

บทที่ 3

วิธีการดำเนินการวิจัย

การวิจัยนี้เป็นการวิจัยเชิงประยุกต์ (Applied Research) โดยใช้วิธีการวิเคราะห์ข้อมูลทุติยภูมิ (Secondary Data Analysis) และการใช้เทคนิคการเรียนรู้ของเครื่อง (Machine Learning) เพื่อพัฒนาโมเดลการทำนายความสำเร็จของสตาร์ทอัพ โดยใช้กระบวนการ Cross-industry Standard Process for Data Mining (CRISP-DM) ในการดำเนินงานวิจัยดังนี้

- 3.1 เครื่องมือที่ใช้ในการวิจัย
- 3.2 ข้อมูลที่ใช้ในการศึกษา
- 3.3 ประชากรและกลุ่มตัวอย่าง
- 3.4 ตัวแปรที่ใช้ในการวิจัย
 - 3.4.1 ตัวแปรต้น (Independent Variables)
 - 3.4.2 ตัวแปรเป้าหมาย (Target Variables)
- 3.5 ขั้นตอนการดำเนินการวิจัย
 - 3.5.1 การทำความเข้าใจธุรกิจ (Business Understanding)
 - 3.5.2 การทำความเข้าใจข้อมูล (Data Understanding)
 - 3.5.3 การเตรียมข้อมูล (Data Preparation)
 - 3.5.4 การสร้างโมเดล (Modeling)
 - 3.5.5 การวัดประสิทธิภาพของโมเดล (Evaluation)
 - 3.5.6 การนำโมเดลไปใช้งานจริง (Deployment)

โดยมีรายละเอียดดังนี้

3.1 เครื่องมือที่ใช้ในการวิจัย

ผู้วิจัยใช้เครื่องคอมพิวเตอร์ Macbook Air มีหน่วยประมวลผลกลาง Apple M1 หน่วยความจำ 8 GB และใช้ระบบปฏิบัติการ Sonoma 14.5 และใช้โปรแกรม Google Colab ด้วยภาษา Python

3.2 ข้อมูลที่ใช้ในการศึกษา

1. ข้อมูลสตาร์ทอัพรวบรวมจากฐานข้อมูล Crunchbase (<https://www.crunchbase.com/>)
2. ข้อมูลประสิทธิภาพเว็บไซต์ ได้จากการใช้เครื่องมือประเมินประสิทธิภาพเว็บไซต์ Lighthouse ใน Google Chrome (<https://developer.chrome.com/docs/lighthouse?hl=en>)
3. ข้อมูลลิงก์ย้อนกลับ ได้จากการใช้เครื่องมือตรวจสอบลิงก์ย้อนกลับจาก <https://ahrefs.com/backlink-checker/>

3.3 ประชากรและกลุ่มตัวอย่าง

ในการศึกษานี้ ประชากรที่ใช้ในการวิจัยคือ บริษัทสตาร์ทอัพที่ก่อตั้งขึ้นในช่วงระหว่างปี ค.ศ. 2016 ถึง ค.ศ. 2024 สำหรับการคัดเลือกกลุ่มตัวอย่าง ได้ดำเนินการโดยใช้วิธีการคัดเลือกแบบเฉพาะเจาะจง (Purposive Selection) โดยมีวัตถุประสงค์ในการเลือกสตาร์ทอัพที่มีสถานะ ‘Acquired’ (ถูกควบรวมหรือซื้อกิจการ) หรือ ‘Closed’ (ปิดกิจการ) ภายในปี ค.ศ. 2024 ทั้งนี้ กลุ่มตัวอย่างที่ได้ประกอบด้วยสตาร์ทอัพจำนวน 750 ราย แบ่งเป็นสตาร์ทอัพที่ประสบความสำเร็จ (สถานะ Acquired) จำนวน 500 ราย และสตาร์ทอัพที่ไม่ประสบความสำเร็จ (สถานะ Closed) จำนวน 250 ราย

3.4 ตัวแปรที่ใช้ในการวิจัย

3.4.1 ตัวแปรต้น (Independent Variables)

3.4.1.1 ข้อมูลพื้นฐานทางธุรกิจ ประกอบด้วย

1. จำนวนผู้ก่อตั้ง (Number of Founders)
2. จำนวนรอบการระดมทุน (Number of Funding Rounds)
3. มูลค่ารวมของเงินทุนในหลายๆ รอบการระดมทุน (Total Funding Amount)
4. จำนวนสิทธิบัตร (Patents Granted)
5. จำนวนเครื่องหมายการค้า (Trademarks Registered)

3.4.1.2 สื่อที่เป็นเจ้าของ ประกอบด้วย

1. จำนวนการเข้าชมเว็บไซต์ในแต่ละเดือน (Monthly Visits)
2. ระยะเวลาเฉลี่ยที่ผู้เข้าชมใช้บนเว็บไซต์ในแต่ละครั้ง (Visit Duration)
3. จำนวนหน้าที่ผู้ใช้งานเข้าชมต่อการเข้าชมหนึ่งครั้ง (Page Views/Visit)

4. อัตราการดึงออกจากเว็บไซต์หลังจากดูหน้าแรกเพียงหน้าเดียว (Bounce Rate)
5. ประสิทธิภาพของเว็บไซต์ (Performance)
6. การเข้าถึงเว็บไซต์ (Accessibility)
7. แนวทางปฏิบัติที่ดีของเว็บไซต์ (Best Practices)
8. การเพิ่มประสิทธิภาพเว็บไซต์เพื่อการค้นหา (SEO)

3.4.1.3 สื่อที่ได้รับ ประกอบด้วย

1. คะแนนความน่าเชื่อถือของโดเมนเว็บไซต์ (Domain Rating)
2. จำนวนลิงก์จากเว็บไซต์อื่นที่ชี้กลับมาที่เว็บไซต์ (Backlinks)
3. จำนวนโดเมนที่แตกต่างกันที่ลิงก์มายังเว็บไซต์ (Linking Websites)

3.4.2 ตัวแปรเป้าหมาย (Target Variables)

ตัวแปรเป้าหมาย จะใช้สถานะของสตาร์ตอัปในชุดข้อมูล โดยสตาร์ตอัปที่ประสบความสำเร็จจากสถานะ Acquired กำหนดให้มีป้ายกำกับเป็น 1 และสตาร์ตอัปที่ล้มเหลวจากสถานะ Closed กำหนดให้มีป้ายกำกับเป็น 0

3.5 ขั้นตอนการดำเนินการวิจัย

ขั้นตอนการดำเนินการวิจัยนี้ใช้กระบวนการ Cross-industry Standard Process for Data Mining (CRISP-DM) ซึ่งเป็นแนวทางมาตรฐานสำหรับการทำเหมืองข้อมูลที่ประกอบด้วยขั้นตอนสำคัญในการพัฒนาโมเดลการวิเคราะห์ข้อมูล ประกอบด้วย 6 ขั้นตอนดังนี้

3.5.1 การทำความเข้าใจธุรกิจ (Business Understanding)

การทำความเข้าใจธุรกิจเป็นขั้นตอนแรกที่สำคัญในการพัฒนาตัวแบบการทำนายความสำเร็จของสตาร์ทอัพ จากการศึกษาก่อนหน้านี้พบว่าร้อยละ 90 ของธุรกิจสตาร์ทอัพไม่ประสบความสำเร็จด้วยเหตุผลหลายประการ โดยร้อยละ 84 ของเหตุผลที่ทำให้สตาร์ทอัพไม่ประสบความสำเร็จนั้นเป็นเรื่องเกี่ยวกับการเงิน ซึ่งสะท้อนให้เห็นถึงความสำคัญของการมีแบบจำลองทางการเงินที่มีประสิทธิภาพ (Ramalakshmi & Kamidi, 2018) (Patel, 2015)

ในยุคของการเปลี่ยนแปลงทางดิจิทัล การปรากฏตัวทางดิจิทัล (Digital Presence) มีบทบาทสำคัญต่อความสำเร็จของสตาร์ทอัพ โดยเฉพาะอย่างยิ่งในด้านการดึงดูดการลงทุน การสร้างความน่าเชื่อถือ และการเข้าถึงกลุ่มลูกค้าเป้าหมาย ซึ่งสามารถวัดได้จากประสิทธิผลของสื่อที่เป็นเจ้าของ (Owned Media) และสื่อที่ได้รับ (Earned Media) ที่จะทำการศึกษานงานวิจัยนี้

3.5.2 การทำความเข้าใจข้อมูล (Data Understanding)

ข้อมูลที่ได้มาจากรฐานข้อมูล Crunchbase ได้รับการวิเคราะห์โดยกำหนดเงื่อนไขเฉพาะเพื่อศึกษาสถานะของสตาร์ทอัพที่ก่อตั้งในปี 2015 โดยแบ่งออกเป็นสองกลุ่มดังนี้:

1. สตาร์ทอัพที่ยังดำเนินกิจการอยู่และถูกซื้อกิจการ (Acquired): ในช่วงปี 2016–2024 พบจำนวนทั้งสิ้น 500 บริษัท
2. สตาร์ทอัพที่หยุดดำเนินกิจการ (Closed): ในช่วงปี 2015–2024 พบจำนวนทั้งสิ้น 250 บริษัท โดยชุดข้อมูลมีคุณลักษณะ (Features) ดังแสดงในตารางที่ 3.1

Organization Name	Website	Number of Founders	Number of Funding Rounds	Total Funding Amount	Total Funding Amount Currency	Total Funding Amount (in USD)	Number of Active	Monthly Sales	Web Duration	Page Views / Month	Source Rate	Patents Granted	Trademarks Registered	Operating Status
Rocket	https://rocket.com	2	6	117500000 USD		117500000	49	19,089	679	1.43 81.35%		3		2 Active
Athelas	https://athelas.com/	2	5	150131620 USD		150131620	4	103,920	215	1.07 95.9%				Active
PathAI	http://www.pathai.com	2	6	355200000 USD		355200000	57	41,863	880	4.25 46.68%		15		3 Active
Tabular	https://tabular.io/	3	2	37000000 USD		37000000	19	5,318	1,891	1.93 78.56%				Active
Armis Security	https://www.armis.com	3	6	537000000 USD		537000000	19	80,630	953	1.87 73.24%		3		5 Active
InsightSoftware	https://www.insightsoftware.com	1	2	100000000 CAD		798604678	32	277,192	202	1.7 67.79%				12 Active
Glovo	http://www.glovoapp.com	4	11	1162249826 USD		1162249826	204	9,413.55	772	1.82 68.58%		1		9 Active
Shine	https://www.theshineapp.com/	2	3	8250000 USD		8250000	60	95,038	115	1.03 99.23%		0		8 Active
RunAI	https://www.run.ai	2	4	118000000 USD		118000000	40	300,274	657	1.93 66.19%		1		0 Active
Convoy	https://convoy.com	2	10	1090500000 USD		1090500000	183	10,712	36	2.06 41.46%				Active
Grail	https://www.grail.com	2	5	2004655040 USD		2004655040	152	76,058	541	1.61 79.75%		91		6 Active
Kustomer	https://www.kustomer.com	2	7	233500000 USD		233500000	97	50,973	321	2.23 69.99%		0		1 Active
Split Software	https://www.split.io/	3	5	169800000 USD		169800000	31	99,569	479	2.44 38.64%		8		3 Active
TruEra	https://truera.com	3	3	42284998 USD		42284998	28	296,250	27	1.04 95.93%		1		2 Active
Spiff	https://spiff.com	8	5	112000000 USD		112000000	58	25,390	826	1.47 4.32%				Active
Petal	https://www.petalcad.com	4	18	992600000 USD		992600000	61	524,330	1,775	1.46 81.39%		0		1 Active
SkyHive	https://www.skyhive.ai	2	6	48000000 USD		48000000	178	2,909	895	3.41 34.86%		3		3 Active
MosaicML	https://www.mosaicml.com	1	2	37000000 USD		37000000	46	14,325	283	1.46 78.88%		0		2 Active
MetaMap	https://www.metamap.com	2	3	86500000 USD		86500000	14	75,824	246	1.57 68.18%		0		1 Active
Perch	https://www.perchhq.com	1	4	908750000 USD		908750000	18	9,725	30	2.32 52.92%		0		4 Active
Care/of	http://www.careof.com	2	5	82869662 USD		82869662	26	568,712	342	1.92 72.94%		0		0 Active
Sayari Labs	https://sayari.com	2	5	94700000 USD		94700000	24	30,239	2,439	1.88 80.93%		0		2 Active
Bipi	http://www.bipicar.com	2	5	31500000 EUR		37030036	17	126,480	303	1.81 75.83%		0		1 Active
Shoreline.io	https://shoreline.io	2	2	56900000 USD		56900000	6	2,123	29	1.06 94.07%		4		1 Active
Alight Solutions	https://www.alight.com	1	6	1642684096 USD		1642684096	60	11,726,4	648	5.48 22.08%		2		36 Active
Catalyst Software	https://catalyst.io	2	7	66735530 USD		66735530	10	24,772	507	6.43 40.94%		0		28 Active
Pismo	https://www.pismo.io	4	2	118000000 USD		118000000	75	38,178	865	2.39 50.8%				Active
Chorus.ai	http://www.chorus.ai	3	4	100300000 USD		100300000	159	90,918	525	1.5 66.6%		1		0 Active
Addverb	https://www.addverb.com	6	1	132000000 USD		132000000	51	2,732	367	1.4 68.67%		0		0 Active
Ultimate	https://www.ultimate.ai/	4	7	27400545 USD		27400545	175	30,554	119	2.37 55.74%				Active
Element AI	http://www.elemental.com	7	4	257483994 USD		257483994	112	1,134	28	1.17 83.25%		46		22 Active
Eden Health	http://edenhealth.com/	2	5	99000000 USD		99000000	35	4,510	299	1.25 87.36%		0		2 Active
Drift	https://www.drift.com	2	5	107000000 USD		107000000	100	286,585	1,422	2.14 63.28%		15		12 Active
Nurx	https://www.nurx.com/	2	8	115920000 USD		115920000	59	2,377.76	283	1.49 78.5%		0		2 Active
Caribu	https://caribu.com	3	5	5050000 USD		5050000	16	15,371	53	1.58 49.05%				Active
Uizard	https://uizard.io	4	3	18600000 USD		18600000	26	1,012.72	494	2.13 65.36%				Active
NODWIN Gaming	http://nodwingaming.com	2	5	48919689 USD		48919689	41	7,208	33	1.03 97.45%				Active
Henchman	https://www.henchman.io	3	4	10500000 EUR		11569702	8	4,865	393	1.14 89.66%		0		1 Active
Perimeter 81	https://www.perimeter81.com	2	6	165000000 USD		165000000	74	174,329	476	2.82 40.8%				Active
Outlier	https://outlier.org/	1	3	46000000 USD		46000000	2,249,412	879	2,31	89.75%		0		3 Active
BigBear.ai	https://www.bigbear.ai/	1	2	225000000 USD		225000000	67	135,243	312	1.47 71.62%		0		5 Active

รูปที่ 3.1 ตัวอย่างข้อมูลดิบที่ได้จากฐานข้อมูล Crunchbase

ตารางที่ 3.1 ตารางแสดงคุณลักษณะ (Features) ของชุดข้อมูลที่ใช้ในการวิเคราะห์

ลำดับที่	คุณลักษณะ	คำอธิบาย
1	Organization Name	ชื่อขององค์กรหรือบริษัทที่จดทะเบียนในฐานข้อมูล
2	Website	URL เว็บไซต์ทางการขององค์กรหรือบริษัท
3	Number of Founders	จำนวนผู้ก่อตั้ง
4	Number of Funding Rounds	จำนวนรอบการระดมทุน
5	Total Funding Amount	มูลค่ารวมของเงินทุนทั้งหมดที่บริษัทได้รับจากนักลงทุนในทุก ๆ รอบการระดมทุน
6	Total Funding Amount Currency	สกุลเงินที่ใช้แสดงมูลค่ารวมของเงินทุน อาทิ USD, EUR หรือสกุลเงินอื่น ๆ
7	Total Funding Amount	มูลค่ารวมของเงินทุนทั้งหมดที่ได้รับ โดยแปลงเป็นดอลลาร์สหรัฐ (USD) เพื่อการเปรียบเทียบระหว่างประเทศ

ตารางที่ 3.1 ตารางแสดงคุณลักษณะ (Features) ของชุดข้อมูลที่ใช้ในการวิเคราะห์ (ต่อ)

ลำดับที่	คุณลักษณะ	คำอธิบาย
8	Number of Articles	จำนวนบทความหรือข่าวที่กล่าวถึงองค์กรในแหล่งสื่ออาทิ ข่าวออนไลน์ บทความ หรือบล็อก ซึ่งสะท้อนถึงความนิยมและการรับรู้ขององค์กรในสื่อ
9	Monthly Visits	จำนวนการเข้าชมเว็บไซต์ขององค์กรในแต่ละเดือน
10	Visit Duration	ระยะเวลาเฉลี่ยที่ผู้เข้าชมใช้บนเว็บไซต์ในแต่ละครั้ง
11	Page Views/Visit	จำนวนหน้าที่ผู้ใช้งานเข้าชมต่อการเข้าชมหนึ่งครั้ง
12	Bounce Rate	อัตราการเดิ๊งออกของผู้ใช้งาน (Bounce Rate) แสดงถึงเปอร์เซ็นต์ของผู้ที่ออกจากเว็บไซต์หลังจากดูหน้าแรกเพียงหน้าเดียว
13	Patents Granted	จำนวนสิทธิบัตรที่องค์กรได้รับการรับรอง
14	Trademarks Registered	จำนวนเครื่องหมายการค้าที่องค์กรได้จดทะเบียน
15	Operating Status	สถานะการดำเนินงานของบริษัท

เพิ่มมิติของข้อมูลโดยใช้เครื่องมือ Lighthouse ด้วยการนำ Website ของแต่ละบริษัทมาวิเคราะห์ด้วย Lighthouse ซึ่งเป็นเครื่องมือโอเพ่นซอร์สที่ให้บริการใน Google Chrome สำหรับการประเมินคุณภาพของเว็บไซต์ โดยจะวิเคราะห์เว็บไซต์และให้คะแนนในด้านต่าง ๆ ซึ่งจะทำให้มี คุณลักษณะด้านสื่อที่เป็นเจ้าของ เพิ่มขึ้นมา 4 คุณลักษณะ ดังนี้

1. ประสิทธิภาพของเว็บไซต์ (Performance)
 - 1.1. ความเร็วในการโหลดเว็บไซต์
 - 1.2. การแสดงผลเนื้อหา
 - 1.3. เวลาที่ใช้ในการโต้ตอบ
2. การเข้าถึงเว็บไซต์ (Accessibility)
 - 2.1. ความสามารถในการเข้าถึงเนื้อหาสำหรับผู้ใช้งานทุกกลุ่ม รวมถึงผู้พิการ
 - 2.2. การใช้แท็ก HTML อย่างเหมาะสม อาทิ alt สำหรับรูปภาพ
3. แนวทางปฏิบัติที่ดีของเว็บไซต์ (Best Practices)
 - 3.1. การใช้ HTTPS (Hypertext Transfer Protocol Secure)
 - 3.2. การหลีกเลี่ยงการใช้ทรัพยากรที่ไม่ปลอดภัย อาทิ ไฟล์สคริปต์ที่ไม่ได้รับการอนุมัติ
4. การเพิ่มประสิทธิภาพเว็บไซต์เพื่อการค้นหา (SEO)
 - 4.1. ความพร้อมของเว็บไซต์สำหรับเครื่องมือค้นหา (Search Engine)
 - 4.2. การใช้แท็ก Meta อย่างเหมาะสม

เพิ่มมิติของข้อมูลโดยใช้เครื่องมือ Ahrefs Backlink Checker เป็นเครื่องมือที่ช่วยวิเคราะห์ข้อมูลลิงก์ย้อนกลับ (Backlink) ของเว็บไซต์หรือ URL หนึ่ง ๆ โดยรายงานค่าที่สำคัญเกี่ยวกับโปรไฟล์ลิงก์ย้อนกลับ เพื่อสร้างคุณลักษณะด้านสื่อที่ได้รับ จำนวน 3 คุณลักษณะ ซึ่งมีรายละเอียดหลัก ๆ ดังนี้:

Domain Rating (DR): เป็นคะแนนที่ Ahrefs ใช้วัดความน่าเชื่อถือของโดเมนเว็บไซต์ โดยมีค่าตั้งแต่ 0 ถึง 100 (คะแนนสูงหมายถึงความน่าเชื่อถือมากกว่า)

Backlinks (ลิงก์ย้อนกลับ): ลิงก์จากเว็บไซต์อื่นที่ชี้กลับมายังเว็บไซต์ ซึ่งอาจอยู่ในรูปแบบข้อความหรือรูปภาพ

Linking Websites (โดเมนที่ลิงก์มายังเว็บไซต์): Linking Websites หรือ Referring Domains คือ จำนวนโดเมนที่แตกต่างกันที่ลิงก์มายังเว็บไซต์

Organization Name	Website	Number of Founders	Number of Funding Rounds	Total Funding Amount	Total Funding Amount Currency	Total Funding Amount USD	Number of Investors	Website Status	Website Location	Page Score	PA	Domain Rating	Trustworthy Registered	Performance	Accountability	Best Practices	SEO	Domain Rating	Backlinks (M)	Linking Websites
Rockset	https://rockset.com	2	6	117500000	USD	117500000	49	19,089	679	1.43	81.35%	3	2	26	84	75	83	65	120	1600
Athelas	https://athelas.com/	2	5	150111600	USD	150111600	4	103,920	215	1.07	98.99%			25	89	54	100	56	4.5	725
PathAI	http://www.pathai.com	2	6	355200000	USD	355200000	57	41,863	880	4.25	48.68%	15	3	56	94	57	92	61	10	1300
Tabular	https://tabular.io/	3	2	37000000	USD	37000000	19	5,318	1,891	1.93	78.56%			61	96	75	85	43	1.1	319
Armis Security	https://www.armis.com	3	6	537000000	USD	537000000	19	80,630	953	1.87	73.24%	3	5	38	89	57	92	75	70	3800
InsightSoftware	https://www.insightsoftware.com	1	2	1000000000	CAD	798046878	32	227,182	202	1.7	67.79%	0	12	36	87	79	92	75	2.6	204
Glovo	https://www.glovoapp.com	4	11	1162249826	USD	1162249826	204	9,413,584	772	1.82	68.58%	1	9	60	86	79	92	84	2.2	317
Shine	https://www.shineapp.com/	2	3	8250000	USD	8250000	60	95,038	115	1.03	99.23%	0	8	53	64	79	92	72	2.4	784
RunAI	http://www.run.ai	2	4	118000000	USD	118000000	40	300,274	657	1.93	66.19%	1	0	76	87	79	92	60	3.7	1500
Convoy	https://convoy.com	2	10	1095000000	USD	1095000000	183	10,712	36	2.06	41.46%			41	79	57	85	68	24	1800
Graii	https://www.graai.com	2	5	2004655040	USD	2004655040	152	76,058	541	1.61	79.75%	91	6	71	77	96	83	69	1	222
Kustomer	https://www.kustomer.com	2	7	232500000	USD	232500000	97	50,973	321	2.23	69.99%	0	1	25	85	75	83	78	24	1900
Split Software	https://www.split.io/	3	5	169800000	USD	169800000	31	99,569	479	2.44	38.64%	8	3	49	90	61	92	71	6.8	1600
TrueEra	https://trueera.com	3	3	42284988	USD	42284988	28	296,250	27	1.04	95.93%	1	2	40	90	72	92	47	2.3	529
Spiff	https://spiff.com	8	5	112000000	USD	112000000	58	25,390	826	1.4	74.32%			36	81	75	92	49	5.8	778
Petal	https://www.petalscard.com	4	18	992600000	USD	992600000	61	524,330	1,775	1.46	81.39%	0	1	57	89	79	92	64	5.9	1700
SkyHive	https://www.skyhive.ai	2	6	48000000	USD	48000000	178	2,909	895	3.41	34.86%	3	3	29	89	57	85	48	1.8	395
MosaicML	https://www.mosaicml.com	1	2	37000000	USD	37000000	46	14,325	283	1.46	78.88%	0	2	29	93	75	92	70	10	1600
MetaMap	https://www.metamap.com	2	3	86500000	USD	86500000	14	75,824	246	1.57	68.18%	0	1	82	78	100	92	48	1.6	512
Perch	https://www.perchhq.com	1	4	908750000	USD	908750000	18	9,725	30	2.32	52.92%	0	4	60	96	75	85	51	0.88	554
Care/of	http://www.takecareof.com	2	5	82999662	USD	82999662	26	568,712	342	1.92	72.94%	0	0	35	92	100	83	72	20	2700
Seyri Labs	https://seyri.com	2	5	94700000	USD	94700000	24	30,239	2,439	1.88	80.93%	0	2	30	90	75	85	52	3.1	797
Biipi	http://www.biipicor.com	2	5	31500000	EUR	37303036	17	126,480	303	1.81	75.83%	0	1	33	81	79	85	47	0.33	82
Shoreline.io	https://shoreline.io	2	2	56900000	USD	56900000	6	2,123	29	1.06	94.07%	4	1	25	96	75	100	56	1.4	298
Alight Solutions	https://www.alight.com	1	6	1642684095	USD	1642684095	60	11,726,422	648	5.48	22.08%	2	36	44	92	100	77	77	7	1900
Catalyst Software	https://catalyst.io	2	7	66735530	USD	66735530	10	24,772	507	6.43	40.94%	0	28	28	88	79	100	59	37	767
Pismo	https://www.pismo.io	4	2	118000000	USD	118000000	75	36,178	865	2.39	50.80%			55	89	75	92	47	2.3	762
Chorus.ai	http://www.chorus.ai	3	4	100300000	USD	100300000	159	90,918	525	1.5	66.60%	1	0	25	79	57	92	72	5.1	1200
Addverb	https://www.addverb.com	6	1	132000000	USD	132000000	51	2,732	367	1.4	68.67%	0	0	67	83	75	83	41	0.57	168
Ultimate	https://www.ultimate.ai/	4	7	27400545	USD	27400545	175	30,554	119	2.37	55.74%			73	88	95	100	66	5.9	1100
Element AI	http://www.elemental.com	7	4	257483994	USD	257483994	112	1,134	28	1.17	83.25%	46	22	47	89	57	100	66	4.7	908
Eden Health	http://edenhealth.com/	2	5	96000000	USD	99000000	35	4,510	299	1.26	87.38%	0	2	54	89	75	92	54	3.4	1000
Drift	https://www.drift.com	2	5	107000000	USD	107000000	100	286,585	1,422	2.14	63.28%	15	12	5	72	75	85	85	154	12000
Nurx	https://www.nurx.com/	2	8	115920000	USD	115920000	59	2,377,764	283	1.49	78.50%	0	2	37	67	100	77	71	18	4200
Caribu	https://caribu.com	3	5	50500000	USD	50500000	16	15,371	53	1.58	48.05%			50	85	75	100	60	2.1	995
Uizard	https://uizard.io	4	3	18600000	USD	18600000	26	1,012,720	494	2.13	65.36%			45	88	54	92	72	56	4300
NODWIN Gaming	https://nodwingaming.com	2	5	48919689	USD	48919689	41	7,208	33	1.03	97.45%			67	89	75	83	43	1.4	435
Henchman	https://www.henchman.io	3	4	10500000	EUR	11569702	8	4,865	393	1.14	89.66%	0	1	37	100	79	85	34	684	109
Perimeter 81	https://www.perimeter81.com	2	6	165000000	USD	165000000	74	174,329	476	2.82	40.80%			81	78	100	77	73	19	2000
Outlier	https://outlier.org/	1	3	46000000	USD	46000000	2	249,412	879	2.31	89.75%	0	3	33	80	57	85	50	4.3	1100
BigBear.ai	https://www.bigbear.ai/	1	2	225000000	USD	225000000	67	135,343	312	1.47	71.62%	0	5	42	90	61	100	54	0.75	47
Clearbit	https://clearbit.com/	5	2	17000000	USD	17000000	29	330,646	152	1.44	72.37%	0	5	41	90	71	92	81	3000	6600
Moonbug	https://www.moonbug.com/	2	2	265000000	USD	265000000	94	45,002	3,040	1.61	83.87%	0	68	35	96	100	92	51	4600	567

รูปที่ 3.2 ตัวอย่างข้อมูลดิบหลังจากมีการเพิ่มมิติของข้อมูลจาก Lighthouse และ Ahrefs

3.5.3 การเตรียมข้อมูล (Data Preparation)

การรวบรวมข้อมูล (Data Collection) และการรวมข้อมูล (Data Integration) โดยทำการเก็บข้อมูลจาก 3 แหล่งข้อมูล ที่กล่าวมาในข้อ 3.5.2 ทำให้ได้ข้อมูล (Dataset) ทั้งหมด 698 แถว 18 คอลัมน์ แบ่งเป็น ข้อมูลที่มีสถานะ Acquired จำนวน 503 แถว และข้อมูลที่มีสถานะ Closed จำนวน 195 แถว จากการเพิ่มสื่อที่เป็นเจ้าของ (Owned Media) และสื่อที่ได้รับ (Earned Media) เข้าไปด้วยการใช้เครื่องมือดังกล่าว จะทำให้ข้อมูลมีคุณลักษณะทั้งหมด 22 คุณลักษณะ รวมตัวแปรเป้าหมาย (Target Variable) หลังจากทำการตัดคุณลักษณะที่ไม่ใช่ออกไป ทำให้เหลือคุณลักษณะที่ใช้งานทั้งหมด 18 คุณลักษณะ และทำการจัดกลุ่มคุณลักษณะที่เป็นตัวแปรต้น 17 คุณลักษณะ เป็น 3 กลุ่ม ดังตารางต่อไปนี้

2	6	117500000	3	2	19,288	679	1.43	81.35%	26	84	41	83	49	65	120	1600	Accu
2	1	14575639	0	82	3,756,966	1,027	1.7	62.75%	1	86	54	77	29	66	52	3300	Accu
1	1	20185149			753,739	679	1.26	88.85%	4	66	50	85	43	51	1.6	778	Accu
2	4	80899457	0	3	6,838,308	522	3.16	18.11%	4	77	57	85	2	65	315	1000	Accu
2	5	107000000	15	12	286,585	1,422	2.14	63.28%	5	72	75	85	100	85	154	12000	Accu
2	4	90000000			267,002	429	2.49	46.51%	7	64	75	92	11	69	107	1500	Accu
1	4	8666698			2,124,671	674	2.32	57.96%	8	88	75	92	3	48	1.3	216	Accu
2	1	48316648			2,217,638	494	1.45	76.64%	10	95	79	100	5	50	1.7	608	Accu
1	5	10200000			1,460,703	679	1.38	77.92%	10	89	100	100	7	74	3600	9200	Accu
2	5	34000000			33,335	622	1.64	82.96%	11	80	54	83	33	5.4		1900	Accu
1	3	95187010			432	48	2.7	43.98%	11	81	54	92	39	46	0.377	232	Accu
2	1	8000000	0	4	25,587	480	1.15	96.56%	12	90	57	86	8	56	4.2	1100	Accu
4	7	3300000	0	6	2,454	2	1.23	76.77%	14	88	61	92	39	58	91	525	Accu
3	3	17700000			1,958	5	1.5	50%	15	77	71	75	3	38	0.664	151	Accu
2	2	65000000	0	2	6,337	52	1.39	69.73%	16	86	54	85	1	40	0.18	51	Accu
2	2	17000000	0	4	550,894	265	2.03	72.28%	18	73	64	85	21	44	737	742	Accu
1	4	50000000			36,892	1,325	5.66	9.27%	20	97	57	85	47	64	5.2	879	Accu
2	2	13000000			2,925	4	1.23	76.62%	20	92	68	77	23	33	0.277	173	Accu
3	1	15000000	0	4	13,698	106	1.42	82.47%	20	93	79	92	29	61	6.2	993	Accu
3	5	35000000	2	0	465,327	205	1.87	61.70%	20	99	61	92	19	70	10	2000	Accu
3	3	4700000			9,139	138	1.42	58.47%	20	91	54	92	22	32	0.969	337	Accu
1	4	51999998			3,348	16	1.94	62.84%	21	82	71	92	23	40	1.4	461	Accu
2	4	92000000	32	3	231	32	3	0%	21	95	57	85	31	41	0.61	194	Accu
3	1	5728604			13,226	333	4.08	35.53%	22	74	68	85	8	36	1.3	295	Accu
3	4	19271872	24	22	82,696	386	3.18	63.47%	23	65	41	77	22	75	101	2.8	Accu
3	5	18629753	0	0	1,816,134	749	1.57	70.12%	23	69	36	85	10	56	3200	1800	Accu
2	4	31830177			25,494	688	1.95	62.30%	24	88	68	77	24	67	3.4	542	Accu
2	1	7113578			7,524	43	1.21	84.97%	24	73	75	92	61	47	3.9	432	Accu
2	5	150131620			103,820	215	1.07	95.90%	25	89	54	100	4	56	4.5	725	Accu
3	7	23500000	0	1	50,973	301	2.25	69.96%	25	85	75	92	83	78	24	1900	Accu
2	2	56900000	4	1	2,123	29	1.06	94.07%	25	96	75	100	6	56	1.4	298	Accu
3	4	100300000	1	0	90,918	525	1.5	66.60%	25	79	57	92	159	72	5.1	1200	Accu
3	6	82000000			2,280	41	1.09	91.36%	25	90	57	85	45	66	2.4	442	Accu
3	5	1700000			1,159	7	1.17	83%	25	75	71	100	8	34	5.3	244	Accu
2	9	14648902	0	5	84,153	2,339	1.75	65.04%	25	95	61	85	35	50	6.2	1200	Accu
3	2	5260000	0	1	6,057	26	1.01	98.89%	25	100	75	100	57	1.8		386	Accu
1	3	15000000			868,733	597	1.78	71.92%	25	87	71	85	150	77	98	7900	Accu
2	3	7300000	0	1	275,939	182	1.72	41.76%	25	67	57	83	14	57	9.9	1200	Accu
2	5	30500000	0	0	30,864	65	1.62	63.60%	25	90	57	92	5	55	0.724	115	Accu
1	2	70000000	18	10	4,837	5	1.22	77.96%	25	82	61	69	36	29	0.221	60	Accu
3	3	26200000			2,449,133	961	3.1	53.34%	25	70	54	92	30	66	150	6100	Accu
3	6	6106325	0	8	101,133	163	1.18	84.62%	25	75	75	77	21	57	2.9	986	Accu
2	6	16000000	0	2	20,138	2,162	2.1	58.20%	25	68	79	85	31	56	1.1	466	Accu
2	6	920000			177,215	449	2.93	47.11%	26	89	57	100	13	35	3.9	820	Accu
2	6	18600000			892,396	797	2.3	69.97%	26	75	71	100	6	71	959	1400	Accu

รูปที่ 3.3 ตัวอย่างชุดข้อมูล (Dataset)

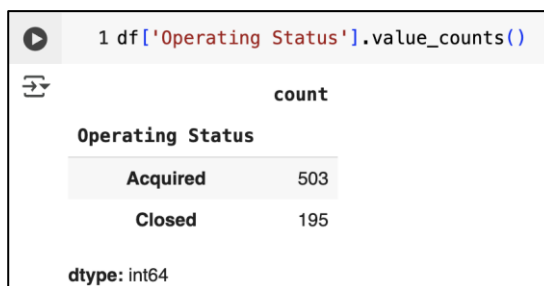
```

1 df.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 698 entries, 0 to 697
Data columns (total 18 columns):
#   Column                                Non-Null Count  Dtype
---  ---                                ---
0    BF-Founder                            698 non-null    int64
1    BF-Round                             698 non-null    int64
2    BF-Funding                           672 non-null    float64
3    BF-Patent                            333 non-null    float64
4    BF-Trademark                         333 non-null    float64
5    OM-Visit                             698 non-null    object
6    OM-Duration                          698 non-null    object
7    OM-View                              698 non-null    float64
8    OM-Bounce                           698 non-null    object
9    OM-Performance                       697 non-null    float64
10   OM-Accessibility                     695 non-null    float64
11   OM-Practice                         698 non-null    int64
12   OM-SEO                              697 non-null    float64
13   EM-Article                          611 non-null    object
14   EM-Rating                           698 non-null    float64
15   EM-Backlink                         698 non-null    float64
16   EM-Linking                          698 non-null    float64
17   Operating Status                     698 non-null    object
dtypes: float64(10), int64(3), object(5)
memory usage: 98.3+ KB

```

รูปที่ 3.4 โครงสร้างชุดข้อมูล (Data Frame) จากโปรแกรม Google Colab ด้วยภาษา Python



รูปที่ 3.5 แสดงการนับจำนวนสถานะของสตาร์ทอัปในชุดข้อมูลจากโปรแกรม Google Colab ด้วยภาษา Python

ตารางที่ 3.2 คุณลักษณะ (Features) ข้อมูลพื้นฐานของธุรกิจ (Business Foundation)

No.	Feature	Type	Description
1	Number of Founders (BF-Founder)	int64	จำนวนผู้ก่อตั้งบริษัท
2	Number of Funding Rounds (BF-Round)	int64	จำนวนรอบการระดมทุน
3	Total Funding Amount (BF-Funding)	float64	มูลค่ารวมของเงินทุนทั้งหมดที่ได้รับเป็นดอลลาร์สหรัฐ
4	Patents Granted (BF-Patent)	float64	จำนวนสิทธิบัตรที่องค์กรได้รับการรับรอง
5	Trademarks Registered (BF-Trademark)	float64	จำนวนเครื่องหมายการค้าที่องค์กรได้จดทะเบียน

ตารางที่ 3.3 คุณลักษณะ (Features) ของปัจจัยด้านสื่อที่เป็นเจ้าของ (Owned Media)

No.	Feature	Type	Description
1	Monthly Visits (OM-Visit)	float64	จำนวนการเข้าชมเว็บไซต์ขององค์กรในแต่ละเดือน
2	Visit Duration (OM-Duration)	float64	ระยะเวลาเฉลี่ยที่ผู้เข้าชมใช้บนเว็บไซต์ในแต่ละครั้ง
3	Page Views / Visit (OM-View)	float64	จำนวนหน้าที่ผู้ใช้งานเข้าชมต่อการเข้าชมหนึ่งครั้ง
4	Bounce Rate (OM-Bounce)	float64	อัตราการเดิ๊งออกของผู้ใช้งาน (Bounce Rate) แสดงถึงเปอร์เซ็นต์ของผู้ที่ออกจากเว็บไซต์หลังจากดูหน้าแรกเพียงหน้าเดียว
5	Website Performance (OM-Performance)	float64	ประสิทธิภาพของเว็บไซต์

ตารางที่ 3.3 คุณลักษณะ (Features) ของปัจจัยด้านสื่อที่เป็นเจ้าของ (Owned Media) (ต่อ)

No.	Feature	Type	Description
6	Website Accessibility (OM-Accessibility)	float64	การเข้าถึงเนื้อหาของเว็บไซต์
7	Website Best Practices (OM-Practice)	int64	แนวทางปฏิบัติที่ดีของเว็บไซต์
8	Website SEO (OM-SEO)	float64	การเพิ่มประสิทธิภาพเว็บไซต์เพื่อการค้นหา

ตารางที่ 3.4 คุณลักษณะ (Features) ของปัจจัยด้านสื่อที่ได้รับ (Earned Media)

No.	Feature	Type	Description
1	Number of Articles (EM-Article)	float64	จำนวนบทความหรือข่าวที่กล่าวถึงองค์กร
2	Domain Rating (EM-Rating)	float64	คะแนนความน่าเชื่อถือของโดเมนเว็บไซต์
3	Backlinks (EM-Backlink)	float64	จำนวนลิงก์จากเว็บไซต์อื่นที่ชี้กลับมาที่เว็บไซต์
4	Linking Websites (EM-Linking)	float64	จำนวนโดเมนที่แตกต่างกันที่ลิงก์มายังเว็บไซต์

การทำความสะอาดข้อมูล (Data Cleaning)

1. ทำการลบอักขระพิเศษออกจากข้อมูลที่เป็นตัวเลขแสดงในโค้ดดิ้ง (Coding)

```

1 df['Owned Media - Bounce Rate'] = df['Owned Media - Bounce Rate'].str.replace('%', '').astype(float)

1 df['Owned Media - Monthly Visits'] = df['Owned Media - Monthly Visits'].str.replace(',', '').astype(float)

1 df['Owned Media - Visit Duration'] = df['Owned Media - Visit Duration'].str.replace(':', '').astype(float)

1 df['Earned Media - Number of Articles'] = df['Earned Media - Number of Articles'].str.replace(',', '').astype(float)

```

รูปที่ 3.6 โค้ดดิ้งการขระอักขระพิเศษ

2. BF-Funding มีข้อมูลหายไปจำนวน 26 แถว ซึ่งจะใช้วิธีเติมค่าที่หายไปด้วยค่ากลาง (Median)

```
1 # ตรวจสอบค่า NaN
2 nan_count = df.isnull().sum()
3 print(nan_count)
```

BF-Founder	0
BF-Round	0
BF-Funding	26
BF-Patent	365
BF-Trademark	365
OM-Visit	0
OM-Duration	0
OM-View	0
OM-Bounce	0
OM-Performance	1
OM-Accessibility	3
OM-Practice	0
OM-SEO	1
EM-Article	87
EM-Rating	0
EM-Backlink	0
EM-Linking	0
Operating Status	0

dtype: int64

รูปที่ 3.7 แสดงการนับข้อมูลที่ไม่สมบูรณ์จากโปรแกรม Google Colab ด้วยภาษา Python

3. BF-Funding มีข้อมูลหายไปจำนวน 26 แถว ซึ่งจะใช้วิธีเติมค่าที่หายไปด้วยค่ากลาง (Median)
4. OM-Performance OM-Accessibility และ OM-SEO มีค่าที่หายไปจำนวน 5 แถว ซึ่งเลือกใช้วิธีลบแถวข้อมูล
5. ข้อมูลที่เหลือทั้งสิ้น 694 แถว แบ่งเป็นข้อมูลที่มีสถานะ Acquired จำนวน 499 แถว และข้อมูลที่มีสถานะ Closed จำนวน 195 แถว

```
1 df['Operating Status'].value_counts()
```

	count
Operating Status	
Acquired	499
Closed	195

dtype: int64

รูปที่ 3.8 แสดงการนับจำนวนสถานะของสตาร์ทอัพในชุดข้อมูลจากโปรแกรม Google Colab ด้วยภาษา Python

สถิติเชิงพรรณนา (Descriptive Statistics)

นี่คือการอธิบายลักษณะข้อมูลที่ได้รับการแก้ไขตามกลุ่มตัวแปรที่กำหนดใหม่ ดังนี้:

ลักษณะของข้อมูล

ข้อมูลที่แสดงในตารางเป็นผลจากการวิเคราะห์เชิงสถิติ (Descriptive Statistics) โดยมีจำนวนตัวอย่างทั้งหมด 694 ตัวอย่าง กลุ่มตัวแปรได้ถูกแบ่งออกเป็น 3 กลุ่มหลัก ได้แก่ ข้อมูลพื้นฐานของธุรกิจ (Business Foundation), ปัจจัยด้านสื่อที่เป็นเจ้าของ (Owned Media), และปัจจัยด้านสื่อที่ได้รับ (Earned Media) รายละเอียดดังนี้:

1. กลุ่มข้อมูลพื้นฐานของธุรกิจ (Business Foundation)

กลุ่มนี้สะท้อนถึงโครงสร้างพื้นฐานของธุรกิจ startup ในด้านของผู้ก่อตั้ง การระดมทุน และสินทรัพย์ทางปัญญา โดยพบว่า:

- จำนวนผู้ก่อตั้ง (BF-Founder) เฉลี่ยอยู่ที่ประมาณ 2.27 คน มีการกระจายไม่สูงนัก (ค่าเบี่ยงเบนมาตรฐาน 1.14 คน) และมีจำนวนผู้ก่อตั้งสูงสุดถึง 10 คน
- รอบการระดมทุน (BF-Round) มีค่าเฉลี่ยที่ 3.39 รอบ โดยสูงสุดอยู่ที่ 18 รอบ
- เงินทุนที่ได้รับ (BF-Funding) เฉลี่ยประมาณ 80.56 ล้านบาท โดยมีค่าความผันแปรสูง (ค่าเบี่ยงเบนมาตรฐานประมาณ 252 ล้านบาท) บ่งบอกถึงความแตกต่างในการได้รับเงินทุนในแต่ละ startup ที่ค่อนข้างสูง
- จำนวนสิทธิบัตร (BF-Patent) และเครื่องหมายการค้า (BF-Trademark) มีค่าเฉลี่ยค่อนข้างต่ำ (1.66 และ 2.00 ตามลำดับ) สะท้อนว่าส่วนใหญ่ startup ยังไม่ค่อยมีการจดทะเบียนทรัพย์สินทางปัญญามากนัก

2. กลุ่มปัจจัยด้านสื่อที่เป็นเจ้าของ (Owned Media)

กลุ่มนี้เป็นข้อมูลที่แสดงประสิทธิภาพของช่องทางที่ startup เป็นเจ้าของโดยตรง โดยเฉพาะเว็บไซต์ พบว่า:

- จำนวนผู้เยี่ยมชมเว็บไซต์ (OM-Visit) เฉลี่ยประมาณ 480,391 ครั้ง มีค่าการกระจายตัวสูง (ค่าเบี่ยงเบนมาตรฐานประมาณ 3,091,476 ครั้ง) สะท้อนความแตกต่างที่มากในระดับการเข้าชมเว็บไซต์
- ระยะเวลาเฉลี่ยในการเข้าชม (OM-Duration) อยู่ที่ประมาณ 435 วินาที หรือประมาณ 7 นาที
- จำนวนหน้าที่ถูกเปิดต่อครั้ง (OM-View) เฉลี่ยอยู่ที่ประมาณ 2.02 หน้า
- อัตราการตีกลับ (OM-Bounce) อยู่ที่ประมาณ 65% ซึ่งค่อนข้างสูง หมายความว่าผู้ใช้จำนวนมากที่ออกจากเว็บไซต์หลังดูเพียงหน้าเดียว
- คะแนนด้านประสิทธิภาพ (OM-Performance) เฉลี่ย 51.95 คะแนน ซึ่งอยู่ในระดับปานกลาง
- คะแนนด้านการเข้าถึงเว็บไซต์ (OM-Accessibility) เฉลี่ยที่ประมาณ 86.53 คะแนน ซึ่งอยู่ในระดับดี
- คะแนนการปฏิบัติตามแนวทางที่ดี (OM-Practice) เฉลี่ย 76.23 คะแนน และคะแนนการปรับแต่งสำหรับการค้นหา (OM-SEO) เฉลี่ยสูงที่ 89.54 คะแนน สะท้อนให้เห็นว่ามีการปรับปรุงเว็บไซต์ตามมาตรฐาน SEO อย่างมีประสิทธิภาพ

3. กลุ่มปัจจัยด้านสื่อที่ได้รับ (Earned Media)

กลุ่มนี้แสดงถึงความสามารถของ startup ในการสร้างการกล่าวถึงจากสื่อภายนอกหรือบุคคลที่สาม โดยพบว่า:

- จำนวนบทความหรือข่าวที่ถูกเผยแพร่ (EM-Article) มีค่าเฉลี่ยอยู่ที่ 26.85 บทความ แต่มีค่าการกระจายสูงมาก (ค่าเบี่ยงเบนมาตรฐานประมาณ 76.68 บทความ) ซึ่งแสดงให้เห็นถึงความแตกต่างในความสามารถในการได้รับการกล่าวถึงจากสื่อ
- คะแนนเรตติ้งหรือการประเมินจากบุคคลภายนอก (EM-Rating) มีค่าเฉลี่ยที่ 46.22 คะแนน
- จำนวน Backlink (EM-Backlink) และการถูกอ้างอิงโดยเว็บไซต์อื่น (EM-Linking) มีค่าเฉลี่ยที่ 300.26 และ 1,204.08 ครั้ง ตามลำดับ แต่มีค่าความผันแปรที่สูงมาก (ค่าเบี่ยงเบนมาตรฐานสูง) ซึ่งชี้ให้เห็นว่ามี startup เพียงบางส่วนที่สามารถสร้างการเชื่อมโยงและการกล่าวถึงได้อย่างมีนัยสำคัญ

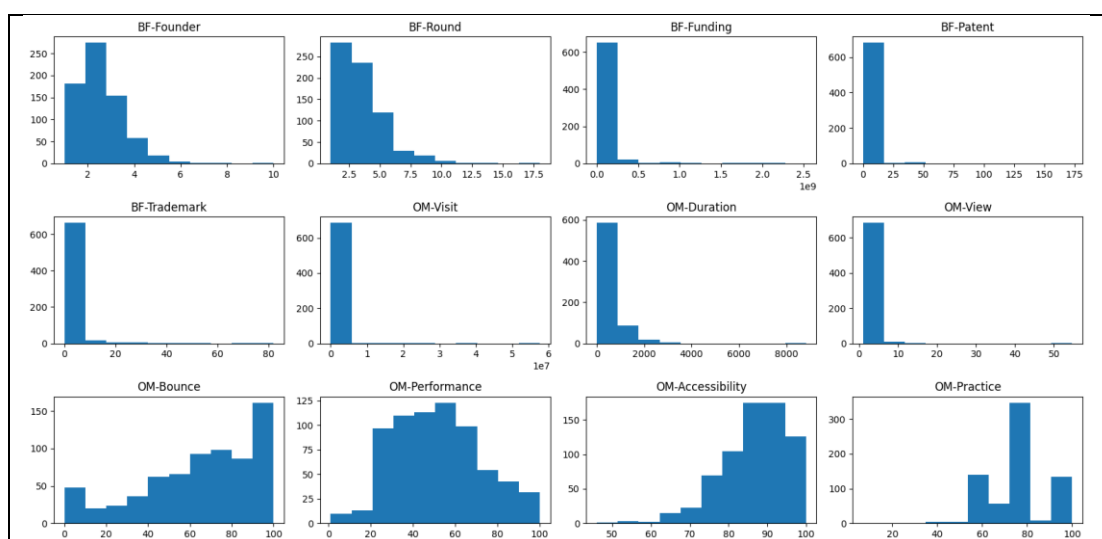
โดยสรุป ข้อมูลนี้สะท้อนให้เห็นถึงความหลากหลายและการกระจายตัวของตัวแปรต่างๆ อย่างชัดเจน ทั้งในมิติของโครงสร้างพื้นฐานของธุรกิจ การใช้งานสื่อที่เป็นเจ้าของ และความสามารถในการดึงดูดความสนใจผ่านสื่อภายนอก ซึ่งสามารถนำไปใช้ในการวิเคราะห์และออกแบบกลยุทธ์ที่เหมาะสมในการพัฒนา startup ต่อไปได้

ตารางที่ 3.5 ตารางสรุปภาพรวมของข้อมูล Data Set

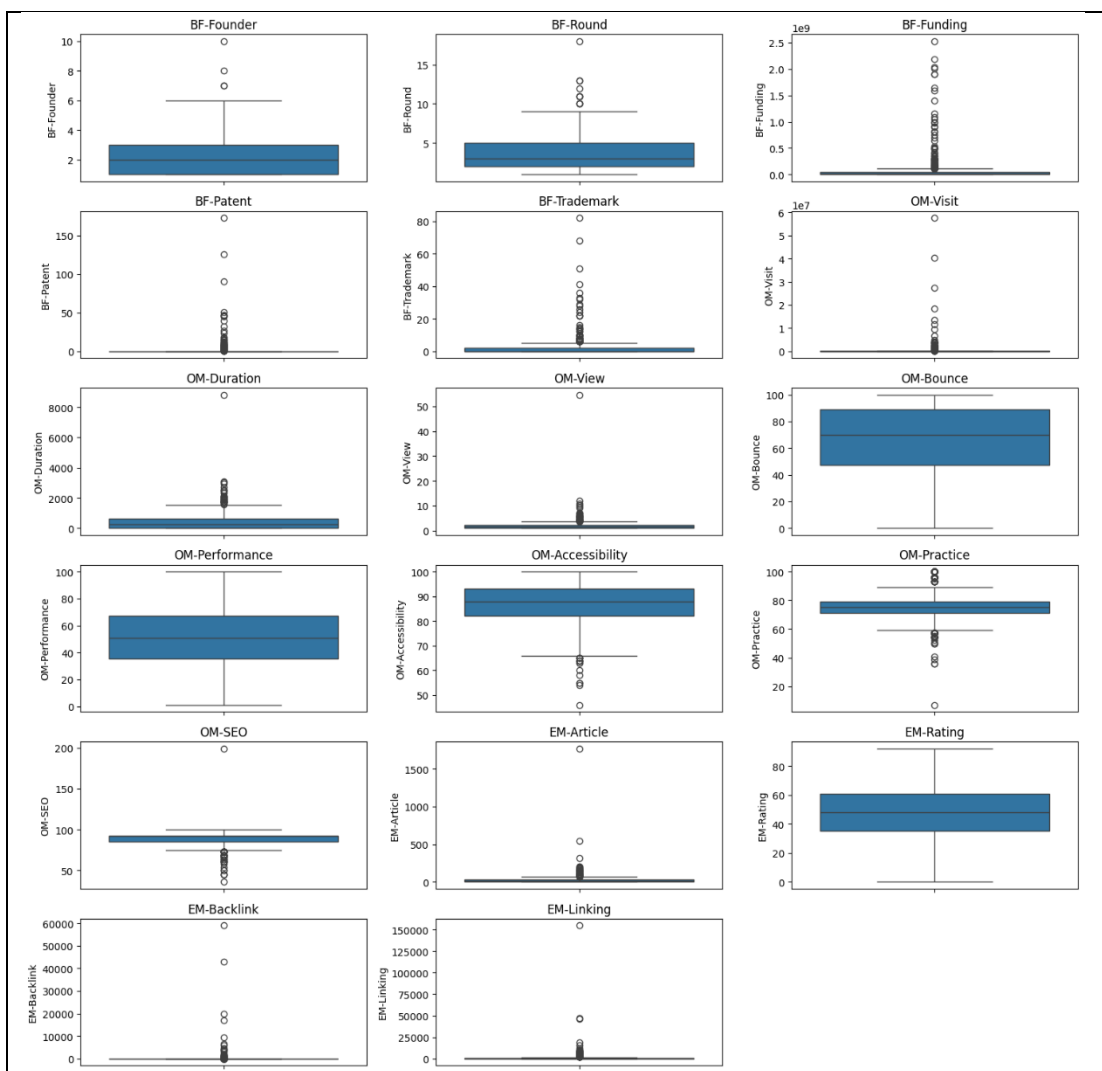
Feature	count	mean	std	min	25%	50%	75%	max
BF-Founder	694	2.27	1.14	1	1	2.00	3.00	10.00
BF-Round	694	3.39	2.15	1	2	3.00	5.00	18.00
BF-Funding	694	80,562,240	252,036,100	15000	5220000	16,525,600	46,321,690	2,529,292,000
BF-Patent	694	1.66	10.09	0	0	0.00	0.00	173.00
BF-Trademark	694	2.00	6.09	0	0	0.00	2.00	82.00
OM-Visit	694	480,390.70	3,091,476	109	1901.5	10,564.50	98,100.50	57,544,020
OM-Duration	694	435.30	616.74	0	25	228.50	630.50	8,828.00
OM-View	694	2.02	2.35	1	1.18	1.61	2.18	54.65
OM-Bounce	694	65.09	28.63	0	47.18	69.95	88.84	100.00
OM-Performance	694	51.95	20.50	1	35.25	51.00	67.00	100.00
OM-Accessibility	694	86.53	8.46	46	82	88.00	93.00	100.00
OM-Practice	694	76.23	13.85	7	71	75.00	79.00	100.00

ตารางที่ 3.5 ตารางสรุปภาพรวมของข้อมูล Data Set (ต่อ)

Feature	count	mean	std	min	25%	50%	75%	max
OM-SEO	694	89.54	10.20	36	85	92.00	92.00	199.00
EM-Article	694	26.85	76.68	0	4	13.00	31.00	1,772.00
EM-Rating	694	46.22	19.32	0	35	48.00	61.00	92.00
EM-Backlink	694	300.26	2,996.77	0	0.435	1.80	6.98	59,000.00
EM-Linking	694	1,204.08	6,560.19	0	115.75	345.00	837.00	155,000.00



รูปที่ 3.9 แผนภาพแสดง ฮิสโทแกรม (Histogram) ของชุดข้อมูล



รูปที่ 3.10 แผนภาพกล่อง (Box plot) ของชุดข้อมูล

การพิจารณาการมีอยู่ของข้อมูลผิดปกติ (Outlier)

ในการวิเคราะห์ข้อมูลเชิงสถิติเกี่ยวกับปัจจัยที่ส่งผลต่อความสำเร็จของ Startup ผู้วิจัยพบว่า ชุดข้อมูลมีการกระจายตัวสูงผิดปกติในบางตัวแปร เช่น ยอดเงินทุนที่ได้รับ จำนวนผู้เยี่ยมชมเว็บไซต์ และจำนวน Backlink ที่ได้รับจากเว็บไซต์อื่นๆ ซึ่งลักษณะดังกล่าวอาจเข้าข่ายข้อมูลผิดปกติหรือ Outlier เนื่องจากมีค่าที่แตกต่างจากข้อมูลส่วนใหญ่อย่างชัดเจน

อย่างไรก็ตาม ผู้วิจัยเลือกที่จะไม่ทำการตัดหรือจัดการข้อมูลที่เป็น Outlier ด้วยเหตุผลสำคัญสองประการ ประการแรก คือข้อมูลที่ปรากฏเป็นข้อมูลที่เกิดจากสถานการณ์จริง (Real-world Scenario) ที่สามารถเกิดขึ้นได้ในบริบทของธุรกิจ Startup เช่น Startup บางแห่งสามารถระดมทุนได้ในจำนวนที่สูงมาก หรือสามารถสร้างการมีส่วนร่วม (Engagement) บนเว็บไซต์ได้เป็นจำนวน

มหาศาล ซึ่งเป็นข้อมูลที่มีคุณค่าในการศึกษาและสะท้อนให้เห็นถึงความเป็นจริงของระบบนิเวศ Startup

ประการที่สอง คือเพื่อป้องกันการเกิดปัญหาการเรียนรู้เกินพอดีหรือ Overfitting ในการสร้างโมเดลเชิงพยากรณ์ เนื่องจากการลบหรือปรับข้อมูลผิดปกติออก อาจทำให้โมเดลที่ได้มีประสิทธิภาพลดลงเมื่อถูกใช้งานจริง เพราะสูญเสียความสามารถในการรองรับสถานการณ์สุดขั้ว (Extreme Case) ที่อาจเกิดขึ้นในอนาคต ดังนั้น การคงไว้ซึ่งข้อมูลผิดปกติเหล่านี้จะช่วยให้โมเดลมีความทนทาน (Robustness) และสามารถรับมือกับความผันผวนของข้อมูลในสภาพแวดล้อมที่เป็นจริง (Real-world Business Context) ได้ดีกว่า

การแปลงข้อมูล (Data Transformation)

การแปลงข้อมูล (Data Transformation) เป็นกระบวนการสำคัญที่สนับสนุนความสามารถของแบบจำลองการเรียนรู้ของเครื่อง ในการวิเคราะห์ข้อมูลอย่างมีประสิทธิภาพ กระบวนการนี้ไม่เพียงช่วยปรับปรุงคุณภาพของข้อมูล แต่ยังช่วยให้ข้อมูลเหมาะสมกับสมมติฐานเฉพาะของโมเดลแต่ละแบบ โดยมุ่งเน้นการแปลงข้อมูลสำหรับแบบจำลองการเรียนรู้ของเครื่อง จำนวน 9 แบบ โดยเน้นย้ำความสำคัญของการเลือกวิธีการที่เหมาะสมที่สุดสำหรับแต่ละโมเดล

1. Random Forest

Random Forest มีความทนทานต่อความหลากหลายของข้อมูล แต่ยังคงต้องการการจัดการค่าผิดปกติและข้อมูลที่ไม่สมบูรณ์อย่างเหมาะสม วิธีการที่ใช้ ได้แก่

1.1. การควบคุมค่าสุดโต่ง ใช้ IQR-Based Filtering เพื่อลดผลกระทบของค่าผิดปกติที่อาจทำให้โมเดลเรียนรู้ผิดพลาด

1.2. แปลงตัวแปรเชิงปริมาณให้เป็นข้อมูลในรูปแบบเลขฐานสอง (Binary) อาทิ การมีหรือไม่มีสิทธิบัตร เพื่อความเรียบง่ายในการประมวลผล

2. XGBoost

XGBoost มีความสามารถสูงในการจัดการข้อมูลที่ไม่สมดุลและมีค่าผิดปกติ โดยแนวทางที่เหมาะสม ได้แก่:

2.1. การแปลงค่าลอการิทึม (Log Transformation) ช่วยลด skewness ของตัวแปรที่มีการกระจายข้อมูลไม่สมมาตร อาทิ ยอดเงินลงทุนรวม

2.2. การควบคุมค่าสุดโต่ง อาทิเดียวกับ Random Forest เพื่อลดผลกระทบจากค่าผิดปกติ

2.3. แปลงตัวแปรเชิงปริมาณให้เป็นข้อมูลในรูปแบบเลขฐานสอง (Binary) อาทิเดียวกับ Random Forest

3. Gradient Boosting

XGBoost มีความสามารถสูงในการจัดการข้อมูลที่ไม่สมดุลและมีค่าผิดปกติ โดยแนวทางที่เหมาะสม ได้แก่:

3.1. ใช้ One-Hot Encoding แปลงตัวแปรหมวดหมู่ อาทิ จำนวนผู้ก่อตั้ง ให้กลายเป็นตัวแปรใหม่ที่มีความเป็นอิสระ

3.2. **Imputation:** เติมค่าที่ขาดหายด้วยค่าเฉลี่ย เพื่อให้ข้อมูลสมบูรณ์มากขึ้น

4. AdaBoost

AdaBoost มีความต้องการข้อมูลที่ถูกปรับขนาดอย่างเหมาะสม แนวทางที่ใช้ได้แก่:

4.1. **RobustScaler และ MinMaxScaler:** ใช้ปรับค่าที่มีขอบเขตจำกัด อาทิ คะแนนเว็บไซต์ เพื่อให้โมเดลทำงานได้อย่างมีประสิทธิภาพ

4.2. **One-Hot Encoding:** อาทิเดียวกับ Gradient Boosting เพื่อเพิ่มความละเอียดในการวิเคราะห์

4.3. **Standard Scaling:** ใช้กับตัวแปรเชิงปริมาณที่มีการกระจายค่ากว้าง

5. MLP (Multi-Layer Perceptron)

MLP มีความต้องการข้อมูลที่มีการกระจายสมมาตรและปราศจากค่าผิดปกติ โดยแนวทางที่เหมาะสม ได้แก่:

5.1. **Log Transformation:** ช่วยปรับสมดุลของตัวแปรที่มี skewness สูง

5.2. **Standardization:** ปรับสเกลตัวแปรเชิงปริมาณเพื่อความเหมาะสมกับการประมวลผลของโมเดล

6. K-Nearest Neighbors (KNN)

KNN อ่อนไหวต่อค่าผิดปกติและข้อมูลที่ไม่สมดุล วิธีการที่เหมาะสม ได้แก่:

6.1. **Capping Outliers:** ใช้ IQR-Based Filtering เพื่อลดผลกระทบจากค่าผิดปกติ

6.2. **Standardization:** ปรับสเกลข้อมูลให้เหมาะสม

7. Logistic Regression

Logistic Regression มีความต้องการข้อมูลที่ปรับสเกลและมีการกระจายสมมาตร:

7.1. **Log Transformation:** ช่วยลด skewness ของตัวแปร อาทิ ยอดเงินทุน

7.2. **Standardization:** ใช้ปรับขนาดตัวแปรเชิงปริมาณทั้งหมดเพื่อเพิ่มความแม่นยำ

8. Decision Tree

Decision Tree มีความยืดหยุ่นต่อการจัดการข้อมูล แต่ยังคงต้องการการปรับเปลี่ยนในบางกรณี:

8.1. **Binary Conversion:** แปลงข้อมูลที่มีหลายศูนย์ให้เป็นไบนารี เพื่อเพิ่มความชัดเจน

8.2. **Label Encoding:** แปลงตัวแปรหมวดหมู่ให้เป็นตัวเลขเพื่อการวิเคราะห์ที่แม่นยำ

9. Support Vector Machine (SVM)

SVM มีความต้องการข้อมูลที่ปรับสเกลอย่างเหมาะสมเพื่อเพิ่มความแม่นยำ:

9.1. **Standard Scaling และ MinMax Scaling:** ใช้ปรับค่าของตัวแปรเชิงปริมาณ

9.2. **Log Transformation:** ลด skewness ของตัวแปรที่มีการกระจายข้อมูลไม่สมมาตร

การเลือกวิธีการแปลงข้อมูลที่เหมาะสมกับลักษณะเฉพาะของแต่ละโมเดล Machine Learning เป็นปัจจัยสำคัญที่ส่งผลต่อความสำเร็จของการพยากรณ์ ตัวอย่างที่กล่าวถึงในบทความนี้แสดงให้เห็นถึงผลกระทบของการปรับแต่งกระบวนการแปลงข้อมูลที่สอดคล้องกับลักษณะเฉพาะของข้อมูลและโมเดล การทำความเข้าใจและเลือกวิธีที่เหมาะสมที่สุดจึงเป็นหัวใจสำคัญในการเพิ่มประสิทธิภาพของระบบ Machine Learning

ตารางที่ 3.6 ลักษณะเฉพาะตัวของแต่ละแบบจำลอง

Model	การแปลงข้อมูลที่เหมาะสม	Normalization	เหตุผล
Random Forest	ไม่จำเป็นต้องทำ	ไม่จำเป็น	สามารถรับมือกับข้อมูลที่ไม่ได้มาตรฐานได้ดี เนื่องจากใช้การสร้างต้นไม้ตัดสินใจจำนวนมาก
Gradient Boosting	ไม่จำเป็นต้องทำ	ไม่จำเป็น	คล้ายกับ Random Forest สามารถรับมือกับข้อมูลที่ไม่ได้มาตรฐานได้ดี
AdaBoost	ไม่จำเป็นต้องทำ	ไม่จำเป็น	คล้ายกับ Random Forest และ Gradient Boosting
K-Nearest Neighbors	จำเป็นต้องทำ	จำเป็นต้องทำ	เนื่องจากโมเดลนี้พิจารณาจากระยะทางระหว่างจุดข้อมูล การทำ Normalization จะช่วยให้ระยะทางมีความหมายมากขึ้น
MLP Classifier (Neural Network)	จำเป็นต้องทำ	จำเป็นต้องทำ	Neural Network มักจะทำงานได้ดีขึ้นเมื่อข้อมูลถูกทำให้เป็นมาตรฐาน เนื่องจากการปรับน้ำหนักของ Neural Network จะง่ายขึ้น
XGBoost	ไม่จำเป็นต้องทำ	ไม่จำเป็น	คล้ายกับ Random Forest สามารถรับมือกับข้อมูลที่ไม่ได้มาตรฐานได้ดี
Logistic Regression	จำเป็นต้องทำ	จำเป็นต้องทำ	การทำ Normalization จะช่วยให้การหาค่าพารามิเตอร์ของ Logistic Regression มีประสิทธิภาพมากขึ้น
Decision Tree	ไม่จำเป็นต้องทำ	ไม่จำเป็น	สามารถรับมือกับข้อมูลที่ไม่ได้มาตรฐานได้ดี เนื่องจากใช้การสร้างต้นไม้ตัดสินใจ

ตารางที่ 3.6 ลักษณะเฉพาะตัวของแต่ละแบบจำลอง (ต่อ)

Model	การแปลงข้อมูลที่เหมาะสม	Normalization	เหตุผล
SVM	จำเป็นต้องทำ	จำเป็นต้องทำ	เนื่องจาก SVM พิจารณาจาก Margin ระหว่างกลุ่มข้อมูล การทำ Normalization จะช่วยให้ Margin มีความหมายมากขึ้น

3.5.4 การสร้างโมเดล (Modeling)

ในงานวิจัยนี้ ได้มีการประยุกต์ใช้กระบวนการ Cross Validation, Feature Selection และ Hyperparameter Tuning เพื่อพัฒนาโมเดลโดยเน้นการจัดการปัญหาความไม่สมดุลของข้อมูล (Imbalance Data) เพื่อให้ผลลัพธ์ที่ได้มีความแม่นยำและสามารถใช้งานได้จริงในสถานการณ์ที่ข้อมูลมีการกระจายตัวของคลาสไม่เท่ากัน

1. กระบวนการ (Cross Validation)

Cross Validation เป็นวิธีการสำคัญที่ใช้ประเมินความสามารถของโมเดล โดยแบ่งข้อมูลออกเป็นชุดสำหรับการฝึกและทดสอบหลายรอบ เพื่อลดโอกาสการเกิดการเรียนรู้ที่ลำเอียง หรือ Overfitting

1.1. มีการใช้ StratifiedKFold เพื่อให้สัดส่วนของคลาสในเป้าหมายยังคงอยู่เท่าเดิมในแต่ละ fold

1.2. จำนวน fold ที่ใช้ในงานนี้มีการกำหนดไว้เป็น 5 หรือ 10 ขึ้นอยู่กับลักษณะของข้อมูล และในบางกรณี การใช้ RepeatedStratifiedKFold จะช่วยเพิ่มความน่าเชื่อถือของผลลัพธ์

1.3. ตัวชี้วัดที่ใช้วัดประสิทธิภาพโมเดล ได้แก่ Accuracy, Precision, Recall, และ F1-Score เพื่อให้ครอบคลุมมิติของการประเมิน

2. การจัดการข้อมูลที่ไม่สมดุล (Imbalanced Data)

2.1. สำหรับข้อมูลที่ไม่สมดุล เทคนิคที่นิยมใช้ได้แก่:

2.1.1. Oversampling โดยใช้ SMOTE (Synthetic Minority Oversampling Technique) เพื่อเพิ่มจำนวนตัวอย่างในคลาสที่มีน้อย

2.1.2. Undersampling เพื่อลดจำนวนตัวอย่างในคลาสที่มีมากเกินไป

2.1.3. Class Weights เพื่อเพิ่มน้ำหนักให้กับคลาสที่มีตัวอย่างน้อยในขั้นตอนการฝึก อาทิ ใช้ class_weight='balanced' สำหรับโมเดลที่รองรับ

2.2. ใช้ Precision-Recall Curve หรือ ROC-AUC เหมาะสมในการวัดผลลัพธ์เมื่อข้อมูลมีความไม่สมดุลสูง

3. การคัดเลือกคุณลักษณะ (Feature Selection)

กระบวนการคัดเลือกคุณลักษณะมีบทบาทสำคัญในการปรับปรุงประสิทธิภาพของแบบจำลอง โดยวิธีที่ใช้ ได้แก่

3.1. Tree-Based Models: ใช้คุณลักษณะที่มีความสำคัญสูงซึ่งสามารถคำนวณได้โดยตรงจากโมเดลต้นไม้ อาทิ Random Forest และ Gradient Boosting

3.2. Permutation Importance: วัดการเปลี่ยนแปลงในประสิทธิภาพของโมเดลเมื่อลดความสำคัญของคุณลักษณะบางอย่าง เหมาะกับโมเดล Logistic Regression หรือ MLP

3.3. การวิเคราะห์ทางสถิติ: ใช้การวิเคราะห์ความสัมพันธ์ระหว่างคุณลักษณะกับเป้าหมาย และตัดคุณลักษณะที่มีความแปรปรวนต่ำหรือความสัมพันธ์ซ้ำซ้อนออก

3.4. การวิเคราะห์จากค่าสัมประสิทธิ์ (Coefficient) ของ Logistic Regression และทำการ Normalize ค่าความสำคัญสัมบูรณ์ (Absolute Importance) ของแต่ละฟีเจอร์ให้รวมกันเป็น 1 เพื่อให้เปรียบเทียบได้ง่ายขึ้น

โดยทำการคัดเลือกคุณลักษณะที่มีคะแนนความสำคัญสูงกว่าค่าเฉลี่ยของคะแนนความสำคัญของคุณลักษณะทั้งหมดของแบบจำลองนั้นๆ การดำเนินการดังกล่าวช่วยลดจำนวนคุณลักษณะที่ไม่จำเป็นในชุดข้อมูล การลดจำนวนคุณลักษณะส่งผลให้แบบจำลองมีความซับซ้อนลดลง และสามารถลดความเสี่ยงของการเกิดภาวะการเรียนรู้เกิน (Overfitting) ซึ่งเป็นปัญหาที่แบบจำลองมีความสามารถในการจดจำข้อมูลฝึกฝนได้ดีเกินไป แต่ไม่สามารถนำไปใช้กับข้อมูลใหม่ได้อย่างมีประสิทธิภาพ โดยแบ่งเป็น 2 แบบ ได้แก่ แบบที่ 1 คือ การใช้ข้อมูลแบบผสมผสาน ซึ่งใช้ข้อมูลพื้นฐานทางธุรกิจและข้อมูลประสิทธิผลการปรากฏตัวทางดิจิทัล แบบที่ 2 คือ การใช้เฉพาะข้อมูลประสิทธิผลการปรากฏตัวทางดิจิทัล ดังแสดงในตารางที่ 3.6

ตารางที่ 3.7 ผลการคัดเลือกคุณลักษณะ (Feature Selection) ของแบบจำลอง

แบบจำลอง	แบบผสมผสาน		แบบที่ใช้เฉพาะการปรากฏตัวทางดิจิทัล	
Random Forest	EM-Article	16.754 %	EM-Article	19.698 %
	BF-Funding	11.628 %	OM-Duration	14.446 %
	OM-Duration	11.200 %	OM-Bounce	10.382 %
	OM-Bounce	10.606 %	OM-View	9.818 %
	EM-Linking	7.038 %	EM-Rating	9.354 %
	EM-Rating	6.892 %		
	EM-Backlink	6.858 %		
	OM-View	6.670 %		
XGBoost	OM-Visit	6.536 %		
	EM-Article	20.257 %	EM-Article	25.330 %
	OM-Duration	17.333 %	OM-Duration	20.563 %
	OM-View	9.200 %	OM-View	9.270 %
	BF-Funding	6.250 %		

ตารางที่ 3.7 ผลการคัดเลือกคุณลักษณะ (Feature Selection) ของแบบจำลอง (ต่อ)

แบบจำลอง	แบบผสมผสาน		แบบที่ใช้เฉพาะการปรากฏตัวทางดิจิทัล	
Gradient Boosting	EM-Article	29.934 %	EM-Article	34.952 %
	BF-Funding	14.664 %	OM-Duration	14.960 %
	OM-Bounce	14.436 %	OM-Bounce	14.572 %
	OM-Duration	13.270 %		
	OM-Performance	4.554 %		
	EM-Backlink	4.144 %		
	OM-Visit	3.944 %		
	EM-Rating	3.706 %		
	EM-Linking	3.290 %		
	OM-View	3.168 %		
AdaBoost	OM-Bounce	22.780 %	OM-Bounce	23.020 %
	BF-Funding	19.250 %	EM-Article	12.054 %
	EM-Article	10.142 %	OM-Duration	10.300 %
	OM-Duration	9.066 %	EM-Rating	9.808 %
	OM-Performance	8.014 %	EM-Backlink	8.802 %
	EM-Backlink	6.016 %		
MLP	EM-Rating	30.348 %	EM-Rating	43.718 %
	EM-Backlink	13.791 %	EM-Linking	15.142 %
	EM-Linking	13.583 %	OM-Performance	12.712 %
	OM-Duration	9.627 %	EM-Backlink	10.000 %
	BF-Funding	7.698 %		
KNeighbors	EM-Rating	15.800 %	EM-Rating	17.150 %
	OM-Bounce	11.565 %	EM-Article	12.638 %
	EM-Article	9.914 %	OM-Bounce	10.845 %
	BF-Funding	8.401 %	OM-Performance	9.405 %
	OM-View	7.607 %	OM-Duration	9.168 %
	OM-Performance	7.445 %		
	OM-Duration	6.997 %		

ตารางที่ 3.7 ผลการคัดเลือกคุณลักษณะ (Feature Selection) ของแบบจำลอง (ต่อ)

แบบจำลอง	แบบผสมผสาน		แบบที่ใช้เฉพาะการปรากฏตัวทางดิจิทัล	
Logistic Regression	EM-Article	14.890 %	EM-Article	29.542 %
	EM-Rating	14.612 %	EM-Rating	19.633 %
	BF-Funding	13.628 %	OM-Visit	13.359 %
	OM-Visit	12.360 %	OM-Performance	9.187 %
	EM-Linking	10.408 %		
	OM-Performance	8.564 %		
	OM-Duration	6.644 %		
	EM-Backlink	6.523 %		
Decision Tree	EM-Article	30.212 %	EM-Article	31.096 %
	OM-Duration	14.056 %	OM-Duration	15.350 %
	OM-Bounce	12.176 %	OM-Bounce	11.992 %
	BF-Funding	9.576 %		
	OM-Visit	7.796 %		
SVM	BF-Funding	31.001 %	EM-Rating	51.929 %
	EM-Rating	26.535 %	EM-Article	24.169 %
	OM-Performance	19.185 %	OM-Performance	17.245 %
	OM-Practice	7.121 %		
	OM-Duration	6.098 %		

4. การปรับค่าพารามิเตอร์ (Hyperparameter Tuning)

กระบวนการปรับค่าพารามิเตอร์ช่วยเพิ่มความแม่นยำของโมเดล โดยการค้นหาค่าที่เหมาะสมที่สุดผ่านวิธีต่าง ๆ ได้แก่:

4.1. กระบวนการปรับแต่ง Hyperparameter

4.1.1. ใช้ RandomizedSearchCV และ GridSearchCV เพื่อค้นหาค่าพารามิเตอร์ที่เหมาะสมที่สุดในช่วงที่กำหนด โดยเลือกวิธีการขึ้นอยู่กับขนาดข้อมูลและทรัพยากรการคำนวณ

4.1.2. ตัวอย่างพารามิเตอร์ที่มักปรับได้แก่ ความลึกของต้นไม้ (max_depth), จำนวนรอบการบูสต์ (n_estimators), และอัตราการเรียนรู้ (learning_rate)

4.2. การปรับแต่งร่วมกับ Cross Validation

4.2.1. ใช้ Cross Validation เพื่อประเมินผลการตั้งค่า Hyperparameter ที่เลือกไว้ และปรับค่าให้เหมาะสมโดยพิจารณาค่าเฉลี่ย (Mean) และค่าความเบี่ยงเบนมาตรฐาน (Standard Deviation) จากผลการทดสอบในแต่ละ fold

การทำ Cross Validation การเลือกคุณลักษณะ และการปรับแต่ง Hyperparameter อย่างเป็นระบบเป็นหัวใจสำคัญของการพัฒนาโมเดลที่มีประสิทธิภาพสูง การออกแบบกระบวนการเหล่านี้ช่วยให้โมเดลสามารถทำนายได้อย่างแม่นยำ รองรับข้อมูลที่ไม่สมดุล และลดโอกาสการเกิด Overfitting อีกทั้งยังสามารถนำไปปรับใช้ในสถานการณ์ที่มีทรัพยากรการคำนวณจำกัดได้อย่างมีประสิทธิภาพ ดังแสดงในตารางที่ 3.7 และตารางที่ 3.8

ตารางที่ 3.8 ผลวิเคราะห์การปรับแต่งไฮเปอร์พารามิเตอร์ของแบบจำลองที่ใช้ข้อมูลแบบผสมผสาน

แบบจำลอง	ไฮเปอร์พารามิเตอร์
Random Forest	'bootstrap': True 'criterion': 'gini' 'max_depth': 50 'min_samples_leaf': 1 'min_samples_split': 5 'n_estimators': 99
XGBoost	'colsample_bytree': 0.9895022075365837 'gamma': 0.23277134043030423 'learning_rate': 0.09970037018749259 'max_depth': 8 'min_child_weight': 3 'n_estimators': 157 'reg_alpha': 0.5142344384136116 'reg_lambda': 0.5924145688620425 'scale_pos_weight': 0.3907815631262525 'subsample': 0.6185801650879991
Gradient Boosting	'learning_rate': 0.26268543237849956 'max_depth': 6 'max_features': 'log2' 'min_samples_leaf': 2 'min_samples_split': 5 'n_estimators': 63 'subsample': 0.8909087983425683
AdaBoost	'learning_rate': 0.146553541187277 'n_estimators': 87

ตารางที่ 3.8 ผลวิเคราะห์การปรับแต่งไฮเปอร์พารามิเตอร์ของแบบจำลองที่ใช้ข้อมูลแบบผสมผสาน (ต่อ)

แบบจำลอง	ไฮเปอร์พารามิเตอร์
MLP	'activation': 'relu' 'alpha': 0.000877781550471965 'hidden_layer_sizes': (100,) 'learning_rate': 'adaptive' 'learning_rate_init': 0.030634622106220845 'solver': 'sgd'
KNN	'metric': 'manhattan' 'n_neighbors': 9 'weights': 'uniform'
KNN	'metric': 'manhattan' 'n_neighbors': 9 'weights': 'uniform'
Logistic Regression	'C': 33.371861113902185 'class_weight': None 'penalty': 'l2' 'solver': 'liblinear'
Decision Tree	'class_weight': None 'criterion': 'entropy' 'max_depth': 15 'min_samples_leaf': 5 'min_samples_split': 2 'splitter': 'random'

ตารางที่ 3.9 ผลวิเคราะห์การปรับแต่งไฮเปอร์พารามิเตอร์ของแบบจำลองที่ใช้เฉพาะข้อมูลประสิทธิผลการปรากฏตัวทางดิจิทัล

แบบจำลอง	ไฮเปอร์พารามิเตอร์
Random Forest	'bootstrap': True 'criterion': 'gini' 'max_depth': 40 'min_samples_leaf': 1 'min_samples_split': 8 'n_estimators': 58

ตารางที่ 3.9 ผลวิเคราะห์การปรับแต่งไฮเปอร์พารามิเตอร์ของแบบจำลองที่ใช้เฉพาะข้อมูล
ประสิทธิภาพการปรากฏตัวทางดิจิทัล (ต่อ)

แบบจำลอง	ไฮเปอร์พารามิเตอร์
XGBoost	'colsample_bytree': 0.9687496940092467 'gamma': 0.0884925020519195 'learning_rate': 0.20402303379495376 'max_depth': 3 'min_child_weight': 8 'n_estimators': 131 'reg_alpha': 0.5396921323890798 'reg_lambda': 0.5867511656638482 'scale_pos_weight': 0.3907815631262525 'subsample': 0.9861021229056552
Gradient Boosting	'learning_rate': 0.013979488347959958 'max_depth': 3 'max_features': 'log2' 'min_samples_leaf': 2 'min_samples_split': 10 'n_estimators': 139 'subsample': 0.6390688456025535
AdaBoost	'learning_rate': 0.023864188780056052 'n_estimators': 166
MLP	'activation': 'tanh' 'alpha': 0.0013359166801651823 'hidden_layer_sizes': (100, 100) 'learning_rate': 'invscaling' 'learning_rate_init': 0.005047786565710607 'solver': 'adam'
KNN	'metric': 'euclidean' 'n_neighbors': 7 'weights': 'uniform'

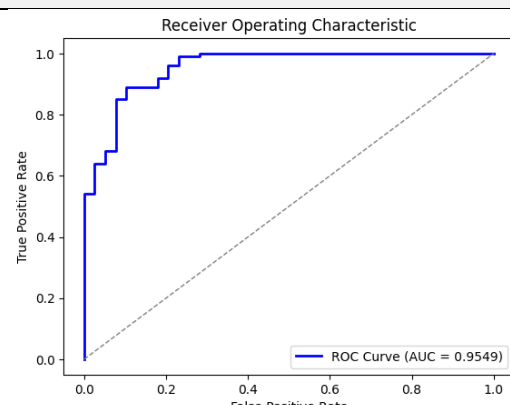
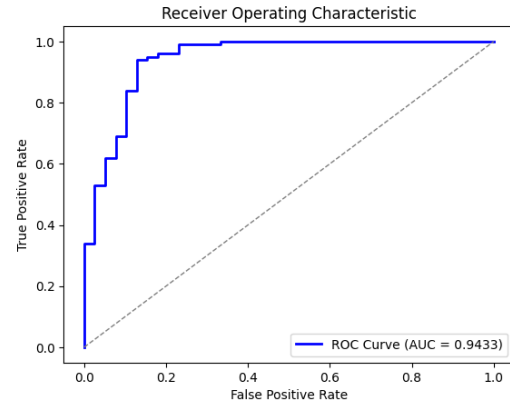
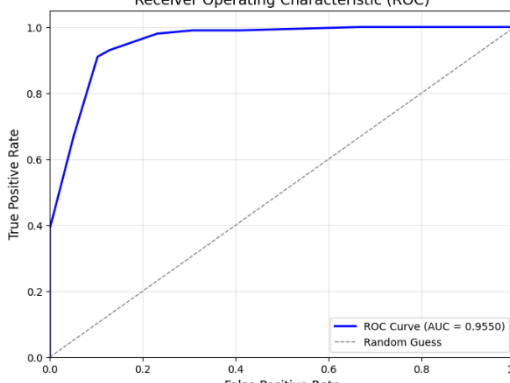
ตารางที่ 3.9 ผลวิเคราะห์การปรับแต่งไฮเปอร์พารามิเตอร์ของแบบจำลองที่ใช้เฉพาะข้อมูล
ประสิทธิภาพการปรากฏตัวทางดิจิทัล (ต่อ)

แบบจำลอง	ไฮเปอร์พารามิเตอร์
Logistic Regression	'C': 33.371861113902185 'class_weight': None 'penalty': 'l2' 'solver': 'liblinear'
Decision Tree	'class_weight': None 'criterion': 'entropy' 'max_depth': 20 'min_samples_leaf': 9 'min_samples_split': 18 'splitter': 'best'
SVM	'C': 2.406242504141576 'degree': 4 'gamma': 0.6128528947223795 'kernel': 'rbf'

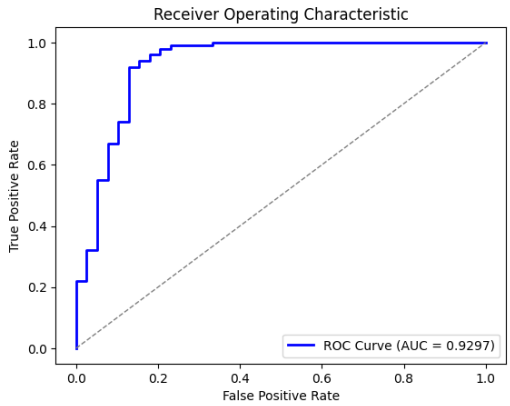
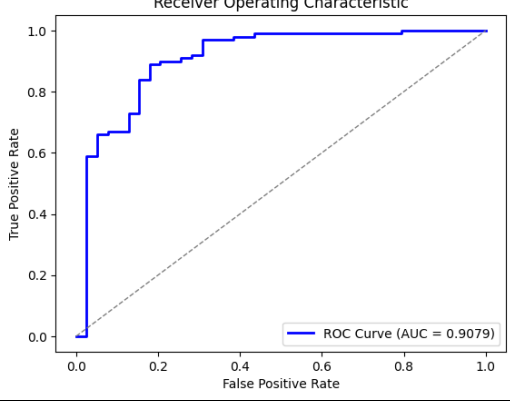
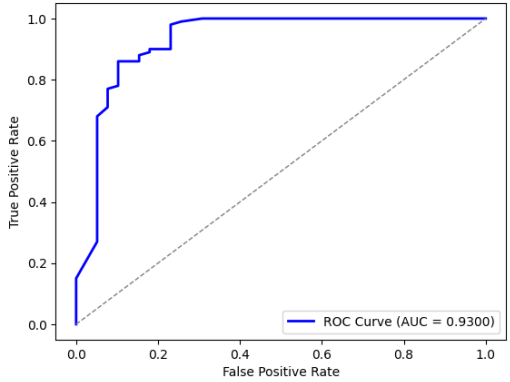
3.5.5 การวัดประสิทธิภาพของโมเดล (Evaluation)

หลังจากสร้างโมเดลแล้ว จะมีการวัดประสิทธิภาพของแต่ละโมเดลโดยใช้ F1-Score ของตัวแปรเป้าหมาย (Target Variables) ที่มีป้ายกำกับ Class เท่ากับ 1 ซึ่งหมายถึงสตาร์ทอัพที่ประสบความสำเร็จ และพิจารณา ROC Curve ซึ่งเป็นกราฟแสดงความสัมพันธ์ระหว่างอัตราการเกิดผลบวกเท็จ (False Positive) และผลบวกจริง (True Positive) และค่า AUC (Area Under Curve) ซึ่งเป็นค่าพื้นที่ใต้โค้ง เป็นตัวชี้วัดประสิทธิภาพของโมเดล ดังแสดงในตารางที่ 3.9 และตารางที่ 3.10

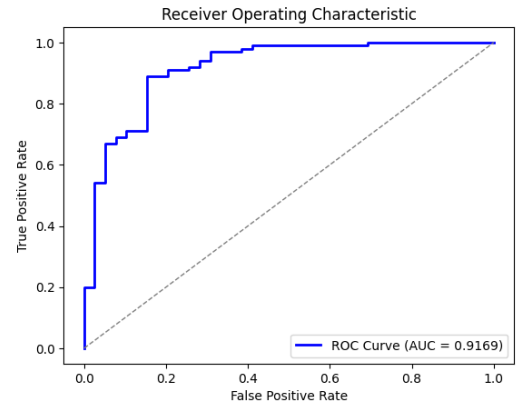
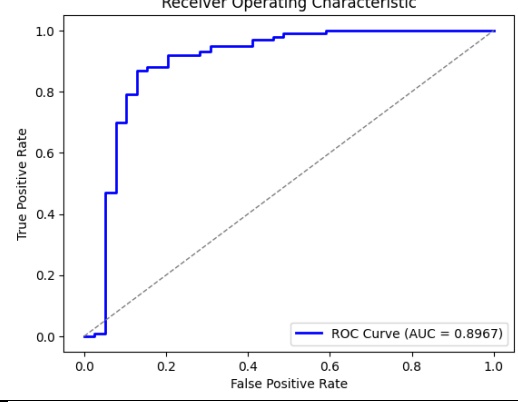
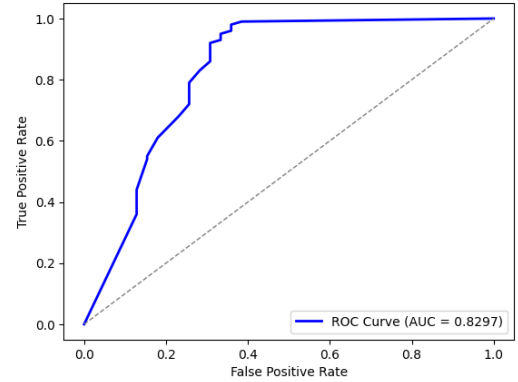
ตารางที่ 3.10 ประสิทธิภาพของแบบจำลองที่ใช้ข้อมูลแบบผสมผสาน

Model	F1-Score Class 1	AUC	ROC Curve
Random Forest	0.9412	0.9549	 Receiver Operating Characteristic True Positive Rate False Positive Rate ROC Curve (AUC = 0.9549)
Gradient Boosting	0.9400	0.9433	 Receiver Operating Characteristic True Positive Rate False Positive Rate ROC Curve (AUC = 0.9433)
KNeighbors	0.9394	0.9550	 Receiver Operating Characteristic (ROC) True Positive Rate False Positive Rate ROC Curve (AUC = 0.9550) Random Guess

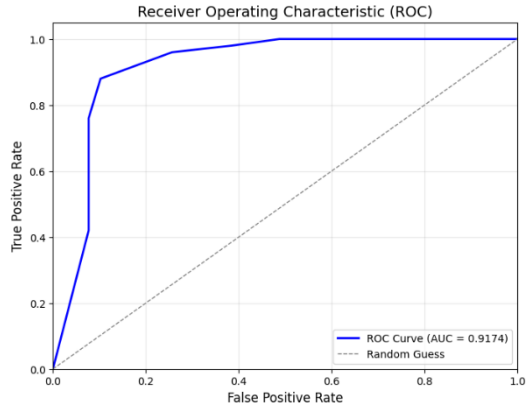
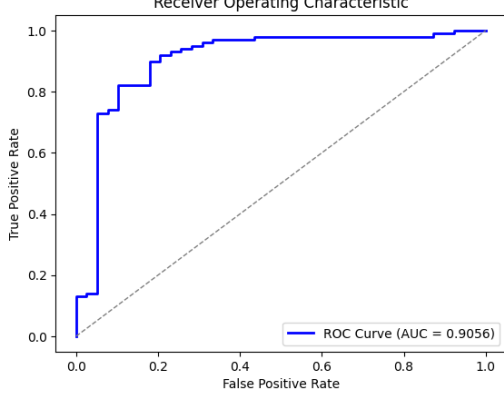
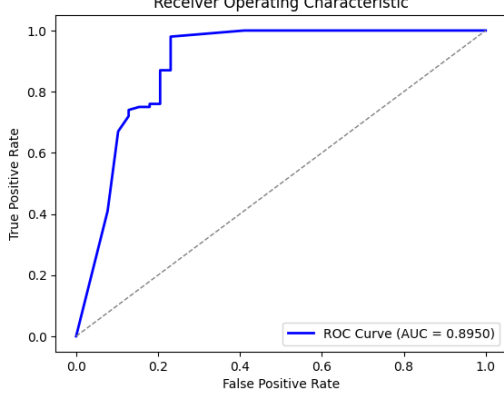
ตารางที่ 3.10 ประสิทธิภาพของแบบจำลองที่ใช้ข้อมูลแบบผสมผสาน (ต่อ)

Model	F1-Score Class 1	AUC	ROC Curve
XGBoost	0.9286	0.9297	
Logistic Regression	0.9179	0.9079	
AdaBoost	0.9154	0.9300	

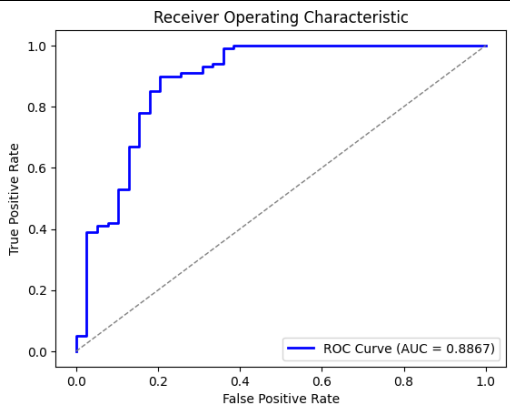
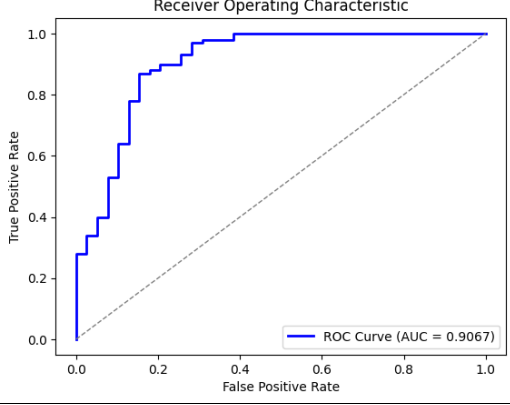
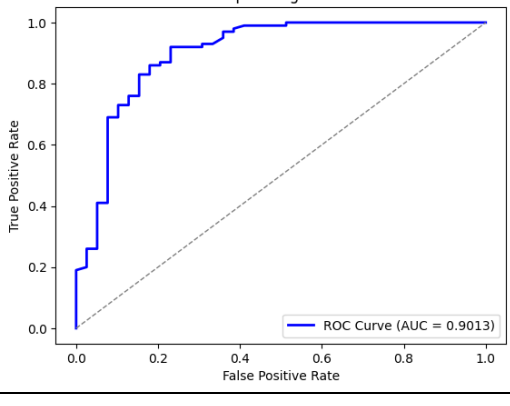
ตารางที่ 3.10 ประสิทธิภาพของแบบจำลองที่ใช้ข้อมูลแบบผสมผสาน (ต่อ)

Model	F1-Score Class 1	AUC	ROC Curve
SVM	0.9091	0.9169	 Receiver Operating Characteristic True Positive Rate False Positive Rate ROC Curve (AUC = 0.9169)
MLP	0.9048	0.8967	 Receiver Operating Characteristic True Positive Rate False Positive Rate ROC Curve (AUC = 0.8967)
Decision Tree	0.9020	0.8297	 Receiver Operating Characteristic True Positive Rate False Positive Rate ROC Curve (AUC = 0.8297)

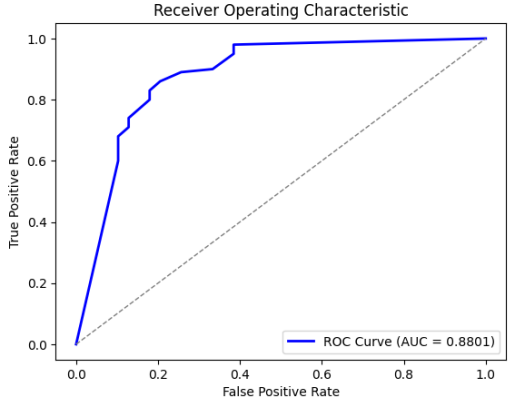
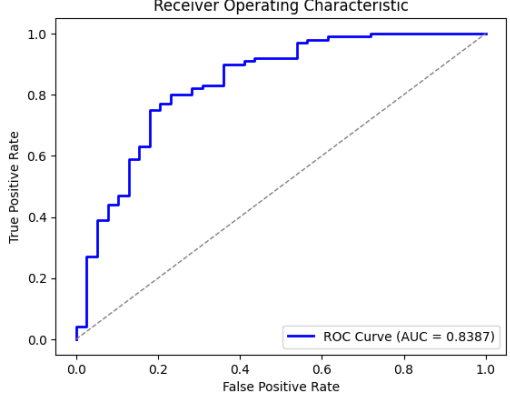
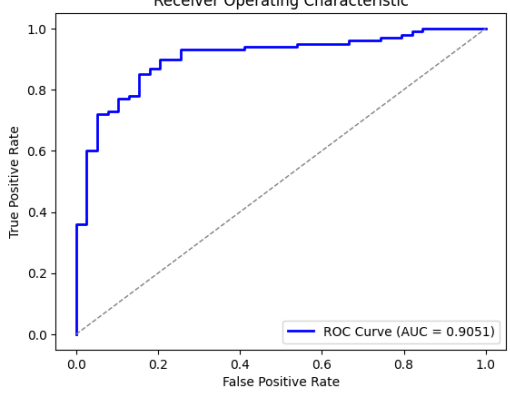
ตารางที่ 3.11 ประสิทธิภาพของแบบจำลองที่ใช้เฉพาะข้อมูลประสิทธิภาพการปรากฏตัวทางดิจิทัล

Model	F1-Score Class 1	AUC	ROC Curve
KNeighbors	0.9320	0.9174	 Receiver Operating Characteristic (ROC) ROC Curve (AUC = 0.9174) Random Guess
Logistic Regression	0.9208	0.9056	 Receiver Operating Characteristic ROC Curve (AUC = 0.9056)
AdaBoost	0.9154	0.8950	 Receiver Operating Characteristic ROC Curve (AUC = 0.8950)

ตารางที่ 3.11 ประสิทธิภาพของแบบจำลองที่ใช้เฉพาะข้อมูลประสิทธิภาพการปรากฏตัวทางดิจิทัล
(ต่อ)

Model	F1-Score Class 1	AUC	ROC Curve
Gradient Boosting	0.9091	0.8867	 Receiver Operating Characteristic True Positive Rate False Positive Rate ROC Curve (AUC = 0.8867)
Random Forest	0.9000	0.9067	 Receiver Operating Characteristic True Positive Rate False Positive Rate ROC Curve (AUC = 0.9067)
XGBoost	0.8878	0.9013	 Receiver Operating Characteristic True Positive Rate False Positive Rate ROC Curve (AUC = 0.9013)

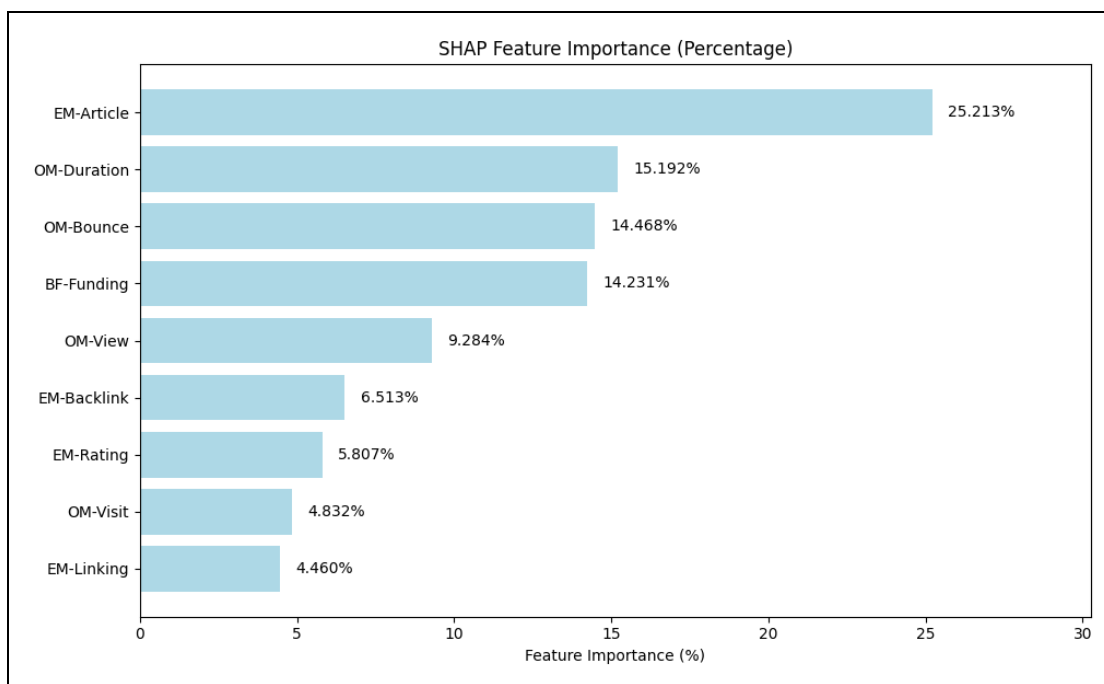
ตารางที่ 3.11 ประสิทธิภาพของแบบจำลองที่ใช้เฉพาะข้อมูลประสิทธิภาพการปรากฏตัวทางดิจิทัล
(ต่อ)

Model	F1-Score Class 1	AUC	ROC Curve
Decision Tree	0.8867	0.8801	
MLP	0.8720	0.8387	
SVM	0.8602	0.9051	

3.5.6 การนำโมเดลไปใช้งานจริง (Deployment)

แบบจำลอง Random Forest เป็นแบบจำลองที่มีประสิทธิภาพสูงสุด ซึ่งสามารถนำไปพัฒนาเป็นแบบจำลองที่ใช้งานจริงในบริบททางธุรกิจได้ โดยอาจพัฒนาเป็นเครื่องมือหรือระบบที่สามารถนำเสนอผลการทำนายได้อย่างสะดวกและรวดเร็ว เพื่อสนับสนุนการตัดสินใจของผู้ประกอบการสตาร์ทอัพและนักลงทุน ซึ่งจะต้องใช้เครื่องมือในการวิเคราะห์ห่อหุ้มลักษณะของข้อมูล

ว่ามีผลค่าที่แบบจำลองทำนายอย่างไร เพื่อผู้ใช้งานจะสามารถให้ความสำคัญกับคุณลักษณะที่มีอิทธิพลสูงต่อค่าที่แบบจำลองทำนาย ซึ่งการที่จะเข้าใจกระบวนการในการทำนายของแบบจำลอง Random Forest จำเป็นที่จะต้องใช้ SHAP (SHapley Additive exPlanations) เป็นเครื่องมือที่ใช้ในการอธิบายผลลัพธ์ของแบบจำลอง จะช่วยให้เราเข้าใจได้ว่าคุณลักษณะ (Feature) ใดในข้อมูลมีผลต่อการทำนายของแบบจำลองอย่างไร ดังแสดงผลในรูปที่ 3.2



รูปที่ 3.11 กราฟแสดงความสำคัญของฟีเจอร์ SHAP Feature Importance Plot

SHAP Feature Importance Plot ซึ่งแสดงถึงอิทธิพลของคุณลักษณะต่าง ๆ ที่มีผลกระทบต่อค่าของการทำนายในแบบจำลอง Random Forest จากการวิเคราะห์ด้วย SHAP (SHapley Additive exPlanations) (Sundararajan & Najmi, 2020)

กราฟนี้แสดงให้เห็นว่าในแบบจำลองนี้ คุณลักษณะ EM-Article เป็นคุณลักษณะที่มีอิทธิพลต่อค่าที่ได้จากการทำนายมากที่สุด โดยมีความสำคัญสูงสุด 25.213% รองลงมาคือ OM-Duration ที่มีอิทธิพล 15.192% และ OM-Bounce ที่มีอิทธิพล 14.468% เป็นอันดับที่สาม