Machine Learning

基於藥物評價的情感分析模型

機器學習專案報告 | 2022年5月24日

目錄 CONTENTS

01

問題解析
&
數據預處理

02

規劃分析
&
模型初探

03

主要問題
&
模型改進

04

總結
&
展望

CONTENTS 目錄

01

問題解析
&
數據預處理

02

規劃分析
&
模型初探

03

主要問題
&
模型改進

04

總結
&
展望

01 研究問題

基於藥物評價的情感分析模型

藥物治療在疾病的治療中起著非常重要的作用和作用。患者對藥物的評價和滿意度也會影響治療進程和醫生的用藥方案。因此，本項目將使用與患者對特定藥物的評論和反饋相關的數據，並將應用機器學習模型來嘗試評估藥物。

研究思路

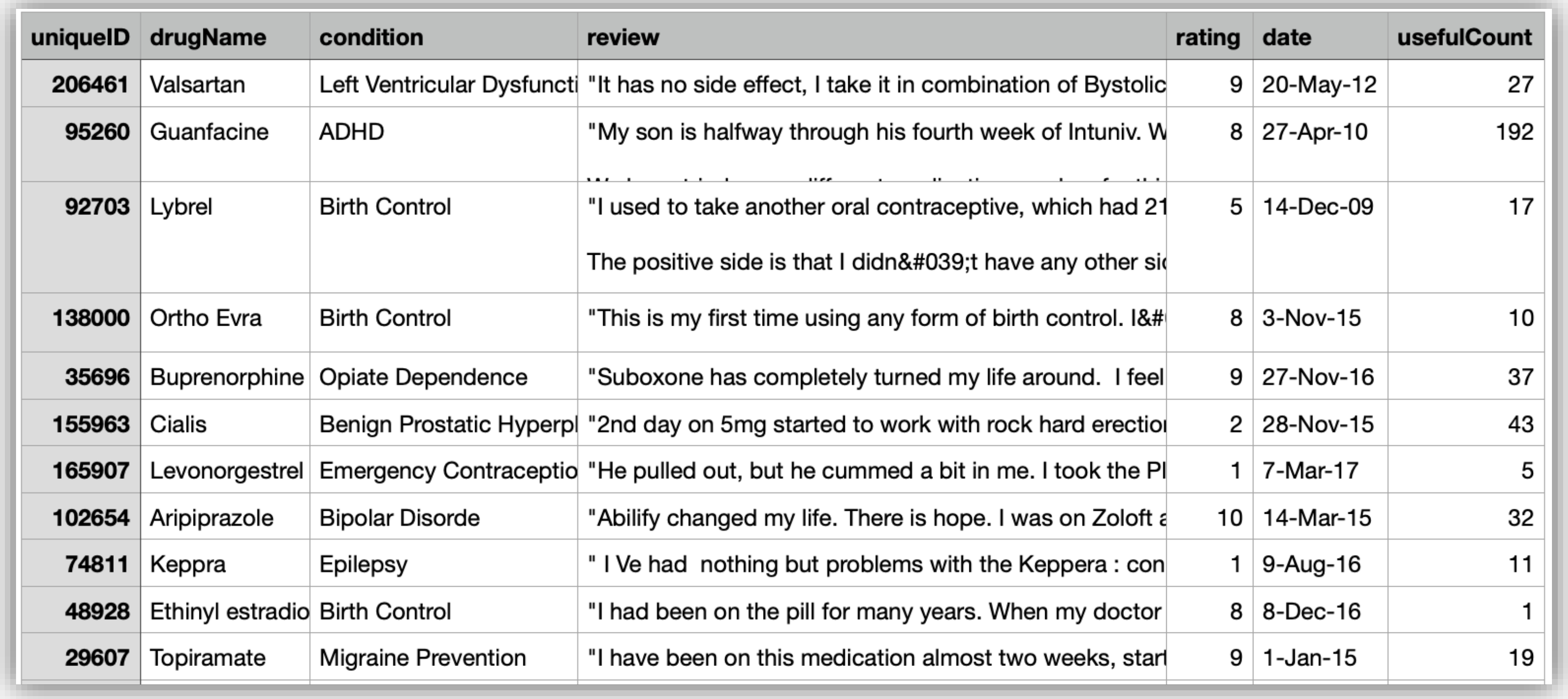
將對藥物評論向量化，通過評價與其對應的情感傾向進行模型訓練，進而得到對藥物的總體評級。

數據來源

Felix Gräßer et al. Aspect-Based Sentiment Analysis of Drug Reviews Applying Cross-Domain and Cross-Data Learning. In Proceedings of the 2018 International Conference on Digital Health (DH ' 18). ACM, New York, NY, USA, 121-125.

01 數據集

本次研究的數據集資訊如下：



uniqueID	drugName	condition	review	rating	date	usefulCount
206461	Valsartan	Left Ventricular Dysfuncti	"It has no side effect, I take it in combination of Bystolic	9	20-May-12	27
95260	Guanfacine	ADHD	"My son is halfway through his fourth week of Intuniv. W	8	27-Apr-10	192
92703	Lybrel	Birth Control	"I used to take another oral contraceptive, which had 21 The positive side is that I didn&#039;t have any other sic	5	14-Dec-09	17
138000	Ortho Evra	Birth Control	"This is my first time using any form of birth control. I&#	8	3-Nov-15	10
35696	Buprenorphine	Opiate Dependence	"Suboxone has completely turned my life around. I feel	9	27-Nov-16	37
155963	Cialis	Benign Prostatic Hyperpl	"2nd day on 5mg started to work with rock hard erectio	2	28-Nov-15	43
165907	Levonorgestrel	Emergency Contraceptio	"He pulled out, but he cummed a bit in me. I took the Pl	1	7-Mar-17	5
102654	Aripiprazole	Bipolar Disorde	"Abilify changed my life. There is hope. I was on Zoloft a	10	14-Mar-15	32
74811	Keppra	Epilepsy	" I Ve had nothing but problems with the Keppera : con	1	9-Aug-16	11
48928	Ethinyl estradio	Birth Control	"I had been on the pill for many years. When my doctor	8	8-Dec-16	1
29607	Topiramate	Migraine Prevention	"I have been on this medication almost two weeks, start	9	1-Jan-15	19

數據集大小

- 訓練集：161,000
- 測試集：53,800

內容資訊

- 藥品名稱(categorial)
- 對應病症(categorial)
- 患者評價(text)
- 患者評分(numerical)
- 評價評價日期(date)
- “按讚數”：認為該評價有用的用戶數量

工作計畫

- 數據預處理：刻畫數據結構，討論模型方法
- 初步模型訓練與算法實現
- 進階模型探索：多分類模型

01 數據刻畫與預處理

01 數據量控制

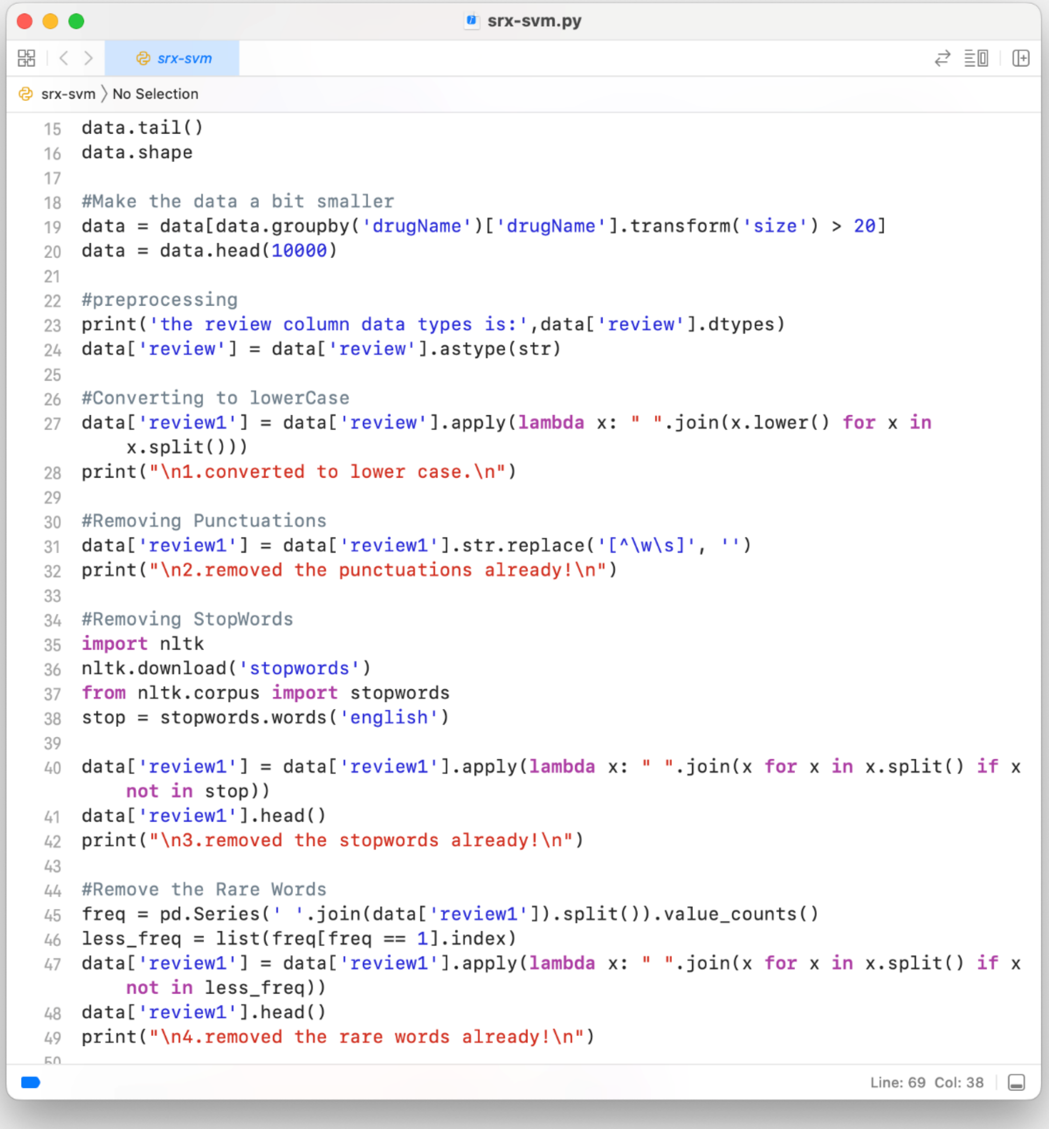
- ◆ 刪除缺失值，並隨機選取了10000個數據作為研究對象
- ◆ 刪除無關列（uniqueID, condition,date,usefulCount列）
- ◆ 刪除少於20個評價的藥物，保證評價具有代表性

02 文本訊息的處理

- ◆ 統一格式：刪除標點符號、大寫字母變為小寫字母
- ◆ 刪除Stop Words（如“the”，“a”，“in"等詞語）
- ◆ 刪除出現頻率過少的詞
- ◆ 提取詞句：

  SnowballStemmer：刪除相似單詞

  PorterStemmer：刪除單詞中常見的形態詞尾和固定詞尾

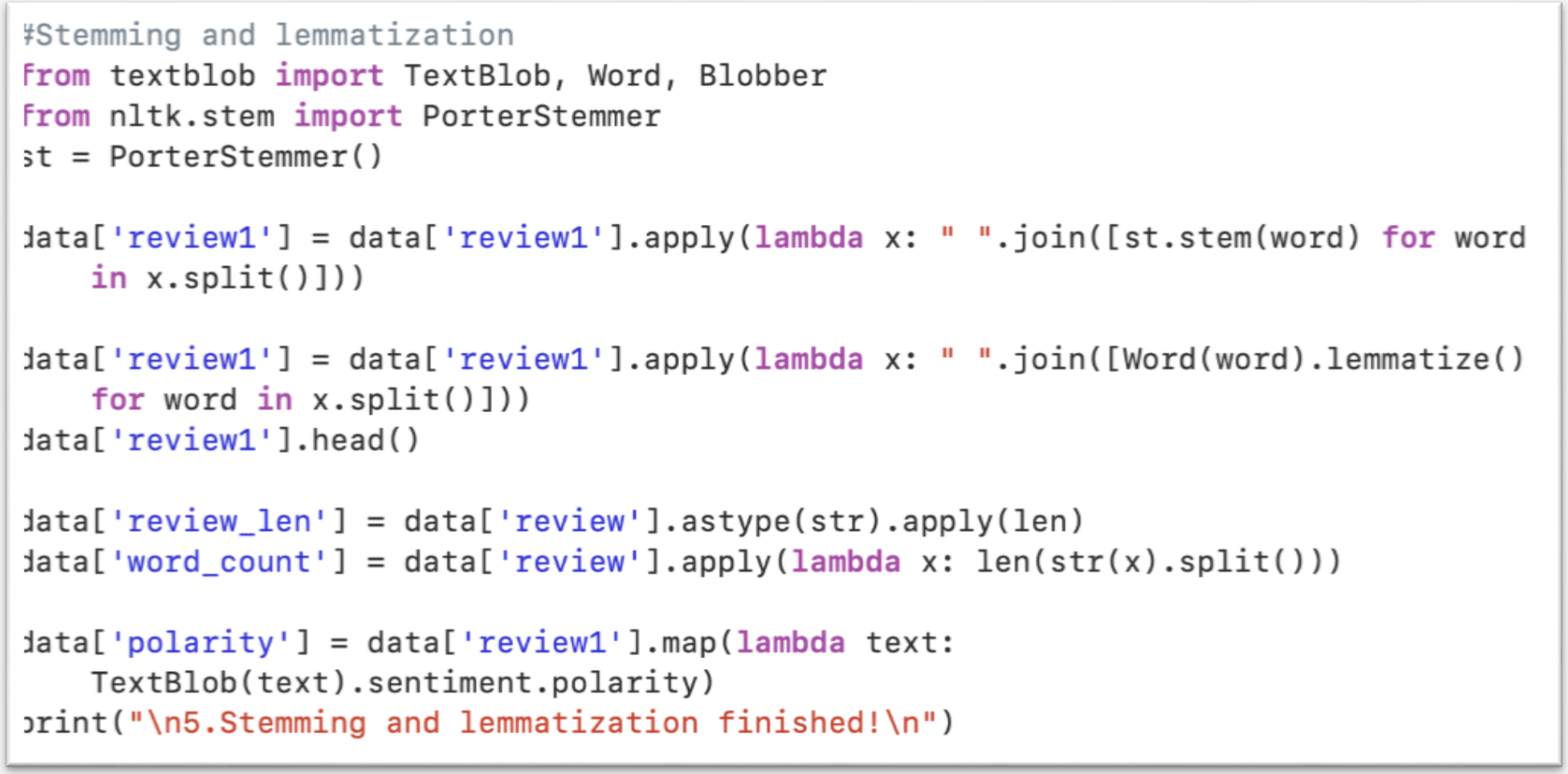


```python
data.tail()
data.shape

#Make the data a bit smaller
data = data[data.groupby('drugName')['drugName'].transform('size') > 20]
data = data.head(10000)

#preprocessing
print('the review column data types is:',data['review'].dtypes)
data['review'] = data['review'].astype(str)

#Converting to lowerCase
data['review1'] = data['review'].apply(lambda x: " ".join(x.lower() for x in x.split()))
print("\n1.converted to lower case.\n")

#Removing Punctuations
data['review1'] = data['review1'].str.replace('[^\w\s]', '')
print("\n2.removed the punctuations already!\n")

#Removing StopWords
import nltk
nltk.download('stopwords')
from nltk.corpus import stopwords
stop = stopwords.words('english')

data['review1'] = data['review1'].apply(lambda x: " ".join(x for x in x.split() if x not in stop))
data['review1'].head()
print("\n3.removed the stopwords already!\n")

#Remove the Rare Words
freq = pd.Series(' '.join(data['review1']).split()).value_counts()
less_freq = list(freq[freq == 1].index)
data['review1'] = data['review1'].apply(lambda x: " ".join(x for x in x.split() if x not in less_freq))
data['review1'].head()
print("\n4.removed the rare words already!\n")
```

01 數據刻畫與預處理

03 情感極性

- ◆ 加入特徵——情感極性（polarity）
- ◆ 情感極性（polarity）：取值範圍為-1 ~ 1，其中-1代表消極情緒，0代表中性，1代表積極情緒。



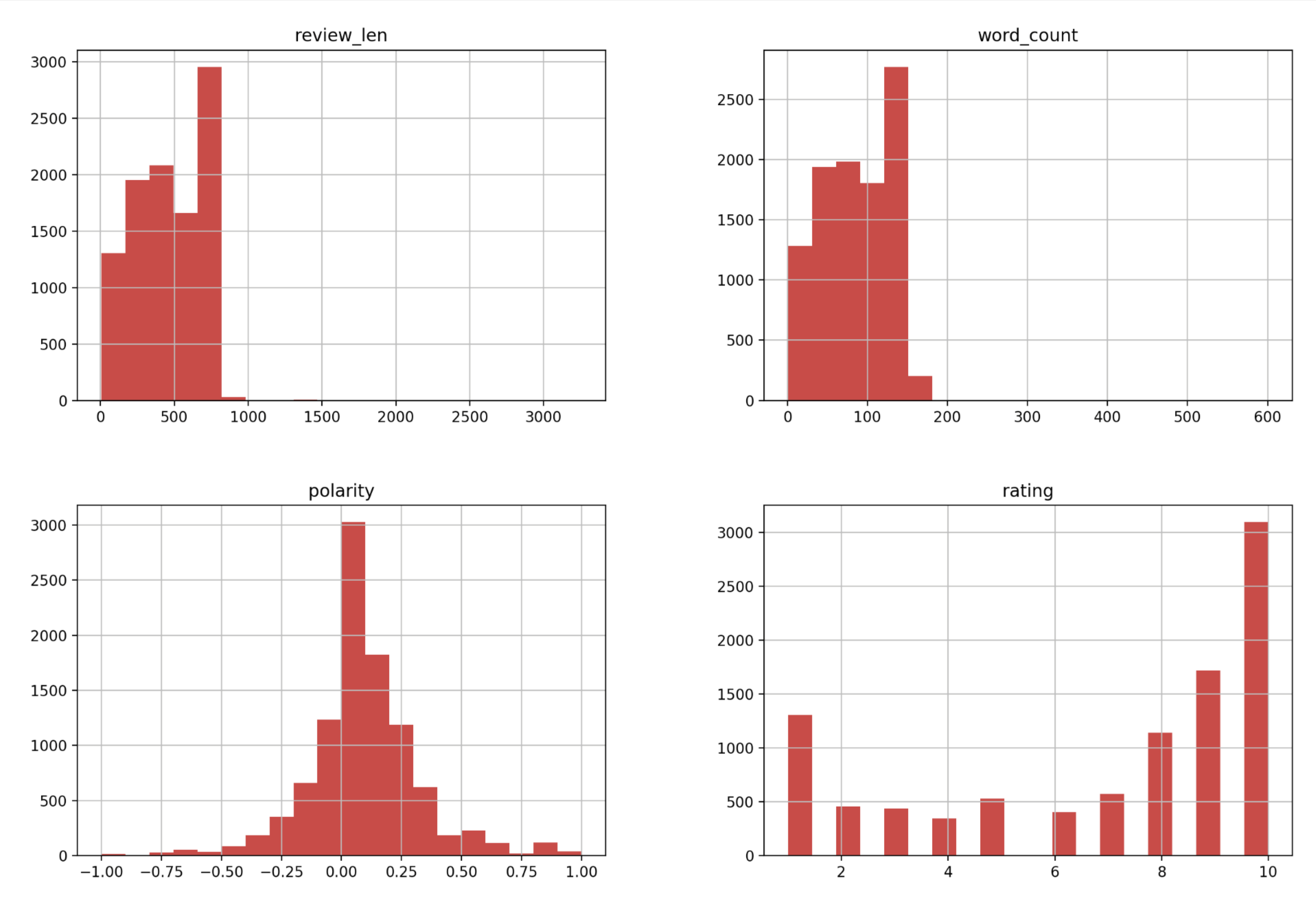
```
#Stemming and lemmatization
from textblob import TextBlob, Word, Blobber
from nltk.stem import PorterStemmer
st = PorterStemmer()

data['review1'] = data['review1'].apply(lambda x: " ".join([st.stem(word) for word
    in x.split()]))

data['review1'] = data['review1'].apply(lambda x: " ".join([Word(word).lemmatize()
    for word in x.split()]))
data['review1'].head()

data['review_len'] = data['review'].astype(str).apply(len)
data['word_count'] = data['review'].apply(lambda x: len(str(x).split()))

data['polarity'] = data['review1'].map(lambda text:
    TextBlob(text).sentiment.polarity)
print("\n5.Stemming and lemmatization finished!\n")
```

01 數據刻畫與預處理

數據特徵

統計得到review的

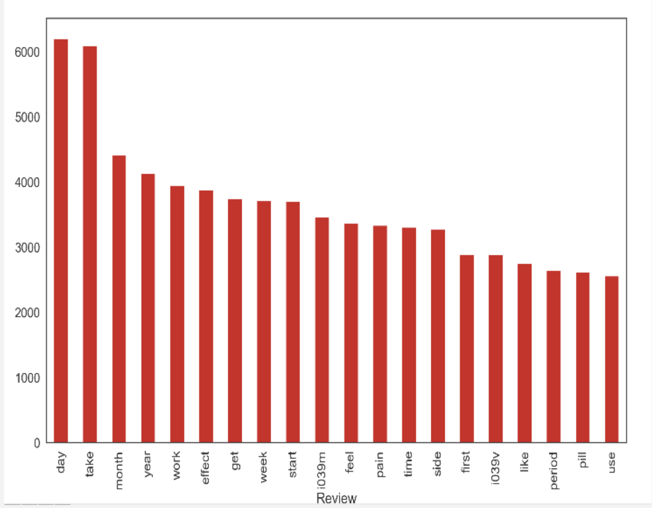
- 長度
- 詞彙數量
- 極性分布
- 評分情況

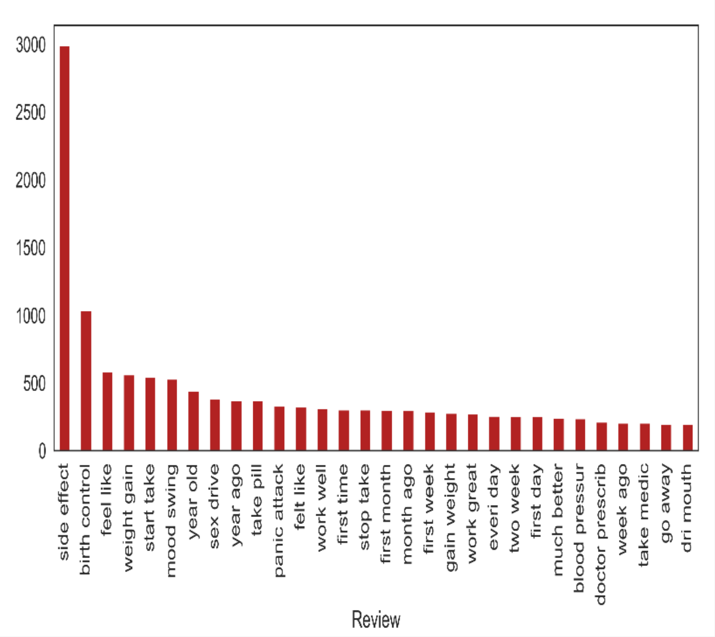


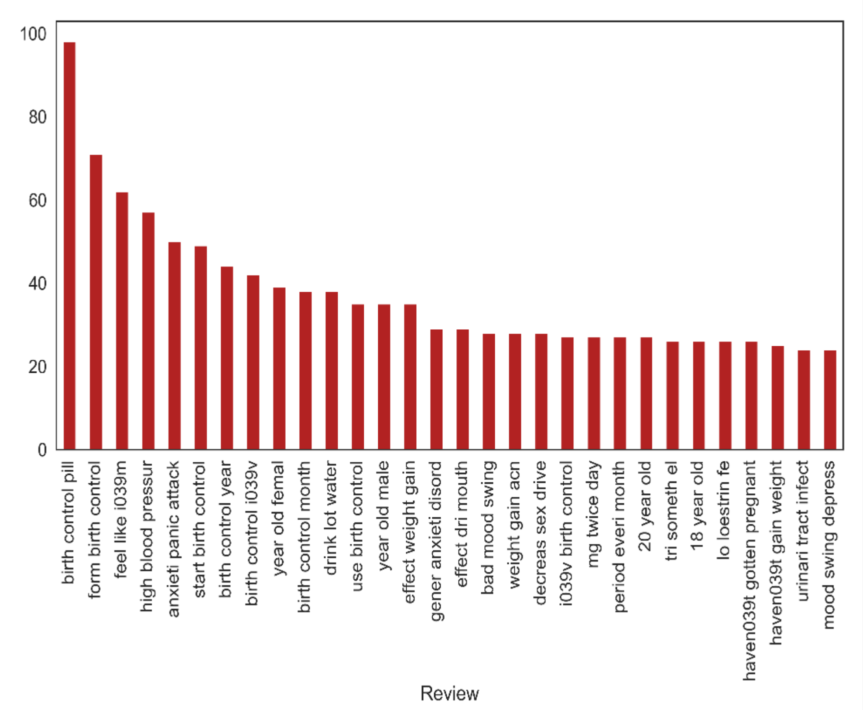
01 數據刻畫與預處理

詞頻分析

- ◆ 1-word 詞頻分析
- ◆ 2-words 詞頻分析
- ◆ 3-words 詞頻分析



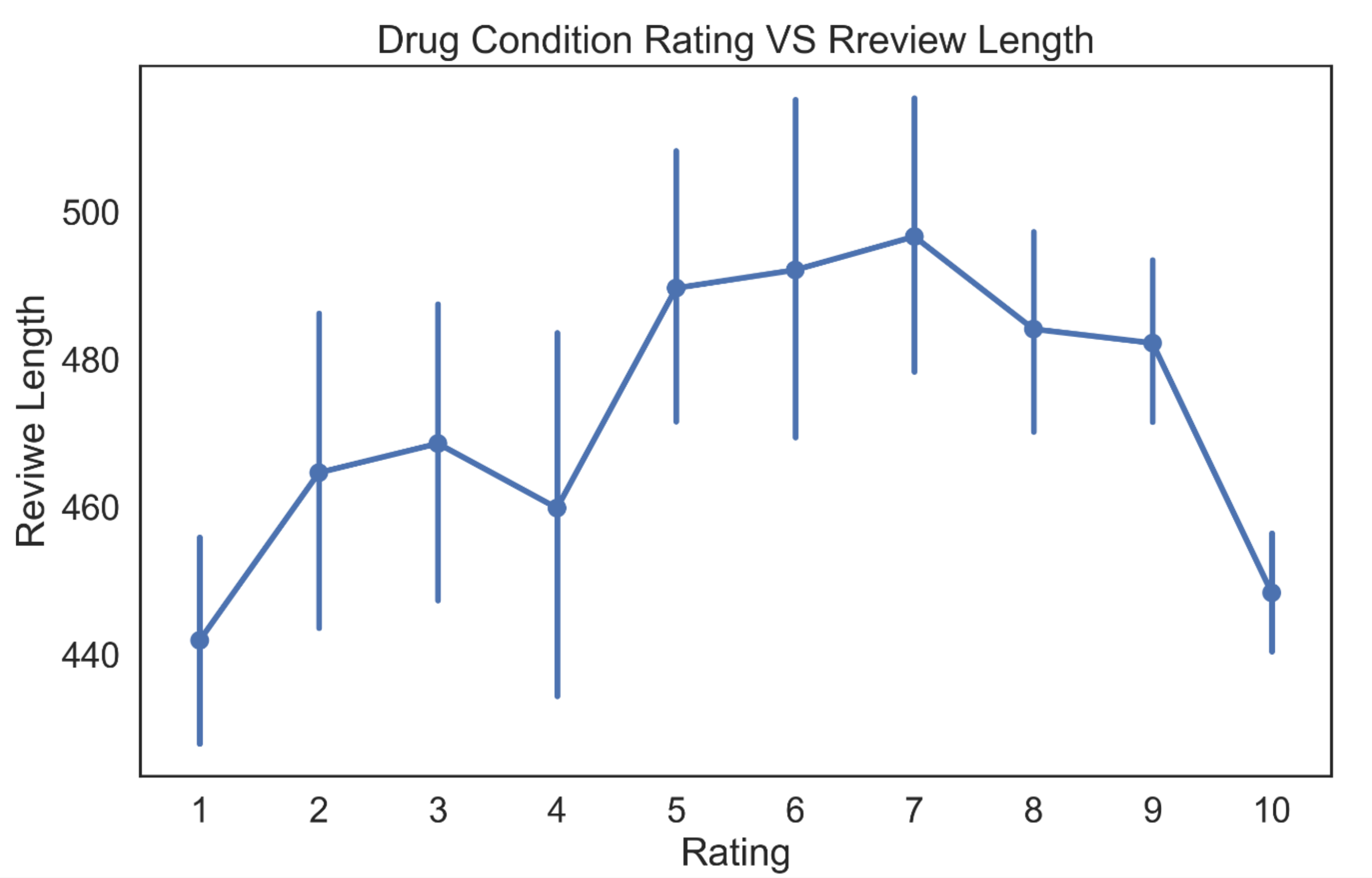




01 數據刻畫與預處理

打分計數 & 評論長度與患者評分間關係

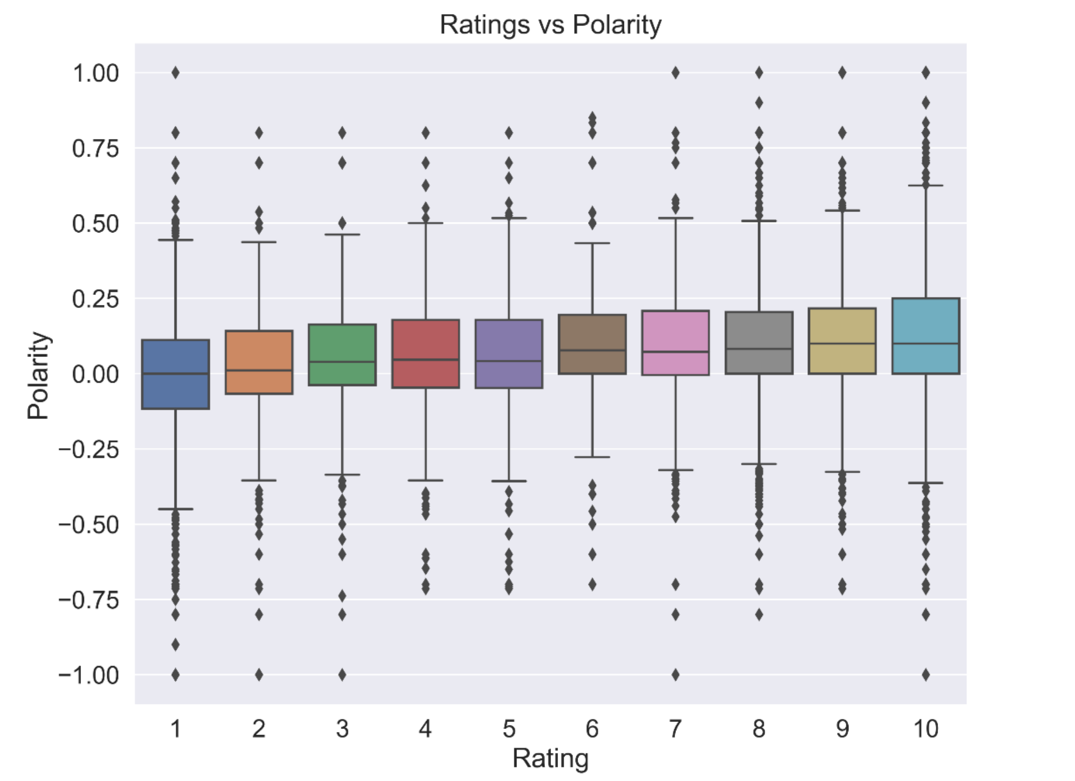
- ◆ 10分review的長度篇短，review在5～9分區間內較長。

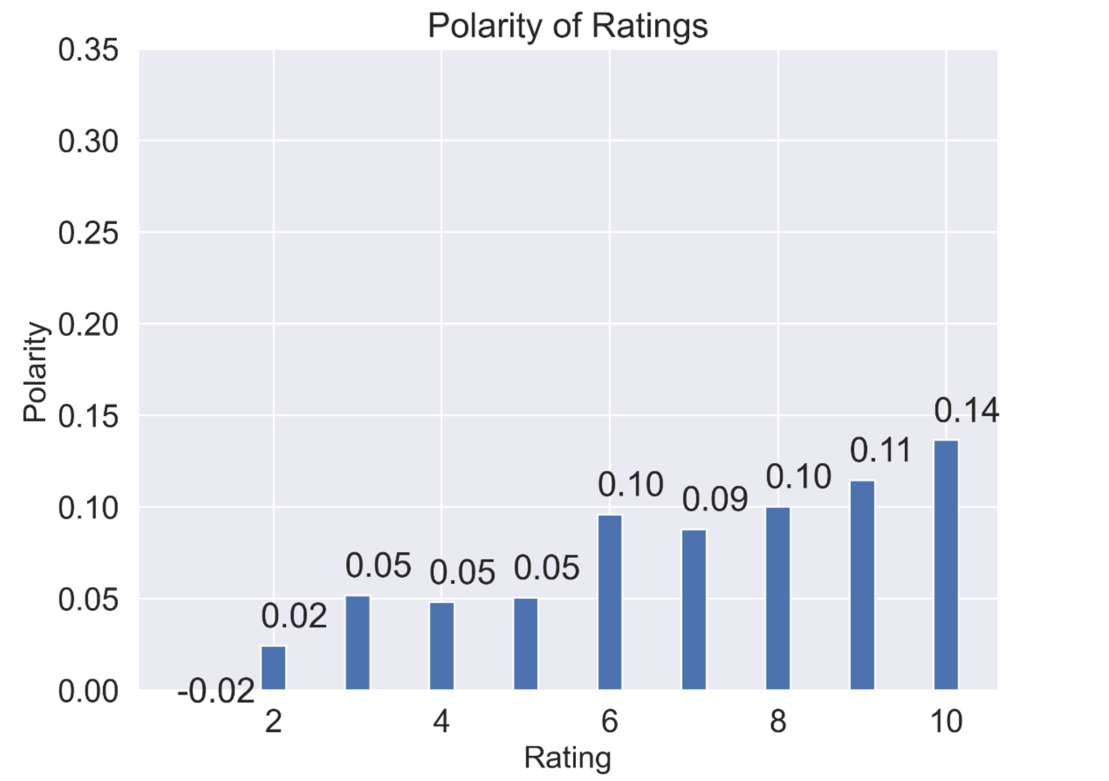


01 數據刻畫與預處理

評分(rating)和情感極性(polarity)關係

- 平均極性會隨著打分的提高而上升，但是在1分評價中異常值較多
- rating>3時，為積極評價； rating<3時，為消極評價-->以rating=3為分界進行初步分類。





目錄

CONTENTS

01

問題解析
&
數據預處理

02

規劃分析
&
模型初探

03

主要問題
&
模型改進

04

總結
&
展望

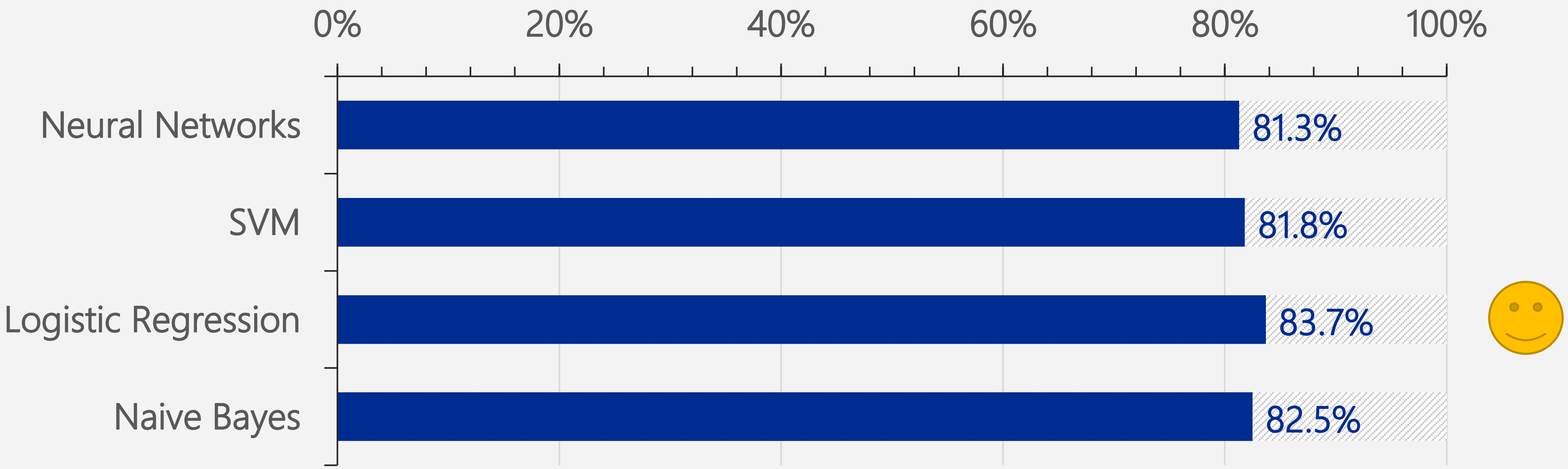
02 初步模型訓練

處理思路

- ◆ 文本向量化
- ◆ 數據集劃分(train/total=7173/10000)
- ◆ 以Positively Rated(0-1)作為y值，向量化文本作為x值進行模型訓練

所選模型

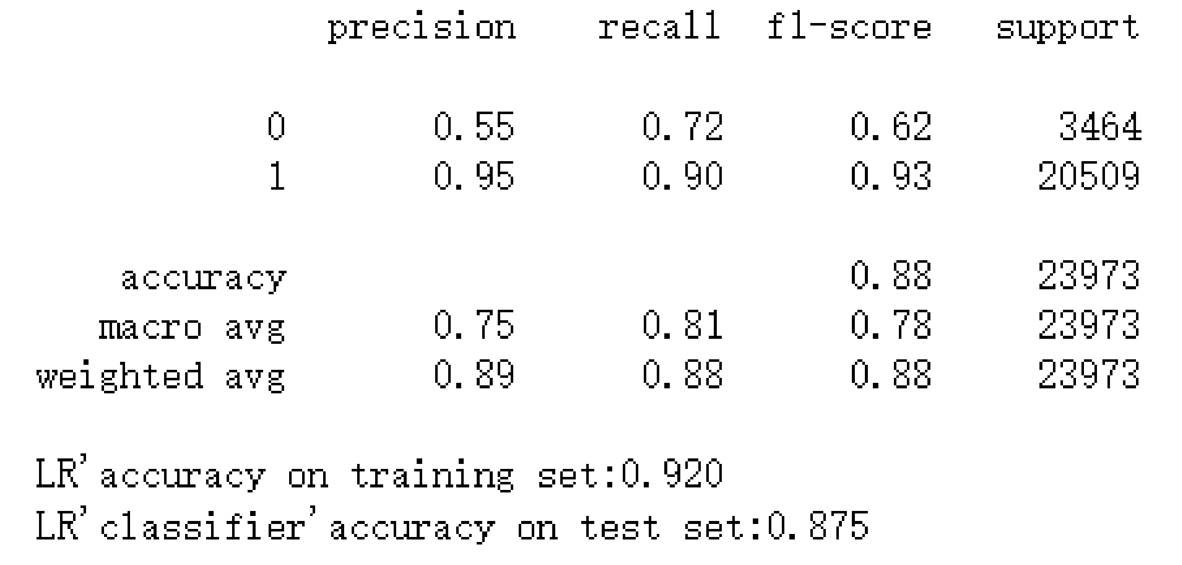
- ◆ Neural Networks
- ◆ SVM
- ◆ Logistic Regression
- ◆ Naive Bayes



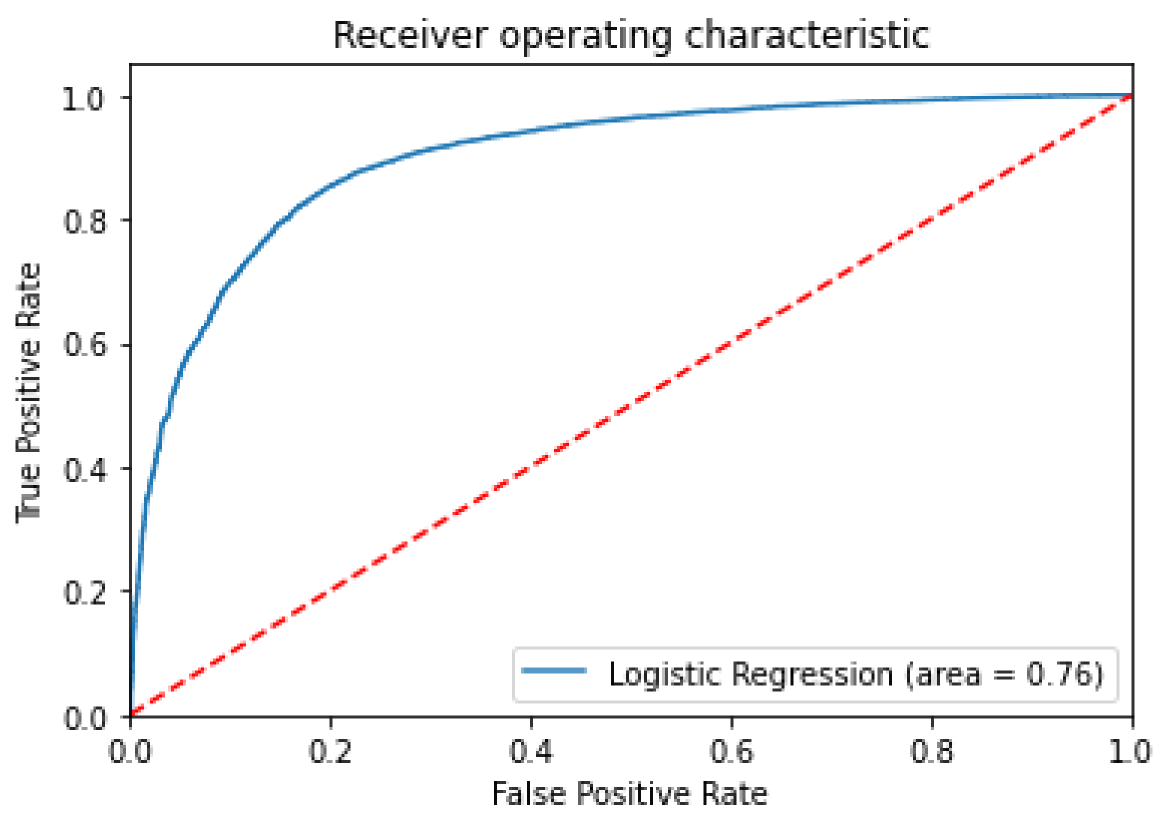
02 初步模型訓練————邏輯回歸模型的改進

過擬合的解決

- ◆ L2正則化
- ◆ 早停（max_iter:10000 -> 1000)



```
              precision    recall  f1-score   support

           0       0.55      0.72      0.62      3464
           1       0.95      0.90      0.93     20509

    accuracy                           0.88     23973
   macro avg       0.75      0.81      0.78     23973
weighted avg       0.89      0.88      0.88     23973

LR’accuracy on training set:0.920
LR’classifier’accuracy on test set:0.875
```



目錄 CONTENTS

01

問題解析 & 數據預處理

02

規劃分析 & 模型初探

03

主要問題 & 模型改進

04

總結 & 展望

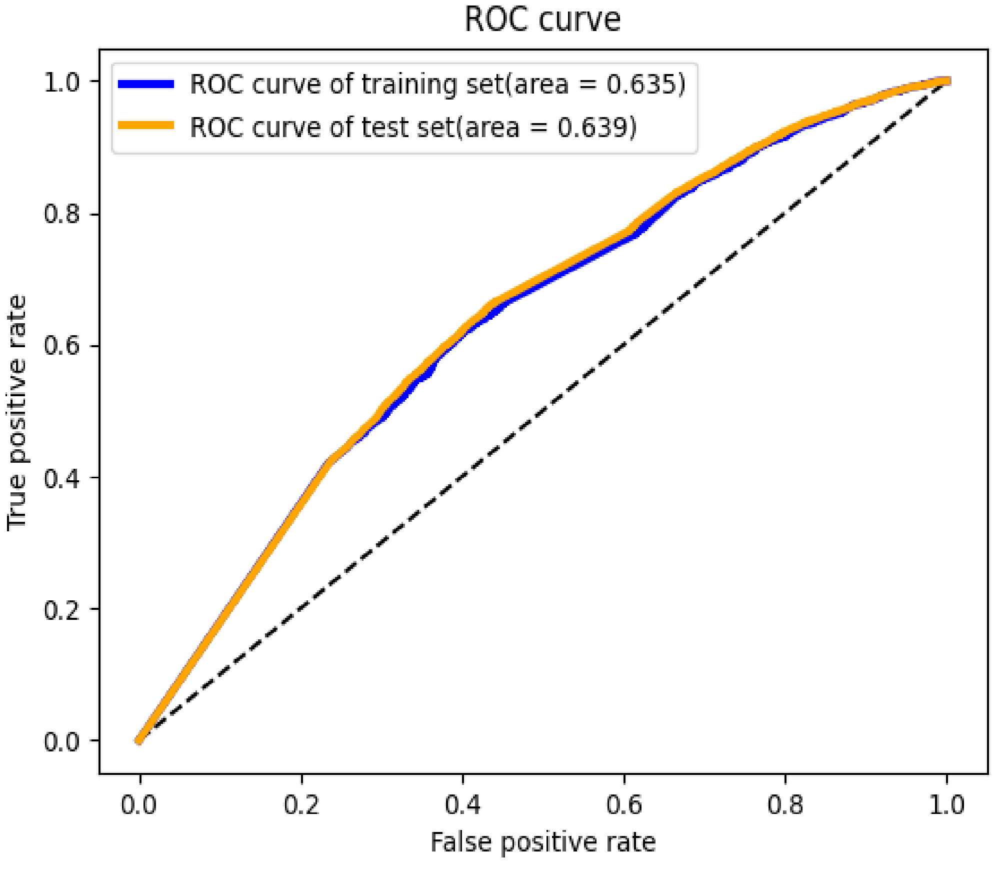
03 模型問題與挑戰

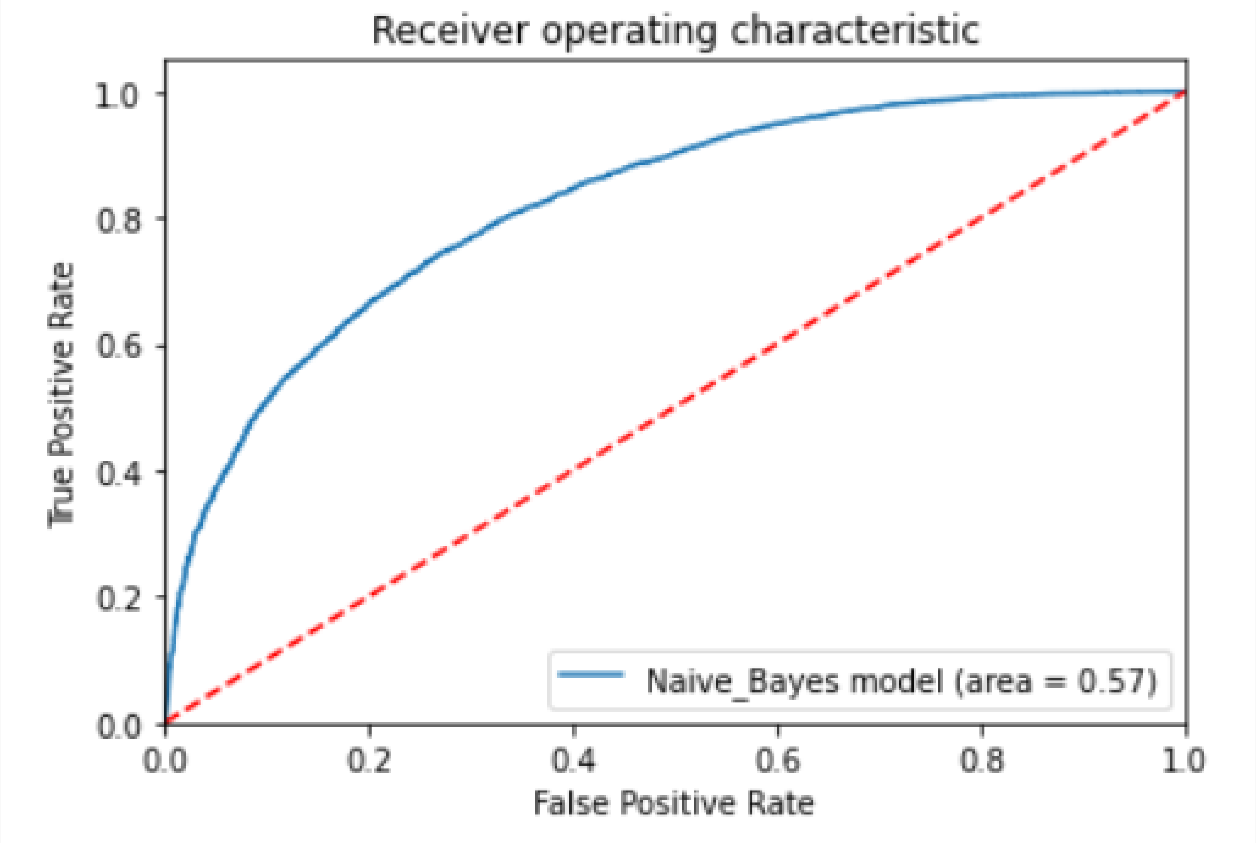
01 初步訓練的模型中，有 **過擬合的問題**

02 情感的傾向是一個更複雜的分類問題，故應該改以 **多分類處理**

03 模型改進

二分類 ROC curve





✓ Neural Networks

- The accuracy of training set: 0.813
- The accuracy of test set: 0.816

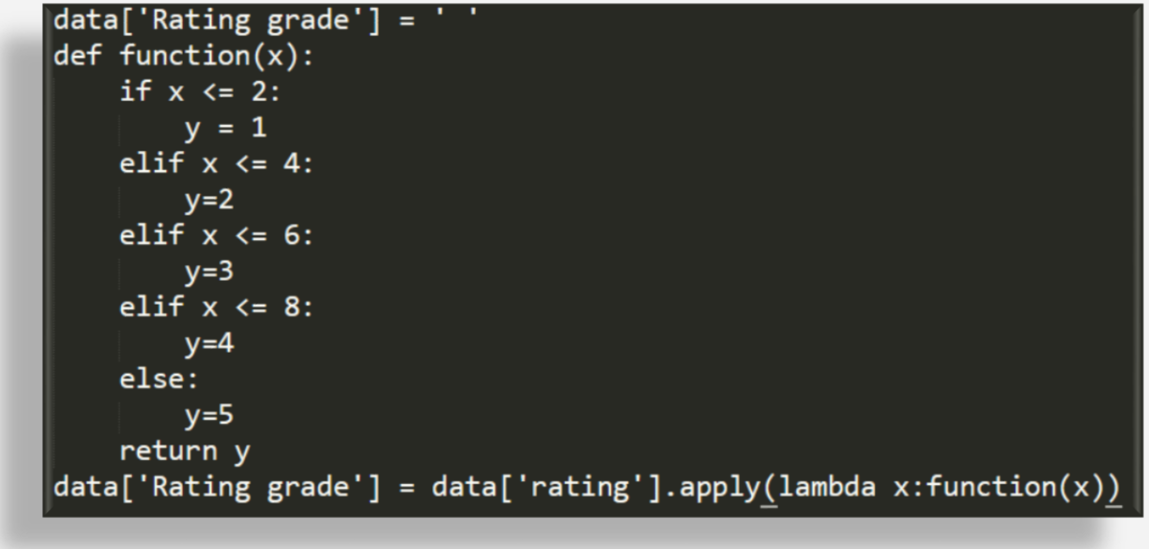
✓ Naïve Bayes

- The accuracy of training set: 0.827
- The accuracy of test set: 0.825

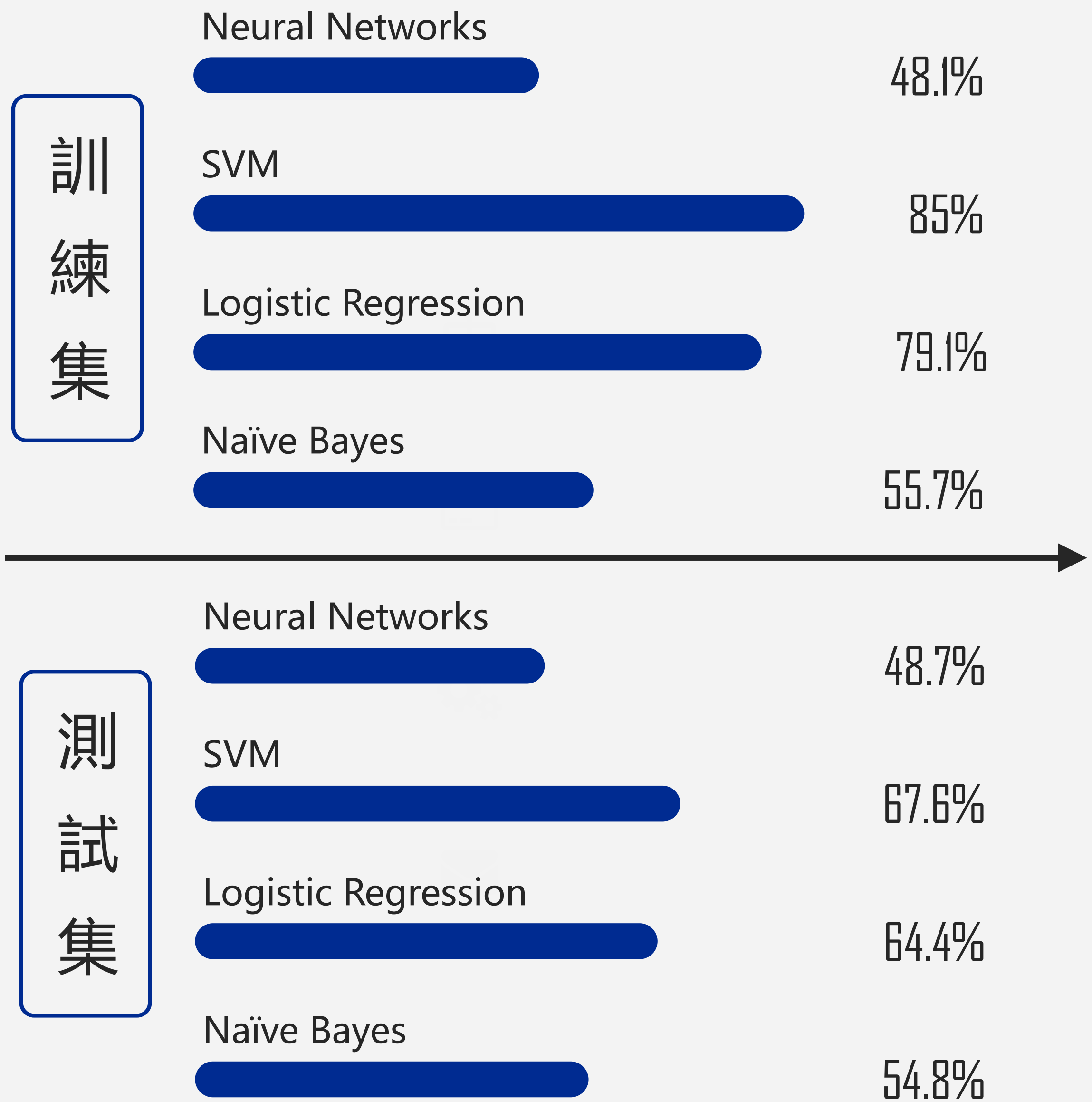
03 模型改進

將二分類轉化為五分類

- ◆ 重新定義評級規則，進行五星級評分。
- ◆ 將原來的1-2分記為1星，3-4分記為2星，5-6分記為3星，7-8分記為4星，9-10分記為5星。



```
data['Rating grade'] = ' '
def function(x):
    if x <= 2:
        y = 1
    elif x <= 4:
        y=2
    elif x <= 6:
        y=3
    elif x <= 8:
        y=4
    else:
        y=5
    return y
data['Rating grade'] = data['rating'].apply(lambda x:function(x))
```



目錄

CONTENTS

01

問題解析
&
數據預處理

02

規劃分析
&
模型初探

03

主要問題
&
模型改進

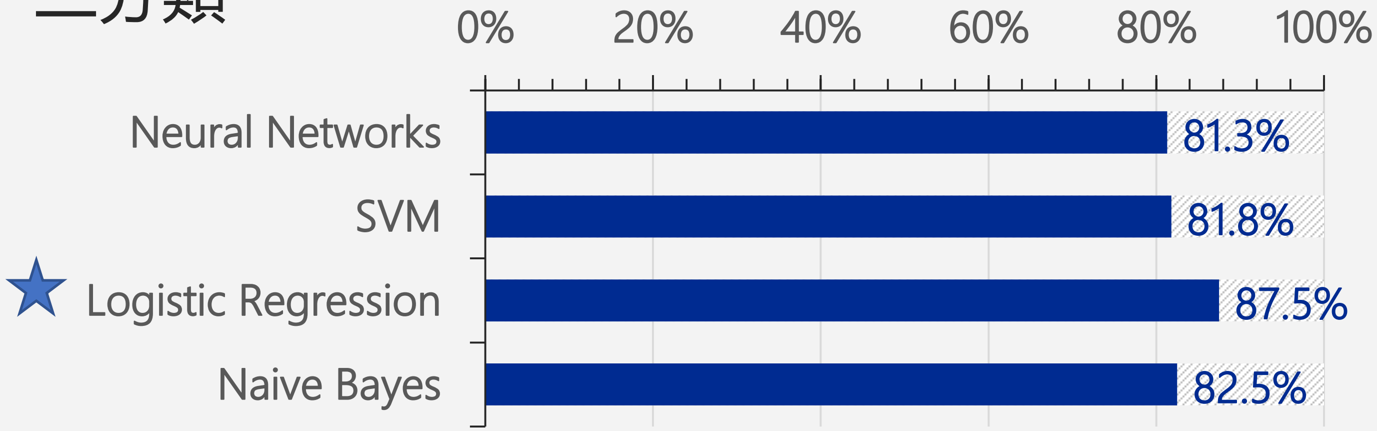
04

總結
&
展望

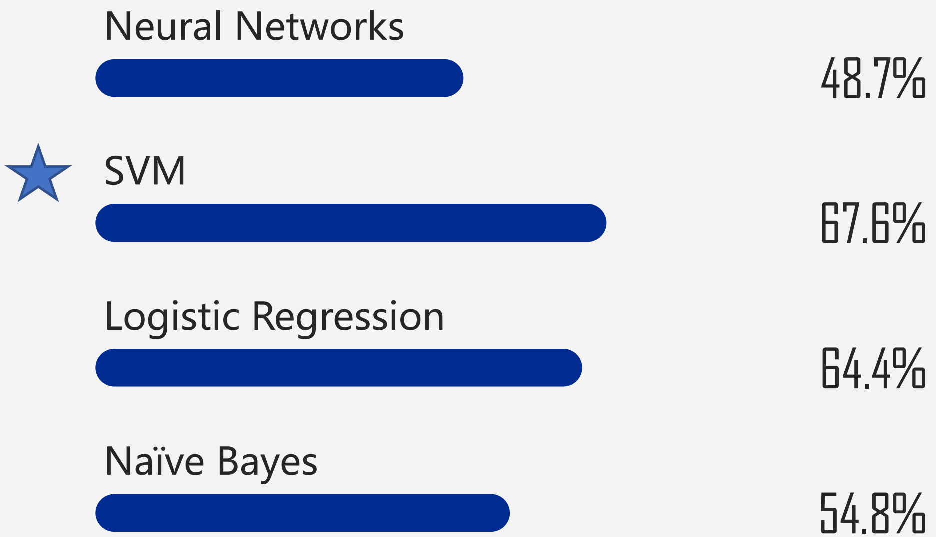
04 總結與展望

- ✓ 原始數據的特徵分析
- ✓ 數據預處理
- ✓ 實現四個模型的二分類
- ✓ 繪製ROC曲線
- ✓ 實現四個模型的多分類
- ✓ 不同模型的比較分析

二分類



五分類



04 总结与展望

對數據中的其他特徵進行訓練，例如建立起患者評價與“點贊數”之間的聯繫，研究通過使用者的主觀評價對網友“點贊數”分類的結果；或者在模型中加入多個特徵進行訓練，探究增加預測準確率的可能性。

進一步優化現有模型，增加其準確率和穩定性，並以此為基礎建立一個完善的藥品預測系統，以實現通過使用者的主觀評價對藥品進行評估，同時也能向醫師提供一個臨床決策的支持工具，進而針對藥物的有效性、安全性等進行研究。另外這也能讓保險公司與藥廠在製造上有所幫助。