Machine Learning Intern
Homework Report

葉乃瑋

Chelsea Yeh

April 2023

Abstract

本次作業主要分成以下 4 個步驟進行：Exploratory Data Analysis (EDA)、Feature Engineering、Modeling 及 Redesign and Optimization。

首先，利用 EDA 對資料集進行初步了解，並選用適合的特徵工程方法進行特徵處理。經過測試與優化後，將資料集以「爬蟲與否」及「看板出現次數」分出四種組合，分別輸入 5 種不同的機器學習模型後，以 Mean Absolute Percentage Error (MAPE) 來進行模型的評估。

根據前述之評估，選擇針對未加入爬蟲資料表現較好的 Support Vector Regression (SVR) 模型，及針對加入爬蟲資料選擇表現較突出的 Random Forest Regression (RFR) 模型得出之四個部分的預測結果結合，得到整體 MAPE 為 35.70%。

最後，再利用 SVR 及 RFR 模型進行後續 Private Test Set 的目標變數預測。最終設計流程如下列 Figure 1 所示。

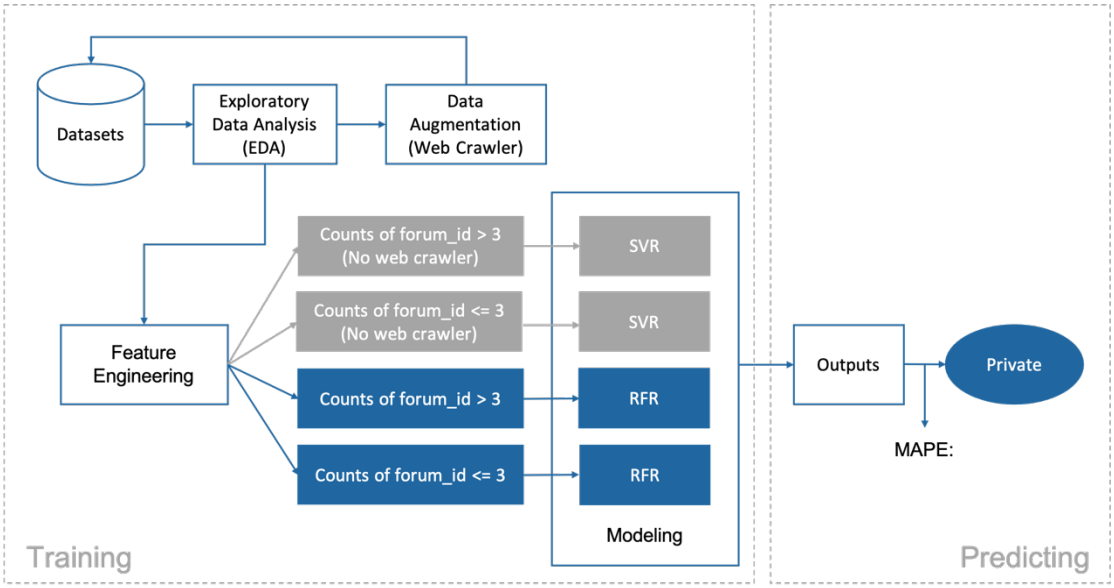

Figure 1

1 Exploratory Data Analysis (EDA)

資料處理及建置模型前，運用視覺化及基本統計工具進行探索式資料分析，幫助對資料集有初步認識，以協助後續進行複雜且嚴謹地分析。

1.1 Basic Information

資料集有 18 個欄位，Train Set、Public Test Set 及 Private Test Set 分別有 50,000、10,000 及 10,000 筆資料，三個資料集中皆無缺失值及重複值。原始資料型態詳見下列 Table 1：

Column name	Data type
title	object
created_at	object
like_count_1~6h	int64
comment_count_1~6h	int64
forum_id	int64
author_id	int64
forum_stats	float64
like_count_24h	int64

Table 1

1.2 Variables details

1.2.1 created_at

擷取發文時間之月份、星期及小時，並將小時分為 Morning (8~14)、Evening (14~20)、Afternoon (20~2) 及 Midnight (2~8) 四區間。

針對月份，運用 histogram 觀察發文時間分佈，不同月份間無太大差異，而是以一星期為週期；針對星期，利用 bar plot 顯示發文時間與星期的關係，得出平日發文數量 > 假日發文數量；針對小時，以 bar plot 觀察四個區間分佈，發現 Evening 為最常發文時段，Midnight 為最少。詳見下列 Figure 2~4。

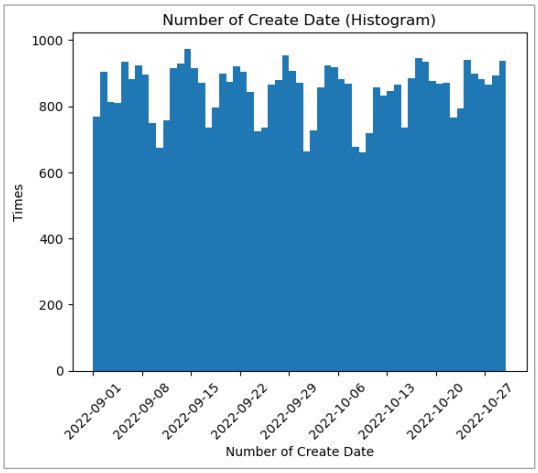

Figure 2

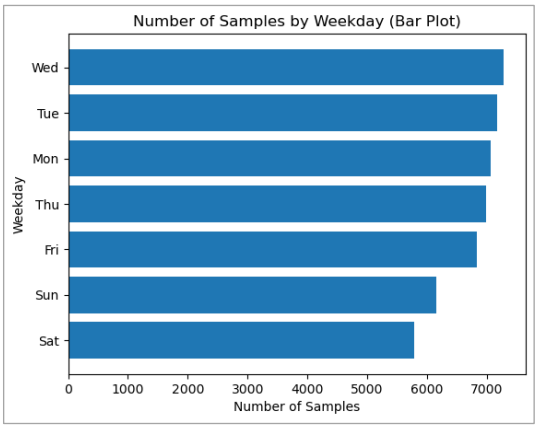

Figure 3

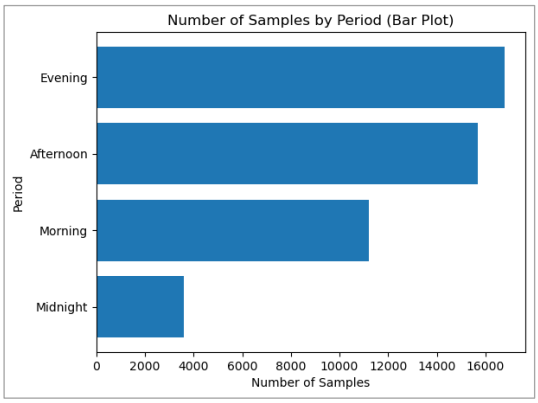

Figure 4

1.2.2 like_count_1~6h

利用 box plot 觀察每小時累積愛心數分佈，發現些許極端值；然而根據對 Dcard 的使用經驗，認為 24 小時的累積愛心數跟過去 6 小時的累積愛心數存在一定線性關係，故不去除極端值，詳見下列 Figure 5。

以 heat map 觀察 1 至 6 小時的累積愛心數與 24 小時的累積愛心數間的相關性，顯示出有強烈正相關，詳見下列 Figure 6。

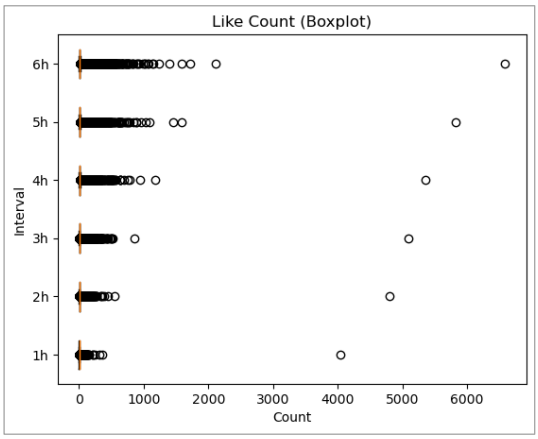

Figure 5

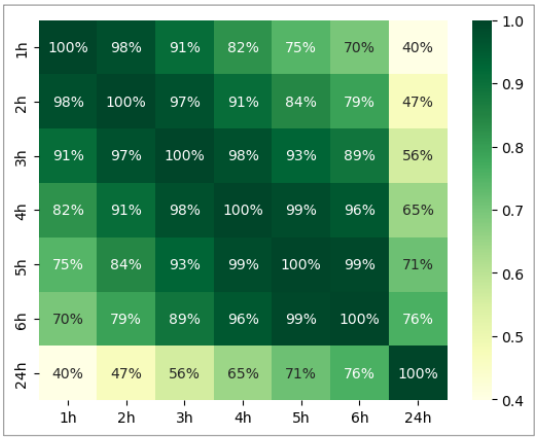

Figure 6

1.2.3 comment_count_1~6h

同上述累積愛心數做法，利用 box plot 及 heat map 觀察累積留言數之分佈及與 24 小時的累積愛心數間之相關性。發現累積留言數極端值較多，而與 24 小時的累積愛心數為弱相關，詳見下列 Figure 7&8。

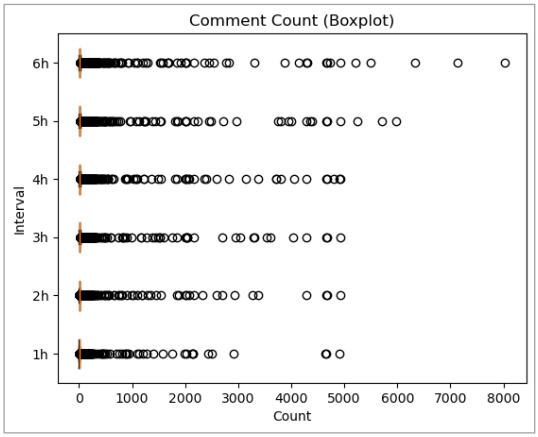

Figure 7

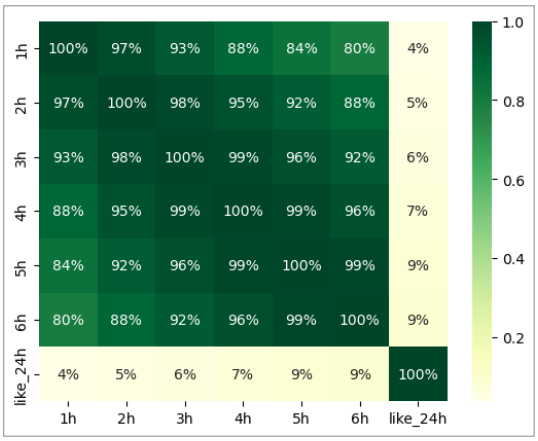

Figure 8

再將 6 小時的累積留言數與 24 小時的累積愛心數的比值 (comment_count_6h / like_count_24h) 排序，發現前 50 名的文章標題大部分為抽獎、廣告文。在此列舉前 10 名於下列 Table 2：

No.	title	cmt/like
1	#頁遊 又双是搶樓🥴（已結束）	1190.000000
2	#頁遊 單人搶樓	936.600000
3	#頁遊 超快速抽羊駝 抽 200 隻 一人一百隻（結束	775.000000
4	#頁遊 深夜(？)搶樓(已結束)	448.000000
5	#頁遊 （結束囉！）今天龍蛋文好多 這篇也是😡	434.583333
6	#頁遊 抽蒂斯（截止）	333.000000
7	#頁遊 小小截樓（截止）	306.071429
8	#頁遊 確診無聊 來搶樓(´▽｀)	288.571429
9	#頁遊 國慶小活動－截樓(已結束)	263.714286
10	（已結束）不講你我他，抽傑尼*1	257.000000

Table 2

1.2.4 forum_id

運用 histogram 及統計數據，發現看板出現次數並不平均，大多數看板只出現過 1 次，少數看板出現次數超過 2,000 次，而出現次數的中位數為 3。根據上述流量差異，推論前者可能為創作者專屬個人看板，後者為官方熱門看板，詳見下列 Figure 9。

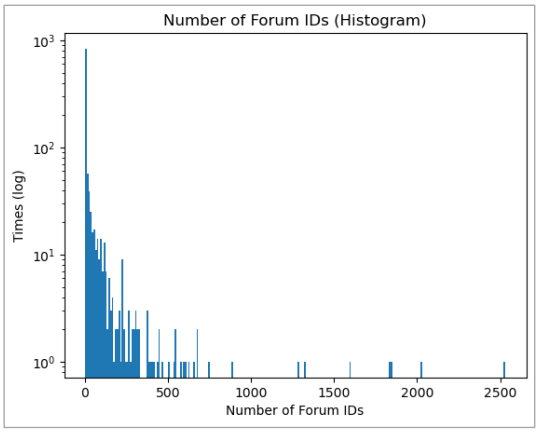

Figure 9

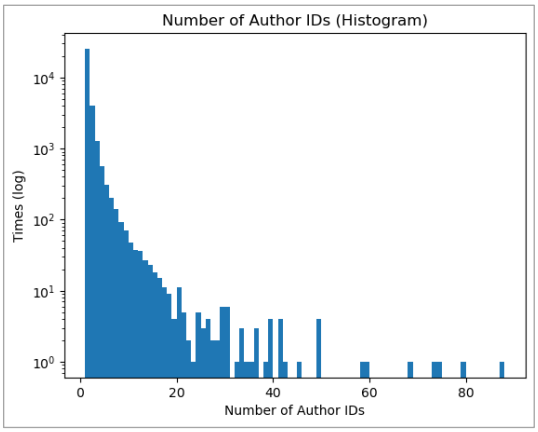

Figure 10

1.2.5 author_id

運用 histogram 及統計數據發現近一半以上作者只發佈 1 篇文章，出現次數

的中位數為 1，詳見上述 Figure 10。

1.2.6 forum_stats

依據自身及對作業 Readme 的理解，推論 forum_stats 為每日看板平均發文數，故與 forum_id 應為一對一之對應關係；但發現 forum_id 為 47568 的看板對應 forum_stats 18.2 及 0.1 兩個不同的值，分別出現 15 及 2 次，在此進行修正。此外，forum_stats 及 like_count_24h 的相關性幾乎為 0。

1.3 Data Augmentation

根據 EDA 後對資料集之初步了解，認為實務上仍有許多因素影響文章之愛心數，如：文章內容、看板名稱等。故將使用網路爬蟲，使用 title 搜尋文章以獲取更多相關資訊，以利精準預測。

整體作業流程如下列 Figure 11 所示：

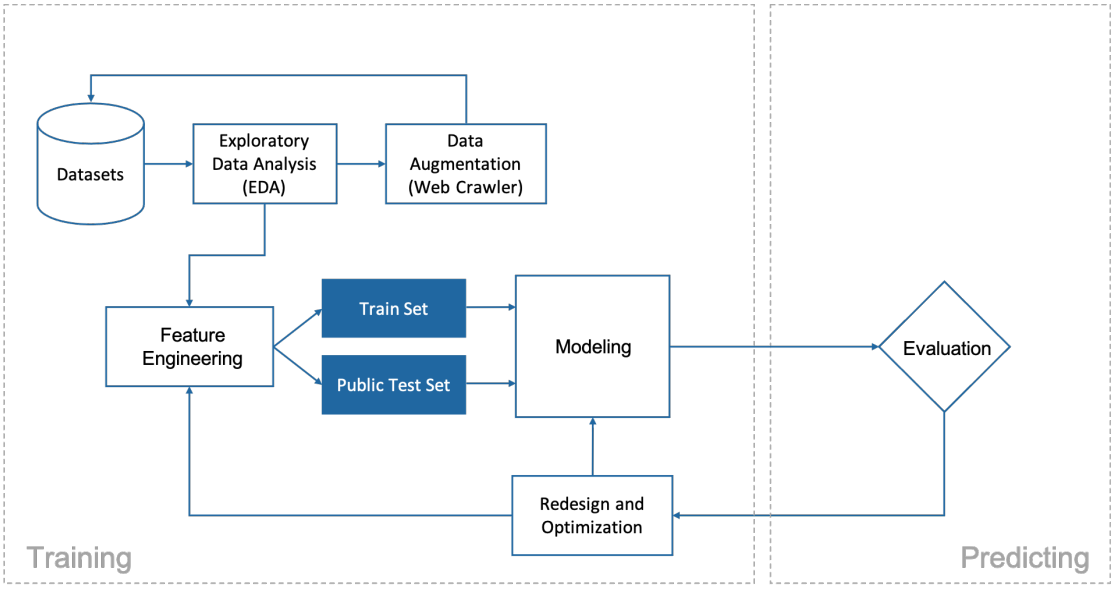

Figure 11

2 Feature Engineering

根據 EDA 後對於資料集的認識與理解，對不同變數進行特徵工程，最大限度地擷取或建立特徵。

2.1 title

由過去使用 Dcard 的經驗中，認為標題內容會影響使用者的點閱率，進而影響愛心數的變化。

故在此利用 NLP 領域被廣泛使用的 transformer 模型中的 tokenizer，並選用適合中文的 MacBERT 預訓練模型，來進行斷詞與轉換文字向量，將標題轉換為可輸入模型的資料型態 (token_1~5)。其中，參數設定輸出最大序列長度 (max_length) 為 7。

2.2 created_at

根據 EDA 結果，認為小時及星期會影響愛心數分佈，故提取小時 (created_hr) 及星期 (created_wkd) 作為新特徵，並轉換為 category variable；並使用 Frequency Encoding 的方式編碼 (created_hr_cnt / created_wkd_cnt)，將類別替換為在 Train Set 中的計數。

2.3 forum_id / author_id

針對 forum_id 及 author_id 類的特徵，因類別間無相對應關係、基數龐大且出現次數不均，由 numeric variable 轉換為 category variable 後，選用適合高基數類別編碼的 K-Fold Target Encoding 方法。

Target Encoding 為將類別變數轉換為預測目標變數的平均值，雖適合高基數特徵，但對極端值敏感；K-Fold Target Encoding 為其優化，利用 cross-validation 求出預測目標變數的統計數據，使編碼結果不過度依賴資料集，故在此選擇後者 (forum_id_mean~kurtosis / author_id_mean~kurtosis)。

最後，對資料集進行標準化及 PCA 降維處理，以進一步優化特徵工程之效果。整體特徵工程流程如下列 Figure 12 所示：

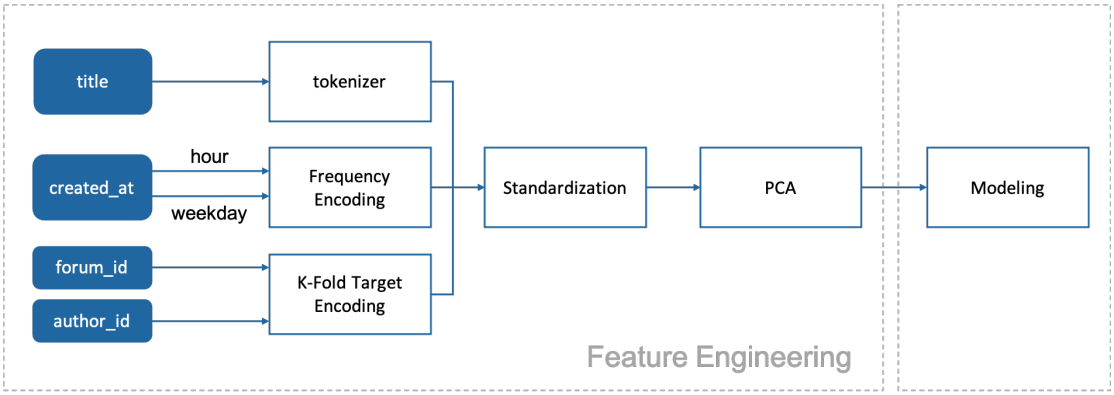

Figure 12

3 Experimental Results and Discussion

3.1 Model

本次目標在於預測 24 小時的累積愛心數數值，屬於回歸問題。故本次皆選用適合處理回歸類型問題的演算法建置模型，詳見下列 Table 3。

Model Name	Ranking of Complexity
Multiple Linear Regression (MLR)	1
K-Nearest Neighbors Regression (KNN)	2
Support Vector Regression (SVR)	3
Random Forest Regression (RFR)	4
eXtreme Gradient Boosting (XGBoost)	5

Table 3

3.2 Experimental Results

使用 Train Set 作為訓練集；Public Test Set 作為測試集。以上述 5 種模型進行訓練，再透過網格搜尋調整為最適合參數後，以本次評量標準 Mean Absolute Percentage Error (MAPE) 判斷模型預測結果，詳見下列 Table 4。

Model	MAPE
MLR	153.65%
KNN	105.29%
SVR	**49.92%**
RFR	107.48%
XGBoost	104.84%

Table 4

3.3 Discussion

由上述 Table 4 結果可知 SVR 模型表現最好；然而，綜合所有模型的 MAPE 分數表現並沒有很好，故藉由重新 EDA 及調整特徵處理方法來優化模型的預測結果。

3.3.1 Reselect Variables

經過不斷測試後發現，若將全部可用的特徵都輸入模型中，易產生太多噪點，導致模型效果不彰。

此外，也重新站在使用者角度反思可能影響文章愛心數的因素，並反覆審視 Train Set 資料集，認為和預測目標 like_count_24h 相關性最高的特徵為 like_count_1~6h，因此在後續的特徵選擇中，皆以可以輔佐 like_count_1~6h 作為考量，以下條列各特徵後續是否選用之原因。

- **title**

  雖然文章標題確實可能影響點閱率，但是轉化為愛心數的轉換率不易計算且可能不高。此外，在嘗試斷詞後發現，雖將文字轉化為向量可輸入模型中；然而，若斷詞輸出的最大序列太多會干擾模型，而太少則不能清楚表達標題。故在後續應用將 title 刪除。

- **created_hr_cnt**

  透過重新觀察發文時間小時分別與 6 小時累積愛心數中位數及 24 小時的累

積愛心數之中位數分佈發現，過去 6 小時的累積愛心數中位數會隨著 Evening > Afternoon > Morning > Midnight 的週期循環；然而，24 小時的累積愛心數中位數在不同小時間的分佈平均。因此認為發文時間的小時可解釋發文後 7~23 小時的愛心數變化，故在此將 created_hr_cnt 特徵留下。分佈詳見下列 Figure 13&14。

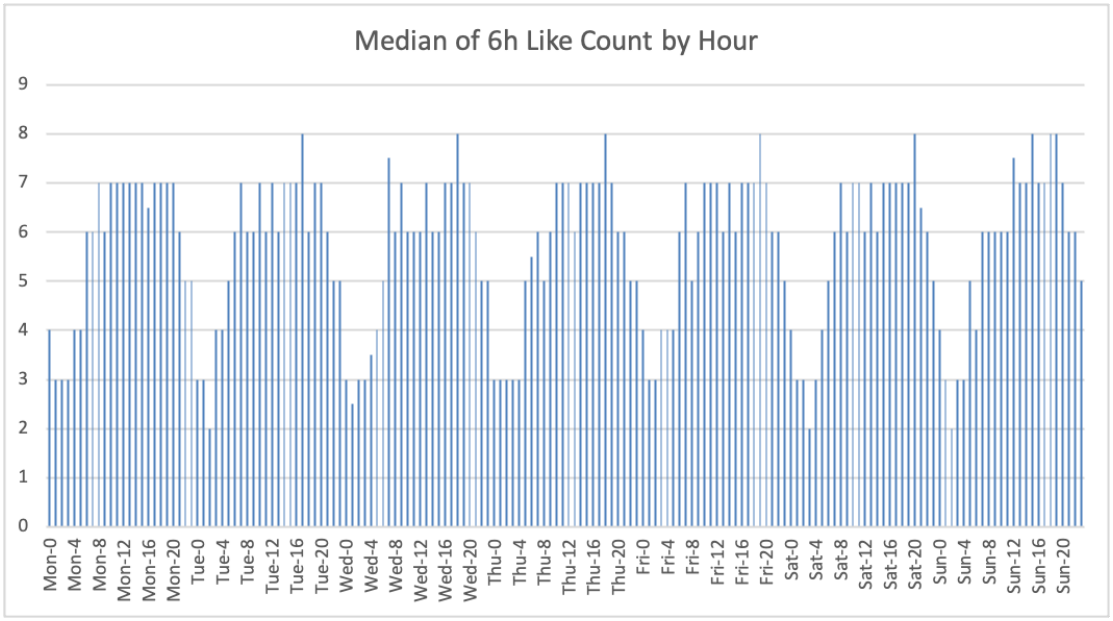

Figure 13

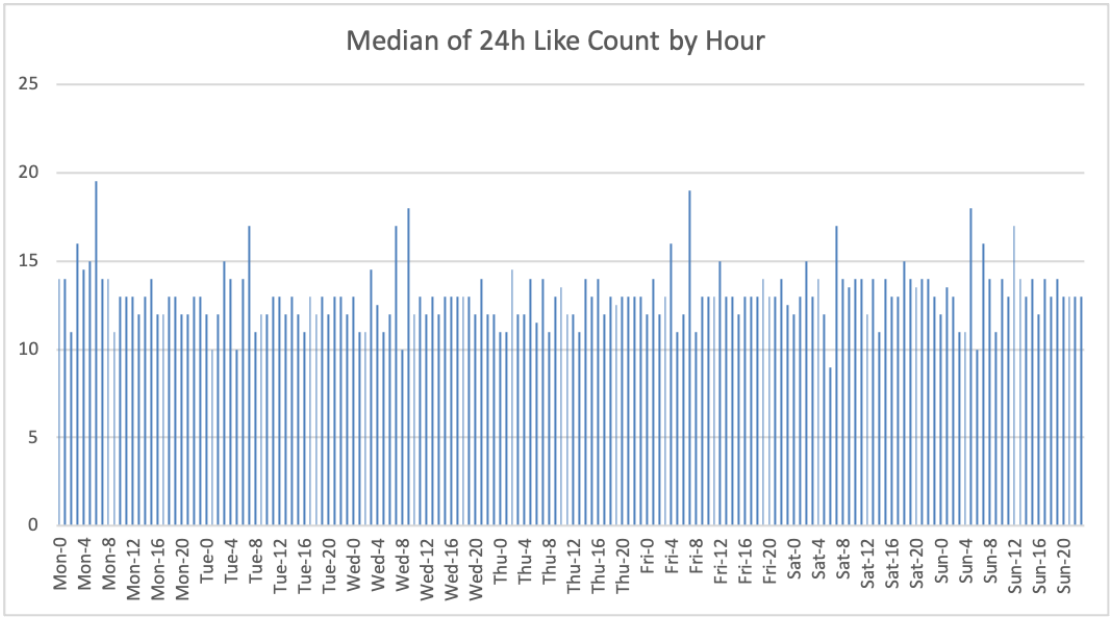

Figure 14

另外，將 created_hr 改變原本的 Frequency Encoding 編碼方式，根據上述分析得知，發文小時對於 24 小時的累積愛心數有一定影響之結果，重新以 Target Encoding 的方式取預測目標之 80 百分位數進行編碼 (created_hr_pct)。

- **created_wkd_cnt**

  透過重新觀察發文時間的星期與 24 小時的累積愛心數之分佈，發現愛心數分佈平均，針對星期無太多差異。故在後續應用將 created_wkd_cnt 刪除。

- **comment_count_1~6h**

  透過一開始的 EDA 得知 comment_count_6h 與 like_count_24h 的比值排序前 50 名的大多數為抽獎、廣告文。在嘗試刪除比值 > 10 的資料後重新計算模型分數，發現 MAPE 分數並無太大變化；雖使 comment_count_1~6h 與 like_count_24h 的相關性提升，但仍屬弱相關範圍。故在後續應用將 comment_count_1~6h 刪除。

- **forum_id**

  forum_id 出現次數之分佈不平均，可能干擾模型之表現，但仍認為看板的出現次數可反映其活躍程度，進而影響文章曝光度，故不直接選用，將在後續做不同應用。

- **author_id**

  藉由 Dcard 的使用經驗發現，發文作者對於愛心數影響較局限於有卡稱的作者發文時會通知追蹤者，進而提高點閱率及愛心數。

  經實測後認為，愛心數成長可直接反映在過去 1 至 6 小時的愛心數上，且多數 author_id 只出現一次，可能干擾模型表現。故在後續應用將 author_id 刪除。

- **forum_stats**

  透過前面的 EDA 推論 forum_id 和 forum_stats 為一對一對應關係，且與

like_count_24h 相關性幾乎為 0。故在看板相關特徵上選擇 forum_id 作為代表，在後續應用將 forum_stats 刪除。

經上述篩選後，留下較能輔助 like_count_1~6h 之特徵，最後共選取 7 個特徵 (like_count_1~6h 及 created_hr_pct)。

3.3.2 Separate Dataset

由於認為看板仍為重要影響因素，在此依照 forum_id 出現次數 > 3 及出現次數 <= 3 兩大類，將 Train Set 及 Public Test Set 進行分類，分別的資料行數如下列 Table 5 所示。

Dataset	Separate by	Rows
Train Set	> 3	49,040
	<= 3	960
Public Set	> 3	9,599
	<= 3	401

Table 5

3.3.3 Web Crawler

利用網路爬蟲最後得到的資訊有：文章 ID (post_id)、距今累積愛心數 (comment_count_now)、距今累積留言數 (like_count_now)、發文作者名稱 (au_name)、文章內容 (content_text)、看板名稱 (forum_name)及圖片數量 (img_count)。

而根據前述所提到選擇特徵的核心概念為輔佐 like_count_1~6h，故在此選擇 like_count_now 加入原始資料集作為新特徵，再計算發文時間距今的天數為新特徵 (created_days)。

另外，由於文章變動會影響爬蟲結果，如刪除文章無法確保 Private Set 都能夠獲取爬蟲資料。故在後續應用再將資料集分為是否可獲取爬蟲資料兩部分輸入模型預測。

3.3.4 Abandon PCA

PCA 原始用意在於降維，較適合用在多特徵資料集上，避免過擬合；但最後選擇的特徵並不超過 10 個，若使用可能造成模型欠擬合，故在後續應用不使用 PCA。

4 Redesign and Optimization

經上述分析，最後的特徵變數共 10 個，資料型態如下列 Table 6：

Column name	Data type
like_count_1~6h	int64
created_hr_pct	float64
now_like_count	float64
created_days	int64
like_count_24h	float64

Table 6

針對無法獲取爬蟲資料之文章，將其依照 forum_id 出現次數分類；而可獲取爬蟲資料之文章，先將其依 forum_id 出現次數進行分類，再加上 2 個由爬蟲獲取的新特徵 (like_count_now / created_days) 。

最後，可將資料集分成以下四種組合，分別輸入模型進行預測，輸入組合詳見下列 Table 7：

No.	Web Crawler	Counts of forum_id
1	No	> 3
2		<= 3
3	Yes	> 3
4		<= 3

Table 7

重新設計之流程如下列 Figure 15 所示：

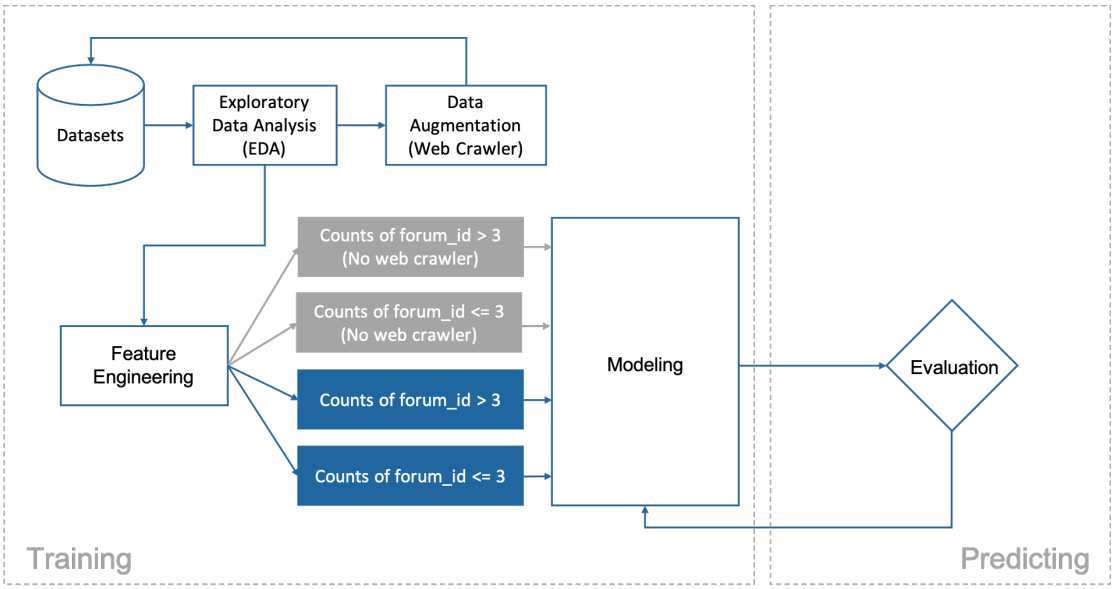

Figure 15

5 Evaluation

由下列 Table 8 可知，重新調整特徵處理方法後，幾乎所有模型分數都大幅下降；針對無法爬蟲獲取資料之文章，SVR 表現最為優秀；而加上爬蟲資料後表現最好的模型為 RFR。因此，最後選擇 SVR 及 RFR 進行後續的預測，詳細流程詳見 Figure 15。

Model	Original	No.1	No.2	No.3	No.4
MLR	172.12%	81.55%	45.23%	57.21%	42.99%
KNN	106.58%	57.37%	36.16%	44.00%	67.06%
SVR	**50.73%**	**35.53%**	**32.89%**	67.98%	70.20%
RFR	107.52%	58.60%	41.01%	**36.21%**	**29.79%**
XGBoost	107.39%	50.57%	45.82%	42.43%	91.47%

Table 8

再將資料集按照上述四個組合，分別以 SVR 與 RFR 模型所預測出來的結果合併，得到的 MAPE 為 35.70%。最後，將 Private Test Set 資料集按照以上步驟四個組合輸出最終結果為 result.csv。

6 Conclusion

根據上述研究結果，針對本次作業主題之預測目標，原始資料集可使用的特徵相對較少，我認為實際與預測目標變數 24 小時的累積愛心數 (like_count_24h) 較相關的特徵變數為過去 6 小時的累積愛心數 (like_count_1~6h)、發文時間小時 (created_hr_pct)；而藉由爬蟲獲取的新特徵距今累積愛心數 (like_count_now) 及發文時間距今天數 (created_days)，也成功協助提升模型表現。

針對模型訓練部分，考量訓練效率、硬體設備效能及資料集適合度，選用 5 個機器學習演算法建置預測模型。透過實測與比較，表現較好的為機器學習中的 SVR 及 RFR 模型。

綜上所述，我認為在資料集的蒐集上若是可以增加更多樣化的特徵或是更多文章相關資訊，能夠幫助預測目標變數更加精確；而未來若能夠有更多訓練時間及提升設備效能，可以嘗試利用深度學習演算法進行預測，相信能夠設計出更好的預測模型。

7 Epilogue

最後，非常感謝 Dcard 團隊給予的機會，在進行作業的過程中使我獲益良多。透過預測 24 小時的累積愛心數，體驗整個產品優化流程，也從中瞭解到將分析與模型建置能力運用於實務可能遇到的問題與困難，更使我增強特徵處理、模型建置及優化能力。

其中，我認為最重要且最具挑戰性的部分在於特徵工程，在進行特徵工程時，不僅需要站在商業行銷的角度思考使用者需求，同時也需要站在機器學習工程的角度反思資料集適合的演算法模型。而經過適合的特徵處理，可以改善並優化模型表現，實現更精準的預測。

我非常享受這次作業的所有過程，期許未來能夠實際參與產品開發過程，協助 Dcard 提供更優質的服務及內容！