Dcard Popular Posts Prediction

Team 6 0816061 王凱俐 0816131 蔡佩君

I. Dcard Popular Posts

1. **Dcard:** Dcard 是一個大型網路論壇，提供台灣民眾分享各式各樣的貼文。
2. **Popular Posts:** Dcard 首頁會呈現當前的熱門文章，並依照熱門程度排序文章，而能成為熱門文章的貼文，為 36 小時內快速取得大家關注的文章。

3. **Objective:** 我們的目的是透過文章的標題、內文、標籤及發文者資訊等，預測該篇文章發布後是否能成為熱門文章，也希望能藉此了解流行趨勢跟時下熱門話題，找出熱門文章通常具備什麼要素。

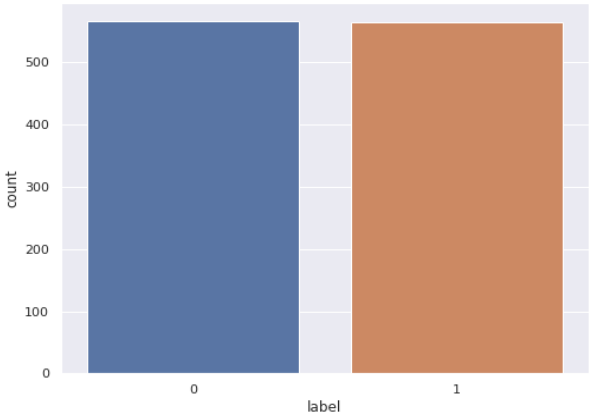
II. Data Collection

1. **Web Crawler:** 我們使用爬蟲抓取 Dcard 上的文章資料。
    a. Popular: 首先從 Dcard 首頁，按照熱門程度排序，抓取前 50~60 篇熱門文章 id，再透過 id 抓取文章內文及詳細資訊。**label 標記為 1**。
    b. Not popular: 首先從 Dcard 首頁依據發布時間排序，抓取已發布超過 36 小時，且愛心數仍小於 50 的文章 id，則判斷此文章為非熱門，再透過 id 抓取文章內文及詳細資訊。**label 標記為 0**。

c. Features: 下圖為資料範例，我們採用的 features 為：標題、內文、是否匿名校名、使否匿名系名、性別、看板名稱及標記。

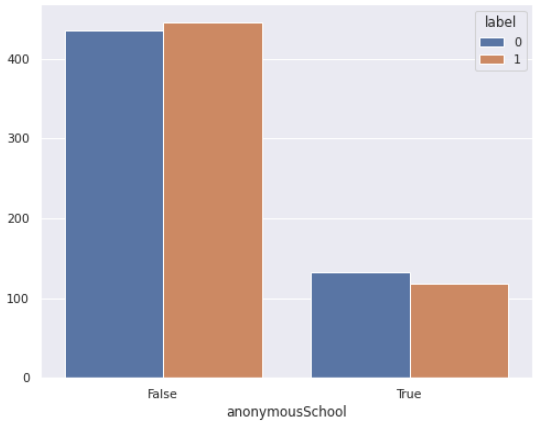
title	content	anonymousSchool	anonymousDepartment	gender	forumName	topics	label
本週連續3位網紅與政府槓上	因為連續看到3位網紅與政府槓上\r\n所以發這篇文\r\n以下按照youtube上傳影片時間...	False	False	M	YouTuber	['博恩', 'cheap', 'bump', '機車', '政府']	1
#贈品 今天抽一位贈送星巴克	今天租到房子雀開心的\n心情好就想送東西??\nhttps://i.imgur.com/qv...	False	False	M	省錢	['贈送', '星巴克', '省錢', '情報', '抽獎']	1
Dr.Wu包裝先別丟??~	目前沒有回收瓶活動喔！！（下文語意修正）\n是包裝留射標籤累積活動??\n\n（購買商品的包...	False	False	F	美妝	['保養', '分享']	1
#詢問多那之工作	想請問推薦去多那之打工嗎～\n爬文過幾乎都是幾年前的留言\n身邊朋友也都沒有相關經歷可以參考...	True	True	F	工作	['工作', '工作經驗', '打工', '面試', '求職']	0
突然覺得好迷茫	前女友是在交友軟體認識的，聊天過程雀開心的，一段時間我們也約出來聊聊天，我對她也雀有好感的，...	True	True	M	感情	['迷茫', '感情', '分手', '失戀', '出軌']	0
#國小 請問桃園英語教甄 數學複試	各位前輩們好???♀?\n請問在進行數學複試時，能否打開課本上課，\n（看著課本裡的數學題目...	False	True	F	教師	['桃園', '英語', '教甄']	0

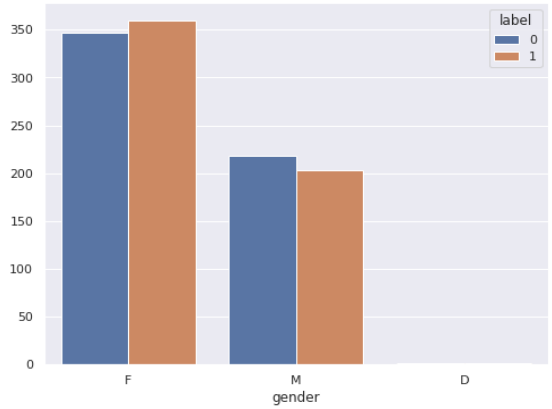
2. **Amount of data:** 為了預測的準確性，熱門與非熱門兩種類別的資料數量，各為百分之 50，各約五百多筆資料。

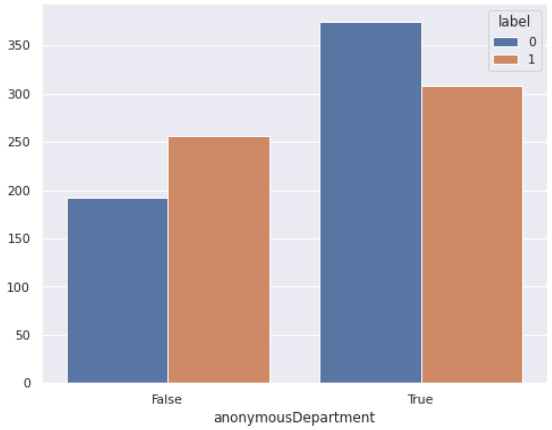


III. Data Analysis

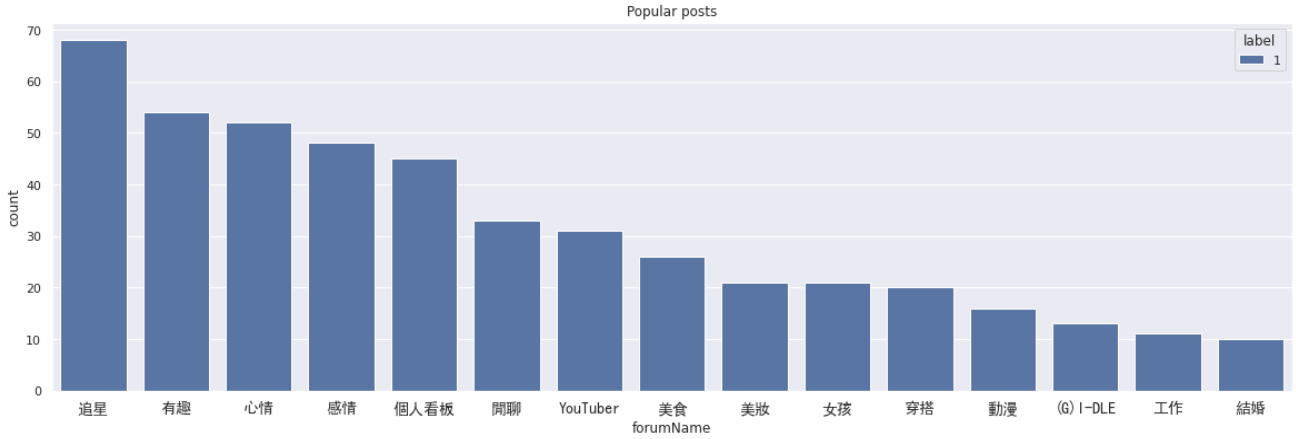
1. **Feature Distribution:** 觀察其中三種 feature：發文者性別、是否匿名校名，及是否匿名系名，在兩個 label 種類中的數量分布。由圖表顯示，這些 feature 在兩個種類中的分布差異都不大。另外也能發現，女性及顯示校名的整體比例都較高。



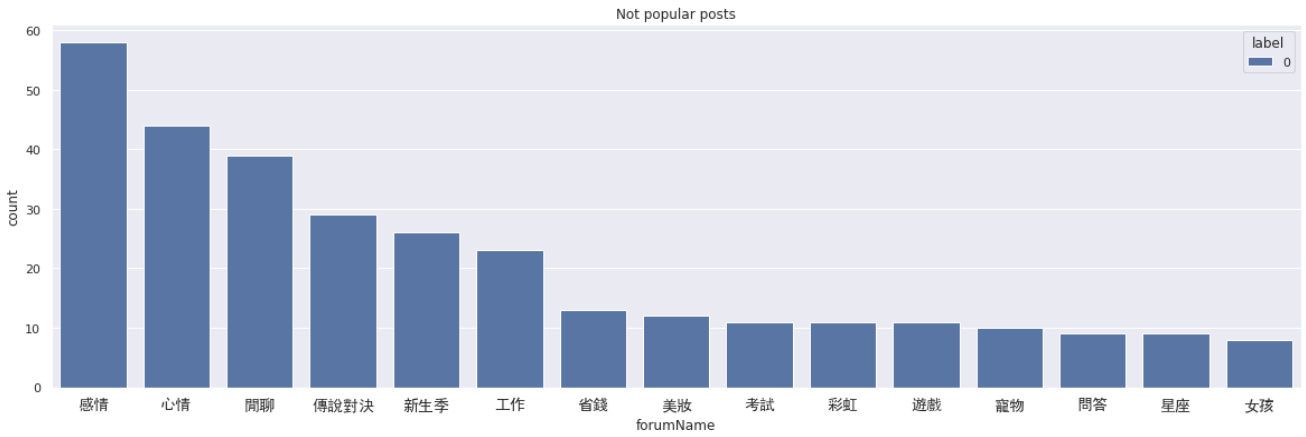




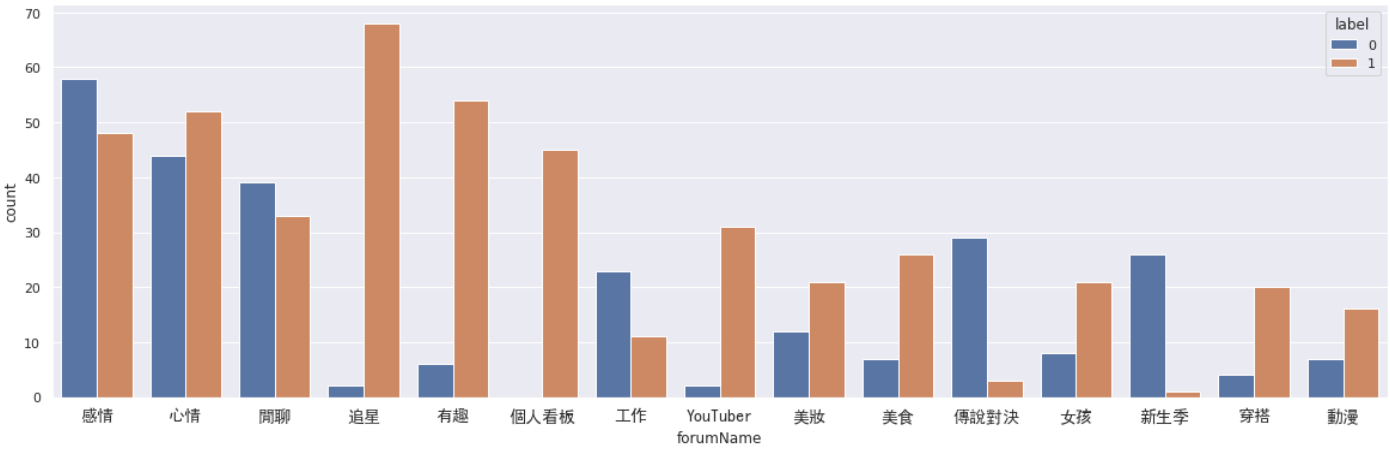
2. **Popular Forum:** 我們抓出在熱門文章中，數量最多的 15 個看板，數量最多的為追星、有趣、心情、感情、個人看板等。



3. **Not Popular Forum:** 非熱門文章部分，分布最多的前幾名看板為感情、心情、閒聊、傳說對決、新生季、工作板等。



4. **Popular and Not Popular Forum:** 綜合兩者觀察，首先在感情、心情、閒聊三個看板中，熱門與非熱門文章都有很多的數量，兩者比例差距不大。而在追星、有趣、youtuber、穿搭版中，可以發現熱門文章幾乎占了大量的比例，由此可以推測在這些看板中的文章，有很大機率可以登上熱門。在傳說對決、新生季兩個看板中，非熱文文章都佔了很大比例，也能由此推測在這兩個看板中的文章，較難以受到大眾關注。



5. Word Cloud

我們對資料中的詞語製作了文字雲，使更清楚觀察不同字詞的影響程度，文字雲中字體越大的詞語，代表出現的頻率越高。

a. **Popular Title:** 在熱門文章標題部分可以發現 IG、男友、穿搭、梗、新聞等占了很大部分，而由於我們蒐集的資料是近期幾周的資料，因此也能發現一些近期的熱門話題，例如：王心凌、確診等。

b. **Not Popular Title:** 非熱門文章標題部分，顯示請益、詢問、請問等，多發表問題的字詞，也推測這些詢問資訊的文章，能得到的關注與回應都較少。

c. **Popular Topics:** 熱門的標記中，穿搭、有趣、梗圖、迷因、搞笑等占了很大部分，能發現這些能帶來生活樂趣的主題，有很大機會能成為熱門文章。

d. **Not Popular Topics:** 非熱門標記則包括工作、感情、傳說、手遊等。

IV. Data Preprocessing

1. **Drop nan:** 移除有缺失的資料。
2. **Lable Encoding:** 由於 dataset 中每一筆資料除了包含文字資料外，還有 categorical feature，譬如: anonymousSchool、anonymousDepartment、gender，所以此處是針對 categorical feature 進行 label encoding，將 True/False 以及 female/male 轉換成 0/1。
3. **Remove Other Symbols:** 由於文章內容會出現數字、中文、字母以外的內容譬如: 表情符號、標點符號、網址等，這些內容會增加模型訓練的困難度，因此我們會將這些內容去除。
4. **Remove Stop Words:** stop words 泛指一些在文章中出現頻率很高但卻沒什麼意義的文字，譬如: 除了、那麼、隨著等。因此我們也會移除這些不具參考價值的字詞，來減少 dataset 的複雜度。
5. **TF-IDF:** 是一種常用於文字處理的統計方法，用來評估詞彙在這篇文章中的重要程度。首先會計算各個詞彙出現的頻率，而頻率越低的詞彙會給予較高的權重，使模型更關注在重要的詞彙上。
6. **Label Binarizer:** 將 categorical feature 經過 label binarizer 轉換成 one hot encoding 的形式。

V. Model

1. **Naive Bayes:** 使用貝氏定理，在已知的條件下，計算各個類別發生的機率，而分類器會輸出發生機率最高的類別。
2. **Logistic Regression:** 適用於二元分類的分類器，目的是要找出一條平滑的線，將兩個類別分開。
3. **K Nearest Neighbor:** 採多數決標準，計算目前該筆資料與其他資料的距離，接著找出 k 個最接近的鄰居來判定該筆資料位於哪一群。在此項作業中，我們設置了不同大小的 k，並進行實驗找出最佳的 k，實驗結果撰寫於後方實驗部分，而最後決定將 k 設置為 10。
4. **MLP:** 我們建立 Multilayer Perceptron 神經網路，作為二元分類器。MLP 為具有多個節點的 layer 所組成，layer 中所有節點都與下一層所有節點相連，形

成全連接層。其中分為輸入層、隱藏層及輸出層，我們實驗了不同層數模型的表現，撰寫於後方實驗部分。

我們使用 Tensor Flow Dense function 來建立全連接層，使用 relu 作為 activation function。由於我們要處理的是二元分類問題，因此輸出層由一個節點組成，並以 sigmoid 作為 activation function，使模型輸出 0~1 之間的預測值，數值越接近 1，代表越有可能是熱門文章。

VI. Experiments

1. All Model

在此作業中，我們一共使用 4 種模型來進行實驗，以下數據皆為 testing data 上的結果，並使用 Accuracy、Precision、F1 score、ROC AUC score、ROC curve、Confusion Matrix 作為 metric。

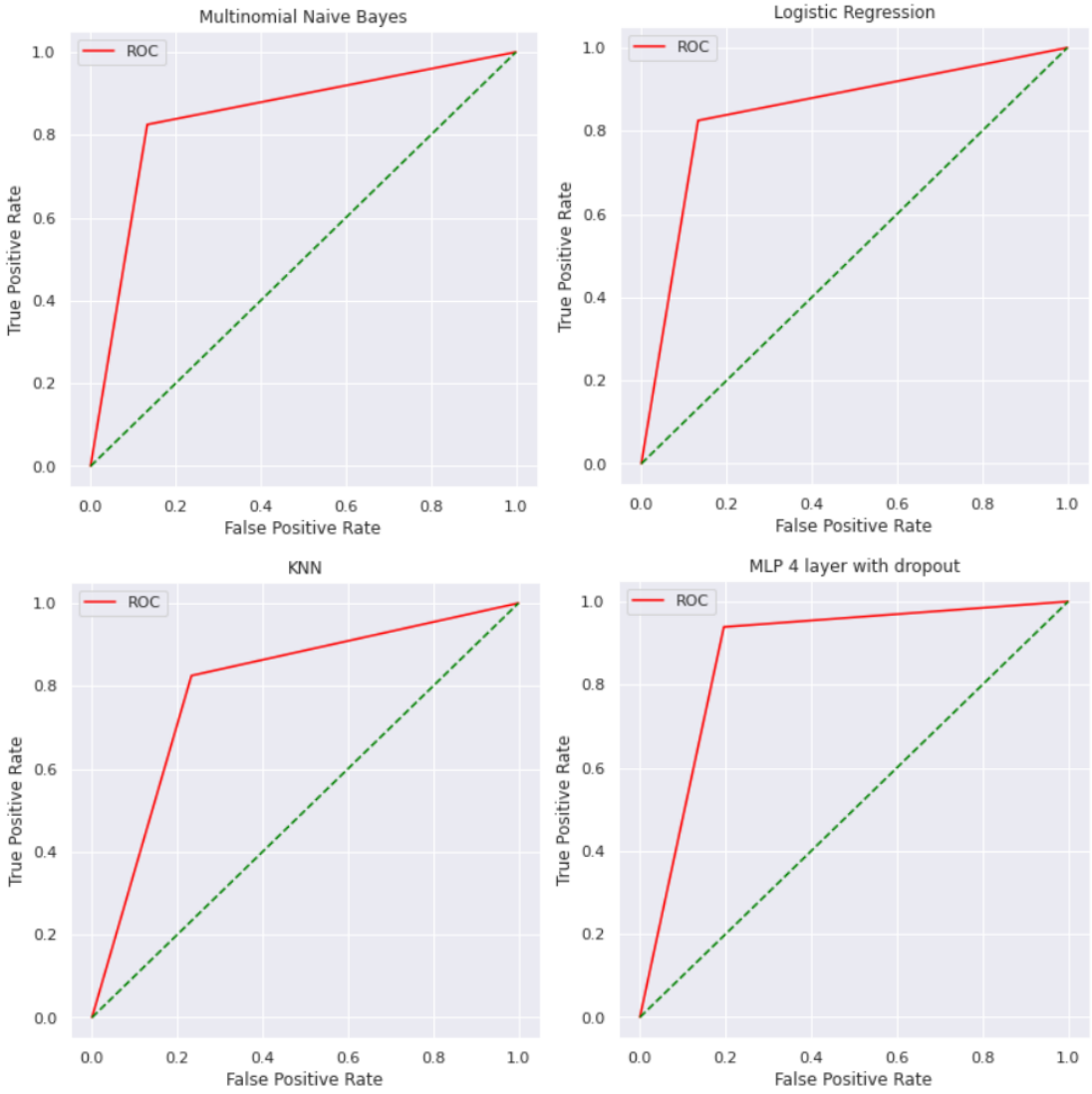
a. Accuracy: 由下表可知，4 種模型的 Accuracy 都具有一定的水準，皆在 0.79 以上，其中表現最好的是 MLP。

Method	Accuracy
Naive Bayes	0.85
Logistic Regression	0.85
KNN(k=10)	0.79
MLP 4 layer with dropout	0.87

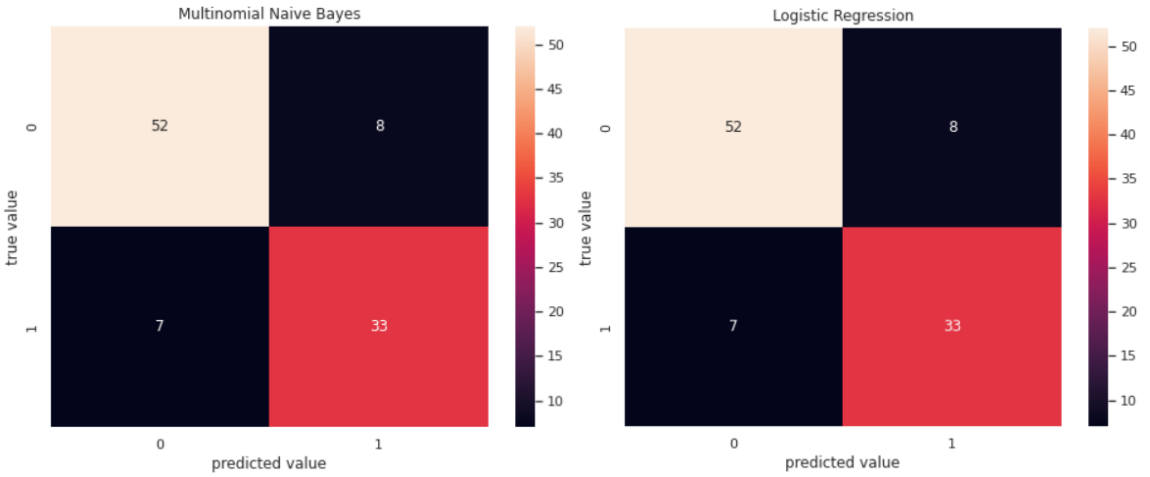
b. Precision、F1 score、ROC AUC score

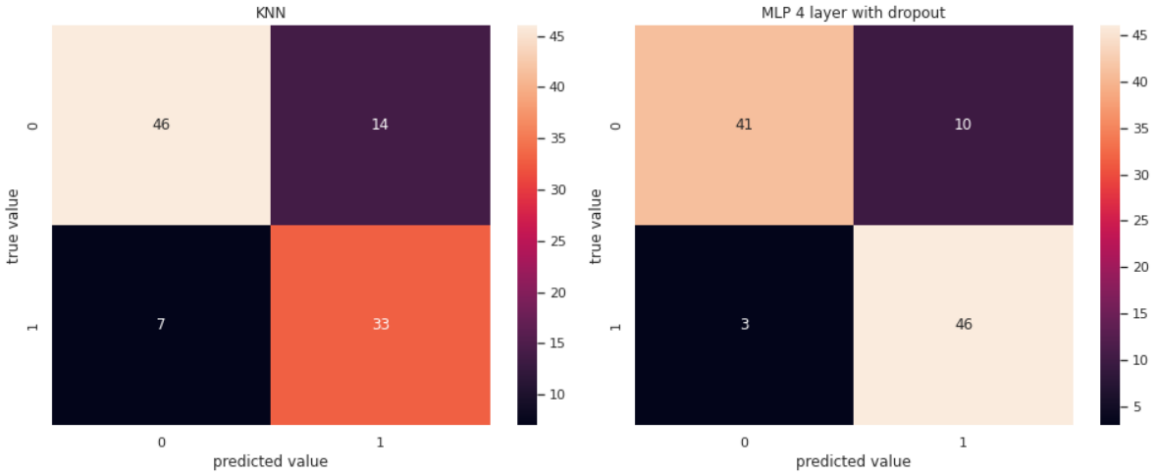
Method	Precision	F1 score	ROC AUC score
Naive Bayes	0.8431	0.8444	0.8458
Logistic Regression	0.8431	0.8444	0.8458
KNN(k=10)	0.785	0.7864	0.7958
MLP 4 layer with dropout	0.8101	0.8696	0.8713

c. ROC curve: 當 ROC curve 以下的面積越大，代表模型的準確率越高，從底下的圖可知，紅線底下的面積很大，代表我們模型的分類精準度都很高。



d. Confusion Matrix





2. KNN

我們實驗了三種不同大小的 k(判斷資料屬於哪個類別，所需的鄰居個數)，並將模型的訓練資料分為兩類: "使用文字資料及 categorical feature"及"單純使用文字資料"。其中表現最好的模型，皆為 k=10 的模型(黃框處)。

k	without categorical feature	with categorical feature
10	0.76	0.79
20	0.73	0.76
30	0.71	0.71

3. MLP

我們實驗了四種不同的 MLP 架構，並監控它們的訓練情形，與在測試資料及上的表現，四種模型分別為：

	input layer	hidden layer 1	hidden layer 2	output layer	dropout
MLP 4 layer without dropout	128	64	32	1	False
MLP 4 layer with dropout	128	64	32	1	True
MLP 3 layer	64	32		1	True
MLP 2 layer	64			1	True

a. Training:

以訓練資料中的 30%作為 validation data，使用 binary crossentropy loss 監督，各訓練 15 個 epochs。

b. Result-accuracy:

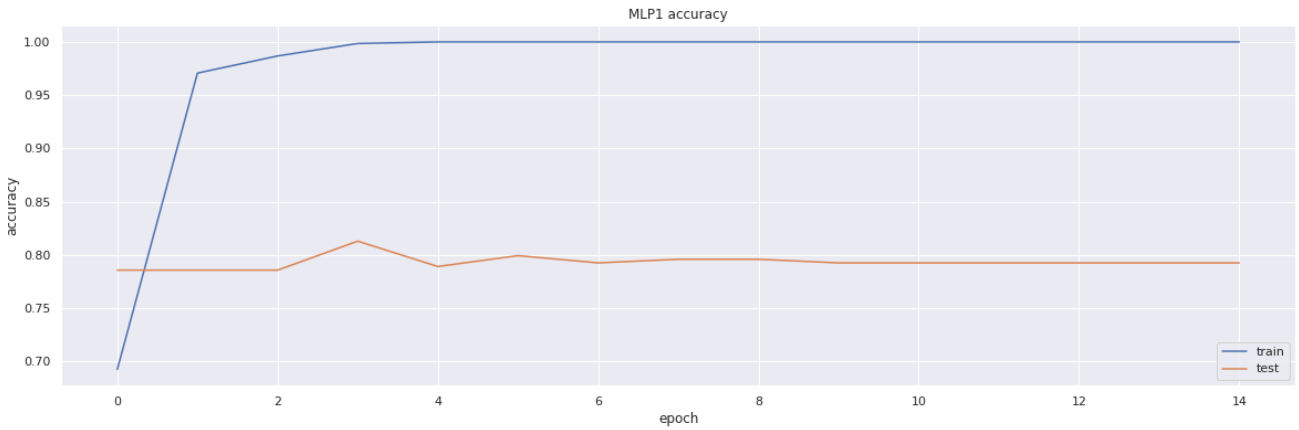
Method	Accuracy
MLP 4 layer without dropout	0.81
MLP 4 layer with dropout	0.87
MLP 3 layer	0.82
MLP 2 layer	0.82

上表為在 test data 上預測結果的 accuracy，其中 MLP 4 layer with dropout 得到了最佳的結果。

c. Result-dropout:

在第一個 MLP 模型中，沒有加入 dropout，而在其餘的模型中，我們在全連接層之間加入了 dropout layer，每個節點有 0.5 的機率會被關閉，防止 overfitting。

下圖為 MLP 4 layer without dropout 在訓練時的準確率變化。可以觀察到，訓練時在 training set 的準確率很快就達到 1.0，但在 validation 及 test 資料上，結果反而較差。可由此推測，在沒有加入 dropout layer 的情況下，容易造成在訓練資料集上過擬合的現象。

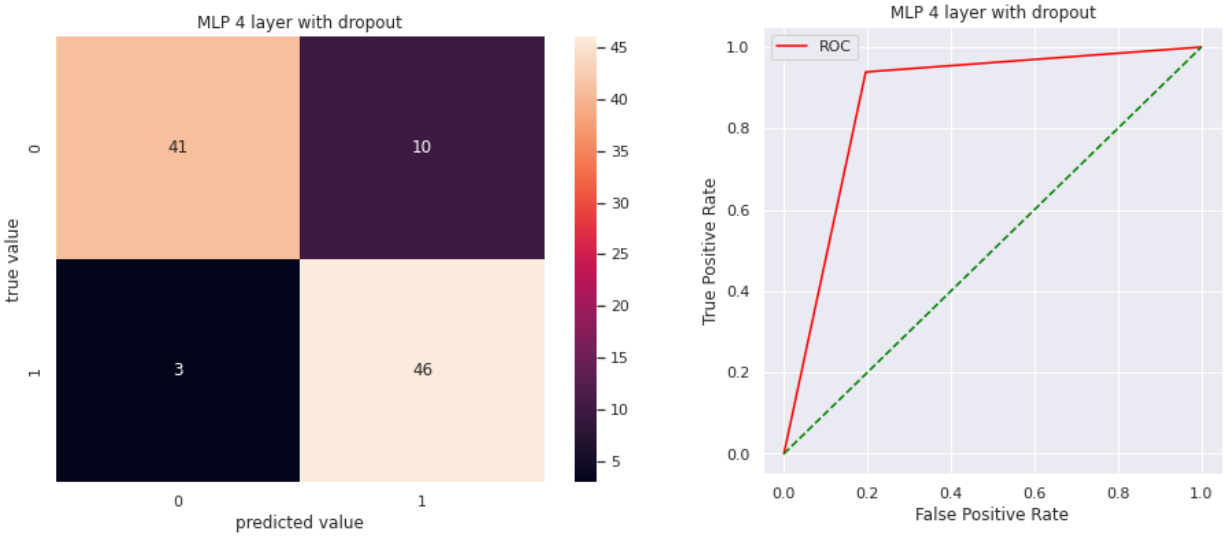


d. Result-Precision, F1 score, ROC AUC score:

Method	Precision	F1 score	ROC AUC score
MLP 4 layer without dropout	0.8100	0.8098	0.8097
MLP 4 layer with dropout	0.8766	0.8697	0.8713
MLP 3 layer	0.8215	0.8197	0.8195
MLP 2 layer	0.8199	0.8199	0.8199

e. Result-Confusion Matrix, ROC Curve:

MLP 4 layer with dropout



3. Categorical Feature:

在資料的 feature 中，我們實驗了不採用 categorical feature，只使用文字資料(標題、內文、標籤)，以及採用 categorical feature(性別、匿名校名、系名、看板名稱)兩種方式，下表為兩種做法在各個模型中的預測準確率。

由結果觀察，加入 categorical feature 後，在各個模型上的表現都能有所提升，顯示 categorical feature 也能提供一定程度的資訊，協助模型分類。

Method	without categorical feature	with categorical feature
Multinomial NB	0.71	0.85
Logistic Regression	0.76	0.85
KNN(k=10)	0.76	0.79
MLP 4 layer without dropout	0.76	0.81
MLP 4 layer with dropout	0.86	0.87
MLP 3 layer	0.81	0.82
MLP 2 layer	0.79	0.82

4. 不同 categorical feature 及文字資料組合:

在使用不同 feature 的實驗上，除了做上述的實驗外(是否加入 categorical feature 作為訓練資料)外，我們還實驗了"使用部分 categorical feature 及文字資料"的實驗，每筆資料的 feature 最多有 anonymousSchool、anonymousDepartment、gender、forumName、title、topics，我們實驗了搭配不同組合的 feature，形成下表的結果。

Anonymous School	Anonymous Department	gender	forumName	title	topics	Accuracy
V	V	V	V	V	V	0.89
V	V	V	X	V	V	0.79
V	V	X	X	V	V	0.81
V	V	X	V	V	V	0.85
V	X	X	V	V	V	0.87
X	X	X	V	V	V	0.88

上表中，表現最好的為，使用所有 feature(黃框處)作為訓練資料的設定。另外我們觀察到，當不加入 forumName 這項 feature 時，會造成 Accuracy 大幅降低，產生兩個表現最差的結果(綠框處)。而當不加入 anonymousSchool 或 anonymousDepartment 或 gender 時，皆只導致 Accuracy 微幅下降。由此可判斷，forumName 對模型來說，為非常重要的判斷依據。

VII. Demo

我們實際去 Dcard 中尋找剛發布的文章，輸入到模型中預測， 並在一天後驗收成果。

1. 熱門文章

首先在文章剛發布時抓取資料，進行預測。

模型預測的結果都在 0.9 以上 代表模型認為他有高機率會成為熱門文章。

```
[[0.91030985]]
```

	content	anonymousSchool	anonymousDepartment	gender	forumName
0	我的鯛魚燒回來了 之前在台中的時候 阿妃對這隻鯛魚燒娃娃情有獨鍾 後來因為搬家不小心把它的娃...	1	1	1	37

```
prediction:  [0.9982193]
```

	title	content	anonymousSchool	anonymousDepartment
1	#閒聊 簡單介紹theqoo專版跟熱門機制	看到Jessica那篇有人在說操作，想說看theqoo蠻久熱門其實不太容易操作，藉這次機會介...	False	True

一天之後，這兩篇文章也的確出現在熱門文章當中。

2. 非熱門文章

兩篇文章中，模型預測的結果都趨近於 0，代表模型認為他成為熱門文章的機率很低。

```
prediction:  [0.00560942]
```

	title	content	anonymousSchool	anonymousDepartment
4	工程業ERP系統	想問大家的公司都有在用ERP系統嗎?\n或是覺得公司人數規模大概多少再用ERP就好?\n我們...	False	True

```
prediction:  [0.00063568]
```

	title	content	anonymousSchool	anonymousDepartment
8	弘光轉學考 書審	今年弘光轉學考的簡章出來了\n目前是想以五專畢業的身分去報考轉學考 科系是語聽\n有看過書審...	False	True

一天之後驗收成果，這兩篇文章的按讚留言數少於 30，非熱門文章。

VIII. Conclusion

1. 根據資料分析結果，我們歸納出上圖中，熱門與非熱門文章具備的要素，若想撰寫出熱門文章時，可依此為參考依據。
2. 我們實驗了多種模型進行分類預測，皆能達到不錯的結果，其中最佳的為 4 層 MLP 模型。
3. 除了文字資料，我們也採用了其他 categorical 資料，並實驗以不同組合的資料訓練模型，當使用所有資料時，能達到最好的結果。
4. 為了測驗模型的實用性，我們也實際抓取剛發布的文章，交由模型預測成為熱門文章的機率，並在數小時後驗證結果，能達到一定的準確率。若是想發布熱門文章，可以在發布前先經由我們的模型預測登上熱門的機率，並依此修改文章，直到模型預測出高的數值，就能有高機率成為熱門文章。
5. 在此 project 中，我們只蒐集到近幾周的文章資料，若想讓模型學習到更完整的流行趨勢，可以再收集更長期的資料，增加模型預測能力。

Github Link

https://github.com/kelly8911/Dcard-Popular-Posts-Prediction