Data Mining Using The CRISP-DM Process

เกี่ยวกับโปรเจค

- ทำนายการย้ายธนาคารของลูกค้า (Customer Churn) ใช้งแบบจำลองกลุ่ม (Ensemble Models) ใน RapidMiner

การนำไปใช้งาน

- นำแบบจำลองไปใช้เพื่อทำนายชุดข้อมูลใหม่ ช่วยในการพัฒนากลยุทธ์รักษาฐานลูกค้า

CRISP-DM

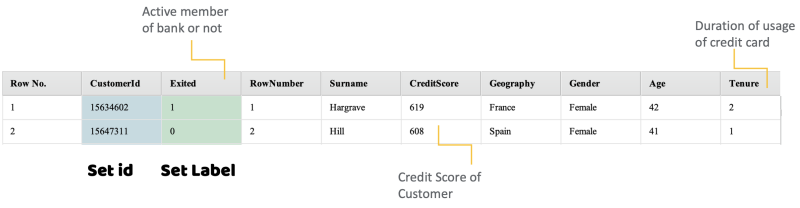
- หรือ Cross Industry Standard Process for Data Mining **เป็นกระบวนการมาตรฐานในการทำเหมืองข้อมูล** (ที่เรียกว่า “ทำเหมืองข้อมูล” เพราะในยุคนี้ข้อมูลก็มีค่าเหมือนแร่ในสมัยก่อน เลยถูกเปรียบเทียบว่าการขุดหา Insight จากข้อมูล ก็เหมือนการทำขุดแร่)
- ซึ่งประกอบไปด้วย 6 ขั้นตอน

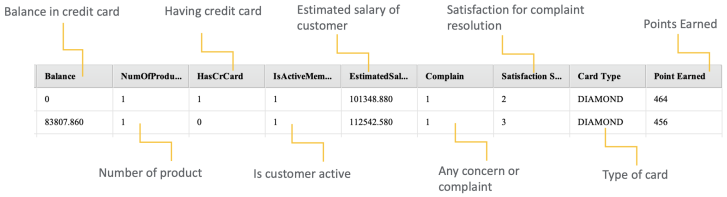
  เข้าใจธุรกิจ

  ในธุรกิจนี้เราเข้าใจกันดีว่าการได้ลูกค้าใหม่มีราคาที่ต้องจ่ายมากกว่าการรักษาลูกค้าเดิมไว้ ฉนั้นเราจะต้องหากลยุทธ์เพื่อรักษาลูกค้าเดิมไว้

  เข้าใจข้อมูล

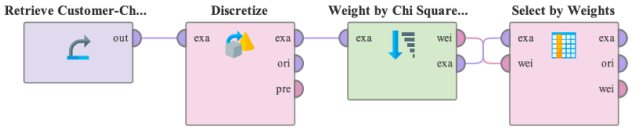
  เข้าใจคอลัมน์ต่าง ๆ ที่เรามีว่าเก็บข้อมูลอะไรไว้ และกำหนดคอลัมน์ที่ใช้สำหรับการทำนาย (Label)





  เตรียมข้อมูล

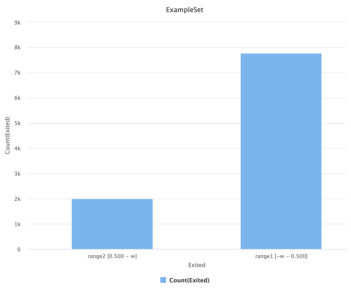
  ทำให้ข้อมูลที่เรามีอยู่พร้อมสำหรับการสร้าง Model โดยขั้นตอนแรกคือการเลือกคอลัมน์ที่มีผลกับคอลัมน์ทำนายของเรา เนื่องจากคอลัมน์ที่เรามีอยู่มีมากเกินไปและบางคอลัมน์ไม่เกี่ยวข้อง ซึ่งกระบวนการเลือกมีหลายวิธีซึ่งเราจะเลือกใช้วิธี Filter appoach → Chi-square เป็นการคำนวณน้ำหนักของคอลัมน์ง่ายต่อการคำนวณและใช้สถิติ



attribute	weight ↓
Complain	9683.357
Surname	2756.877
NumOfPr...	1492.983
Age	1379.422

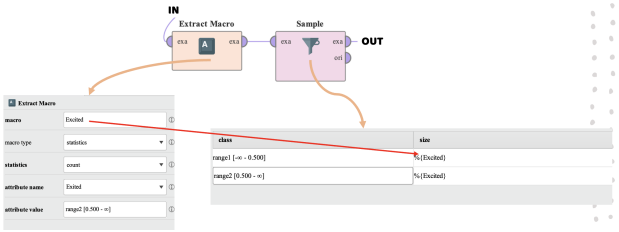
จากรูป เราเลือกคอลัมน์ที่มีน้ำหนักมากที่สุด คือ **Complain, Surname, NumOfProducts และ Age**

จากนั้นจะสังเกตว่าคอลัมน์ที่ใช้ทำนายมีข้อมูลที่ไม่สมดุลกัน



เพราะฉนั้นเราต้องทำให้ข้อมูลเท่ากัน โดยใช้วิธีการเข้าไปแก้ไขที่ตัวข้อมูลเอง (Resampling Approach) และกระบวนการที่ใช้คือการดึงข้อมูลที่มากให้ลงมาเท่ากับตัวแปรที่น้อย (Undersampling)

ใช้ Operator `Extract Macro`

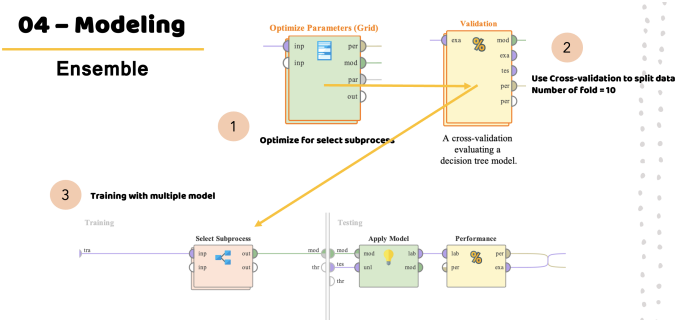


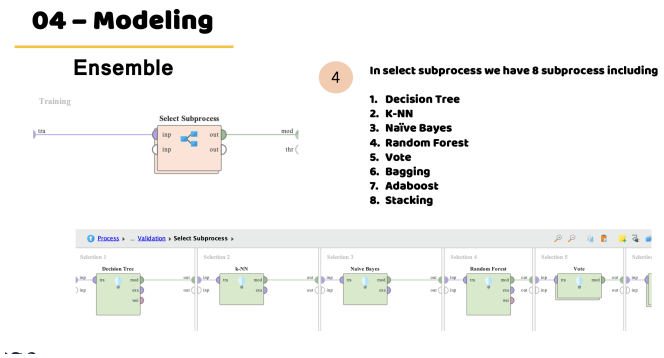
สร้างแบบจำลอง

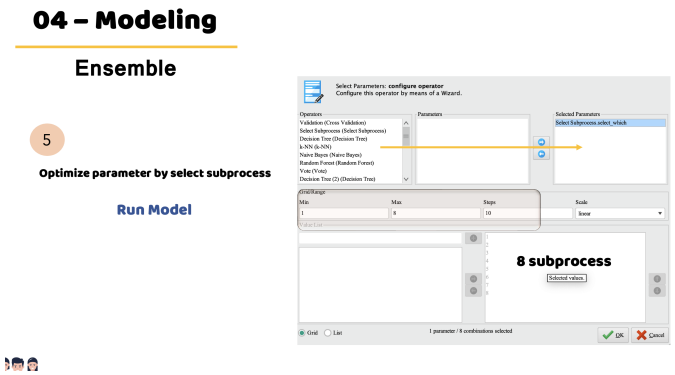
Ensemble เป็นเทคนิคที่ใช้โมเดลหลาย ๆ รูปแบบมาผสมกันเพื่อปรับปรุงประสิทธิภาพของโมเดลทำนายโดยรวมแทนที่จะพึ่งพาโมลเดลเดียว

มีหลายประเภท เช่น

- Ensemble by Vote รวมผลโดยการเฉลี่ยหรือ Vote เสียงข้างมาก
- Bagging เรียนรู้ชุดข้อมูลฝึกที่แตกต่างกันโดยการสุ่มแบบแทนที่ รวมผลโดยการเฉลี่ยหรือ Vote เสียงข้างมาก
- Random Forest เป็นการสร้างโมเดล Decision Tree หลายๆ ตัว
- Boosting เป็นการสร้างข้อมูลหลายๆ รอบโดยที่ข้อมูลที่ทำนายไม่ถูกจะมีน้ำหนักเพิ่มขึ้นในการสุ่มรอบถัดไป
- Stacking รวมผลการทำนายของโมเดลพื้นฐานหลายรูปแบบโดยใช้โมเดลเมต้า







เมื่อกดรันโมเดลผลที่ได้คือ Accuracy ของแต่ละโมเดลซึ่งจะเห็นว่าโมเดลที่โปรแกรมจัดเป็นอันดับแรกคือ **Random Forest**

Optimize Parameters (Grid) (8 rows, 3 columns)

iteration	Select Subprocess.select_which	accuracy
4	4	0.998
5	5	0.998
1	1	0.998
3	3	0.998
6	6	0.998
7	7	0.998
8	8	0.997
2	2	0.991

ประเมินผล

- **F1 score:** คะแนน F1 score อยู่ระหว่าง 0 ถึง 1 คะแนนที่สูงกว่า แสดงว่าโมเดลมีประสิทธิภาพดี
- **Recall:** ค่า Recall อยู่ระหว่าง 0 ถึง 1 ค่าที่สูง แสดงว่าโมเดลสามารถระบุข้อมูลที่เป็นจริงได้ดี
- **Precision:** ค่า Precision อยู่ระหว่าง 0 ถึง 1 ค่าที่สูง แสดงว่าโมเดลทำนายข้อมูลเป็นบวกได้ถูกต้อง

accuracy: 99.80% +/- 0.20% (micro average: 99.80%)

	true range1 [-∞ - 0.500]	true range2 [0.500 - ∞]	class precision
pred. range1 [-∞ - 0.500]	1990	3	99.85%
pred. range2 [0.500 - ∞]	5	1992	99.75%
class recall	99.75%	99.85%	

จากโมเดล **Random Forest ถือว่าโมเดลประสิทธิภาพดีมาก**

Accuracy = 99.8 % , Precision = 99.75 % , Recall = 99.85 % , F1- score = 99.79 %

นำไปใช้

นำแบบจำลองไปใช้เพื่อทำนายชุดข้อมูลใหม่ ช่วยในการพัฒนากลยุทธ์รักษาฐานลูกค้า