Kaggle 資料學習 Bosch Production Line Performance 資料分析競賽

林子軒

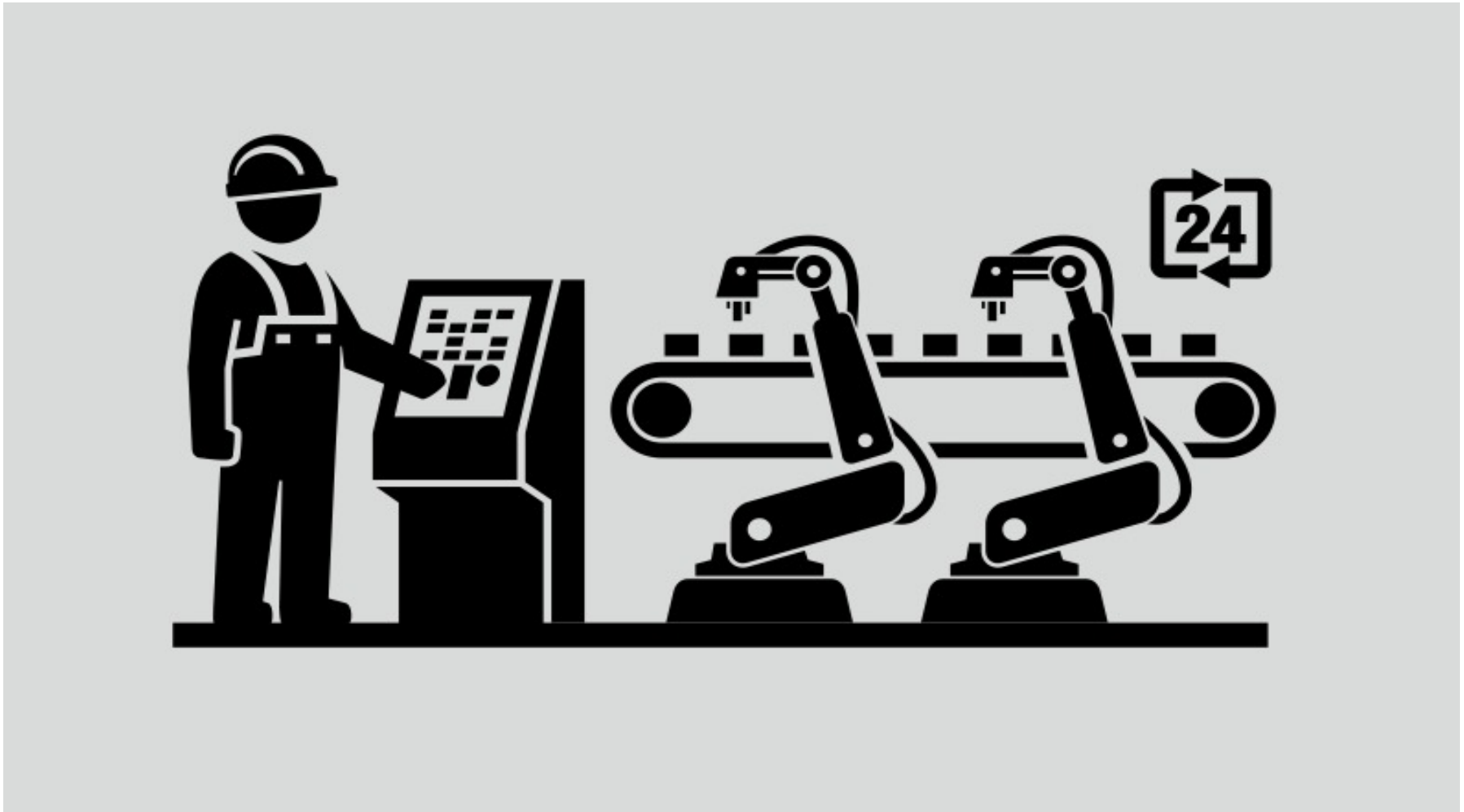
國立東華大學

2017/06/15

Outline

- 緒論
- 資料介紹
- 特徵製造
- Fitted model
- 結論與討論

Production Line



緒論

- 人工生產線
- 自動化
- 機器人
- 提高良率

結論

變數數量：$4000 -> 50$

可能出問題的製程：

L3_S32_F3850

L1_S24_F1723

L3_S33_F3859

資料介紹

- 緒論
- 資料介紹
- 特徵製造
- Fitted model
- 結論與討論

資料介紹

- 生產線分析
- 設備老舊，人為疏失
- 4000道製程
- 提高良率 (良品 vs 不良品)

資料介紹

Imblance :

Response	1	0
	1176868	6879

rate of Response == 1 : 0.0058

資料介紹

Kaggle 提供以下資料：

data	size	n (資料比數)	p (變數數量)
train_numeric	2.1GB	100萬筆	970個
train_date	2.9GB	100萬筆	1157個
train_categorical	2.7GB	100萬筆	2141個

note： train_numeric 讀進 R 中需要 8.5 GB ram

資料介紹

主要變數如下：

變數名稱	意義
Response	目標值, 0： 良品, 1： 不良品
Id	產品代號
Lx_Sx_Fx	L：line；S：station；F：feature number

L3_S36_F3939，代表第 3 條生產線上，第 36 個設備中的第 3939 個特徵值

資料介紹

部分 train_numeric :

Id	L0_S0_F0	L0_S0_F2	L0_S0_F4	Response
11	-0.055	-0.086	0.294	0
13	0.003	0.019	0.294	0
14	NA	NA	NA	0
16	NA	NA	NA	0
18	-0.016	-0.041	-0.179	0

資料介紹

部分 train_date :

Id	L0_S0_F0	L0_S0_F2	L0_S0_F4
11	602.64	602.64	602.64
13	1331.66	1331.66	1331.66
14	NA	NA	NA
16	NA	NA	NA
18	517.64	517.64	517.64

特徵製造

- 緒論
- 資料介紹
- 特徵製造
- Fitted model
- 結論與討論

特徵製造

- 緒論
- 資料介紹
- 特徵製造
- Fitted model
- 結論與討論

feature engineering 1

在生產線上，可能在某一時段機器故障，導致不良品

feature	解釋
first	進入該製成時間點
min	製成時間點最小值
last	離開該製成時間點

feature engineering 1

feature	解釋
max	製成時間點最大值
class.amount	該製成時間點種類數量
na.amount	na數量，代表未經過製程的數量

分別對 "所有生產線"、L0、L1、L2、L3 進行特徵工程。

ex : all_first, L0_first, L1_first, L2_first, L3_first

feature engineering 2

同一時間製造多個產品，它們的表現可能有高度相關

feature	解釋
next	下一個產品是否為同時製造的產品
prev	上一個產品是否為同時製造的產品
cost.time	該製程耗時
prev.cost.time	與上一個產品相比，製程耗時差距
next.cost.time	與下一個產品相比，製程耗時差距

變數選擇 — train_numeric

計算 不良品 比率：

train_numeric

ID	res1.per	var.name
1	0.0451	L3_S32_F3850
2	0.0093	L1_S24_F1768
3	0.0093	L1_S24_F1763
⋮	⋮	⋮
968	0.0003	L1_S25_F2512

feature selection

- 變數：450 個 $->$ 50 個
- By xgb.importance

Gain　　在加入該變數後，準確度提高多少

Cover

Frequency

feature selection

450 個變數		pred	
		0	1
real	0	1176353	515
	1	4216	2663

$$MCC = 0.568$$

feature selection

50 個變數		pred	
		0	1
real	0	1176304	564
	1	4360	2519

$MCC = 0.545$

feature selection

$$\text{MCC} = \frac{TP*TN-FP*FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}}$$

		Prediction	
		T	F
Real	T	TP	FN
	L	FP	TN

other

1. imbalance 非常嚴重：分類− >迴歸，0.25 作為分界
2. 使用 RMSE 逼近 MCC

Fitted model

- 緒論
- 資料介紹
- 特徵製造
- Fitted model
- 結論與討論

Fitted model

- feature engineering 1
- feature engineering 2
- 變數選擇 — train_numeric
- xgb.imporance 挑選出 50 個 feature
- 進步了將近 2 倍 (MCC : 0.18 $->$ 0.46)

結論與討論

- 獨特的特徵工程，自動化尋找 feature
- feature — 需要改善的製程 — 提高良率
- feature by numeric and date data

結論與討論

50 個變數		pred	
		0	1
real	0	1176304	564
	1	4360	2519

red vs blue：兩者意義不同

Reference

Bosch Production Line Performance. (2016) .

Daniel FG. (2016).

MY github

MY github

https://github.com/f496328mm

THANK YOU