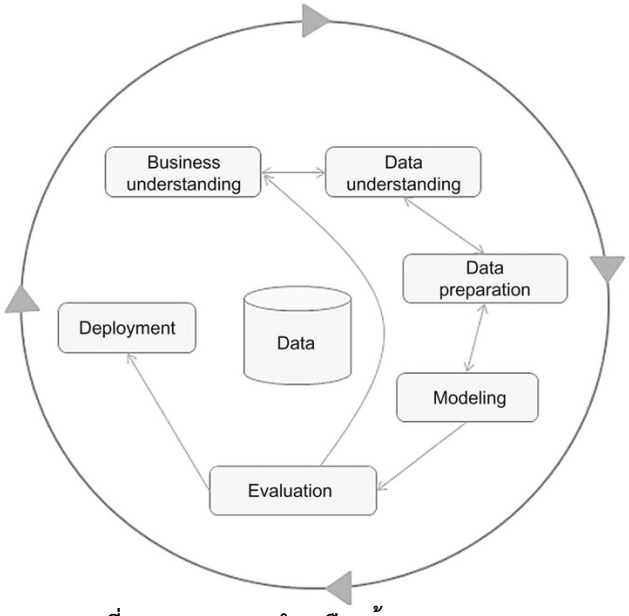
บทที่ 2 กระบวนการทางวิทยาศาสตร์ข้อมูล

หัวข้อหลัก

- กระบวนการทางวิทยาศาสตร์ข้อมูล คือแนวทางการแก้ไขปัญหาที่ระบุสิ่งที่ต้องทำ ลำดับขั้นตอน รวมไปถึงรูปแบบการดำเนินงาน เพื่อให้ได้ผลลัพธ์ที่ตอบโจทย์ปัญหาทางวิทยาศาสตร์ข้อมูลได้ดีที่สุด
- Cross Industry Standard Process for Data Mining (CRISP-DM) คือ กระบวนการทางวิทยาศาสตร์ข้อมูลรูปแบบหนึ่งที่ได้รับการยอมรับมากที่สุดในปัจจุบัน
- กระบวนการทางวิทยาศาสตร์ข้อมูล เป็นกระบวนการแบบทำซ้ำ (iterative process) ซึ่งโดยทั่วไปประกอบด้วยกิจกรรมหลัก 5 อย่าง คือ (1) การทำความเข้าใจปัญหาเชิงธุรกิจและข้อมูล (2) เตรียมข้อมูล (3) พัฒนาโมเดลโดยใช้อัลกอริทึมการเรียนรู้ (4) ทดสอบประสิทธิภาพของโมเดลที่พัฒนาขึ้น (5) นำโมเดลไปใช้งานจริง

การค้นหารูปแบบและความสัมพันธ์ที่มีประโยชน์ที่แฝงอยู่ในชุดข้อมูล เป็นกระบวนการที่ประกอบด้วยการทำกิจกรรมหลัก 5 อย่างแบบวนซ้ำ จนกว่าจะได้ผลลัพธ์เป็นโมเดลที่สามารถนำไปใช้ตอบปัญหาทางวิทยาศาสตร์ข้อมูลได้ กระบวนการวิทยาศาสตร์ข้อมูลมีกิจกรรมหลักที่เกี่ยวข้องทั้งหมด 5 อย่าง ดังนี้ คือ (1) การทำความเข้าใจปัญหา (2) การเตรียมข้อมูล (3) การพัฒนาโมเดลการเรียนรู้ (4) การทดสอบประสิทธิภาพของโมเดล (5) การนำโมเดลไปใช้งานจริง

ในช่วงหลายปีที่ผ่านมา มีกระบวนการวิทยาศาสตร์ข้อมูล ที่ถูกคิดค้นขึ้นมาหลากหลายรูปแบบ เช่น CRISP-DM [2], KDD [3] และ SEMMA [4] เป็นต้น ในบรรดากระบวนการเหล่านี้ รูปแบบที่ดูจะได้รับความนิยมที่สุดก็คือ CRISP-DM ดังแสดงในรูปที่ 2.1



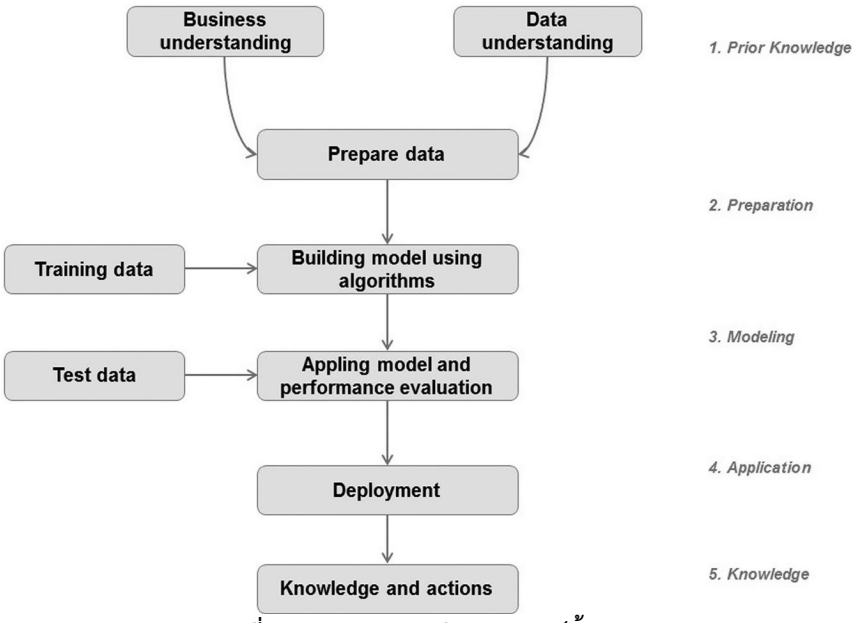
รูปที่ 2.1 กรอบการทำเหมืองข้อมูล CRISP-DM

จากรูปที่ 2.1 จะเห็นว่า ข้อมูล (Data) ถือได้ว่าเป็นหัวใจหลักของการดำเนินกิจกรรมต่าง ๆ ของโครงการ กิจกรรมหลักใน CRISP-DM ได้แก่ การทำความเข้าใจธุรกิจ (Business understanding) การทำความเข้าใจข้อมูล (Data understanding) การเตรียมข้อมูล (Data preparation) การสร้างโมเดล (Modeling) การประเมินประสิทธิภาพของโมเดล (Evaluation) และ

การนำโมเดลไปใช้งานจริง (Deployment) ซึ่งในการดำเนินกิจกรรมเหล่านี้ อาจจะต้องมีการทำซ้ำ (iterative executions) ในบางขั้นตอนหรือทุกขั้นตอน จนกว่าจะได้โมเดลที่มีประสิทธิภาพอยู่ในระดับที่ต้องการ

ในปัจจุบัน หากกล่าวถึง วิทยาศาสตร์ข้อมูล คนส่วนใหญ่จะให้ความสนใจเกือบทั้งหมดไปที่ขั้นตอนของการสร้างโมเดล โดยใช้เทคนิคการทำเหมืองข้อมูล การเรียนรู้ของเครื่องจักร หรือปัญญาประดิษฐ์ ซึ่งอยู่ในขั้นตอนที่ 3 ของกระบวนการทั้งหมดของการทำวิทยาศาสตร์ข้อมูล ซึ่งอาจจะทำให้เกิดความคลาดเคลื่อนเกี่ยวกับธรรมชาติของโครงการด้านวิทยาศาสตร์ข้อมูลได้ เพราะในความเป็นจริง นักวิทยาศาสตร์ข้อมูลที่มีประสบการณ์ต่างทราบดีว่า ขั้นตอนที่ใช้เวลามากที่สุดของการทำงาน ไม่ใช่กิจกรรมการสร้างโมเดล แต่เป็นกิจกรรมการเตรียมข้อมูล ส่วนขั้นตอนที่ใช้เวลามากเป็นอันดับสองรองลงมาก็คือ การทำความเข้าใจปัญหาของธุรกิจและข้อมูล

ในรายวิชานี้ เราจะใช้กระบวนการวิทยาศาสตร์ข้อมูลแบบทั่วไป ที่ไม่ขึ้นกับโดเมนหรือเครื่องมือทางวิทยาศาสตร์ข้อมูลใด ๆ ดังรูปที่ 2.2



รูปที่ 2.2 กระบวนการวิทยาศาสตร์ข้อมูล

จากรูปที่ 2.2 กระบวนการวิทยาศาสตร์ข้อมูลโดยทั่วไป จะประกอบด้วยขั้นตอน ดังต่อไปนี้คือ

1. รวบรวมความรู้ตั้งต้น (Prior Knowledge)
    ก. การทำความเข้าใจธุรกิจ (Business understanding)
    ข. การทำความเข้าใจข้อมูล (Data understanding)
2. การเตรียมข้อมูล (Data preparation)
3. การสร้างโมเดล (Modeling)
    ก. การสร้างโมเดลโดยใช้อัลกอริทึมเรียนรู้รูปแบบและความสัมพันธ์ต่าง ๆ ในชุดข้อมูลฝึกฝน (Building model using algorithms on training data)
    ข. การทดสอบประสิทธิภาพของโมเดลบนชุดข้อมูลทดสอบ (Applying model and performance evaluation on test data)
4. การนำไปใช้งานริง (Deployment)
5. ความรู้และการกระทำ (Knowledge and actions)

เพื่อให้เข้าใจได้ง่ายขึ้น เราจะทำความเข้าใจกระบวนการวิทยาศาสตร์ข้อมูลโดยทั่วไป จากกรณีตัวอย่าง เกี่ยวกับธุรกิจการให้สินเชื่อสำหรับลูกค้ารายย่อย ซึ่งต้องการหาอัตราดอกเบี้ยการให้สินเชื่อที่เหมาะสมกับลูกค้าแต่ละราย

2.1 การรวบรวมความรู้ตั้งต้น

สิ่งที่จะต้องทำในขั้นตอนแรกของวิทยาศาสตร์ข้อมูล ก็คือการรวบรวมความรู้ตั้งต้นที่จำเป็น ได้แก่ ปัญหาทางธุรกิจที่ต้องการแก้ไข บริบทในเชิงธุรกิจที่เกี่ยวข้อง และข้อมูลที่จำเป็นต้องใช้ในการแก้ปัญหา

2.1.1 กำหนดวัตถุประสงค์

การกำหนดปัญหา ถือว่าเป็นขั้นตอนที่สำคัญที่สุดในกระบวนการวิทยาศาสตร์ข้อมูล และจำเป็นอย่างยิ่งที่จะต้องกำหนดปัญหาให้ชัดเจนถูกต้อง ก่อนที่จะเริ่มดำเนินการในขั้นต่อไป

จากกรณีตัวอย่าง ของการให้สินเชื่อสำหรับลูกค้ารายย่อย หลังจากที่นักวิทยาศาสตร์ข้อมูลได้พูดคุยกับผู้ใช้แล้วพบว่าปัญหาที่ธุรกิจต้องการแก้ไข สามารถอธิบายได้ ดังนี้คือ

> *ถ้าเรามีข้อมูลเกี่ยวกับอัตราดอกเบี้ยสินเชื่อและคะแนนเครดิตของผู้กู้ยืมในอดีต, เราสามารถสร้างโมเดลสำหรับทำนายอัตราดอกเบี้ยที่เหมาะสมของผู้กู้ยืมรายใหม่จากคะแนนเครดิตได้หรือไม่*

2.1.2 บริบทในเชิงธุรกิจ

ในระหว่างกระบวนการวิทยาศาสตร์ข้อมูล จะมีรูปแบบและความสัมพันธ์แฝงต่าง ๆ มากมาย ถูกค้นพบ แต่รูปแบบแฝงที่ค้นพบส่วนใหญ่จะไม่ใช่รูปแบบที่มีนัยสำคัญ เป็นหน้าที่ของนักวิทยาศาสตร์ข้อมูลในการคัดกรองรูปแบบแฝงที่ค้นพบ โดยคงเหลือไว้แต่รูปแบบที่มีนัยสำคัญในทางสถิติและเกี่ยวข้องกับปัญหาที่ต้องการแก้ไข ดังนั้นนักวิทยาศาสตร์ข้อมูล จึงจำเป็นต้องศึกษาหาความรู้เกี่ยวกับธุรกิจ บริบทของงานและกระบวนการทางธุรกิจที่สร้างข้อมูลขึ้นมา

จากกรณีตัวอย่าง นักวิทยาศาสตร์ข้อมูลจะต้องทำความเข้าใจภาพกว้างของธุรกิจการให้สินเชื่อ และรายละเอียดเกี่ยวกับกระบวนการในการสมัคร และกำหนดดอกเบี้ยที่เหมาะสม

2.1.3 ข้อมูล

ในขั้นตอนนี้ นักวิทยาศาสตร์ข้อมูลจะต้องสำรวจข้อมูลทั้งหมดที่สามารถใช้ในการตอบโจทย์ของโครงการได้ โดยจะต้องกำหนดและพิจารณาปัจจัยที่เกี่ยวข้อง เช่น แหล่งข้อมูล คุณภาพของข้อมูล ปริมาณข้อมูล และสิทธิ์ในการนำข้อมูลมาใช้ เป็นต้น

จากกรณีตัวอย่างการให้สินเชื่อ นักวิทยาศาสตร์ข้อมูล ได้สำรวจและสรุปข้อมูลที่เกี่ยวข้องกับการทำนายอัตราดอกเบี้ยจากคะแนนเครดิตของผู้สมัครขอรับสินเชื่อได้ 3 อย่าง คือ รหัสผู้ขอสินเชื่อ (Borrow ID), คะแนนเครดิต (Credit Score), อัตราดอกเบี้ย (Interest Rate %) โดยสามารถเก็บรวบรวมตัวอย่างข้อมูลได้ 10 ตัวอย่าง ดังในตารางที่ 2.1

ตารางที่ 2.1 ชุดข้อมูลล (Dataset) ของผู้ขอสินเชื่อ

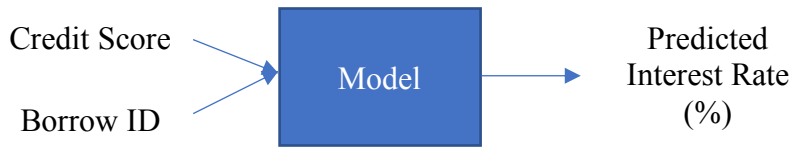
input features, attributes
อินพุทฟีเจอร์

label, class label, target variable
ลาเบล
ค่าตัวแปรเป้าหมาย

Borrow ID	Credit Score	Interest Rate (%)
01	500	7.31
02	600	6.70
03	700	5.95
04	700	6.40
05	800	5.40
06	800	5.70
07	750	5.90
08	550	7.00
09	650	6.50
10	825	5.70

instance, sample, data point
ตัวอย่าง

ชุดข้อมูล (Dataset) ในตารางที่ 2.1 จะถูกนำไปใช้ในการสร้างและทดสอบโมเดล ในขั้นตอนต่อไป เมื่อได้โมเดลที่ต้องการแล้ว ในขั้นตอนการนำไปใช้งานจริง ค่าของคะแนนเครดิตจะถูกป้อนเป็นฟีเจอร์อินพุท (input features) ของโมเดล เมื่อโมเดลได้รับค่าอินพุท ก็จะให้เอาท์พุทเป็นค่าอัตราดอกเบี้ยสินเชื่อที่เหมาะสมออกมา ดังแสดงในรูปที่ 2.3



รูปที่ 2.3 โมเดลที่ได้จะสามารถทำนายอัตราดอกเบี้ยที่เหมาะสม จากคะแนนเครดิตของข้อมูลใหม่ได้

2.2 การเตรียมข้อมูล

การเตรียมข้อมูล เป็นขั้นตอนที่ใช้เวลานานที่สุดในกระบวนการวิทยาศาสตร์ข้อมูล เนื่องจาก โดยปกติชุดข้อมูลที่รวบรวมมาได้จะอยู่ในรูปแบบที่ไม่เหมาะกับการประมวลผลของอัลกอริทึมทางวิทยาศาสตร์ข้อมูล ซึ่งส่วนใหญ่ต้องการอินพุทที่มีโครงสร้างแบบตาราง โดยแต่ละแถวคือหนึ่ง instance และแต่ละคอลัมน์คือ attribute

กรรมวิธีที่ใช้ในการเตรียมข้อมูลมีหลายวิธี เช่น การเติมค่าที่หายไปด้วยค่าเฉลี่ย ค่าแปลงค่าให้อยู่ในช่วงมาตรฐาน การจัดการค่าผิดปกติ (outliers), การคัดเลือกฟีเจอร์ (feature selection), และการสุ่มตัวอย่างข้อมูล (data sampling)

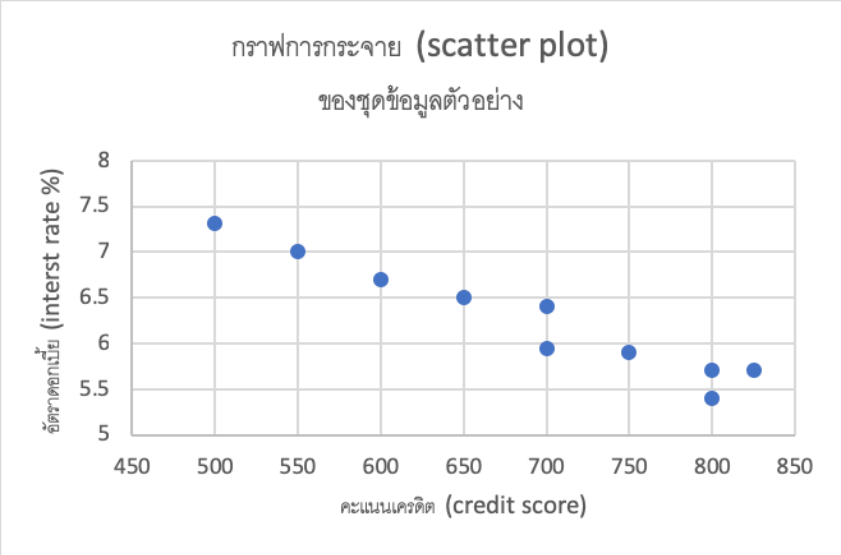
2.2.1 การสำรวจชุดข้อมูล (Data Exploration)

ก่อนที่จะเริ่มเตรียมข้อมูล ส่วนใหญ่เราจำเป็นต้องทำการสำรวจข้อมูลในเชิงลึกเพื่อทำความเข้าใจเกี่ยวกับชุดข้อมูลให้ดียิ่งขึ้น การสำรวจข้อมูล (data exploration หรือ exploratory data analysis) คือการทำความเข้าใจเกี่ยวกับข้อมูลเบื้องต้น โดยการประยุกต์ใช้เครื่องมือพื้นฐานสำหรับการวิเคราะห์ข้อมูล เช่น สถิติพรรณนา (descriptive statistics) และการทำให้เห็นเป็นภาพ (data visualization) เป็นต้น

ตัวอย่างของสถิติพรรณนาที่นิยมนำมาใช้ในการทำความเข้าใจเกี่ยวกับคุณลักษณะของชุดข้อมูล ได้แก่ ค่าเฉลี่ย (mean), ค่ามัธยฐาน (median), ฐานนิยม (mode), ค่าเบี่ยงเบนมาตรฐาน (standard deviation), พิสัย (range)

ตัวอย่างของการทำให้เห็นเป็นภาพที่นิยมใช้ในการสำรวจชุดข้อมูล ได้แก่ กราฟแท่ง (bar chart), แผนภูมิการกระจาย (scatter plot), แท่งความถี่ (histogram)

จากกรณีตัวอย่าง เราสามารถใช้ scatter plot มาช่วยในการทำความเข้าใจเกี่ยวกับชุดข้อมูลตัวอย่างได้ดังรูปที่ 2.4 จากรูปแผนภูมิ จะเห็นได้ว่า ความสัมพันธ์ระหว่างคะแนนเครดิต กับอัตราดอกเบี้ยมีลักษณะแปรผกผัน กล่าวคือ ยิ่งคะแนนเครดิตสูง อัตราดอกเบี้ยที่เหมาะสมก็จะต่ำลง



รูปที่ 2.4 การใช้แผนภูมิการกระจาย (scatter plot) เพื่อทำความเข้าใจคุณลักษณะของชุดข้อมูลเบื้องต้น

2.2.2 การเตรียมข้อมูลก่อนเริ่มดำเนินการสร้างโมเดล (Pre-processing)

คุณภาพของข้อมูล มีผลต่อประสิทธิภาพของโมเดลที่ได้จากการเรียนรู้มาก หากเราป้อนข้อมูลที่มีคุณภาพต่ำ (เช่น มีข้อมูลซ้ำซ้อน ไม่ครบถ้วน) ให้กับอัลกอริทึมการเรียนรู้ ก็จะเป็นไปได้ยากมากที่โมเดลที่ได้จะมีประสิทธิภาพสูง ดังนั้น ก่อนที่จะทำการสร้างโมเดล เราจึงจำเป็นต้องใช้กระบวนการทำความสะอาดข้อมูล ทำให้ข้อมูลมีคุณภาพและอยู่ในรูปแบบที่เหมาะกับการนำไปใช้สร้างโมเดลเสียก่อน การ pre-process ข้อมูลที่มักนำมาใช้ มีดังนี้คือ

- การกำจัดเรกคอร์ดซ้ำ (elimination of duplicate records)
- การแยกค่าผิดปกติ (outliers)
- การทำค่าของแอทริบิวต์ให้อยู่ในรูปแบบ/ช่วงมาตรฐาน (standardization of attribute values)
- การแทนค่าที่ขาดหายไป (substitution of missing values)
- การคัดเลือกฟีเจอร์ (feature selection)

2.3 การสร้างโมเดล (Modeling)

โมเดล คือ ตัวแทนอย่างย่อ (abstract representation) ของข้อมูลและความสัมพันธ์ต่าง ๆ ในชุดข้อมูล

โมเดลถูกสร้างขึ้น โดยการรันอัลกอริทึมการเรียนรู้บนชุดข้อมูล เพื่อค้นหาและสกัดรูปแบบที่มีนัยสำคัญ จากชุดข้อมูล ในปัจจุบัน มีอัลกอริทึมการเรียนรู้ต่าง ๆ ให้เลือกใช้มากมาย ทั้งในรูปแบบของโปรแกรมสำเร็จรูป และในรูปแบบโปรแกรมมิ่งไลบารี่ ในรายวิชานี้ เราจะศึกษาหลักการ กลไกการทำงาน วิธีการปรับแต่ง และรูปแบบของปัญหาที่เหมาะกับการนำไปใช้งานของอัลกอริทึมที่เป็นที่นิยมใช้ในทางปฏิบัติ

ขั้นตอนในการสร้างโมเดล มีดังต่อไปนี้ คือ

(1) สร้างชุดข้อมูลฝึกฝน (training dataset) และ ชุดข้อมูลทดสอบ (testing dataset)

(2) เลือกอัลกอริทึมการเรียนรู้ที่เหมาะสมกับปัญหา

(3) ประเมินประสิทธิภาพของโมเดล

2.3.1 สร้างชุดข้อมูลฝึกฝน และ ชุดข้อมูลทดสอบ

เพื่อให้การประเมินประสิทธิภาพของโมเดลที่สร้างขึ้น มีความเที่ยงตรงแม่นยำ ข้อมูลที่ใช้สำหรับประเมินประสิทธิภาพของโมเดลจะต้องเป็นข้อมูลที่ไม่เคยถูกป้อนให้กับโมเดลมาก่อน ดังนั้น ก่อนเริ่มการฝึกฝน (model training) เราต้องแยกข้อมูลออกเป็นสองส่วนคือ ข้อมูลส่วนแรกใช้สำหรับการฝึกฝน และ ส่วนที่สองใช้สำหรับการทดสอบประเมินประสิทธิภาพ ซึ่งในทางปฏิบัติ มีหลักการแบ่งคือ ให้สุ่มเลือก 2 ใน 3 (หรือประมาณ 70%) ของข้อมูลทั้งหมด เป็นข้อมูลสำหรับฝึกฝน และข้อมูลที่เหลืออีก 1 ใน 3 เป็นข้อมูลสำหรับทดสอบประสิทธิภาพของโมเดล

จากกรณีตัวอย่าง เราจะแบ่งชุดข้อมูลของเราออกเป็น ชุดข้อมูลฝึกฝนขนาด 7 เรกคอร์ด และชุดข้อมูลทดสอบขนาด 3 เรกคอร์ด ดังแสดงในตารางที่ 2.2 และ 2.3 ตามลำดับ

ตารางที่ 2.2 ชุดข้อมูลฝึกฝน (training dataset)

Borrow ID	Credit Score X	Interest Rate (%) y
01	500	7.31
02	600	6.70
03	700	5.95
05	800	5.40
06	800	5.70
08	550	7.00
09	650	6.50

ตารางที่ 2.3 ชุดข้อมูลทดสอบ (testing dataset)

Borrow ID	Credit Score X	Interest Rate (%) y
04	700	6.40
07	750	5.90
10	825	5.70

2.3.2 เลือกอัลกอริทึมการเรียนรู้ที่เหมาะสมกับปัญหา

คำถามทางธุรกิจและข้อมูลที่มี จะเป็นตัวกำหนดงานทางวิทยาศาสตร์ข้อมูล (การจัดกลุ่ม, การแบ่งประเภท, การวิเคราะห์การถดถอย และอื่น ๆ) ที่สามารถนำมาใช้ได้ เมื่อกำหนดชนิดงานทางวิทยาศาสตร์ข้อมูลที่ต้องทำได้แล้ว นักวิทยาศาสตร์ข้อมูลจะต้องเลือกอัลกอริทึมที่เหมาะสมจากหลากหลายอัลกอริทึมที่มีอยู่ของประเภทงานดังกล่าว เช่น งานการแบ่งประเภท (classification task) มีอัลกอเริทึมการเรียนรู้อยู่หลายชนิด เช่น ต้นไม้ของการตัดสินใจ (decision trees) เครือข่ายประสาทเทียม (neural networks), และ k-NN เป็นต้น

สำหรับกรณีตัวอย่าง เราต้องการทำนายค่าอัตราดอกเบี้ยที่เหมาะสม เมื่อทราบคะแนนเครดิตของผู้ขอสินเชื่อ ซึ่งในทางคณิตศาสตร์ โมเดลที่เราต้องการหาจะรับค่าอินพุทเป็นตัวเลขคะแนนเครดิต และให้ค่าเอาท์พุทเป็นตัวเลขอัตราดอกเบี้ย ($f: X \rightarrow y$) งานลักษณะนี้ เรียกว่า การวิเคราะห์การถดถอย (regression problem) ดังนั้น เราจะต้องเลือกอัลกอริทึมสำหรับสร้างโมเดลการถดถอย ซึ่งมีหลากหลายตัว เช่น linear regression, neural network เป็นต้น

เพื่อให้เข้าใจได้ง่ายขึ้น สมมติว่า เราเลือกสร้างโมเดลด้วย linear regression หรือ การถดถอยเชิงเส้น ซึ่งมีรูปแบบสมการทั่วไปของโมเดลคือ

$$y = wX + b$$

เมื่อ

y คือค่าเอาท์พุท หรือ ตัวแปรตาม (dependent variable)

X คือค่าอินพุท หรือ ตัวแปรต้น (independent variable)

w คือค่าสัมประสิทธิ์ของสมการ (coefficients)

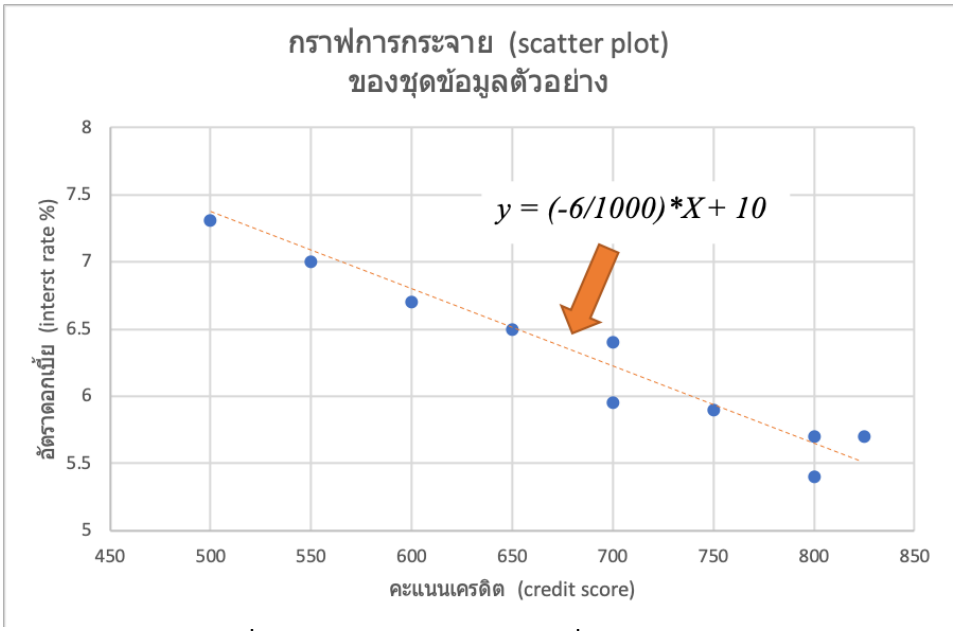
b คือจุดตัดแกน y (y-intercept)

หากพิจารณาจากสมการของโมเดล linear regression จะพบว่า การทำนายค่า y (อัตราดอกเบี้ย) เราต้องทราบค่า w, X และ b ซึ่ง X เราทราบแล้วเพราะเป็นค่าอินพุทฟีเจอร์ ดังนั้นสิ่งที่เรายังไม่ทราบก็คือ ค่าสัมประสิทธิ์ w และ ค่า b นั่นเอง

ค่า W และ b สามารถหาได้โดยการหาเส้นตรงที่เข้ากับข้อมูลได้ดีที่สุด (มีความคลาดเคลื่อนในการทำนายค่า y น้อยที่สุด) ซึ่งสำหรับชุดข้อมูลกรณีศึกษาของเรา พบว่า เส้นตรงที่ต้องการหา มีค่า w = 6/100000 และ ค่า b = 0.1 ดังในสมการ

$$y = \frac{-6}{1000}X + 10$$

หากนำสมการเส้นตรงที่ได้ไปวาดบนกราฟร่วมกับจุดข้อมูล จะได้ดังรูปที่ 2.5 จากรูปเห็นได้อย่างชัดเจนว่า เส้นตรงซึ่งก็คือโมเดลของเราได้สกัดเอาคุณลักษณะสำคัญของชุดข้อมูลฝึกฝนออกมาได้ดีพอสมควร กล่าวคือ สมการเส้นตรงที่ได้บอกเราว่า ค่าของอัตราดอกเบี้ยซึ่งเป็นตัวแปรตาม มีค่าแปรผกผันกับค่าของคะแนนเครดิต



รูปที่ 2.5 โมเดลการถดถอยเชิงเส้น ที่ได้จากการเทรน

2.3.3 ประเมินประสิทธิภาพของโมเดล

โมเดลที่มีคุณภาพดี คือโมเดลที่สามารถทำนายค่าของจุดข้อมูลที่ไม่เคยเห็นมาก่อนได้ ในทางเทคนิคเรียกคุณสมบัตินี้ว่า (*generalization*) วิธีประเมินว่าโมเดลที่สร้างขึ้นสามารถทำนายค่าของจุดข้อมูลใหม่ได้หรือไม่ ทำได้โดยใช้ชุดข้อมูลทดสอบ (test datasets) ที่ได้แบ่งไว้ก่อนที่จะเริ่มเทรนโมเดล ในการหาค่าความคลาดเคลื่อน ซึ่งวิธีการวัดค่าความคลาดเคลื่อนมีหลายวิธี เช่น RMSE (Root Mean Squared Errors) MAE (Mean Absolute Errors)

วิธีการวัดค่าความคลาดเคลื่อนที่เหมาะสม จะขึ้นอยู่กับเป้าหมายของผู้ใช้งาน ดังนั้นในการทำงานจริง นักวิทยาศาสตร์ข้อมูลจะต้องศึกษาเป้าหมายทางธุรกิจของผู้ใช้ให้เข้าใจอย่างถ่องแท้ เพื่อจะได้เลือกใช้ตัววัดค่าที่เหมาะสม

จากกรณีตัวอย่าง เราจะประเมินประสิทธิภาพโดยการวัดค่าความผิดพลาด (prediction errors) บนชุดข้อมูลทดสอบในตารางที่ 2.3 ผลการวัดค่าความผิดพลาดด้วย ค่า RMSE (ค่าเฉลี่ยของรากที่สองของผลรวมของกำลังสองของค่าความผิดพลาด) แสดงในตารางที่ 2.4

ตารางที่ 2.4 การประเมินประสิทธิภาพของโมเดลบนชุดข้อมูลทดสอบ

Borrow ID	Credit Score X	Interest Rate (%) y	คำทำนายอัตราดอกเบี้ยที่ได้จากโมเดล y= (-6/1000)X+ 10	Errors	Squared Errors
04	700	6.40	5.8	-0.6	0.36
07	750	5.90	5.5	-0.4	0.16
10	825	5.70	5.05	-0.65	0.4225

RMSE (Root Mean Squared Error)

$$= \frac{1}{3}\sqrt{(0.36 + 0.16 + 0.4225)}$$

$$= 0.324$$

2.4 การนำไปใช้งานจริง (Deployment)

ผลลัพธ์ที่ได้จากกระบวนการทางวิทยาศาสตร์ข้อมูล จะต้องถูกนำไปหลอมรวมเข้ากับกระบวนการทางธุรกิจ (business process) ซึ่งส่วนมากจะอยู่ในรูปแบบแอพลิเคชั่นซอฟต์แวร์ สิ่งที่ต้องคำนึงถึงในขั้นตอนนี้ ได้แก่ ความพร้อมในการนำไปใช้ในระบบจริง (production system) การผสานเข้ากับระบบงานอื่น (technical integration) เวลาตอบสนอง (response time) การอัพเดตโมเดล (model refresh) การส่งต่อผลลัพธ์ไปยังผู้ใช้งาน (assimilation)

2.5 ความรู้และการกระทำ (Knowledge and Actions)

กระบวนการทางวิทยาศาสตร์ข้อมูลเริ่มต้นด้วย ความรู้ตั้งต้น (Prior Knowledge) และจบลงด้วยความรู้แจ้งที่เพิ่มเติมขึ้น ซึ่งได้มาจากกระบวนการเรียนรู้จากข้อมูลแบบทำซ้ำ นักวิทยาศาสตร์ข้อมูลจะต้องคัดสรรความรู้ใหม่ที่มีนัยสำคัญ และนำไปใช้ในการตัดสินใจ หรือการกระทำอื่น ๆ ที่เป็นประโยชน์ในเชิงธุรกิจ

แบบฝึกหัด

1. กระบวนการทางวิทยาศาสตร์ข้อมูลคืออะไร และมีกี่ขั้นตอน
2. นักวิทยาศาสตร์ข้อมูล ใช้วิธีอะไรในการทำความเข้าใจกระบวนการทางธุรกิจและข้อมูล
3. ในทางปฏิบัติ ขั้นตอนใดในกระบวนการทางวิทยาศาสตร์ข้อมูล ที่ต้องใช้เวลาในการดำเนินการมากที่สุด เพราะเหตุใด
4. จงยกตัวอย่างเทคนิคที่ใช้ในการทำความสะอาดข้อมูลก่อนการประมวลผล
5. ทำไมจึงต้องแบ่งชุดข้อมูลออกเป็น ชุดข้อมูลฝึกฝน (training dataset) และชุดข้อมูลทดสอบ (testing dataset)
6. จากกรณีตัวอย่างการสร้างโมเดลเพื่อทำนายค่าอัตราดอกเบี้ยที่เหมาะสม นอกจาก linear regression algorithm แล้ว เราสามารถใช้อัลกอริทึมใดในการเรียนรู้จากข้อมูลสำหรับปัญหานี้ (ระบุอย่างน้อย 2 อัลกอริทึม)
7. การประเมินประสิทธิภาพของโมเดล ควรใช้เกณฑ์ใดในการประเมิน และใครควรเป็นผู้กำหนดเกณฑ์การประเมิน

เอกสารอ้างอิง

[1] Bala Deshpande, Vijay Kotu. *Data Science*. 2nd Edition, Morgan Kaufmann, 2018.

[2] Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., & Wirth, R. (2000). CRISP-DM 1.0: Step-by-step data mining guide. SPSS Inc. ดึงเอกสารจาก ftp://ftp.software.ibm.com/software/analytics/spss/support/Modeler/Documentation/14/UserManual/CRISP-DM.pdf

[3] Fayyad U, Piatetsky-Shapiro G, Smyth P. From data mining to knowledge discovery in databases. *AI Magazine*. 1996;17(3):37–54.

[4] SAS Institute. (2013). Getting started with SAS enterprise miner 12.3.