玉山人工智慧公開挑戰賽2019夏季賽

台灣不動產AI神預測

建模方法分享

隊伍：WWW-YCC-IDV-TW

Loss Function 的選擇

- Black-Scholes 選擇權定價模型中假設金融資產價格成 log normal distribution，即 log(價格) 成 normal distribution

- 基於此假設，我們選擇在 log(價格) 上做 Mean Square Error (MSE) 作為 loss function

- Public Score: 5480 (沒加 log) -> 5807 (+327)

讓 Loss Function 更準確，神秘逆轉換

- 因資料中為神秘轉換後的價格，無法直接套用定義的 loss function
- 利用以下觀察，我們得到神秘逆轉換公式：
  - 某些轉換後價格經常出現，應該是整齊的價格 (e.g. 1000萬)
  - 大部分的轉換後價格是無理數，但某些卻是整數?
  - ~~可能還需要一點通靈體質~~

$$P = \left(\frac{P' - 196452}{200}\right)^{\frac{2}{3}} \cdot 10000$$

- Public Score: 6093 -> 6133 (+40)

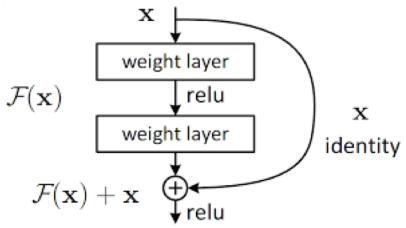
Feature Engineering

- 神秘轉換價格 -> 真實價格 -> 單位面積價格
- 缺值補值
- One-hot encoding
- 利用取log、開方根… 把 feature 轉換成線性分佈
  - 取log：價格、面積、etc
  - 方根：方圓內設施數量、etc

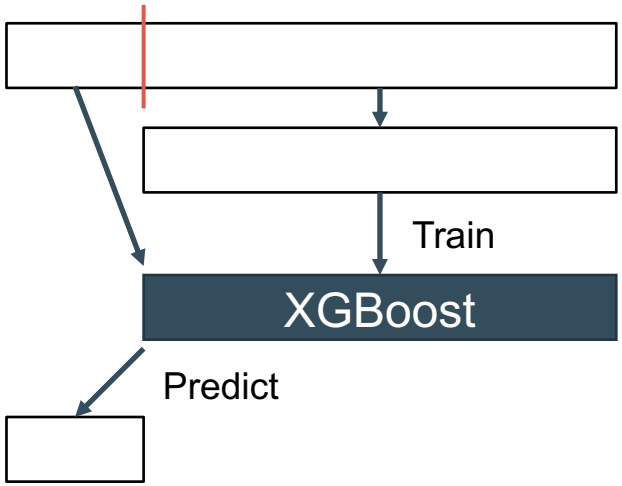
3 個模型

- 2x Gradient Boosted Decision Tree: **XGBoost, LightGBM**
  - 使用 Bayesian Optimization 調 hyper parameters，記得做 CV

- 1x DNN 深度神經網路
  - 仿造 Residual Network 結構，兩個 block，各包含兩層 fully-connected layer
  - 加入 Dropout layer 來避免 overfitting
  - 弄很久但其實沒有 XGBoost 準



Ensemble 流程



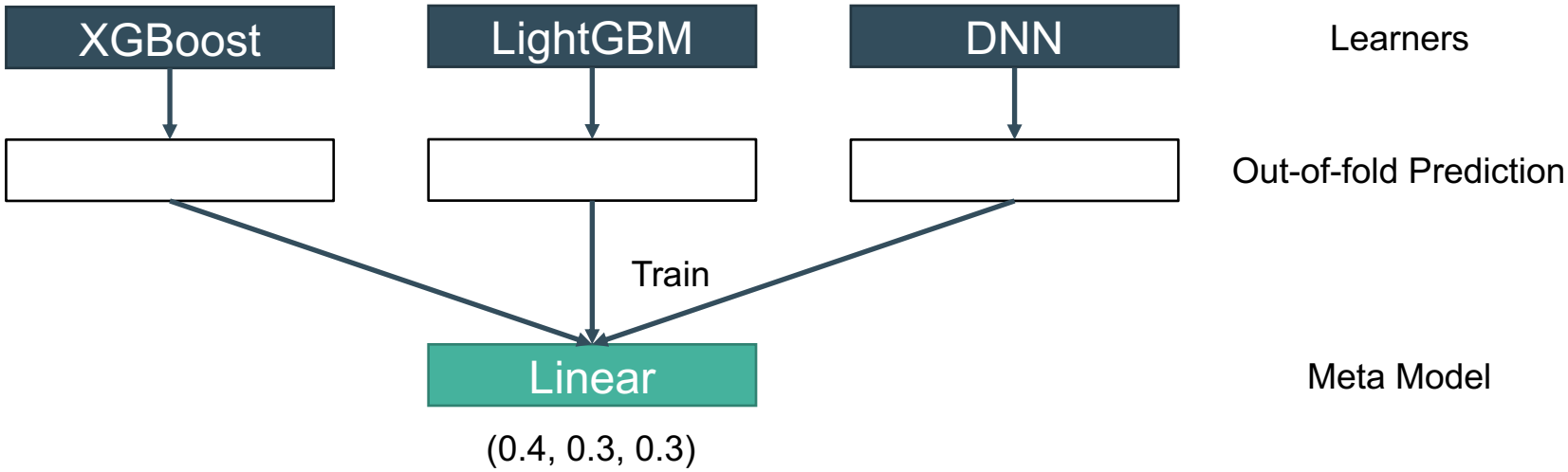
Full Training Data

Sampled Training Data

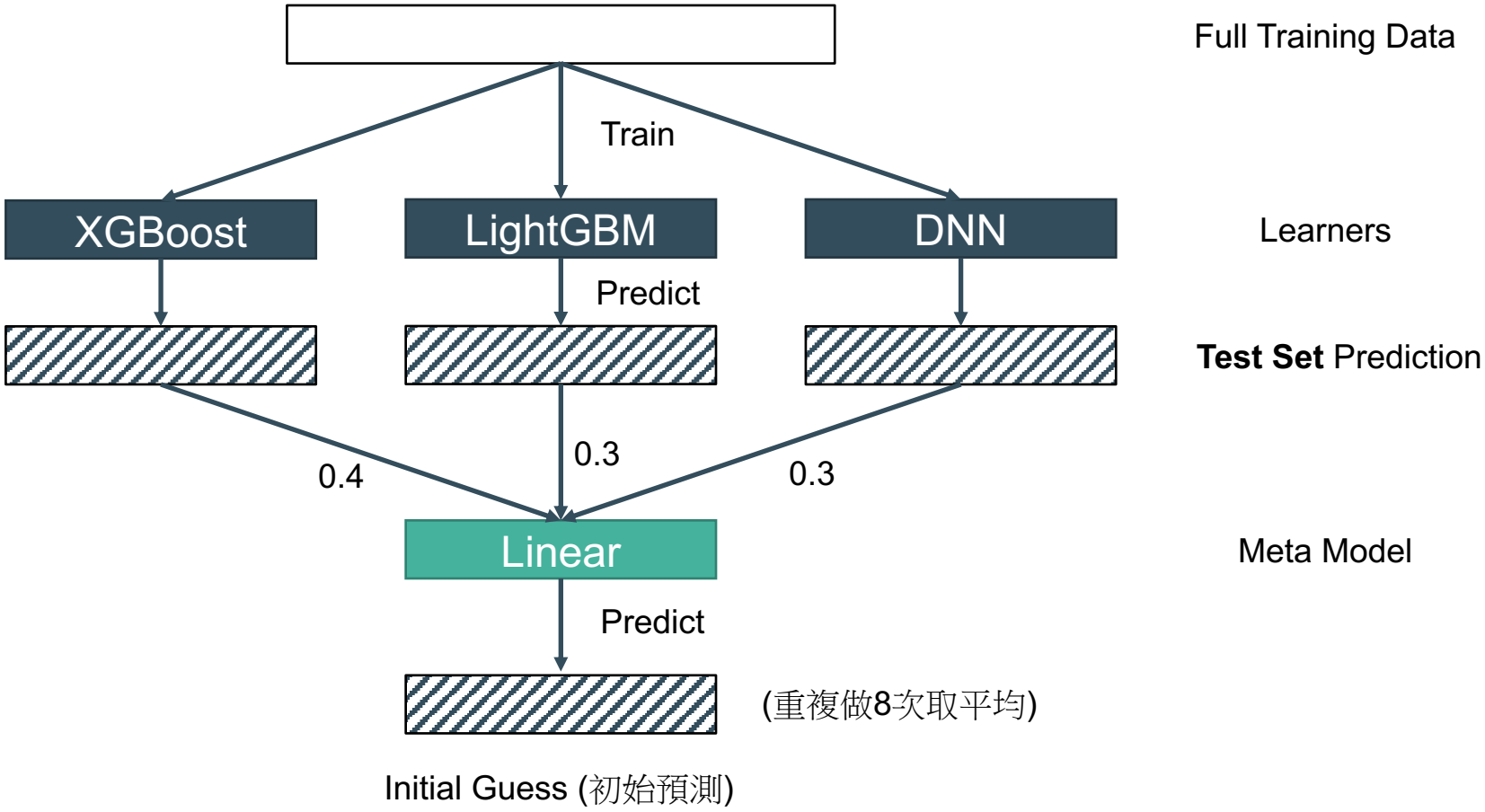
Learner

Out-of-fold Prediction

Ensemble 流程 (cont.)



Ensemble 流程 (cont.)

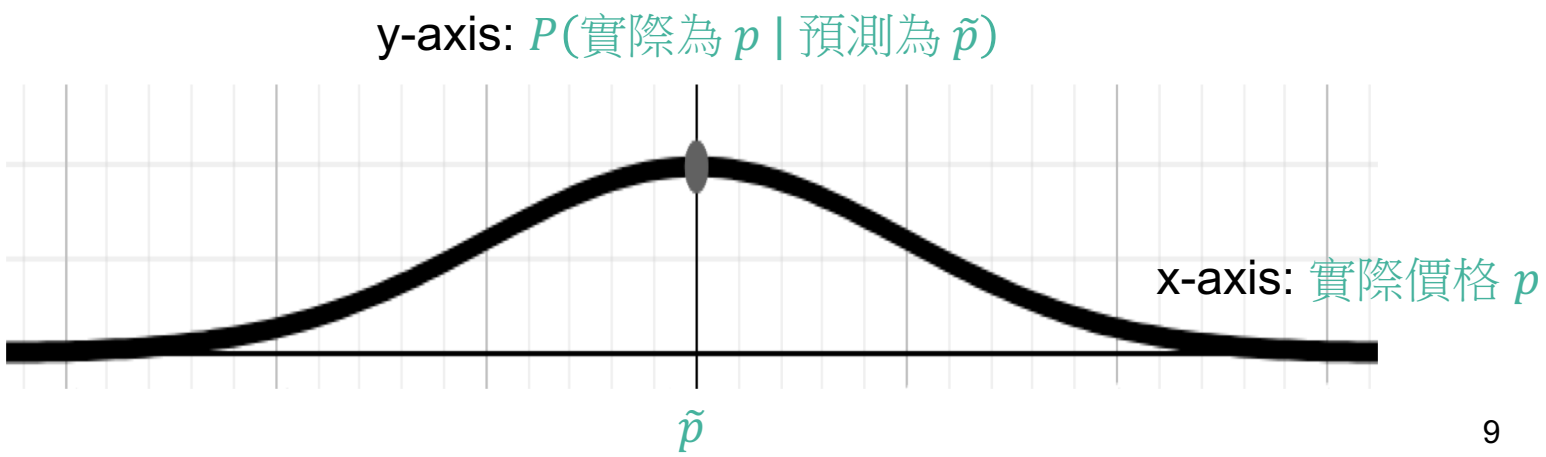


Public Score: 6133(XGBoost) -> 6257 (+124)

Warning of Math

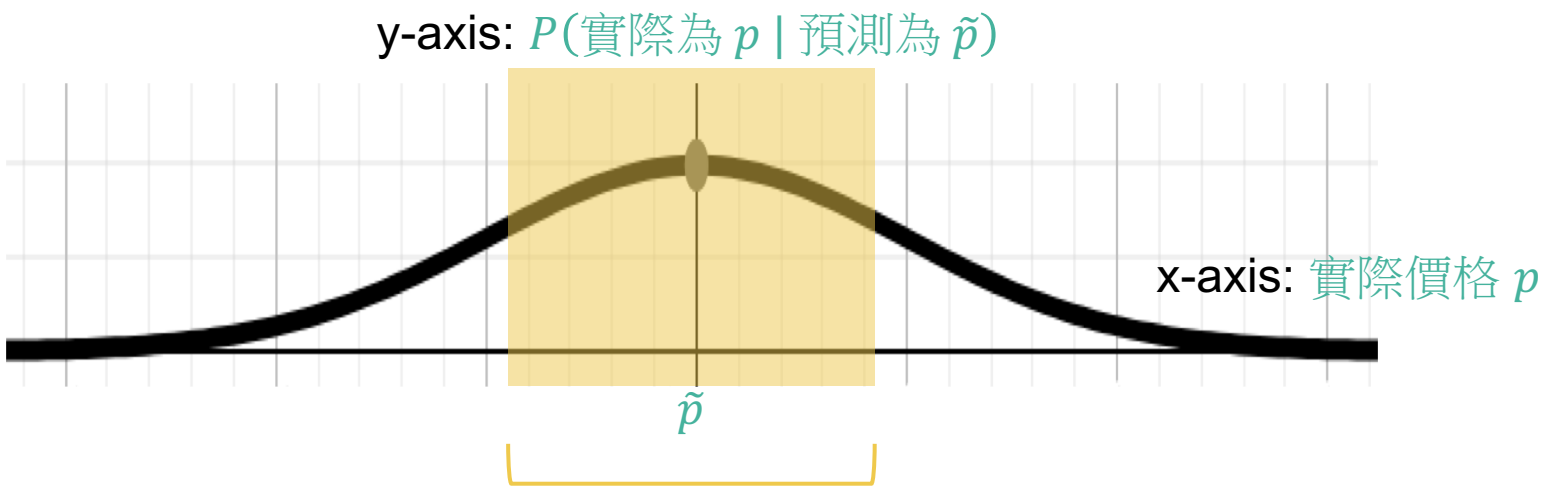
初始預測與 hit rate 的關係

- 初始預測 $\tilde{p}$、實際(對數)價格 p
- 假設
  1. 實際價格成 normal distribution (一開始的假設)
  2. 初始預測能完美最大化 likelihood
- 則 $P(\text{實際為 } y \mid \text{預測為 } \tilde{y}) \sim N(\tilde{y}, \sigma^2)$



初始預測與 hit rate 的關係 (cont.)

- 初始預測 $(\tilde{p})$、實際價格 (p)

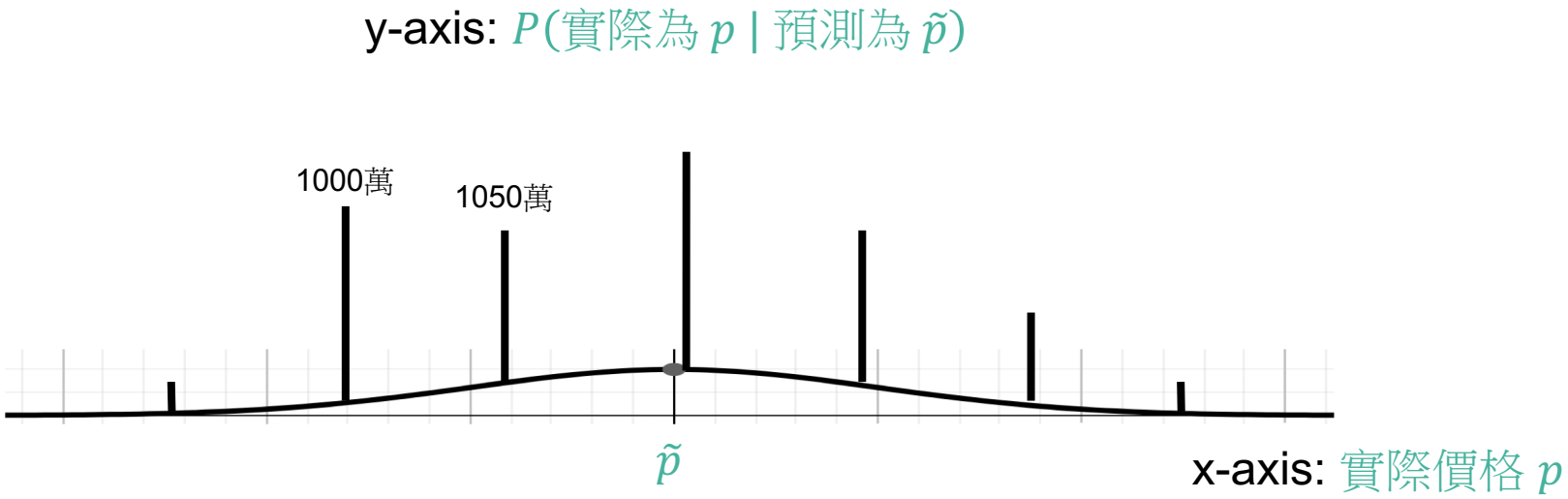


上傳 $\tilde{p}$ 所產生的正負10%範圍，實際價格在此區間內則 hit rate 上升

- 小結：上傳初始預測即可最大化 hit rate 期望值。真的嗎?

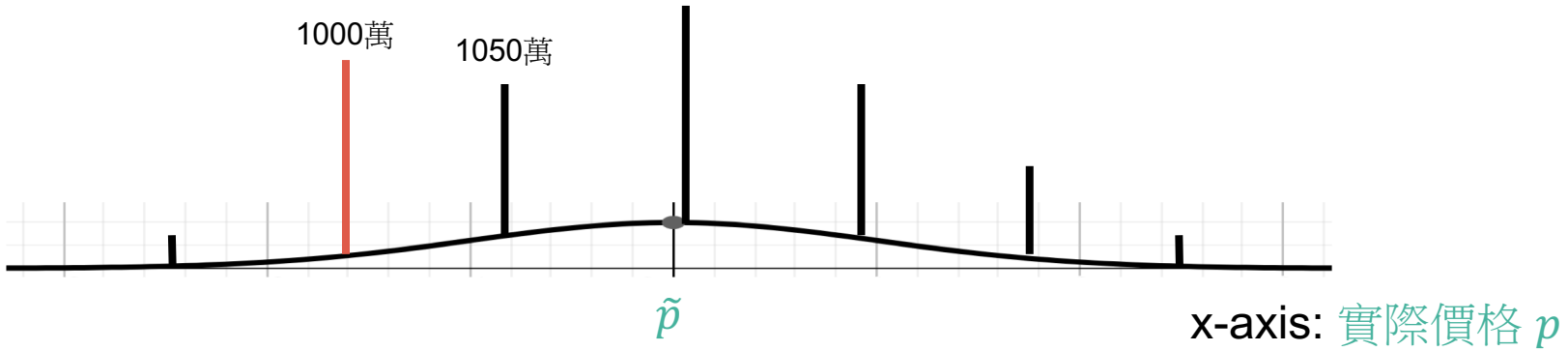
假設錯誤

- 實際對數價格並非是 normal distribution，
  而是在整數倍價格上有spike



- 初始預測需要修正，否則無法最大化 hit rate 期望值

用貝氏定理求離散機率分佈

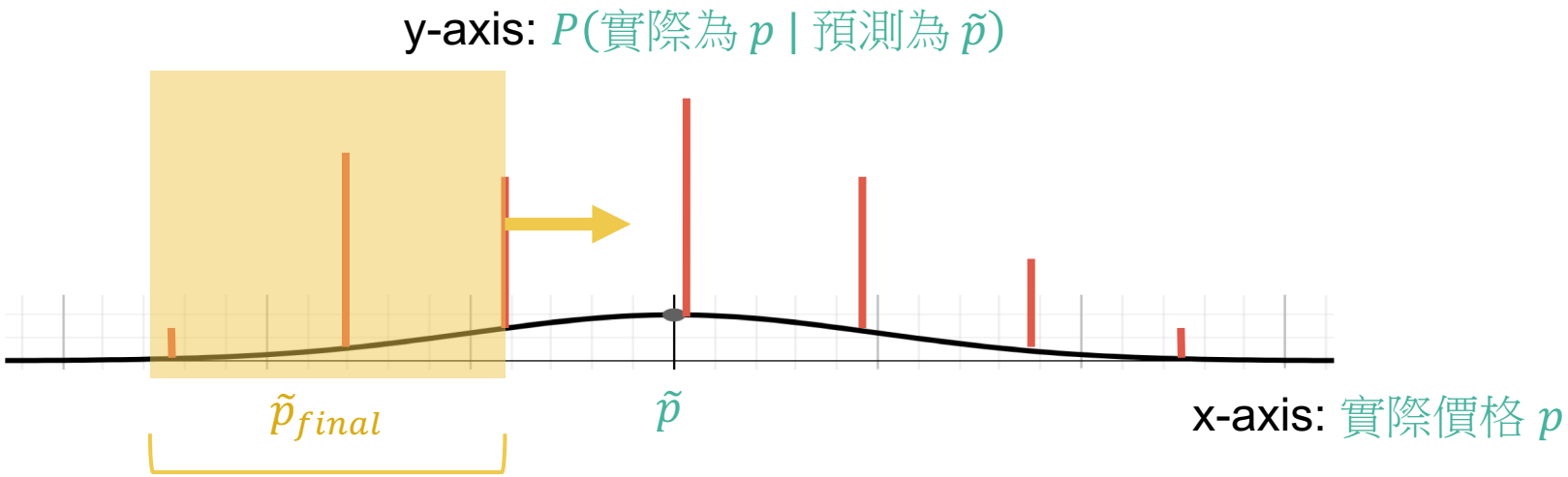


- 求在初始預測為 $\tilde{p}$ 的情況下，實際價格為1000萬的機率

$$P(\text{實際為1000萬} \mid \text{預測為 } \tilde{p}) = \frac{P(\text{預測為 } \tilde{p} \mid \text{實際為1000萬}) \cdot P(\text{實際為1000萬})}{P(\text{預測為}\tilde{p})}$$

  - $P(\text{預測為 } \tilde{p} \mid \text{實際為1000萬})$ 可用 normal distribution pdf 求得
  - $P(\text{實際為1000萬})$ 可用1000萬在 training data 中出現的次數估計
  - $P(\text{預測為}\tilde{p})$ 可直接忽略

最終預測



上傳 $\tilde{p}_{final}$ 所產生的正負10%範圍

- 找到能包含最大機率加總的區間，此時 $\tilde{p}_{final}$ 即最終預測
- Public Score: 6257 -> 6420 (+163)