基於機器學習之筆電價格預測

授課教師：曾士桓 博士
組員：林垣志、許楷俊、許至佑

摘要

本研究使用三種機器學習技術來預測筆電價格資料，本實驗的特色在於針對資料集進行多項前處理功能，將資料集分類得更詳細，訓練模型方面則是選擇了兩種樹狀結構的模型 XGBoost 與 Random forest，以及傳統的迴歸模型 Linear Regression 進行比較，實驗結果說明樹狀結構的模型對於筆電價格的預測能力較佳，最後模型使用常見的評估標準 R 平方指標呈現，再使用散點圖來說明實際值與預測值的比較。

1. 研究動機與目的

1.1. 研究動機

隨著科技的進步和消費者對便攜式電腦的需求不斷增加，現今的人們越來越需要選擇一款價格合理且符合個人需求的筆記型電腦 (簡稱筆電)，不論是用於研究相關專業領域、遊玩遊戲或是處理文書工作，選擇一台適合的筆記型電腦變得格外重要。

筆記型電腦的價格隨著廠商的硬體配製而有所不同，消費者如何正確地挑選出最合適的規格需花費大量時間了解與探索，而傳統市場的分析方法通常基於人工經驗和專業知識，然而，這些方法往往難以捕捉到市場中大量的資料和複雜的關聯性，消費者在擔心受騙、不了解市場行情的情況下，將更難以去決定理想的硬體規格。

現今機器學習模型發展速度越來越快，已經能根據大量的歷史資料和相關特徵進行分析，若透過使用這些模型進行筆電的價格預測，不僅能幫助消費者做出更明智的購買決策，同時也為生產商制定適當的價格策略和市場定位提供重要參考。綜上所述，本研究期盼利用機器學習技術預測各家筆電的價格，找到最理想且合理價錢的筆電，甚至能在實際場域上準確預測未來筆記型電腦的價格趨勢。

1.2. 研究目的

為了輔助消費者決定購買理想的筆電，我們將應用機器學習技術，以預測筆記型電腦的價格。具體而言，本次實驗使用的資料集是在 kaggle 網站上 “Laptop Price Prediction using specifications” 所提供的資料集，而使用的機器學習模型分為三種，包含線性迴歸 (Linear Regression)、隨機森林 (Random forest)、eXtreme Gradient Boosting (XGBoost)，當模型訓練完成，將測試集輸入至模型進行預測與評估實驗結果，再以視覺化圖表呈現預測值與實際值的比較。

2. 文獻探討

2.1. 回歸預測模型

2.1.1. Linear Regression [1]

迴歸分析 (Regression Analysis) 是運用一個或多個變數間的關係來解釋與預測另一個變數的統計技術總稱，目的是探討自變數 (Independent Variable，常以 X 表示) 和依變數 (Dependent Variable，常以 Y 表示) 之間的關係，利用最小平方法 (Least Squares Method) 找出最適切數學公式，並建立預測模型。在線性關係假設成立的情況下，迴歸分析以直線方程式來進行統計決策與應用，稱為線性迴歸 (Linear Regression)。

線性迴歸因自變數多寡分為兩種：(1) 只有單一自變數的簡單線性回歸 (Simple Linear Rregression)；(2) 具多個自變數的多元線性迴歸 (Multiple Linear Regression)。當只有單一自變數者的簡單線性回歸模型其公式如下：

$$Y = \beta_0 + \beta_1 X + \varepsilon \quad (1)$$

其中 Y 為依變數，X 為自變數，β_0 為截距，β_1 為迴歸係數，ε 為誤差項。若有 N 筆實際值，透過此迴歸方程式，可將自變數進行線性整合得出 Y 的預估值，估計迴歸公式如下：

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + e \quad (2)$$

$\hat{Y}$、$\hat{\beta}_0$、$\hat{\beta}_1$ 為自變數代入公式所分別得到的 Y、β_0、β_1 估計數值，e 為誤差項，表示 Y - $\hat{Y}$ 的差值，表示預測值 $\hat{Y}$ 與實際觀察值 Y 對 X 進行迴歸分析時的線性樣本估計值的差距，是利用迴歸方程式無法準確預測產生的誤差。多個自變數的多元線性回歸模型公式如下：

$$Y = \beta_0 + \beta_1 X_1 + \beta_1 X_2 + \ldots\ldots + \beta_k X_k + \varepsilon \quad (3)$$

其中 Y 為依變數，$X_1, X_2, \cdots, X_k$ 為自變數，β_0 為截距，$\beta_1, \beta_2, \cdots, \beta_k$ 為迴歸係數，ε 為誤差項。若有 N 筆實際值，透過此迴歸公式，可將自變數進行線性整合得出 Y 的估計值，估計迴歸方程式如式 (2.X)：

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_1 X_2 + \ldots\ldots + \hat{\beta}_1 X_k + e \quad (4)$$

$\hat{Y}$、$\hat{\beta}_0, \hat{\beta}_1, ..., \hat{\beta}_k$ 為自變數帶入方程式所分別得到的 Y、$\beta_0, \beta_1, \cdots, \beta_k$ 估計數值，e 為殘差，k 為第 k 個自變數。

2.1.2. Random Forest

隨機森林 (Random Forest) [2] 係以決策樹 (Decision tree) 為基礎所發展而成的技術，其可被視為是一種利用許多不同決策樹對資料樣本進行訓練並預測的學習分類器。隨機森林於建構分類預測模型的過程中引用了袋裝法 (Bagging) 的技術 [3]，其基本作法是從訓練資料樣本中多次以隨機方式抽取資料樣本並藉此訓練多個基分類器 (樣本取完後放回)。而決策樹是一種用來處理分類問題的樹形結構，每個內部節點表示一個資料特徵或評估欄位，每個葉節點表示一種分類結果，每個分支表示一個測試輸出的問題。建立決策樹的過程某種程度上就像一連串的 if-else 的問句，因此具有 (1) 可讀性高； (2) 可處理的屬性較多；(3) 對於有缺失屬性的樣本較不敏感； (4) 能在短時間內對大量資料樣本完成分類評估等的上述優點，然而，決策樹也有一些不足之處，像是 (1) 容易對訓練資料過度學習，會過度依賴訓練資料中的局部特徵，導致對測試資料的泛化能力不佳。(2) 因基於貪婪法，可能導致對全局最佳分類的忽略，從而影響決策樹的準確性。(3) 無法處理高維度資料等上述問題。

在建立模型時，隨機森林會使用未知資料樣本進行評估，讓每棵決策樹分別對其進行預測，直到每棵決策樹產生結果後，再藉由投票的方式，決定整個隨機森林的預測結果。在分類問題中，隨機森林的預測結果為多數決，而在回歸問題中，隨機森林的預測結果為所有決策樹預測值的平均值。此運作機制使得整體模型的結果具有較高的準確度和泛化能力。隨機森林除了擁有決策樹原本的優點之外，還有可以學習非線性關係的特徵、處理高維度資料、對於不平衡資料集來說，隨機森林能平衡誤差，但最重要的是降低了過度學習 (Overfitting) 的發生機率。此外可以發現，若能有效選出關鍵的特徵，則可以降低其建立模型時的運算成本，提高隨機森林的效能以及評估準確率。

2.1.3. XGBoost

全名為 eXtreme Gradient Boosting [4]，以 Gradient Boosting 為基礎下去實作。除了可以做分類也能進行迴歸連續型數值的預測，其優點包含（1）利用了二階梯度來對節點進行劃分（2）利用區域近似演算法對分裂節點進行優化（3）在損失函數中加入了 L1/L2 項（泰勒展開式），控制模型的複雜度（4）提供 GPU 平行化運算，但也面臨相對多的困難。例如：雖然利用排序和近似演算法可以降低尋找最佳分裂點的計算量，但在節點分裂過程中仍需要遍歷資料集，而當資料集屬於高維度資料時，因每次都需儲存的特徵值和其排序結果，將消耗訓練資料兩倍的記憶體，花費的成本相當高。

XGBoost 藉由不斷添加新的決策樹，透過殘差多次迭代擬合一個值，提高學習器的效率和性能。其目標函數＝損失函數＋正規化項。模型的預測準確度由模型的偏差和方差共同決定，而損失函式代表了模型的偏差（衡量預測值與實際值的誤差）。使用梯度提升演算法的好處是在 Boosting tree 被建立後，可以相對直接地得到每個屬性的重要性得分，其衡量了特徵在模型中的提升決策樹建構中的價值。

XGBoost 與隨機森林不同的地方在於前者係以 (實際結果 - 前 m - 1 顆樹的預測結果加總) 當作第 m 顆樹的預測目標，而後者是從樣本中隨機抽出 n 筆資料，重複執行迴圈 m 次，建立出 m 顆完全不相關的決策樹，且每一棵樹預測目標皆相同，最後取所有預測結果的平均當作最後的輸出結果。

2.2. 特徵選取

特徵選取通常使用於機器學習領域，又稱子集選取 (subset selection)，用來過濾掉沒有效用、不具有關鍵影響力，以及有著重複或類似鑑別能力的雜訊特徵，最後僅保留下真正對效能指標有影響的特徵，藉以決定最佳的特徵子集合，使其效能指標達到最佳化，也達到降低特徵空間 (feature space) 維度 (即，特徵數量) 的目的。運用特徵選取方法排除低關聯或是重複性值的欄位屬性一般認為對於資料分析作業具有顯著的助益 [5]。其原因在於：特徵選取能夠 (1) 提高學習模型的預測準確性；(2) 降低模型的複雜度；(3) 減少數據成本；以及 (4) 提高學習效率。舉例來說，文件分類通常使用字詞作為分類的特徵，而其效能評估指標則為分類結果的準確度。在此，文件分類的特徵選取便是刪除對於分類準確度沒有影響的字詞特徵。

在本研究中使用皮爾森相關係數給予每個特徵重要度 [6]，此方法是衡量兩個資料變數之間線性關係的強度，其值介於 -1 ~ 1 之間。若兩變數之間的相關係數絕對值較大，則表示彼此之間的關聯性程度較大。一般而言，若兩變數之間為正相關 (以 1 表示)，則當 X 提升時，Y 也會隨之提升；反之，若兩變數之間為負相關 (以－1 表示)，則當 X 提升時，Y 會隨之下降。其運算式可表示如下：

$$Corr(X,Y)=\frac{\sum_{i=1}^{N}(X_i-\bar{X})(Y_i-\bar{Y})}{\sqrt{\sum_{i=1}^{N}(X_i-\bar{X})^2}\sqrt{\sum_{i=1}^{N}(Y_i-\bar{Y})^2}} \qquad (5)$$

其中，$Corr(X, Y)$ 為相關係數，N 為共有 N 筆資料，X_i 為 X 類別的第 i 筆資訊，$\bar{X}$ 為 X 的平均數；Y_i 為 Y 類別的第 i 筆資訊，$\bar{Y}$ 為 Y 的平均數。皮爾森相關係數假設：兩個變數是常態分布，兩個變數間有線性關係存在，資料是平均分配在回歸線周圍。

3. 研究方法

本研究中，我們根據工作性質的不同而將計畫執行內容區分為五大階段進行推動，其分別為：(1) 資料蒐集；(2) 資料清洗與轉換；(3) 特徵選取；(4) 資料切割；以及 (5) 創建模型。圖 1 為本研究執行內容之概念模型，至於各工作項目的推動與執行方式說明於以下小節內容中。

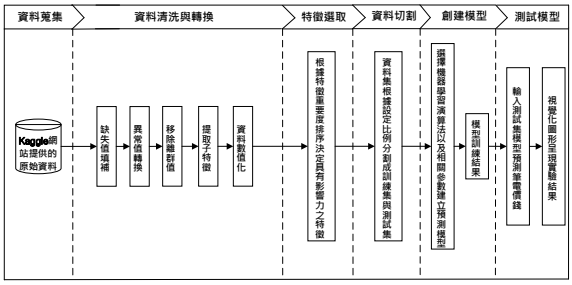


圖 1. 本研究執行內容之概念模型

3.1. 資料蒐集

本研究所分析之筆電價格資料記錄係由 Kaggle 網站所提供，為確保本研究執行所得成果可被實際應用在真實購買筆電網站，我們深入探討預測各家筆電不同硬體規格下的價錢的具體可行措施。此資料目前共有 19 家廠商，其分別為：Acer、Apple、Asus、Chuwi、Dell、Fujitsu、Google、HP、Huawei、Lenovo、LG、Mediacom、Microsoft、MSI、Razer、Samsung、Toshiba、Vero、Xiaomi。資料集總數為 1302 筆，每一筆筆電硬體規格資料紀錄的欄位資料包含：Manufacturer、Model Name、Category、Screen Size、Screen、CPU、RAM、Storage、GPU、Operating System、Operating System Version、Weight、Price 等內容。

3.2. 資料清洗與轉換

有鑑於資料集中若存在著遺漏、異常值或離群值的資料樣本，都會對資料處理與分析結果的正確性造成負面影響 [7]。因此，資料處理係資料探勘過程中不可或缺的步驟。以本研究所提供的資料內容為例，該資料集內所出現的問題包括：欄位 [Operating System Version] 的值為 NAN、欄位 [Screen Size, RAM, Weight] 的值認定為物件 (object)、欄位 [Price] 的值是國外幣值，需做幣值轉換成台幣，且需移除離群值等情況產生。透過資料前處理排除上述的異常問題。一旦將異常的問題排除，即可進行資料特徵提取與型態轉換。

以本研究為例，我們在資料清洗階段先將異常值、缺失值與幣值轉換處理完成，接著從資料集中提取特徵欄位為 Screen 的內容值，將此特徵拆分成螢幕解析度、種類，以及是否有觸控功能等子特徵；提取特徵欄位為 CPU 的內容值，將此特徵提取出處理器速度 (freq) 成新的特徵，且留下原 CPU 特徵值；由於筆電有分單/雙硬碟，因此提取特徵欄位為 Storage 的內容值，將此特徵拆分統計硬碟大小與種類。

而資料轉換的處理則是將特徵內容進行 One-Hot Encoding 與 Label Encoding。Label encoding 是把每個類別 mapping 到某個整數，不會增加新欄位，而 One hot encoding 則為每個類別新增一個欄位，用 0/1 表示是否。這兩個編碼方式的目的是為了將類別 (categorical) 或是文字 (text) 的資料轉換成數值，讓模型能夠更好的去理解及運算。圖 2 為轉換示例。

Label Encoding

Food Name	Categorical #	Calories
Apple	1	95
Chicken	2	231
Broccoli	3	50

→ One Hot Encoding

Apple	Chicken	Broccoli	Calories
1	0	0	95
0	1	0	231
0	0	1	50

圖 2. Encoding 轉換示例圖

而在執行特徵選取前，先以視覺畫圖表觀察 Price 與其他特徵的關係，幫助我們在進行特徵選取前先了解特徵間的關係程度。圖 3 為統計所有作業系統的筆電價格，圖 4 可得知不同種類的筆電也會影響價格，圖 5 則是使用相關熱圖分析數值變量與價格之間的關係。

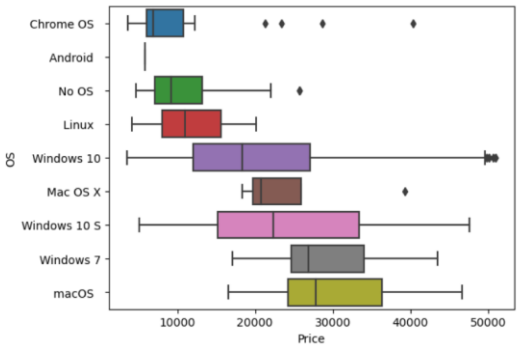


圖 3. 統計各家作業系統與筆電價格比較

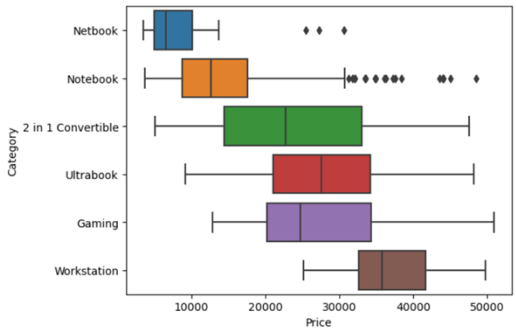


圖 4. 不同種類的筆電與價格比較

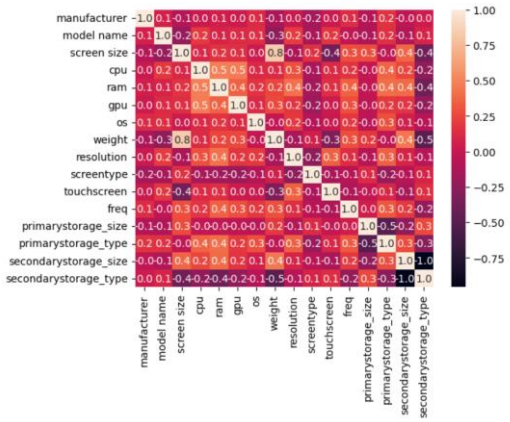


圖 5. 相關熱圖分析數值變量與價格之間的關係

3.3. 特徵選取

欲預測之資料集特徵的多寡、重要性以及其型態應用於關聯分析、分類預測、分群歸類等技術層面之結果，皆會對模型訓練時間與預測效果產生重大影響。因此，如何從資料集中挑選出對預測有利之特徵極其重要。普遍來說，會採用特徵選取技術 (feature selection) 作為輔助的工具。我們在本研究中採用皮爾森相關係數 (Pearson Correlation) 當作主要的特徵選取手段，產生出特徵重要度並進行排名。依據重要度排序和研究需要，選取分數較高之特徵作為後續創建模型使用。

3.4. 資料切割

為了能夠將資料用於製作模型，必須先將資料集切割為用於創建模型之訓練集，以及用於測試模型效能之測試集。預測的問題不同，訓練集與測試集之間比例也不盡相同，且有可能會根據獲取資料數量多寡而有所變化。因此，本階段將針對前處理完後的資料集進行切割，並設定欲切割的資料集比例，切割出訓練集和測試集，作為預測模型的資料來源。

3.5. 創建模型

我們在創建模型時可採用三種方法，第一種為使用線性回歸模型 (Linear Regression)。建立模型時，會將特徵選取後的訓練集之特徵輸入至線性回歸模型中對其進行訓練。訓練完成後，將對測試集進行價格的預測。

第二種方法為使用隨機森林 (Random Forest)，在建立隨機森林模型時，先從訓練集中隨機選擇 N 個資料樣本 (取完後放回) 構成一子訓練集，再利用子訓練集建構出一棵決策樹，接著，從決策樹裡的所有待選特徵中隨機選取一定的特徵，其做法為從 M 個特徵內隨機選取 m 個特徵 (m < M)，並使用方法 (如 Gini 係數) 比較哪種特徵最適合作為節點，如此反覆建構出 K 棵決策樹，以完成一座隨機森林。當建構完成後，將使用建構出的隨機森林模型對測試集進行預測。

最後一種方法則是使用 XGBOOST 模型。其預測過程是以 (實際結果 - 前 m - 1 顆樹的預測結果加總) 當作第 m 顆樹的預測目標，而超參數定設定包含 learning_rate = 0.19、n_estimators = 500、max_depth = 8、min_child_weight = 1。當建構完成後，將使用建構出的 XGBOOST 模型對測試集進行預測。

4. 實驗結果

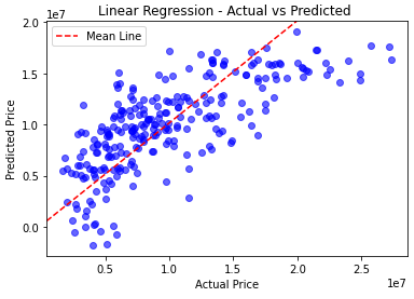
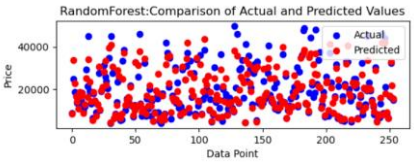
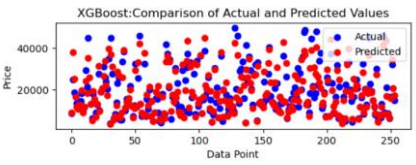
我們利用 R 平方(R squared) 當作評估模型的指標，是一種衡量回歸模型表現的指標，代表從獨立變數 X 可以解釋依變數 Y 變異的比例。當殘差平方和越小，表示不能解釋的部分越小，則 R 平方自然越高，如果找到一條完美回歸線穿越所有資料點，使殘差變成 0，R 平方就會變成 1，表示模型解釋所有變異量。因此，正常情況下 R 平方的值會落在[0,1]。其公式如下：

$$R^2 = 1 - \frac{\sum_{i=1}(y_i - f_i)^2}{\sum_{i=1}(y_i - \bar{y})^2} \qquad (6)$$

公式裡的分子為殘差平方和 (residual sum of squares) 就是『不能解釋的部分』，分子為總平方和 (total sum of squares) 為『欲解釋的總變異量』，y_i 為實際值，f_i 為預測值，$\bar{y}$ 為平均值。

在實驗結果中，最佳的模型為有設定參數後的 XGBoost，其 R-squared Score = 83.04 %，其次為隨機森林，其 R-squared Score = 85.07 %，最後為 Linear Regression，其 R-squared Score = 66.54%。而三個模型之預測結果與實際值的比較如表 1 所示。

表 1. 三種模型的預測值與實際值比較圖

線性迴歸	
隨機森林	
XGBOOST	

5. 結論與未來展望

5.1. 結論

本次實驗使用 Linear Regression、Random forest、XGBoost 三種機器學習模型對 Kaggle 各家筆電硬體規格之資料集進行價格預測，其結果是 XGBoost 的效果最佳，在測試集中達到 85.07% 的預測準確率，而同樣是樹狀結構的 RF 模型在相同資料集與參數的情況下也可以達到 83.04% 的預測準確率，因此可認為樹狀結構的機器學習模型有利於各家廠商的筆電價錢預測，雖最後無使用特徵選取方法，但預測準確率卻比有使用特徵選取的資料集的預測結果還高，藉此可推斷在前處理資料集後的所有特徵都一定程度的重要性。

5.2. 未來展望

本研究應可針對資料集更進一步的處理，尤其是對各家廠牌的價格進行分析，將文書筆電與遊戲筆電分類，藉此確認實際各家廠商筆電裡的硬體規格，以及標定價格，而在特徵選取方面，應再找比皮爾森相關係數更佳的演算法，提高預測準確率，避免與此次最後無使用特徵選取功能的情況發生。最後，未來若有機會，可使用其他機器學習模型進行預測，像是傳統 SVM 或較新的 LightGBM。

參考文獻

[1] 線性迴歸相關資料：上課投影片與 Chatgpt

[2] L. Breiman, “Random forests,” *Machine Learning*, vol. 45, no. 1, pp. 5-32, 2001.

[3] L. Breiman, “Bagging Predictors,” *Machine Learning*, vol. 24, no. 2, pp. 123-140, 1996.

[4] T. Chen and C. Guestrin, “XGBoost: ： A Scalable Tree Boosting System,” *In Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pp. 785-794, 2016.

[5] H. Liu, H. Motoda, and L. Yu, “A selective sampling approach to active feature selection,” *Artificial Intelligence*, vol. 159, no. 1-2, pp. 49-74, 2004.

[6] I. M. Nasir, M. A. Khan, M. Yasmin, J. H. Shah, M. Gabryel, R. Scherer, and R. Damaševičius, “Pearson Correlation-Based Feature Selection for Document Classification Using Balanced Training,” *Sensors*, vol. 20, no. 23, pp. 6793-6811, 2020.

[7] S. F. Crone, S. Lessmann, and R. Stahlbock, “The impact of preprocessing on data mining: An evaluation of classifier sensitivity in direct marketing,” *European Journal of Operational Research*, vol. 173, no. 3, pp. 781-800, 2006.

組員工作分配表

林垣志	資料前處理 XGBoost 模型 Python 程式實作 期中、期末報告
許楷俊	資料蒐集 隨機森林模型 Python 程式實作 期末報告
許至佑	特徵選取 線性回歸模型 Python 程式實作 期末報告