好萊塢電影利潤預測

與台積電股價預測

資管四 409402207 黃姸馨

統資四 409421285 廖彥甄

目錄

[第一章 緒論 1](#_Toc155822098)

[1.1 研究背景與動機 1](#_Toc155822099)

[1.2 研究目的 1](#_Toc155822100)

[第二章 研究方法 3](#_Toc155822101)

[2.1 資料來源 3](#_Toc155822102)

[2.2 研究架構及流程 3](#_Toc155822103)

[第三章 研究過程 6](#_Toc155822104)

[3.1 資料預處理 6](#_Toc155822105)

[3.2 變數之敘述性統計 8](#_Toc155822106)

[3.3 分類分析 11](#_Toc155822107)

[3.4 迴歸分析 11](#_Toc155822108)

[3.5 股票分析 12](#_Toc155822109)

[第四章 研究結果 15](#_Toc155822110)

[5.1 分類分析 15](#_Toc155822111)

[5.2 迴歸分析 19](#_Toc155822112)

[5.3 股票分析 22](#_Toc155822113)

[第五章 結論 27](#_Toc155822114)

[5.1 分類分析 27](#_Toc155822115)

[5.2 迴歸分析 27](#_Toc155822116)

[5.3 股票分析 28](#_Toc155822117)

[第六章 參考文獻 29](#_Toc155822118)

第一章 緒論

## 研究背景與動機

好萊塢為全球最大的電影產業中心，其影響力遍及全球。每年製作的電影不僅在票房上創造巨大收益，同時也對文化、娛樂產業乃至於全球經濟產生深遠影響。因此，分析好萊塢電影的收益模式，對於電影製作公司、投資者、市場分析師以及學術研究者都具有重要的意義。

台灣積體電路製造股份有限公司為台灣最大的專業積體電路製造服務公司，在台灣半導體產業中佔有舉足輕重的地位。台積電股票表現不僅反映了公司自身的經營狀況，同時也是科技產業和經濟健康的重要指標。其股價受到多種因素的影響，包括市場總體經濟狀況、行業趨勢、公司經營績效、技術創新等。本研究欲通過分析歷史數據和市場趨勢，準確預測台積電股價趨勢。

## 研究目的

在此研究中，將以分類模型和迴歸模型來分析電影與股票資料，以下為本研究之目的:

1. 利用分類和迴歸模型來分析好萊塢電影利潤，找出各分析中表現較好的模型適合用於分析好萊塢電影的收益，以找出哪些超參數設定能夠提供最佳的預測效果，以及對於電影利潤之具有影響力的變數。
2. 使用迴歸模型分析2330台灣積體電路製造股票之走勢，並提出股票預測模型中較好的模型後，給予股票未來走勢之預測。

第二章 研究方法

## 資料來源

Kaggle提供了大量的公開數據集，為數據科學家和機器學習提供服務的平台，在好萊塢電影利潤分析中，選擇的資料是Kaggle網站上的Top 1000 Highest Grossing Movies。另外在股票分析上，經由yahoo finance網站下載，使用了台灣積體電路製造的股票，股票代碼為2330。

## 研究架構及流程

圖2-2-1為分類分析之研究架構流程圖，第一步是將資料收集並做資料前處理，接下來將使用CART、C4.5、SVM、KNN演算法來建構模型，並評估模型，找出較好的模型與各模型之變數重要性，最後討論研究之結果。

一張含有 文字, 螢幕擷取畫面, 圖表, 字型 的圖片

自動產生的描述

圖2-2-1、分類分析之研究架構流程圖

迴歸分析之研究架構流程圖如圖2-2-2所示，如同分類分析流程，差別在於建構模型上，在迴歸分析上使用CART、SVR和KNN演算法建立模型。

一張含有 文字, 螢幕擷取畫面, 行, 字型 的圖片

自動產生的描述

圖2-2-2、迴歸分析之研究架構流程圖

圖2-2-3為股票分析流程圖，將資料處理後，將資料訓練集內分為Expanding和Rolling方法，並各自使用Linear Regression、CART、SVR與KNN演算法建構模型，最後評估並找出相對較好的模型後，討論研究結果。

一張含有 文字, 螢幕擷取畫面, 圖表, 字型 的圖片

自動產生的描述

圖2-2-3、股票分析之研究架構流程圖

第三章 研究過程

## 資料預處理

本研究所使用的Top 1000 Highest Grossing Movies原始資料總筆數共1000筆，共14個變數，5個連續型變數，9個類別變數。台積電股票的原始資料則不需要做資料預處理。

1. **資料格式轉換**

將欄位Running Time的資料單位轉換成分鐘，並重新命名為Running\_Time，如表3-1-1所示。

表3-1-2、變數格式轉換

|  |  |
| --- | --- |
| 狀態 | df[‘Running\_Time’] |
| 處理前 | 一張含有 文字, 螢幕擷取畫面, 字型, 數字 的圖片  自動產生的描述 |
| 處理後 | 一張含有 文字, 螢幕擷取畫面, 字型, 代數 的圖片  自動產生的描述 |

1. **資料分類**

將變數Genre中每一種電影類型自成一個欄位，若一部電影為此類型則為1，不是為0。將變數Distributor中最佔比數最多的前四名經銷商提出來分別自成一個欄位，其他經銷商則歸類為other。將變數Year分為Year\_0奇數年及Year\_1偶數年兩個欄位變數。以下為表3-1-2:

表3-1-2、變數資料分類

|  |  |  |  |
| --- | --- | --- | --- |
|  | df[‘Genre’] | df[‘Distributor’] | df[‘Year’] |
| 處理前 | 一張含有 文字, 字型, 螢幕擷取畫面, 數字 的圖片  自動產生的描述 | 一張含有 文字, 螢幕擷取畫面, 字型, 收據 的圖片  自動產生的描述 | 一張含有 文字, 螢幕擷取畫面, 字型 的圖片  自動產生的描述 |

1. **移除遺失值及欄位**

移除欄位License、Distributor、Budget中有遺失值的資料，並將與目標變數有直接相關的欄位World Wide Sales、Domestic Opening、Domestic Sales、International Sales刪除。

1. **建立預測目標**

迴歸分析的目標變數為全球總銷售額-Budget，分類分析的目標變數為全球總銷售額-Budget並判斷是否大於一億五千萬，等於1為是，等於0為否 (Quader et al., 2017)。

## 變數之敘述性統計

本研究資料經資料預處理後總筆數共772筆，共29個變數，4個連續型變數，25個類別變數，以表3-2-1和表3-2-2表示:

表3-2-1、目標變數介紹

|  |  |
| --- | --- |
| **Y -目標變數** | |
| **Independent Variable (分類)** | **N = 772; n (%)** |
| Profit |  |
| (1) Yes | 63.1% |
| (2) No | 36.9% |
| **Independent Variable (迴歸)** | **Mean ± SD** |
| Profit | 3250498**±**269159900 |

表3-2-2、自變數介紹

|  |  |  |  |
| --- | --- | --- | --- |
| **X -自變數 連續** | | **X -自變數 類別** | |
| **Independent Variable** | **Mean ± SD** | **Independent**  **Variable** | **N = 772;**  **n (%)** |
| Running \_Time | 117.91**±**21.08 | Distributor |  |
| **X -自變數 類別** | | (1) Warner Bros. | 16.6% |
| **Independent Variable** | **N = 772;**  **n (%)** | (2) Twentieth  Century Fox | 12% |
| Year |  | (3) Walt Disney  Studios Motion  Picture | 16% |
| (1) odd | 51.5% | (4) Universal  Pictures | 13% |
| (2) even | 48.5% | (5) others | 42.4% |
| Action |  | License |  |
| (1) NO | 49.2% | (1) G | 1.8% |
| (2) Yes | 50.8% | (2) PG | 23.1% |
| Adventure |  | (3) PG-13 | 51.1% |
| (1) NO | 48.7% | (4) R | 23.5% |
| (2) Yes | 51.3% | Comedy |  |
| Animation |  | (1) NO | 61.2% |
| (1) NO | 85.4% | (2) Yes | 38.8% |
| (2) Yes | 14.6% | Crime |  |
| Biography |  | (1) NO | 87.1% |
| (1) NO | 95.6% | (2) Yes | 12.9% |
| (2) Yes | 4.4% | Documentary |  |
| Drama |  | (1) NO | 99.9% |
| (1) NO | 69.6% | (2) Yes | 0.1% |
| (2) Yes | 30.4% | Family |  |
| Fantasy |  | (1) NO | 76.7% |
| (1) NO | 72.5% | (2) Yes | 23.3% |
| (2) Yes | 27.5% | Horror |  |
| Music |  | (1) NO | 93.7% |
| (1) NO | 92.8% | (2) Yes | 6.3% |
| (2) Yes | 7.2% | Musical |  |
| Mystery |  | (1) NO | 94.8% |
| (1) NO | 90.3% | (2) Yes | 5.2% |
| (2) Yes | 9.7% | Romance |  |
| Sci-Fi |  | (1) NO | 84.2% |
| (1) NO | 71.4% | (2) Yes | 15.8% |
| (2) Yes | 28.6% | Sport |  |
| Thriller |  | (1) NO | 98.1% |
| (1) NO | 71.4% | (2) Yes | 1.9% |
| (2) Yes | 28.6% | War |  |
| Western |  | (1) NO | 97.1% |
| (1) NO | 98.7% | (2) Yes | 2.9% |
| (2) Yes | 1.3% |  |  |

## 分類分析

在分類分析中，我們建立預測模型，將資料以80/20的比例切分訓練集與測試集，並在訓練集中分別以Holdout+Holdout、Cross Validation+、Holdout、Nested的方式訓練模型。以下表3-3-1為不同演算法之超參數與其數值(Marco et al., 2022; Han et al., 2012; Grama et al., 2017; Nikdel et al., 2018):

表3-3-1、分類分析使用模型及超參數介紹

|  |  |
| --- | --- |
| **Model** | **Hyperparameters** |
| CART | max\_depth: 2, 6, 8  max\_leaf\_nodes: 5, 20, 100  min\_samples\_leaf: 20, 40, 100 |
| C4.5 |
| SVM | kernel: rbf  C: 4, 16, 64  gamma: 0.2, 0.5, 8 |
| kernel: sigmoid  C: 4, 16, 64  gamma: 0.2, 0.5, 8  r: 0.25, 0.35, 2 |
| KNN | k: 1, 3, 5  algorithm: ‘ball\_tree’, ‘kd\_tree’, ‘brute’  p: 1, 2 |

## 迴歸分析

在迴歸分析中，我們將資料以80/20的比例切分訓練集與測試集，並建立預測模型。本研究在迴歸分析中使用了CART、SVR、KNN分析，其中表3-4-1說明各演算法使用的參數 (Marco et al., 2022; Han et al., 2012; Grama et al., 2017; Nikdel et al., 2018)。

表3-4-1迴歸分析使用模型及超參數介紹

|  |  |
| --- | --- |
| **Model** | **Hyperparameters** |
| CART | max\_depth: 2, 6, 8  max\_leaf\_nodes: 5, 20, 100  min\_samples\_leaf: 20, 40, 100 |
| SVR | kernel: rbf  C: 4, 16, 64  gamma: 0.2, 0.5, 8 |
| kernel: sigmoid  C: 4, 16, 64  gamma: 0.2, 0.5, 8  r: 0.25, 0.35, 2 |
| KNN | k: 1, 3, 5  algorithm: ‘ball\_tree’, ‘kd\_tree’, ‘brute’  p: 1, 2 |

## 股票分析

1. **股票基本資訊**

在股票分析方面，此研究使用 2330台灣積體電路製造，研究時間為2022年12月12號至2023年12月15號，總共筆數為236筆，圖3-5-1為股票走勢圖。將2330股票資訊下載後，留下收盤日期與收盤價，並透過收盤價來算出前一天、前兩天、前三天收盤價，並且將當日收盤價為此股票研究之預測目標 (Y)。

一張含有 文字, 螢幕擷取畫面, 行, 筆跡 的圖片

自動產生的描述

圖3-5-1、2330股票走勢圖

1. **建構預測模型**

將資料分成80%的訓練集以及20%的測試集，分別為190筆和46筆。接著對於訓練集分別採用Expanding和Rolling來建立迴歸模型，使用先前的迴歸分析中表3-4-1的模型。圖3-5-2和圖3-5-3為Expanding Window 和 Rolling Window圖。

一張含有 螢幕擷取畫面, 行, Rectangle, 平行 的圖片

自動產生的描述

圖3-5-2、Expanding Window 圖

一張含有 螢幕擷取畫面, 文字, 行, Rectangle 的圖片

自動產生的描述

圖3-5-3、Rolling Window 圖

第四章 研究結果

## 分類分析

1. **模型評估與比較**

表4-1-1顯示將每個演算法的超參數帶入模型後，得出對應的各個評估指標數值。因CART與C4.5的最佳超參數相同外，且在各項指標的數值、ROC圖與混淆矩陣皆相同，因此將兩者放在一起。從表中得知在四種模型中，以AUC作為標準 (Huang & Ling, 2005)，CART&C4.5的表現是最好的，AUC為0.7。另一方面，從所有的指標來看，CART&C4.5除了在Specificity數值最差，其他指標都表現比較好。其次表現較好的模型為SVR的rbf kernel，第三名為KNN，最後則是SVR的sigmoid kernel，後三者的AUC差別只差距0.01。

表4-1-1、分類分析模型評估與比較

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Algorithm** | **Best**  **Hyperparameters** | **Accuracy** | **F1**  **Score** | **Precision** | **Sensitivity** | **Specificity** | **AUC** |
| CART  & C4.5 | max\_depth: 6  max\_leaf\_nodes: 20  min\_samples\_leaf: 20 | 81.69 | 89.92 | 81.69 | 81.90 | 0.00 | 0.7 |
| SVM | kernel: rbf  C: 4  gamma: 0.5 | 60.75 | 73.65 | 64.56 | 85.71 | 16.42 | 0.55 |
| kernel: sigmoid  C: 4  gamma: 0.5  r: 0.25 | 55.91 | 64.35 | 66.67 | 62.18 | 44.78 | 0.57 |
| KNN | C: 5  algorithm: ‘ball\_tree’  p: 1 | 62.90 | 72.51 | 68.94 | 76.47 | 38.81 | 0.56 |

下表4-1-2為每個模型之ROC圖和混淆矩陣，在ROC圖中表現最好的是CART&C4.5，然而在混淆矩陣中，CART&C4.5的TN為0，代表模型沒有正確的預測電影利潤低於一億五千萬的電影，相對於其他模型來說表現較差; 其他三者的ROC圖因為AUC相差不大而差不多，但在混淆矩陣上表現較好的模型為KNN，有117筆資料被正確分類。

表4-1-2、分類模型ROC圖和混淆矩陣

|  |  |  |
| --- | --- | --- |
| **Algorithm** | **ROC圖** | **混淆矩陣** |
| CART  & C4.5  (Nested) | **一張含有 文字, 行, 螢幕擷取畫面, 繪圖 的圖片  自動產生的描述** | **一張含有 文字, 螢幕擷取畫面, 鮮豔, Rectangle 的圖片  自動產生的描述** |
| SVM  (kernel=rbf)  (Holdout+Holdout) | **一張含有 文字, 圖表, 行, 繪圖 的圖片  自動產生的描述** | **一張含有 螢幕擷取畫面, 文字, 鮮豔, 圖表 的圖片  自動產生的描述** |
| SVM  (kernel:sigmoid)  (Holdout+Holdout) | **一張含有 文字, 圖表, 行, 螢幕擷取畫面 的圖片  自動產生的描述** | **一張含有 文字, 螢幕擷取畫面, 鮮豔, 圖表 的圖片  自動產生的描述** |
| KNN  (Holdout+Holdout) | **一張含有 文字, 行, 繪圖, 螢幕擷取畫面 的圖片  自動產生的描述** | **一張含有 文字, 螢幕擷取畫面, 鮮豔, 圖表 的圖片  自動產生的描述** |

圖4-1-1為CART&C4.5模型的決策樹，由電影的運行時間為第一個節點，並往下的節點分別為電影分級為建議家長指導級和電影類型為冒險。

**一張含有 文字, 筆跡, Rectangle, 字型 的圖片

自動產生的描述**

圖4-1-1、CART&C4.5決策樹

1. **重要變數**

將各個模型之前十名重要變數提出，由於KNN沒有變數重要性，因此只列出四個模型，並且以平均值計算總排名，如表4-1-3所示。

表4-1-3、分類分析之變數重要性

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | CART | SVR (rbf) | SVR (sigmoid) | 總排名 |
| 1 | Running\_Time | Comedy\_1 | Year\_1 | Comedy\_1 |
| 2 | Distributor\_other | Adventure\_1 | Year\_0 | Fantasy\_1 |
| 3 | War\_0 | Distributor\_ Walt Disney Studios Motion Pictures | License\_R | Mystery\_1 |
| 4 | Mystery\_1 | Action\_1 | Thriller\_0 | Year\_0 |
| 5 | Animation\_1 | Mystery\_1 | Fantasy\_1 | Drama\_0 |
| 6 | Fantasy\_1 | Sci-Fi\_1 | Adventure\_1 | Running\_Time |
| 7 | Crime\_0 | Crime\_0 | Family\_1 | Crime\_0 |
| 8 | Year\_0 | Romance\_0 | Distributor\_ Walt Disney Studios Motion Pictures | Animation\_1 |
| 9 | License\_PG-13 | Biography\_0 | Drama\_0 | Adventure\_1 |
| 10 | Comedy\_1 | Thriller\_0 | Animation\_1 | Distributor\_ Walt Disney Studios Motion Pictures |

## 迴歸分析

1. **模型評估與比較**

表4-2-1顯示各演算法的最佳參數組合以及其對應的各個評估指標數值。這六項指標數值越小代表其模型越好，從所有的指標來看，KNN的模型表現最好，在所有模型指標的數值都最小。

表4-2-1、迴歸分析模型評估與比較

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Algorithm** | **Best Hyperparameters** | **RMSE** | **MSE** | **MAPE** | **SMAPE** | **MdRAE** | **MASE** |
| CART | max\_depth: 6  max\_leaf\_nodes: 20  min\_samples\_leaf: 20 | 24563882814.  66 | 6.0338e+18 | 10173.  61 | 4537.56 | 3790.46 | 59.14 |
| SVR | kernel: rbf  C: 4  gamma: 0.5 | 30708298905.  94 | 9.4300e+18 | 8009.49 | 5133.88 | 4582.38 | 65.89 |
| kernel: sigmoid  C: 4  gamma: 0.5  r: 0.25 | 30708299381.  09 | 9.4300e+18 | 8009.49 | 5133.88 | 4582.38 | 65.89 |
| KNN | C: 5  algorithm: ‘ball\_tree’  p: 1 | 21479337075.  94 | 4.6136e+18 | 7965.25 | 4202.84 | 3535.79 | 52.95 |

圖4-2-1為KNN模型的預測散佈圖，紅線為迴歸線，表示自變數和目標變數之間的預測關係。

一張含有 行, 螢幕擷取畫面, 繪圖, 圖表 的圖片

自動產生的描述

圖4-2-1、KNN模型預測散佈圖

1. **變數重要性**

因KNN演算法沒有計算重要變數的模組，因此在迴歸分析中本研究以CART、SVR計算各演算法最佳模型之前十名重要變數，並以平均值計算總排名。以下為表4-2-2:

表4-2-2、迴歸分析之變數重要性

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | CART | SVR (rbf) | SVR (sigmoid) | 總排名 |
| 1 | Action | Animation | Animation | Animation |
| 2 | Western | Adventure | Sci-Fi | Year\_1 |
| 3 | Year\_0 | War | License\_PG-13 | Western |
| 4 | Animation | Family | Distributor\_Warner Bros | Sci-Fi |
| 5 | Biography | Comedy | Year\_1 | Year\_0 |
| 6 | Year\_1 | Distributor\_other | Drama | Distributor\_Warner Bros |
| 7 | Documentary | Year\_1 | Western | License\_PG-13 |
| 8 | Sci-Fi | Year\_0 | Mystery | Crime |
| 9 | Running\_Time | Distributor\_Warner Bros | Romance | Drama |
| 10 | License\_PG-13 | Western | Crime | Action |

根據表4-2-3變數重要性之總排名，本研究對這些重要變數更深入探討其與電影利潤的相關性，如圖4-2-2及圖4-2-3。可以發現電影類型為卡通片及科幻片的電影平均利潤較非此類型之電影高; 而西方片、戲劇片及犯罪的電影平均利潤較非此類型之電影低。

一張含有 圖表, 螢幕擷取畫面, 行, Rectangle 的圖片

自動產生的描述

圖4-2-2、卡通片及科幻片對電影利潤的箱型圖

一張含有 圖表, 螢幕擷取畫面, 行, 文字 的圖片

自動產生的描述

圖4-2-3、西方片、戲劇片及犯罪對電影利潤的箱型圖

## 股票分析

1. **Expanding 方法**

下表4-3-1為各個演算法中表現較好的超參數，以及各個評估指標的數值。在表中Linear Regression 表現最好，在所有評估指標中的數值都是最低的。其次是CART與KNN，最後則為SVR的rbf kernel與 sigmoid kernel。

表4-3-1、Expanding方法之評估模型與比較

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Algorithm** | **Best Hyperparameters** | **RMSE** | **MSE** | **MAPE** | **SMAPE** | **MdRAE** | **MASE** |
| Linear Regression |  | 6.25 | 39.05 | 0.85 | 0.85 | 0.66 | 0.81 |
| CART | max\_depth: 6  max\_leaf\_nodes:5 min\_samples\_leaf: 20 | 7.33 | 53.67 | 0.99 | 0.99 | 0.73 | 0.93 |
| SVR | kernel: rbf  C: 64  gamma: 0.2 | 28.93 | 836.70 | 4.32 | 4.45 | 3.92 | 4.17 |
| kernel: sigmoid  C: 4  gamma: 0.5  r: 0.25 | 32.64 | 1065.46 | 5.01 | 5.17 | 6.15 | 4.84 |
| KNN | n\_neighbors: 5  algorithm='brute'  p: 1 | 7.26 | 52.66 | 1.04 | 1.04 | 0.84 | 0.99 |

表4-3-2為各個演算法各自的走勢圖與預測散佈圖，Linear Regression 在剛剛的各個評估指標的表現上為最好，但在走勢圖上看似有給予預測的走勢，跟原本的股票走勢幅度相同，預測結果幾乎只有將原本走勢向右平移，預測散佈圖上的預測點幾乎落在線上; 在CART方面，走勢圖和預測散佈圖的表現沒有很好，走勢只有些許的幅度，而預測點也呈直線，幾乎沒有落在迴歸線上; SVR的 rbf kernel中，預測散佈圖的預測點只有極少數的點在迴歸線附近，在走勢圖上雖然高度上有差距，但有表現出股票波動; kernel為sigmoid中，不論是走勢圖或預測散佈圖皆是表現最差的，因為資料沒有訓練起來，因此預測的走勢和預測點皆顯示成一條水平線; 最後是 KNN，在五張走勢圖中是表現最好的，在預測散佈圖則是幾乎所有的點都落在迴歸線上。

表4-3-2、Expanding方法之各模型股價走勢圖與預測散佈圖

|  |  |  |
| --- | --- | --- |
| **Algorithm** | **走勢圖** | **預測散佈圖** |
| Linear Regression | 一張含有 螢幕擷取畫面, 繪圖, 行, 圖表 的圖片  自動產生的描述 | 一張含有 行, 螢幕擷取畫面, 繪圖, 圖表 的圖片  自動產生的描述 |
| CART | 一張含有 文字, 圖表, 行, 繪圖 的圖片  自動產生的描述 | 一張含有 螢幕擷取畫面, 行, 繪圖, 文字 的圖片  自動產生的描述 |
| SVR  (kernel: rbf) | 一張含有 繪圖, 行, 圖表, 螢幕擷取畫面 的圖片  自動產生的描述 | 一張含有 行, 螢幕擷取畫面, 繪圖, 圖表 的圖片  自動產生的描述 |
| SVR  (kernel: sigmoid) | 一張含有 繪圖, 行, 螢幕擷取畫面, 圖表 的圖片  自動產生的描述 | 一張含有 行, 螢幕擷取畫面, 繪圖, 圖表 的圖片  自動產生的描述 |
| KNN | 一張含有 行, 繪圖, 螢幕擷取畫面, 圖表 的圖片  自動產生的描述 | 一張含有 行, 螢幕擷取畫面, 繪圖, 圖表 的圖片  自動產生的描述 |

1. **Rolling 方法**

與Expanding作法相同，超參數帶入每個演算法的模型後，得出對應的各個評估指標數值，在此找出的各個演算法超參數中，只有KNN的超參數與先前的不同，以表4顯示。在結果中，Linear Regression 表現最好，不同的是其次表現較好的是KNN，再來則是CART，最後一樣為SVR的rbf kernel與 sigmoid kernel。表4-3-3為評估模型與比較。

表4-3-3、Rolling方法之評估模型與比較

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Algorithm** | **Best Hyperparameters** | **RMSE** | **MSE** | **MAPE** | **SMAPE** | **MdRAE** | **MASE** |
| Linear Regression |  | 6.25 | 39.05 | 0.85 | 0.85 | 0.66 | 0.81 |
| CART | max\_depth=6  max\_leaf\_nodes=5 min\_samples\_leaf=20 | 8.22 | 67.54 | 1.11 | 1.12 | 0.94 | 1.05 |
| SVR | kernel: rbf  C: 64  gamma: 0.2 | 28.93 | 836.70 | 4.32 | 4.45 | 3.92 | 4.17 |
| kernel: sigmoid  C: 4  gamma: 0.5  r: 0.25 | 32.64 | 1065.46 | 5.01 | 5.17 | 6.15 | 4.84 |
| KNN | n\_neighbors=5  algorithm='brute'  p=1 | 7.07 | 50.00 | 1.01 | 1.01 | 0.89 | 0.96 |

表4-3-4為不同演算法的走勢圖與預測散佈圖，Linear Regression、CART、SVR的結果與Expanding的結果相同。KNN則因超參數不同而有所差別，走勢圖的最後走勢是開始往上漲，而在先前的結果則是往下跌; 預測散佈圖的預測點有些許位置不同，但整體依舊是落點在迴歸線上。

表4-3-4、Rolling方法之各模型股價走勢圖與預測散佈圖

|  |  |  |
| --- | --- | --- |
| **Algorithm** | **走勢圖** | **預測散佈圖** |
| Linear Regression | 一張含有 螢幕擷取畫面, 繪圖, 行, 圖表 的圖片  自動產生的描述 | 一張含有 行, 螢幕擷取畫面, 繪圖, 圖表 的圖片  自動產生的描述 |
| CART | 一張含有 文字, 圖表, 行, 繪圖 的圖片  自動產生的描述 | 一張含有 螢幕擷取畫面, 行, 繪圖, 文字 的圖片  自動產生的描述 |
| SVR  (kernel: rbf) | 一張含有 繪圖, 行, 圖表, 螢幕擷取畫面 的圖片  自動產生的描述 | 一張含有 行, 螢幕擷取畫面, 繪圖, 圖表 的圖片  自動產生的描述 |
| SVR  (kernel: sigmoid) | 一張含有 繪圖, 行, 螢幕擷取畫面, 圖表 的圖片  自動產生的描述 | 一張含有 行, 螢幕擷取畫面, 繪圖, 圖表 的圖片  自動產生的描述 |
| KNN | 一張含有 文字, 螢幕擷取畫面, 行, 繪圖 的圖片  自動產生的描述 | 一張含有 行, 螢幕擷取畫面, 繪圖, 斜率、斜坡 的圖片  自動產生的描述 |

第五章 結論

## 分類分析

由以上的研究結果得知，在各項指標數據上表現較好的模型為CART&C4.5，然而在混淆矩陣中表現較差; 而KNN則是在混淆矩陣中為四個模型中表現較好的。

在變數重要性排名上，當電影類型為喜劇、奇幻、懸疑、動畫、冒險時，對收益較具有影響力; 另外播放時長也在前十名內，可以了解到播放時長會影響電影收益; 在經銷商方面，迪士尼為第十名之變數重要性，而迪士尼主要製作動畫片。

## 迴歸分析

藉由研究結果中得知，在各項評估指標數據上表現最好的模型為KNN，且其最佳超參數組合為k=5、 algorithm=‘ball\_tree’與 p=1。

在變數重要性分析中，可以得知當電影類型為卡通、科幻時，電影收益相對較高，且變異程度也較高;然而當電影類型為西方、劇情、犯罪片時，電影收益則相對較低。

另外經銷商為華納兄弟、不適合13歲以下兒童觀賞之電影也都在迴歸分析的變數重要性排名當中，對電影收益也具有影響力。

## 股票分析

在訓練集使用Expanding和Rolling方法中，Linear Regression 在各項評估指標上皆為表現較好的模型，而在走勢圖中則是KNN表現較佳， 另外預測散佈圖上，Linear Regression與KNN表現不相上下。

而在股票趨勢圖可以看到不同模型之預測走向: Linear Regression對於未來的走向為持續穩定上升，股價有機會回到歷史最高價; 而Expanding方法的KNN模型結果則為股價開始從最高點下跌，但還是維持在一定的高度; 最後在Rolling方法的KNN模型之股票走勢圖上，股價開始上升，然而其浮動機率也比較大。整體而言，此股票雖然目前的浮動較大，但與前半年相比股價已經成長許多。

第六章 參考文獻

1. Quader, N., Gani, M. O., & Chaki, D. (2017, December). Performance evaluation of seven machine learning classification techniques for movie box office success prediction. In *2017 3rd International Conference on Electrical Information and Communication Technology (EICT)* (pp. 1-6). IEEE.
2. Marco, R., Ahmad, S. S. S., & Ahmad, S. (2022). Bayesian hyperparameter optimization and Ensemble Learning for Machine Learning Models on software effort estimation. *International Journal of Advanced Computer Science and Applications*, *13*(3).
3. Grama, L., Tuns, L., & Rusu, C. (2017, June). On the optimization of SVM kernel parameters for improving audio classification accuracy. In *2017 14th International Conference on Engineering of Modern Electric Systems (EMES)* (pp. 224-227). IEEE.
4. Nikdel, H., Forghani, Y., & Mohammad Hosein Moattar, S. (2018). Increasing the speed of fuzzy k‐nearest neighbours algorithm. *Expert Systems*, *35*(3), e12254.
5. Han, S., Qubo, C., & Meng, H. (2012, June). Parameter selection in SVM with RBF kernel function. In *World Automation Congress 2012* (pp. 1-4). IEEE.
6. Huang, J., & Ling, C. X. (2005). Using AUC and accuracy in evaluating learning algorithms. *IEEE Transactions on knowledge and Data Engineering*, *17*(3), 299-310.