



Chapter 11 Recap: Machine Learning for Data Science

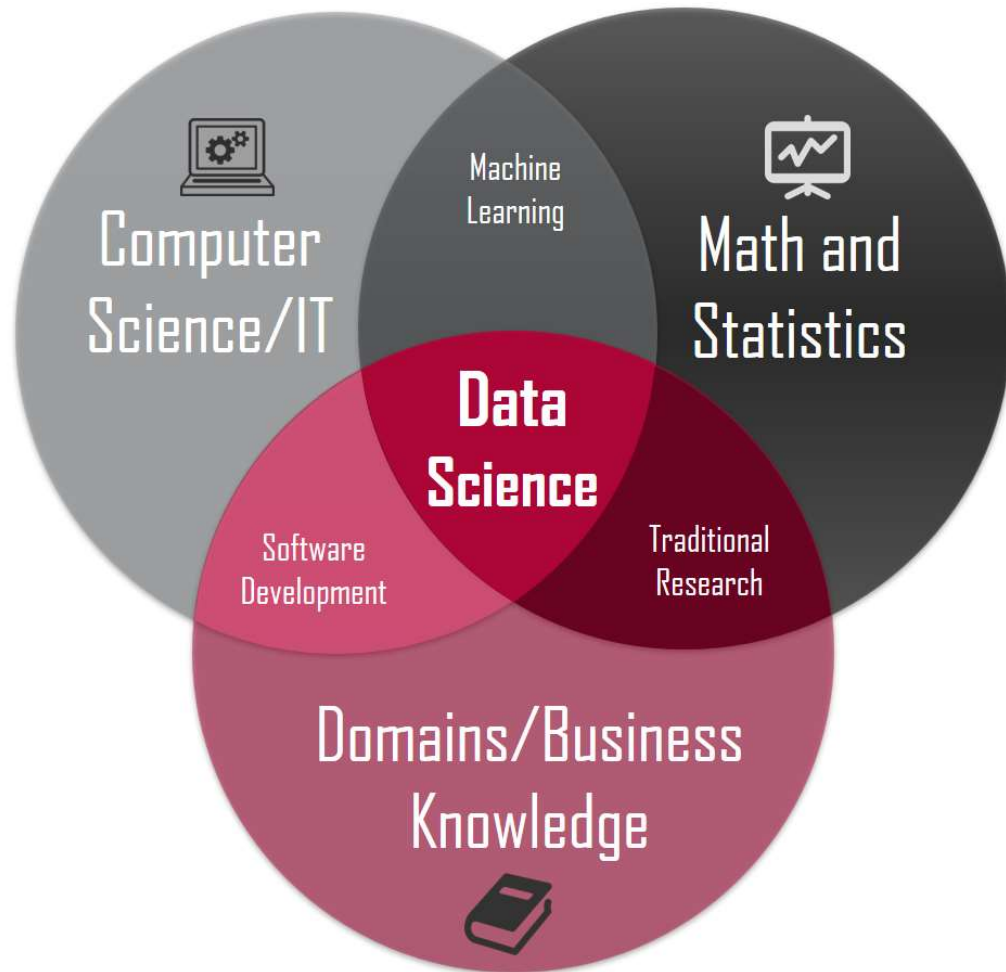
CSS 341 Introduction to
Data Science

Chukiat Worasuchep

Contents

- ▣ Recap data science and machine learning
- ▣ Classification
- ▣ Clustering
- ▣ Regression
- ▣ Summary and Case studies

Data science



Process of data science

1. Business and problem analysis
2. Data acquisition
3. Data preparation (data wrangling)
4. Exploratory data analysis (EDA)
5. Data visualization
6. Data modeling (with machine learning)
7. Deployment and maintenance

Common tasks of data exploration and preparation

1. Data exploration (or inspection)

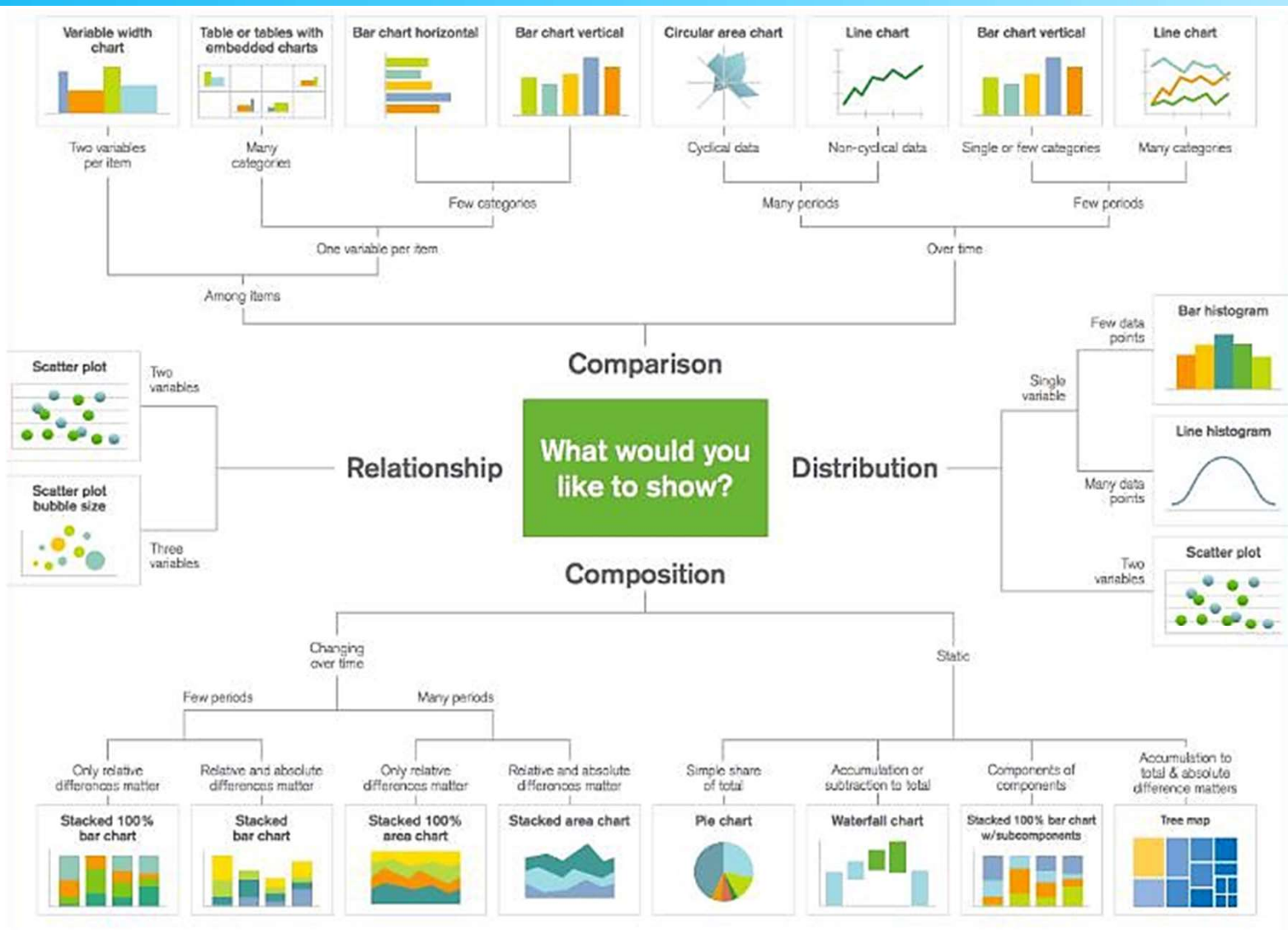
- Observe data characteristics – `describe()`, `info()`, `head()`, `tail()`, `len()`
- Deal with missing values – `isnull()`, `dropna()`, `fillna()`
- Handle duplicate data – `duplicated()`, `drop_duplicates()`
- Handle outliers – `quantile()`, IQR

2. Data manipulation (*feature engineering*)

- Column operations (slicing, `rename`, `drop`, `sort` by column)
- Create new variables
- Row operations (`slice`, `filter`)
- Combine dataframes – `merge()`, `join()`
- Replace values in a dataframe with condition
- Discretization (bucket variables) – `cut()`, `qcut()`
- Data encoding – `LabelEncoder()`, `pd.get_dummies()`
- Normalization and standardization – `MaxMinScaler()`, `StandardScaler()`

Data visualization

Source: [KDnuggets](#)



Major categories of machine learning

1. การเรียนรู้แบบมีผู้ฝึกสอน (supervised learning)

- เป็นการเรียนรู้ข้อมูลต่าง ๆ โดยใช้ข้อมูล (ที่ทราบคำตอบ) ในการฝึกฝน เพื่อช่วยให้ตัวโมเดลสามารถเรียนรู้ผลลัพธ์ได้อย่างแม่นยำยิ่งขึ้น
- Classification, regression

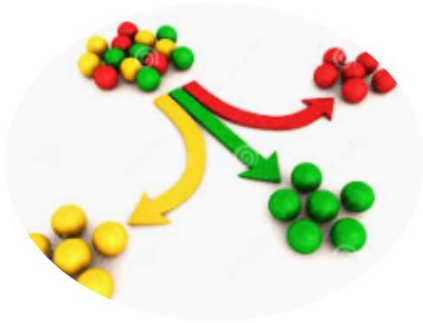
2. การเรียนรู้แบบไม่มีผู้ฝึกสอน (unsupervised learning)

- ใช้วิธีการทางคณิตศาสตร์และสถิติในการวิเคราะห์หาโครงสร้างที่ซ่อนอยู่ในข้อมูลได้โดยตรง ไม่ต้องมีการฝึกสอนก่อน
- Clustering, Association rule mining

2. การเรียนรู้แบบเสริมกำลัง (reinforced learning)

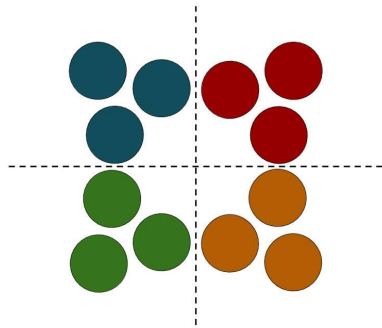
- หาลำดับของการกระทำ (actions) ที่ดีที่สุดที่ทำให้ได้ผลลัพธ์ (outcomes) ที่เหมาะสมที่สุด (Dynamic environment)

Major machine learning techniques



Classification

การจำแนกประเภท



Clustering

การจัดกลุ่ม



Regression analysis

การวิเคราะห์ถดถอย



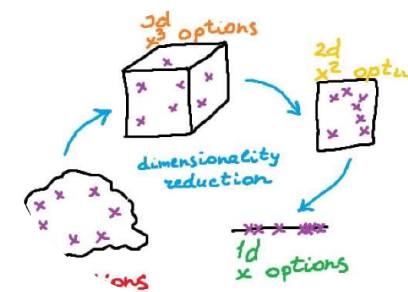
Association rule mining

การค้นหากฎความสัมพันธ์



Recommendation engine

ระบบแนะนำ



Dimensionality reduction

การลดมิติข้อมูล

Contents

- ❑ Recap data science and machine learning

- ❑ Clustering

- ❑ Classification

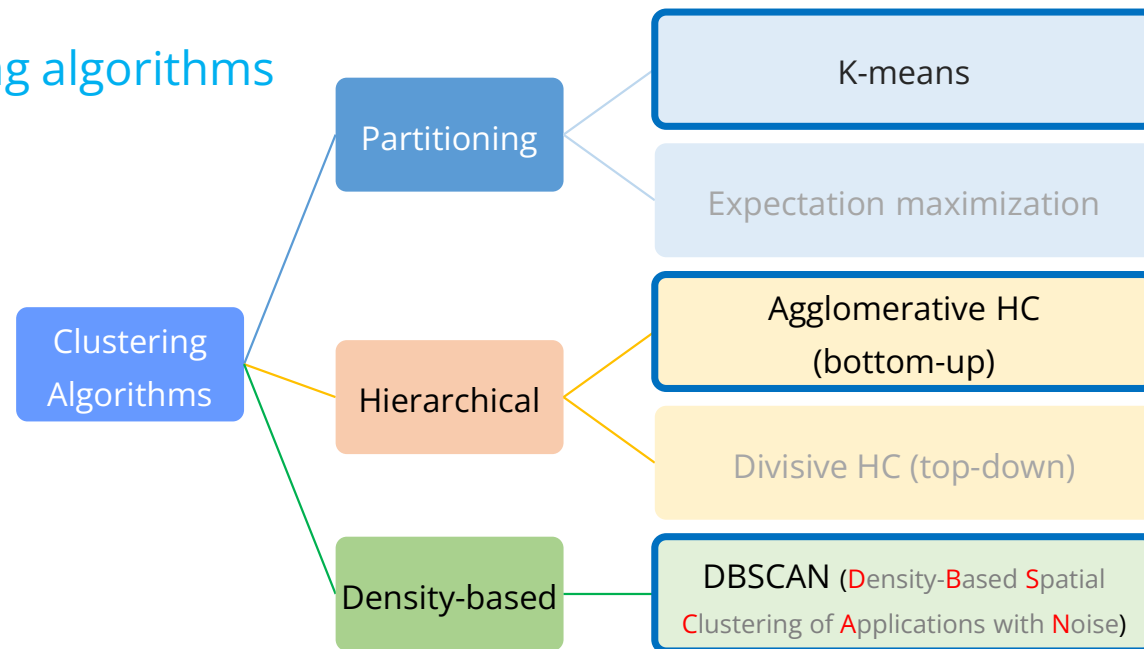
- ❑ Regression

- ❑ Summary and Case studies

Clustering algorithm (ขั้นตอนวิธีการจัดกลุ่ม)

- Clustering เป็นการจัดกลุ่มของข้อมูลจำนวนหนึ่งออกเป็นกลุ่มย่อย ๆ โดยที่...
 - สมาชิกในแต่ละกลุ่มย่อยควรมีลักษณะคล้ายกันมาก ๆ และ แตกต่างจากกลุ่มอื่น
- จัดเป็น *Unsupervised learning* คือ ไม่มีการนำข้อมูลไปฝึกสอน (training) ก่อนนำไปใช้งาน
- แต่ใช้วิธีทางคณิตศาสตร์และสถิติมาคำนวณเพื่อจัดกลุ่ม

- Well-known clustering algorithms



Chukiat Worasuchee

Main application of clustering for business

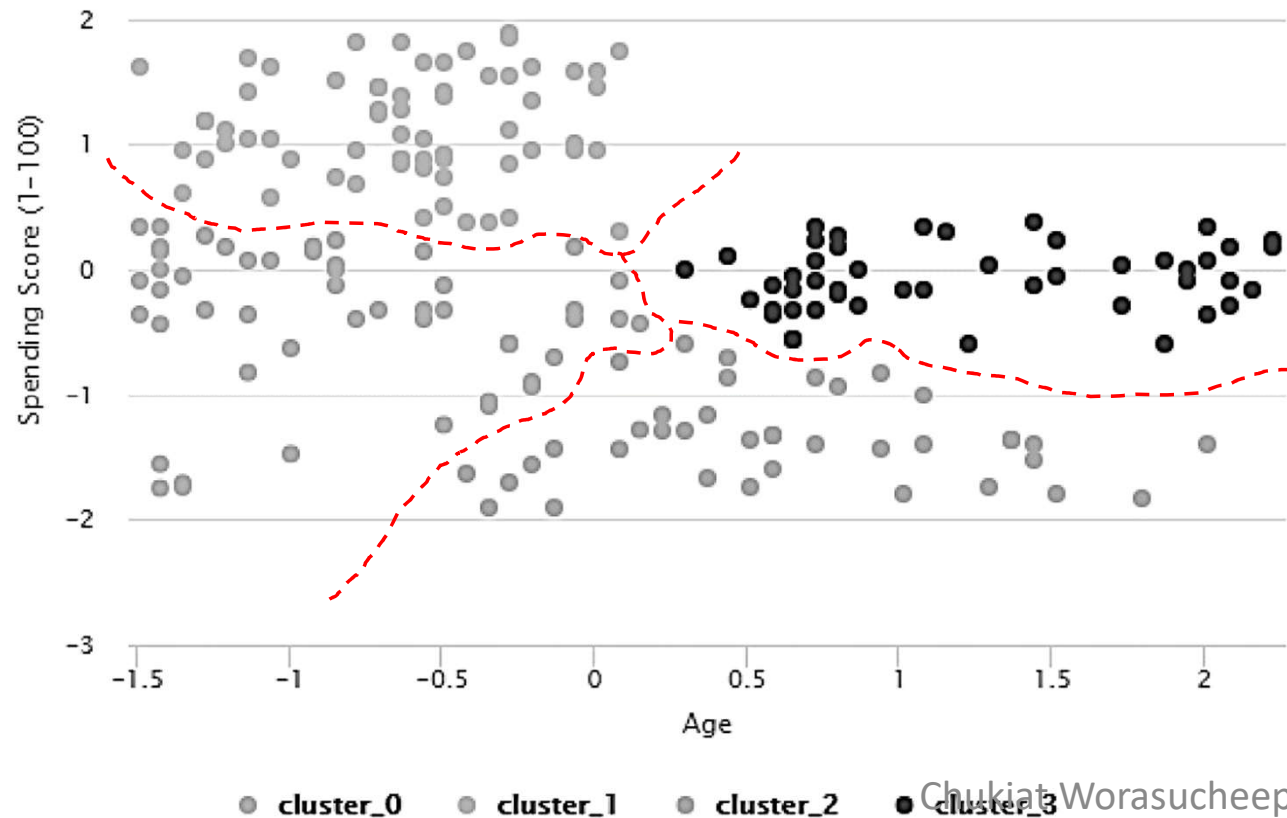
- *Customer segmentation* การแบ่งกลุ่มลูกค้า
 - based on personal profile, behavior, interests, or activity monitoring.
- Common data for *customer segmentation*

1. Customer profile

- Age อายุ
- Gender เพศ
- Marital Status สถานะสมรส
- Income รายได้

2. Behavioral Information

- *Recency* จำนวนวันจากที่ซื้อครั้งล่าสุด
- *Frequency of purchase* ความถี่การซื้อ
- *Monetary* (amount purchase) มูลค่าการซื้อ
- *Balance* ยอดสินทรัพย์คงเหลือ
- *# of items purchase* จำนวนสินค้าที่ซื้อ
- *Time or Day of purchase* เวลาหรือวันที่ซื้อ
- *Products purchase* สินค้าที่ซื้อ
- *Spending score* คะแนนจากการซื้อ



Chukiat Worasuchep

Outputs of clustering mall customers data

	CustomerID	Gender	Age	Annual Income (k\$)	Spending Score (1-100)	label
0	1	Male	19	15	39	0
1	2	Male	21	15	81	3
2	3	Female	20	16	6	4
3	4	Female	23	16	77	3
4	5	Female	31	17	40	0

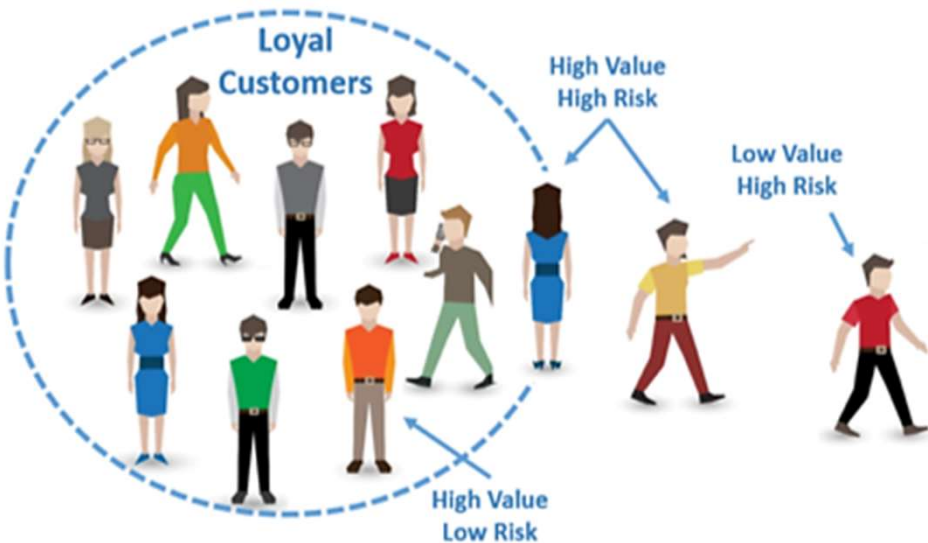


Contents

- ☐ Recap data science and machine learning
- ☐ Clustering
- ☒ Classification
- ☐ Regression
- ☐ Summary and Case studies

Classification problems

- ปัญหาการจำแนกประเภท (Classification) หมายถึงปัญหาการจำแนกข้อมูลต่าง ๆ ออกเป็น **ประเภท (classes)** ต่าง ๆ ที่กำหนดไว้ (เช่น buy/not buy, churn/royal, NLP default/good)
- โดยใช้คุณลักษณะจำเพาะ (features) บางอย่างเป็นเกณฑ์ในการจำแนก
- Supervised learning*



Features						Label
	A	B	C	D	E	F
1	Gender	Name	Payment Method	Age	LastTransaction	Churn
2	male	Nicolas Garrett	credit card	64	98	loyal
3	male	Isaac Reyes	cheque	35	118	churn
4	female	Jaime Sullivan	credit card	25	107	loyal
5	male	Curtis Frazier	credit card	39	90	loyal
6	female	Jeannie Palmer	cheque	28	189	churn

Example use of *classification* for *churn modeling*

	A	B	C	D	E	F
1	Gender	Name	Payment Method	Age	LastTransaction	Churn
2	male	Nicolas Garrett	credit card	64	98	loyal
3	male	Isaac Reyes	cheque	35	118	churn
4	female	Jaime Sullivan	credit card	25	107	loyal
5	male	Curtis Frazier	credit card	39	90	loyal
6	female	Jeannie Palmer	cheque	28	189	churn
7	female	Phyllis Romero	credit card	21	102	loyal
8	male	Lionel Mendoza	credit card	48	141	loyal
9	female	Maureen Norman	credit card	70	153	churn
10	male	Santiago Cruz	credit card	36	46	loyal
11	male	Nelson Davis	credit card	22	51	loyal
12	male	Clarence Vaughn	cash	27	137	loyal
13	male	Jon Griffin	cash	22	147	loyal
14	female	Nettie Neal	credit card	49	158	churn
15	female	Belinda Reeves	cash	24	162	churn
16	male	Taylor Murphy	credit card	45	55	loyal
17	male	Emmett James	credit card	45	160	loyal
18	female	Paula Murray	cash	66	156	churn
19	female	Penny Reese	cash	82	177	churn
20	female	Janis Hernandez	credit card	35	176	loyal
21	female	Dianne Wolfe	credit card	17	133	loyal

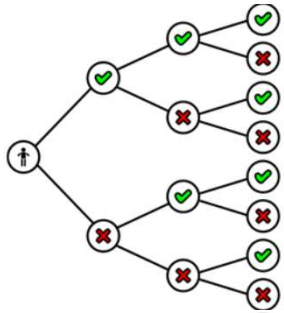
Common results of classification

- Predicted class
- Accuracy, F1-score, recall *if data is imbalanced*
- Feature importance
 - **Easy** for decision, random forest, and logistic regression
 - **Not easy** for ANN, kNN
- Profiling of misclassified groups

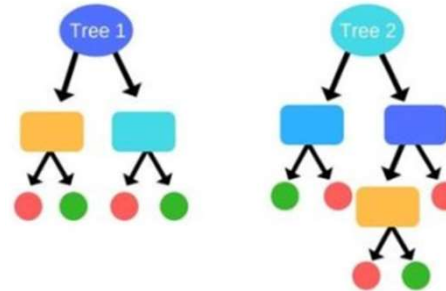
		Predicted values	
		royal	churn
Actual values	royal	TN = 7039	FP = 68
	churn	FN = 45	TP = 245

Chukiat Worasucheeep

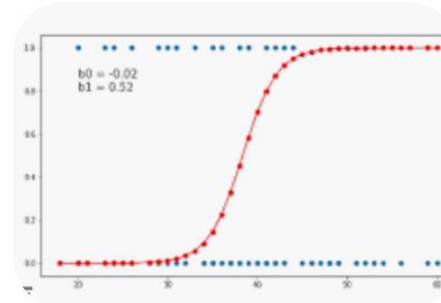
Commonly used classification algorithms



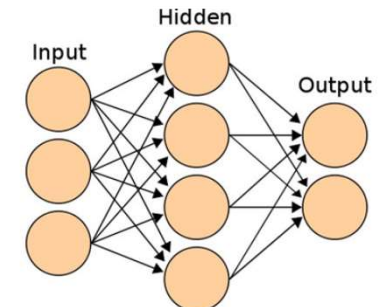
Decision Tree (DT)



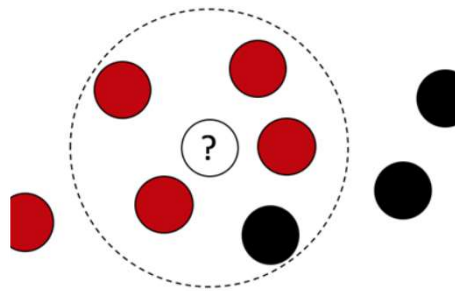
Random Forest (RF)



Logistic Regression (LR)



Artificial Neural Network (ANN)



K-Nearest Neighbor (kNN)

Likelihood of the evidence given that the Hypothesis is True

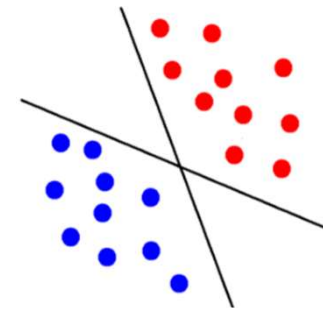
Prior Probability of the Hypothesis

$$P(H|E) = \frac{P(E|H) * P(H)}{P(E)}$$

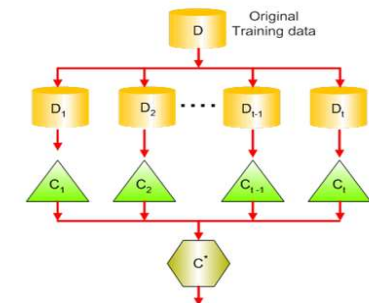
Prior Probability of the evidence given the Evidence is True

Prior Probability that the evidence is True

Naive Bayes (NB)



Support Vector Machine (SVM)



Ensemble (or Combined)

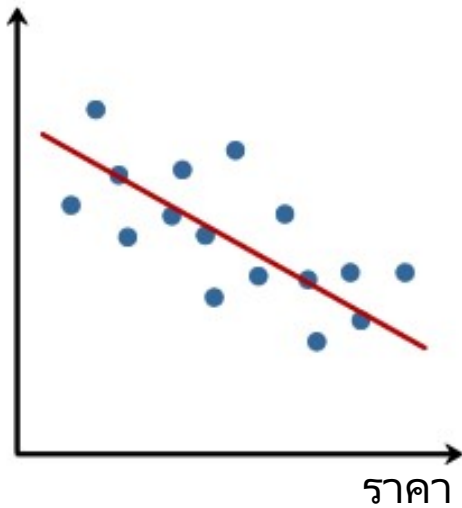
Contents

- ☐ Recap data science and machine learning
- ☐ Clustering
- ☐ Classification
- ☒ Regression
- ☐ Summary and Case studies

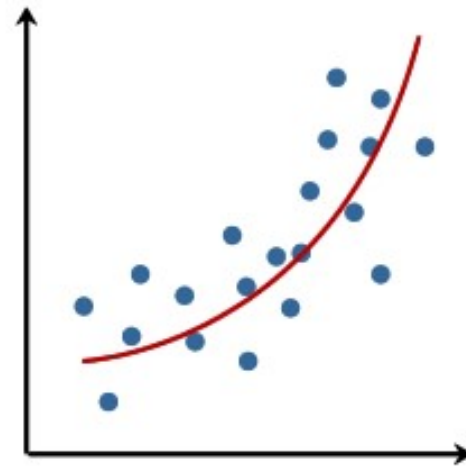
Regression analysis (การวิเคราะห์การถดถอย)

- Regression analysis เป็นการสร้างโมเดลเพื่อหาความสัมพันธ์ของ *ตัวแปรตาม* (a *dependent* or *target* variable) กับ one or more *ตัวแปรอิสระ* *independent* variables or *factors*) ที่เป็น ตัวเลข.

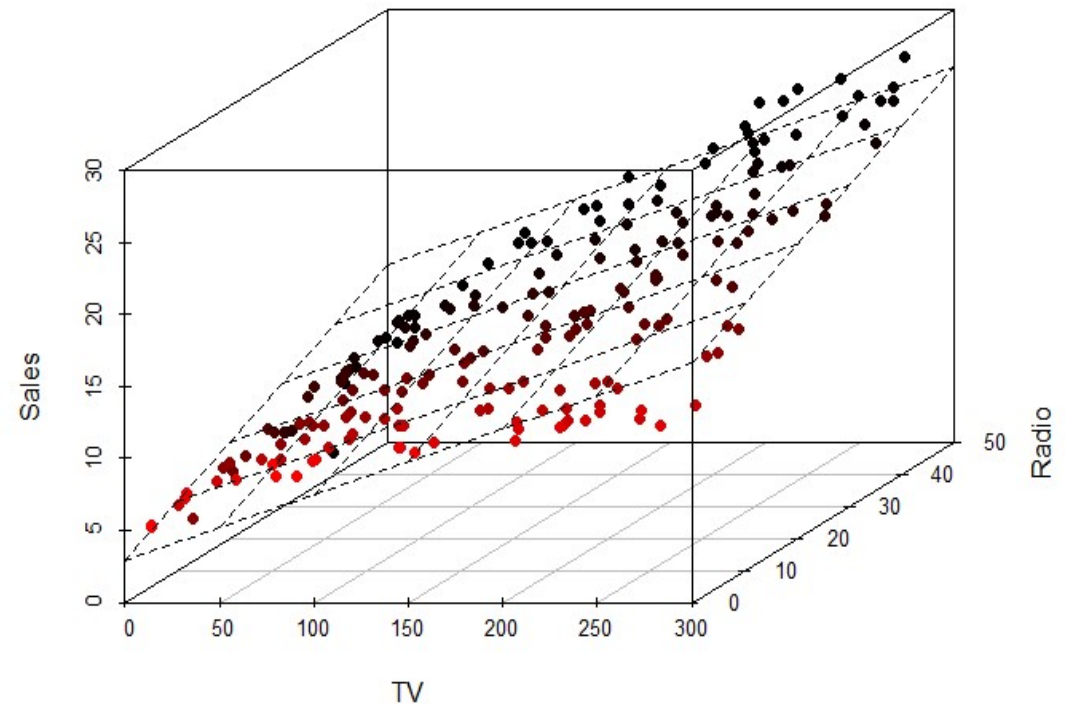
จำนวนขาย



Simple linear regression



Simple
polynomial regression



Multiple Linear Regression

Applications of regression analysis

- พยากรณ์ยอดขาย, ต้นทุน
- ยอดเคลมประกันภัย, ยอดหนี้สงสัยจะสูญ
- จำนวนพนักงานลาออก, ค่าไฟฟ้าประปา
- $\text{Sales} = b_0 + b_1 * \text{TV-ads} + b_2 * \text{Radio-ads}$

TV ads (x_1)	Radio ads (x_2)	Sales (mil.) (y)
89	4	7
66	1	5.4
78	3	6.9
111	6	7.4
44	1	4.8
77	3	6.4
80	3	7
66	2	5.6
109	5	7.3

Simple
linear
regression

$$y = b_0 + b_1 * x_1$$

Multiple
linear
regression

Dependent variable (DV) Independent variables (IVs)

$$y = b_0 + b_1 * x_1 + b_2 * x_2 + \dots + b_n * x_n$$

- b_i เป็นค่าน้ำหนักของแต่ละปัจจัย
- ใช้ `sklearn.preprocessing.PolynomialFeatures` เพื่อปรับไปใช้เป็น polynomial regression

Example – predict CO2 from Weight and Volume

	A	B	C	D	E
1	Car	Model	Volume	Weight	CO2
2	Toyoty	Aygo	1000	790	99
3	Mitsubishi	Space Star	1200	1160	95
4	Skoda	Citigo	1000	929	95
5	Fiat	500	900	865	90
6	Mini	Cooper	1500	1140	105
7	VW	Up!	1000	929	105
8	Skoda	Fabia	1400	1109	90
9	Mercedes	A-Class	1500	1365	92
10	Ford	Fiesta	1500	1112	98
11	Audi	A1	1600	1150	99
12	Hyundai	I20	1100	980	99
13	Suzuki	Swift	1300	990	101
14	Ford	Fiesta	1000	1112	99
15	Honda	Civic	1600	1252	94
16	Hundai	I30	1600	1326	97
17	Opel	Astra	1600	1330	97
18	BMW	1	1600	1365	99
19	Mazda	3	2200	1280	104
20	Skoda	Rapid	1600	1119	104
21	Ford	Focus	2000	1328	105
22	Ford	Mondeo	1600	1584	94
23	Opel	Insignia	2000	1428	99

```
import pandas
from sklearn import linear_model

df = pandas.read_csv("cars.csv")

X = df[['Weight', 'Volume']]
y = df['CO2']

regr = linear_model.LinearRegression()
regr.fit(X, y)

#predict the CO2 emission of a car where the weight is 2300kg, and the volume is 1300cm³:
predictedCO2 = regr.predict([[2300, 1300]])

print(predictedCO2)
```

Result:

[107.2087328]

More: https://www.w3schools.com/python/python_ml_multiple_regression.asp
<https://www.geeksforgeeks.org/linear-regression-python-implementation/>

Chukiat Worasucheeep

Contents

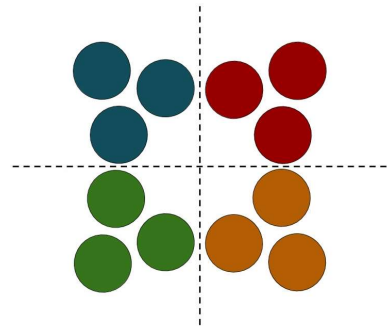
- ☐ Recap data science and machine learning
- ☐ Clustering
- ☐ Classification
- ☐ Regression
- ☒ Summary and Case studies

Major machine learning techniques



Classification

การจำแนกประเภท



Clustering

การจัดกลุ่ม



Regression analysis

การวิเคราะห์ถดถอย



Association rule mining

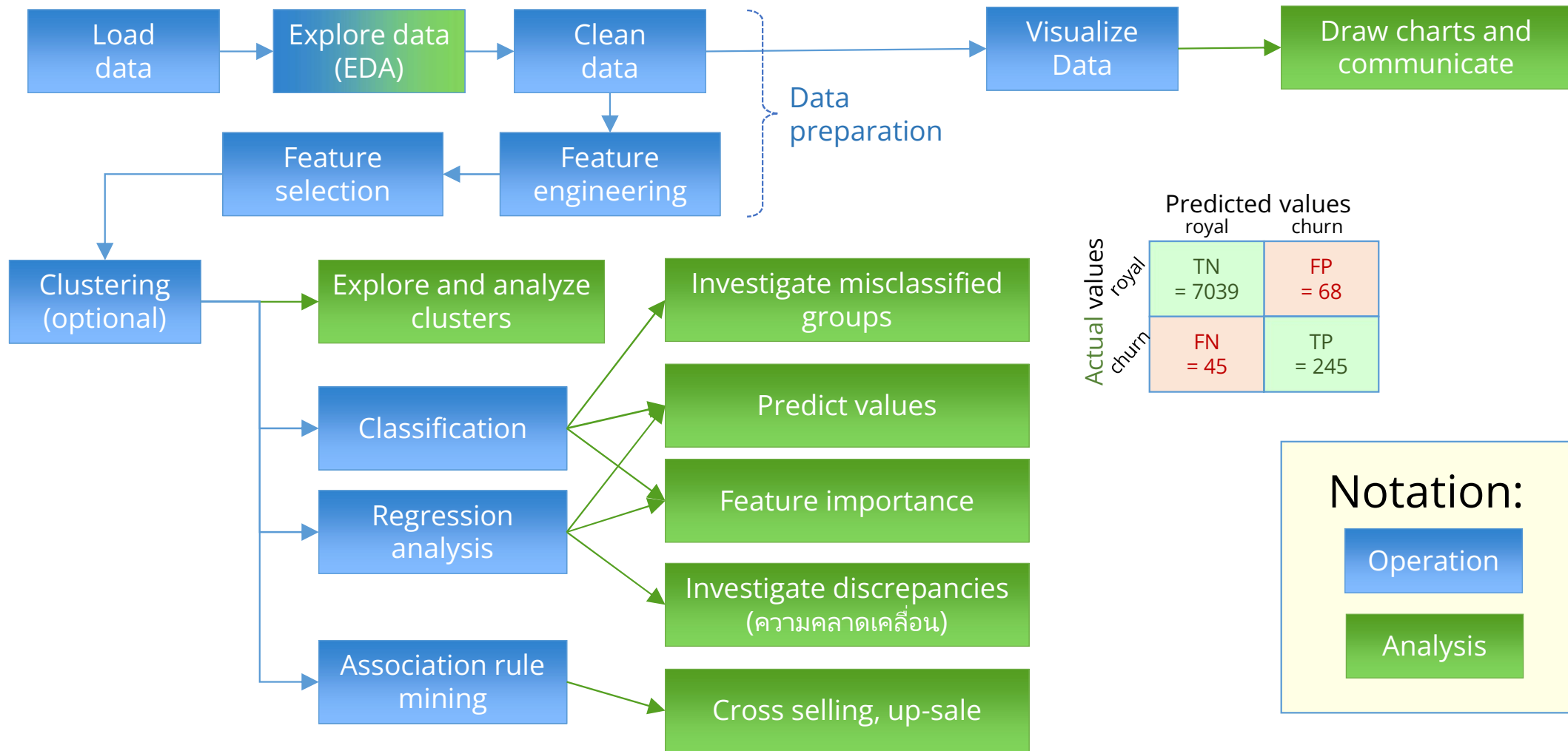
การค้นหากฎความสัมพันธ์



Recommendation engine

ระบบแนะนำ

Common practices (but flexible though)



Chukiat Worasuchep

Example: Churn prediction of a motor insurance firm

- Step 8: Analysis of results

Predict: CHURN	1 = True Positive			2 = False Positive		
	Pred.Churn Churn		94,922	Pred.Churn Renew		27,848
	Product	SUPER LITE	40%	Product	Type 1	43%
	Channel	AL	69%	Channel	AL	52%
	Renew Time	1	99%	Renew Time	1	100%
	Status Claim	No Claim	89%	Status Claim	No Claim	91%
			%Premm Change	-2%	35%	19%
Predict: RENEW	3 = False Negative			4 = True Negative		
	Pred.Renew Churn		38,558	Pred.Renew Renew		124,073
	Product	Type 1	55%	Product	Type 1	67%
	Channel	Branch	46%	Channel	Tbroker	44%
	Renew Time	2	32%	Renew Time	2	25%
	Status Claim	No Claim	91%	Status Claim	No Claim	93%
			%Premm Change	0%	23%	35%

Cases for Group-based Workshop

Cases:

1. Insurance
2. Credit card
3. Debt collection
4. APR for successful loan application
5. (Satisfaction score of bank service)
6. (Security firm)

What to do?

- Study and discuss problem, data, and questions (or goals)
- Propose process of data science (with machine learning)
 - Feel free to use, *but not limited to*, the operations listed here →.

Data preparation

Create features:
..., ..., ..., ?

Feature selection

Classification

Clustering

Regression
analysis

Association rule
mining

Recommender
system

Case 1: An insurance

- Problems

- ลูกค้าประกันกรมธรรม์ภัยรถยนต์ภาคสมัครใจ ไม่ค่อยต่ออายุ

- Data (800,000 rows)

- Customers: gender, age, occupation, est. income, region
- Policy: expire date, product type, premium, sales channel
- Claim: claim amount, date
- Vehicle: type, brand, model, engine size, registration year

- Question/Goal

- ปัจจัยอะไรบ้างที่ส่งผลมากที่สุดต่อการไม่ต่ออายุกรมธรรม์ (churn) และมีผลมากน้อยเท่าใด?
- ลูกค้ากลุ่มใดที่น่าจะไม่ต่ออายุกรมธรรม์ เพื่อเร่งจัด marketing campaign ล่วงหน้า

- Process

		Predicted values	
		royal	churn
Actual values	royal	TN	FP
	churn	FN	TP

Case 2: A credit cards company

- Problems

- ลูกค้าบัตรเครดิตมียอดใช้บัตรน้อย

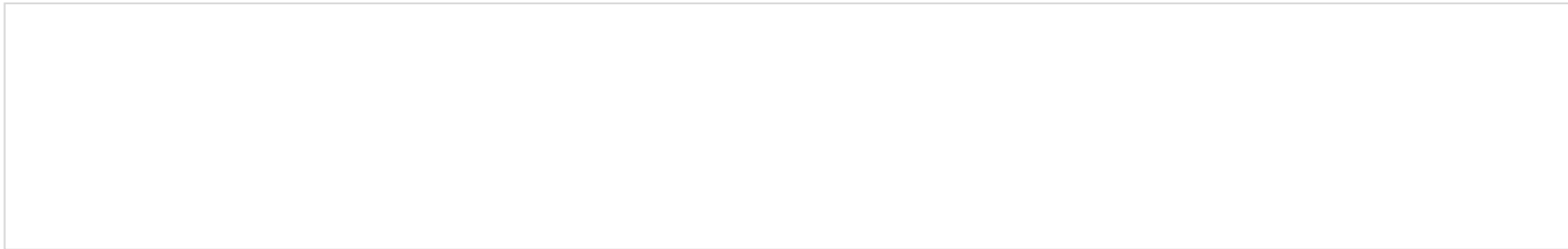
- Data (1,500,000 rows)

- Customer profile: gender, age, occupation, est. income, region
- Credit card transactions: date, amount, merchant type (shopping, gas, dining, retails, health, sports, entertainment, education, etc.)

- Question/Goal

- จะแนะนำสินค้ากลุ่มใหม่ตัวใดให้กับลูกค้าได้เหมาะสมที่สุด

- Process



Case 3: Debt collection of hire purchase

- Problems

- การตามหนี้ลูกค้าเช่าซื้อ แบ่งโดยใช้ระยะเวลาตามหนี้ (easy, difficult (7d), unpaid (30d))
- หนี้สูง ต้องตั้งสำรองสูง ส่งผลต่อกำไร
- กลุ่มลูกค้าสินเชื่รถยนต์ใหม่, สินเชื่รถยนต์ใช้แล้ว, สินเชื่รถแลกเงินที่เคยมีประวัติค้างชำระอย่างน้อย 1 งวด

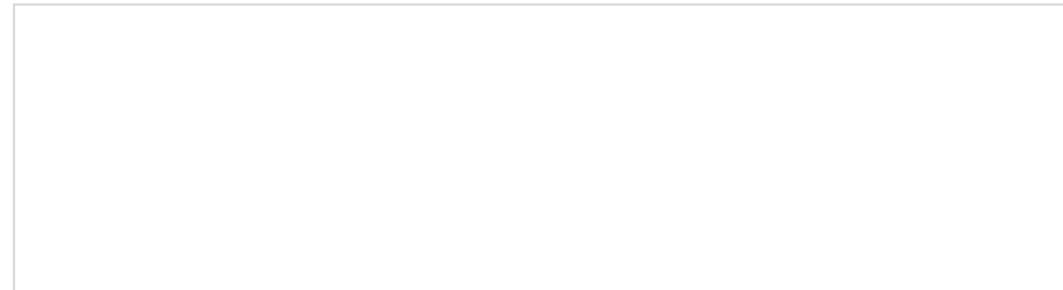
- Question/Goal

- แฉ่งเตือนลักษณะของลูกค้าที่มีโอกาสเปลี่ยนสถานะ easy → difficult → unpaid
- จะใช้วิธีตามหนี้ (SMS, call, outsource, etc.) วิธีใดถึงจะเหมาะสม

- Data (400,000+)

- Customer profile: age, gender, est. income, profession, marital status
- Debt: [loan-to-value](#) (LTV), balance, no. of weeks in debt bucket 1, 2, 3, 4, no. of unsuccessful calls, days of payment after calls, etc.
- Vehicle: type, price, age, brand, class

- Process



Case 4: APR Calculation for successful loan application

- Problems

- การคิด [annual percentage rate](#) (APR) ให้ลูกค้าเพื่อให้ตกลงข้อเสนอ เป็นเรื่องไม่ง่าย คิดแพงไปลูกค้าไม่รับ คิดถูกไปก็จะกำไรน้อยเกินไป

- Data (50000+): loan proposal

- Age, income, profession, province, credit scores, debt-to-income, loan-to-value (LTV), loan term, interest rate, officer ID, APR, success/failure
- *Auto loan*: vehicle age, buy rate, credit score, dealer margin, age, income, profession, province, success/failure.

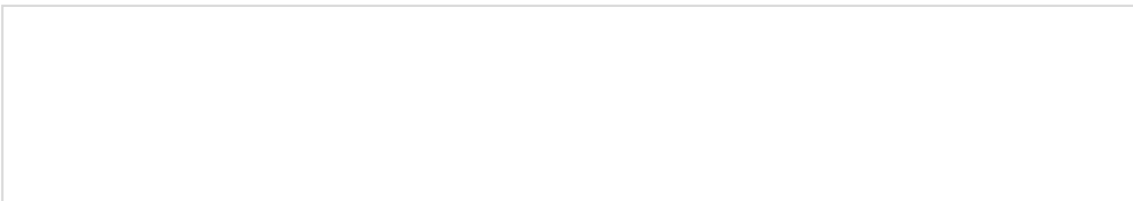
- Question/Goal

- จะกำหนด APR กับลูกค้าเงินกู้ บัตรเครดิต สินเชื่อส่วนบุคคล ฯลฯ เท่าใดถึงเหมาะสม

- Process

Case 5: Satisfaction score of bank service

- Problems
 - Analyze the satisfaction of bank service (of 30 different banks) by using the collected questionnaire data.
- Data (8000+)
 - Clerk service attitude, customer-oriented service, flexibility in handling customer inquiries, service efficiency, transaction errors, convenience of branch locations, service traffic flow, service counter design, etc.
- Question/Goal
 - Want to predict overall satisfaction score of a bank, accuracy and predicted model to formulate policy to improve satisfaction.
- Process



Case 6: A security firm

- Problems

- ลูกค้าปัจจุบันซื้อผลิตภัณฑ์ไม่หลากหลาย หรือมียอดซื้อน้อย
- ต้องการจะแนะนำผลิตภัณฑ์ใหม่เพิ่มเติม (equities, derivatives, unit trusts, warrants, structure notes, block trade, private wealth)

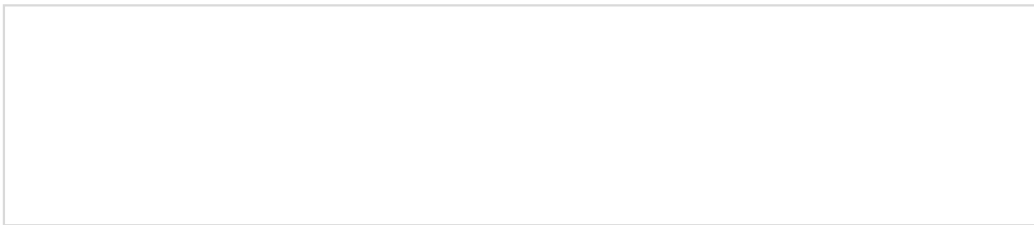
- Data (500,000 rows)

- Customer profile (gender, age, occupation, est. income, region)
- Customer transactions (product type, amount, date)
- Suitability tests

- Question

- จะแนะนำสินค้ากลุ่มใหม่ตัวใดให้กับลูกค้าได้เหมาะสมที่สุด

- Process



What are excluded from this course

- ✗ Feature engineering and feature selections
- ✗ Complete ML techniques
- ✗ (Hyper)parameter tuning
- ✗ NLP
- ✗ Computer vision

