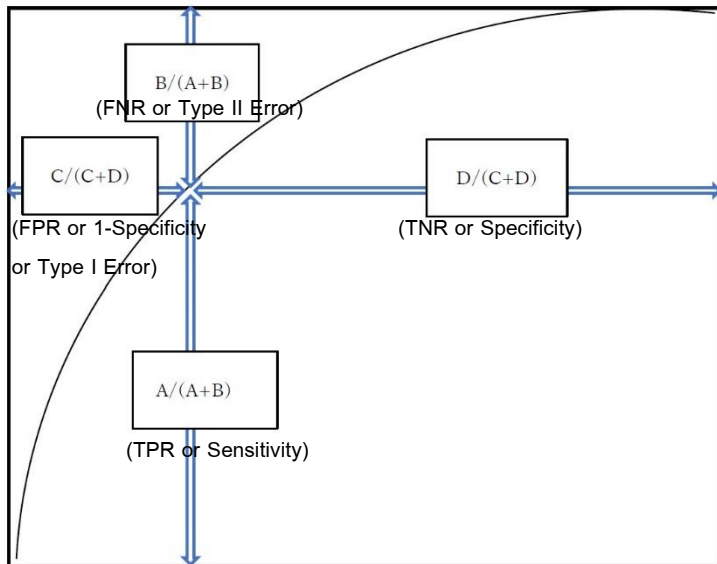
Model	AUC (%)
Decision tree (continuous vars)	72.43
Random forest (continuous vars)	80.45

นำไปทำ strategy curve
เพื่อสอบทาน scorecard



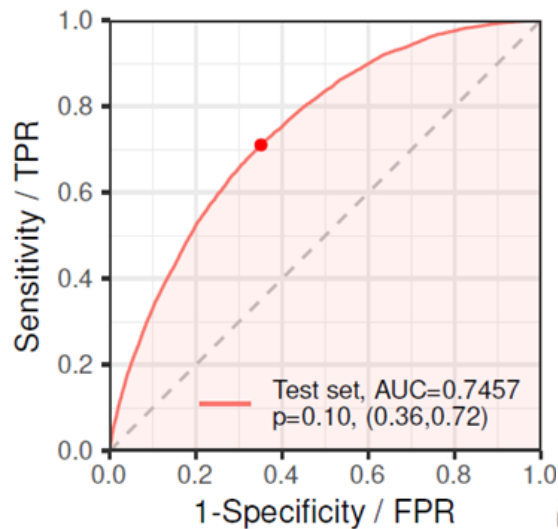
ธนาคารแห่งประเทศไทย
BANK OF THAILAND

ROC Curve Concept

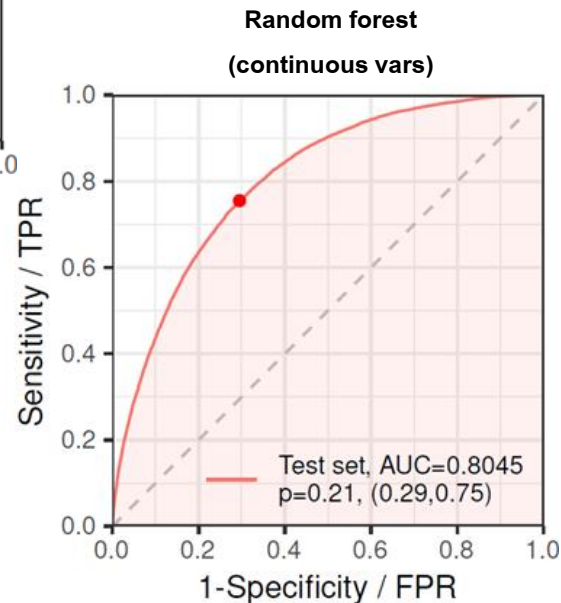


		Actual	
		Default	Non-Default
Predicted	Default	A (TP)	C (FP)
	Non-Default	B (FN)	D (TN)

Comparison of models performance (ROC Curve)



Logistic regression
(weight of evidence)



Q&A

