

Credit Card Approval Analysis and Predictive Modeling

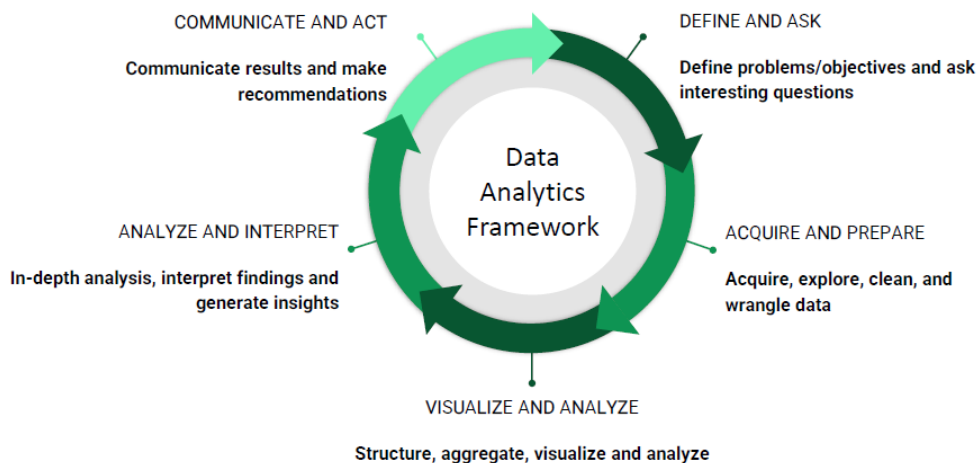
การวิเคราะห์และการสร้างแบบจำลองคาดการณ์การอนุมัติบัตรเครดิต

1. Introduction

สถาบันการเงินอนุมัติ/ปฏิเสธการสมัครบัตรเครดิต โดยพิจารณาจากข้อมูลผู้สมัคร โดยกระบวนการพิจารณาอาจใช้เวลานานและขึ้นอยู่กับดุลยพินิจของนักวิเคราะห์สินเชื่อ ซึ่งอาจนำไปสู่ความเสี่ยงหรือการสูญเสียโอกาสทางธุรกิจ การวิเคราะห์ปัจจัยที่สำคัญเพื่อค้นหาคุณลักษณะที่สำคัญของผู้สมัครมีความสัมพันธ์กับการอนุมัติมากที่สุด เพื่อลดความเสี่ยงทางการเงิน (อัตราการผิดนัดชำระหนี้ - Default Rate) และเพิ่มประสิทธิภาพในการดำเนินงาน (Operational Efficiency) ของฝ่ายอนุมัติสินเชื่อของธนาคารและสถาบันการเงิน (ฝ่ายบริหารความเสี่ยง, ฝ่ายการตลาด, ฝ่ายปฏิบัติการ)

2. Problem Statement/Background

การวิเคราะห์ปัจจัยที่สำคัญเพื่อค้นหาคุณลักษณะที่สำคัญของผู้สมัครมีความสัมพันธ์กับการอนุมัติมากที่สุด เพื่อลดความเสี่ยงทางการเงิน (อัตราการผิดนัดชำระหนี้ - Default Rate) โดยสนใจในคุณลักษณะใดของผู้สมัคร (เช่น อายุ, รายได้, ประเภทงาน) ที่มีความสัมพันธ์กับการอนุมัติมากที่สุด และเหมาะสมที่จะอนุมัติสินเชื่อมากที่สุด โดยแนวทางการวิเคราะห์จะดำเนินการตาม Data Analysis Framework ต่อไปนี้



2. Data Analytics Methodology

2.1 Analytical Questions

เพื่อค้นหาคุณลักษณะของผู้สมัคร (เช่น อายุ, รายได้, ประเภทงาน) ที่มีความสัมพันธ์กับการอนุมัติมากที่สุด และหาากลุ่มลูกค้า (Segments) ที่มีความเสี่ยงสูง/ต่ำ ที่แตกต่างกันชัดเจน

2.2. Data Sources/Attributes

Kaggle Dataset: "Credit Card Approvals (Clean Data)" (ซึ่งจำลองมาจากข้อมูลการสมัครบัตรเครดิต)

ข้อมูลในทางปฏิบัติ: สำนวนข้อมูลภายในของธนาคาร/สถาบันการเงิน (ข้อมูลการสมัคร, ประวัติสินเชื่อ, ข้อมูลเครดิตบูโร)

Gender	Age	Debt	Married	BankCustomer	Industry	Ethnicity	YearsEmployed	PriorDefault	Employed	CreditScore	DriversLicense	Citizen	ZipCode	Income	Approved
1	30.83	0	1	1	Industrials	White	1.25	1	1	1	0	ByBirth	202	0	1
0	58.67	4.46	1	1	Materials	Black	3.04	1	1	6	0	ByBirth	43	560	1
0	24.5	0.5	1	1	Materials	Black	1.5	1	0	0	0	ByBirth	280	824	1
1	27.83	1.54	1	1	Industrials	White	3.75	1	1	5	1	ByBirth	100	3	1
1	20.17	5.625	1	1	Industrials	White	1.71	1	0	0	0	ByOtherMeans	120	0	1
1	32.08	4	1	1	CommunicationServices	White	2.5	1	0	0	1	ByBirth	360	0	1
1	33.17	1.04	1	1	Transport	Black	6.5	1	0	0	1	ByBirth	164	31285	1
0	22.92	11.585	1	1	InformationTechnology	White	0.04	1	0	0	0	ByBirth	80	1349	1
1	54.42	0.5	0	0	Financials	Black	3.96	1	0	0	0	ByBirth	180	314	1
1	42.5	4.915	0	0	Industrials	White	3.165	1	0	0	1	ByBirth	52	1442	1
1	22.08	0.83	1	1	Energy	Black	2.165	0	0	0	1	ByBirth	128	0	1
1	29.92	1.835	1	1	Energy	Black	4.335	1	0	0	0	ByBirth	260	200	1
0	38.25	6	1	1	Financials	White	1	1	0	0	1	ByBirth	0	0	1
1	48.08	6.04	1	1	Financials	White	0.04	0	0	0	0	ByBirth	0	2690	1
0	45.83	10.5	1	1	Materials	White	5	1	1	7	1	ByBirth	0	0	1
1	36.67	4.415	0	0	Financials	White	0.25	1	1	10	1	ByBirth	320	0	1
1	28.25	0.875	1	1	CommunicationServices	White	0.96	1	1	3	1	ByBirth	396	0	1
0	23.25	5.875	1	1	Materials	White	3.17	1	1	10	0	ByBirth	120	245	1
1	21.83	0.25	1	1	Real Estate	Black	0.665	1	0	0	1	ByBirth	0	0	1

ภาพที่ 1 ตัวอย่าง Data Set

Attribute	Description	Data Type	Valid Range/Example
Gender	Gender	Nominal	0=Female, 1=Male
Age	Age (Years)	Interval	[20, 80]
Debt	Outstanding debt (Feature has been scaled)	Ratio (Continuous)	[0, Infinity)
Married	Married Status	Nominal	Married, 0=Single/Divorced/etc, 1=Married
BankCustomer	Bank Customer Status	Nominal	0=Does not have a bank account, 1=Has a bank account)
Industry	Industry (Job sector of current or most recent job)	Nominal	[0, 14]
Industry_Name	Industry Name	Nominal	[CommunicationServices, ConsumerDiscretionary, ConsumerStaples, Education, Energy, Financials, Healthcare, Industrials, InformationTechnology, Materials, Real Estate, Research, Transport, Utilities]
Ethnicity	Ethnicity	Nominal	White, Black, Asian, Latino, Other
YearsEmployed	Years employed	Ratio (Discrete)	[0, Infinity)
PriorDefault	Prior default	Asymmetric Binary	0=No prior defaults, 1=Prior default
Employed	Employed Status	Asymmetric Binary	0=Not employed, 1=Employed
CreditScore	Credit score (Feature has been scaled)	Interval	[0, 100]
DriversLicense	Drivers license	Symmetric Binary	0=No license, 1=Has license
Citizen	Citizenship	Nominal	ByBirth, ByOtherMeans, Temporary
ZipCode	ZipCode	Nominal	5 digit number
StateName	StateName	Nominal	[100 = Michigan, etc.]
Income	Income (Feature has been scaled)	Ratio (Continuous)	[0, Infinity)
Approved	Approved Status	Asymmetric Binary	0=Not approved, 1=Approved

ภาพที่ 2 ตัวอย่าง Data Dictionary

2.3 Data cleaning and preprocessing

เนื่องจากเป็น "Clean Data" จะเน้นการตรวจสอบขั้นสุดท้าย การจัดการกับค่าที่หายไป (Missing Values) ที่ยังอาจมีอยู่

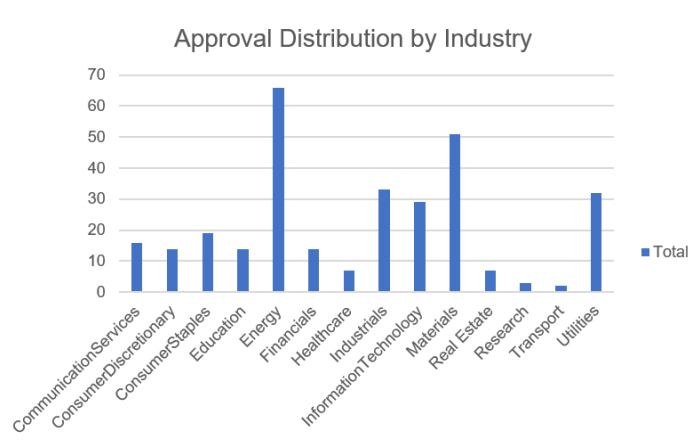
3. Analytics Methodology

3.1 EDA techniques: การวิเคราะห์ความถี่และค่าสถิติเชิงพรรณนา เพื่อทำความเข้าใจความสัมพันธ์ของแต่ละคุณลักษณะกับผล
การอนุมัติ

	Gender	Age	Debt	Married	BankCustomer	Industry	YearsEmployed	PriorDefault	Employed	CreditScore	DriversLicense	ZipCode	Income	Approved
count	690.000000	690.000000	690.000000	690.000000	690.000000	690.000000	690.000000	690.000000	690.000000	690.000000	690.000000	690.000000	690.000000	690.000000
mean	0.695652	31.514116	4.758725	0.760870	0.763768	6.960870	2.223406	0.523188	0.427536	2.400000	0.457971	180.547826	1017.385507	0.444928
std	0.460464	11.860245	4.978163	0.426862	0.425074	3.802822	3.346513	0.499824	0.495080	4.86294	0.498592	173.970323	5210.102598	0.497318
min	0.000000	13.750000	0.000000	0.000000	0.000000	1.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
25%	0.000000	22.670000	1.000000	1.000000	1.000000	3.000000	0.165000	0.000000	0.000000	0.000000	0.000000	60.000000	0.000000	0.000000
50%	1.000000	28.460000	2.750000	1.000000	1.000000	7.000000	1.000000	1.000000	0.000000	0.000000	0.000000	160.000000	5.000000	0.000000
75%	1.000000	37.707500	7.207500	1.000000	1.000000	10.000000	2.625000	1.000000	1.000000	3.000000	1.000000	272.000000	395.500000	1.000000
max	1.000000	80.250000	28.000000	1.000000	1.000000	14.000000	28.500000	1.000000	1.000000	67.000000	1.000000	2000.000000	100000.000000	1.000000

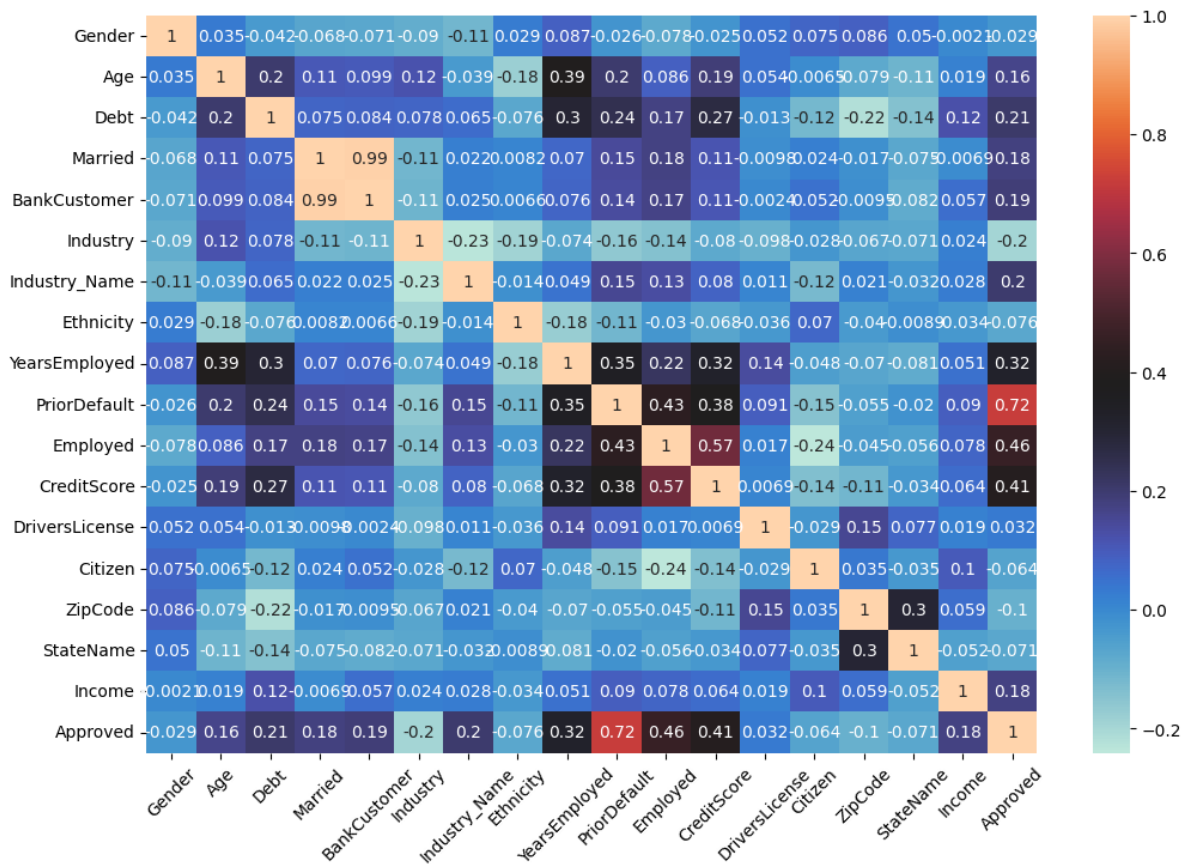
ภาพที่ 4 สถิติเชิงพรรณนา (Descriptive Statistics)

Industry Name	Number of Approved
CommunicationServices	16
ConsumerDiscretionary	14
ConsumerStaples	19
Education	14
Energy	66
Financials	14
Healthcare	7
Industrials	33
InformationTechnology	29
Materials	51
Real Estate	7
Research	3
Transport	2
Utilities	32
Grand Total	307



ภาพที่ 5 Pivot Table and Graph

3.2 Visualization strategy: Heatmap สำหรับ Correlation Matrix



ภาพที่ 6 Correlation Heatmap (แผนที่ความรื้อนสหสัมพันธ์)

ค่าสัมประสิทธิ์สหสัมพันธ์แบบเพียร์สัน (Pearson Correlation) ที่มีค่า มากกว่า 0.3 (ทั้งค่าบวกและลบ)
 หมายความว่ามีความสัมพันธ์เชิงเส้นระหว่างตัวแปรทั้งสองในระดับ ปานกลางถึงค่อนข้างมาก โดยมีค่าดังนี้

Married and BankCustomer = 0.992
PriorDefault and Approved = 0.720
Employed and CreditScore = 0.571
Employed and Approved = 0.458
PriorDefault and Employed = 0.432
CreditScore and Approved = 0.406
Age and YearsEmployed = 0.391
PriorDefault and CreditScore = 0.380
YearsEmployed and PriorDefault = 0.346
YearsEmployed and Approved = 0.322
YearsEmployed and CreditScore = 0.322
ZipCode and StateName = 0.302

ภาพที่ 7 ค่าสัมประสิทธิ์สหสัมพันธ์แบบเพียร์สัน (Pearson Correlation) ที่มีค่า มากกว่า 0.3

ความสัมพันธ์ที่โดดเด่น	ค่า Correlation	ความหมายเชิงธุรกิจ
PriorDefault and Approved	0.720	สำคัญที่สุด: ผู้สมัครที่มีประวัติ ไม่เคยผิดนัดชำระหนี้ (PriorDefault=0) มีโอกาสได้รับการอนุมัติสูงมาก (ความเสี่ยงต่ำ)
Employed and Approved	0.458	ผู้สมัครที่มี งานทำ (Employed=1) มีแนวโน้มได้รับการอนุมัติสูงกว่า
CreditScore and Approved	0.406	คะแนนเครดิตสูง เป็นปัจจัยหลักในการอนุมัติ
YearsEmployed and Approved	0.322	ระยะเวลาการทำงานที่ยาวนาน มีความสัมพันธ์กับการอนุมัติ ซึ่งบ่งชี้ถึงความมั่นคงทางการเงิน

ภาพที่ 8 ค่าสัมประสิทธิ์สหสัมพันธ์แบบเพียร์สัน (Pearson Correlation) ที่มีสัมพันธ์กับการอนุมัติสินเชื่อ

บทสรุปเชิงวิเคราะห์: ปัจจัยด้าน ประวัติความเสี่ยง (PriorDefault) และ ความมั่นคง (Employed, CreditScore, YearsEmployed) เป็นคุณลักษณะหลักที่ขับเคลื่อนผลการตัดสินใจอนุมัติบัตรเครดิต ซึ่งตอบคำถามเชิงวิเคราะห์ที่ว่า "คุณลักษณะใดของผู้สมัครมีความสัมพันธ์กับการอนุมัติมากที่สุด"

3.3 Segmentation approach: การจัดกลุ่มลูกค้าตามโปรไฟล์ความเสี่ยง (Risk Profiles)

การเปรียบเทียบ Credit Score (CreditScore) ระหว่างผู้ที่ได้รับการอนุมัติ (Approved=1) และผู้ที่ไม่ได้รับการอนุมัติ (Approved=0) สามารถทำ t-test ได้ เนื่องจาก CreditScore เป็นตัวแปรเชิงปริมาณ (Numerical/ Continuous) และเราต้องการเปรียบเทียบค่าเฉลี่ยของตัวแปรนี้ระหว่างสองกลุ่มอิสระ (Approved=1 และ Approved=0) การใช้ Independent Samples T-test (การทดสอบทีสำหรับกลุ่มตัวอย่างอิสระ) จึงเป็นวิธีที่เหมาะสมที่สุด

Independent Samples T-test

สมมติฐาน

H_0 (สมมติฐานว่าง): ค่าเฉลี่ย CreditScore ของผู้ที่ได้รับการอนุมัติ ไม่แตกต่าง จากผู้ที่ไม่ได้รับการอนุมัติ

H_a (สมมติฐานทางเลือก): ค่าเฉลี่ย CreditScore ของผู้ที่ได้รับการอนุมัติ แตกต่าง จากผู้ที่ไม่ได้รับการอนุมัติ

ค่าเฉลี่ย CreditScore กลุ่ม Approved (1): 4.6059

ค่าเฉลี่ย CreditScore กลุ่ม Not Approved (0): 0.6319

ค่า T-statistic: 10.6384

ค่า P-value: 0.0000

P-value < 0.05: ปฏิเสธ H_0 สรุปได้ว่า Credit Score เฉลี่ยแตกต่างกันอย่างมีนัยสำคัญ

ภาพที่ 9 Independent Samples T-test

ผลจากการวิเคราะห์ความสัมพันธ์ของพีเจอร์ (Correlation Analysis) เป็นไปตามสมมติฐานที่ว่า มีกลุ่มลูกค้า (Segments) ที่มีความเสี่ยงสูง/ต่ำ ที่แตกต่างกันชัดเจนหรือไม่ กล่าวโดยสรุปคือ ค่าเฉลี่ย Credit Score ของผู้ที่ได้รับการอนุมัติ แตกต่าง จากผู้ที่ไม่ได้รับการอนุมัติ

4. Findings and Business Insights

คุณลักษณะหลักที่ขับเคลื่อนผลการอนุมัติบัตรเครดิต (Feature Importance) คือ ปัจจัยด้าน ประวัติความเสี่ยง (PriorDefault) และ ความมั่นคง (Employed, CreditScore, YearsEmployed) โดยสามารถ ระบุกลุ่มลูกค้าที่มีความเสี่ยงสูง/ต่ำ ที่ชัดเจนได้โดย Credit Score ซึ่งเป็นปัจจัยขับเคลื่อนหลัก (Primary Driver) ในกระบวนการอนุมัติ ทำให้เกิดการแบ่งกลุ่มลูกค้าเป็นสองกลุ่มได้อย่างมีประสิทธิภาพ

ดังนั้นแนวโน้มและรูปแบบคุณลักษณะของผู้สมัครที่ถูกรับอนุมัติ/ปฏิเสธ คือ ประวัติความเสี่ยง (PriorDefault) และ ความมั่นคง (Employed, CreditScore, YearsEmployed) เป็นคุณลักษณะหลักที่ขับเคลื่อนผลการตัดสินใจอนุมัติบัตรเครดิต ซึ่งตอบคำถามเชิงวิเคราะห์ที่ว่า "คุณลักษณะใดของผู้สมัครมีความสัมพันธ์กับการอนุมัติมากที่สุด"

จากการเปรียบเทียบ Credit Score (CreditScore) ระหว่างผู้ที่ได้รับการอนุมัติ (Approved=1) และผู้ที่ไม่ได้รับการอนุมัติ (Approved=0) ค่าเฉลี่ย Credit Score ของผู้ที่ได้รับการอนุมัติ แตกต่าง จากผู้ที่ไม่ได้รับการอนุมัติ ซึ่งตอบคำถามเชิงวิเคราะห์ที่ว่า "มีกลุ่มลูกค้า (Segments) ที่มีความเสี่ยงสูง/ต่ำ ที่แตกต่างกันชัดเจนหรือไม่"

5. Conclusion and Recommendation

จากการวิเคราะห์ที่กล่าวมา จะทำให้เราทราบถึงปัจจัยสำคัญที่นำมาวิเคราะห์การอนุมัติสินเชื่อที่เหมาะสม ทำให้ลดการเกิดหนี้เสีย (ลดความเสี่ยง) เนื่องจากมีความแม่นยำในการคัดกรองมากขึ้น และเพิ่มโอกาสในการอนุมัติลูกค้าที่มีคุณภาพเร็วขึ้น (เพิ่มกำไร) และผู้สมัครได้รับผลการตัดสินใจรวดเร็วขึ้น ทำให้ประสบการณ์การใช้บริการดีขึ้น