Machine Learning

基于药物评价的情感分析模型

2022年5月24日

目录

CONTENTS

01
问题解析
数据预处理

02
模型初探

03
模型改进

04
总结展望

目录

CONTENTS

01

问题解析
数据预处理

02

模型初探

03

模型改进

04

总结展望

01 研究问题

基于药物评价的情感分析模型

药物治疗在疾病的治疗中起着非常重要的作用和作用。患者对药物的评价和满意度也会影响治疗进程和医生的用药方案。因此，本项目将使用与患者对特定药物的评论和反馈相关的数据，并将应用机器学习模型来尝试**评估药物**。

研究思路

将对药物评论向量化，通过评价与其对应的情感倾向进行模型训练，进而得到对药物的总体评级。

数据来源

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 数据集

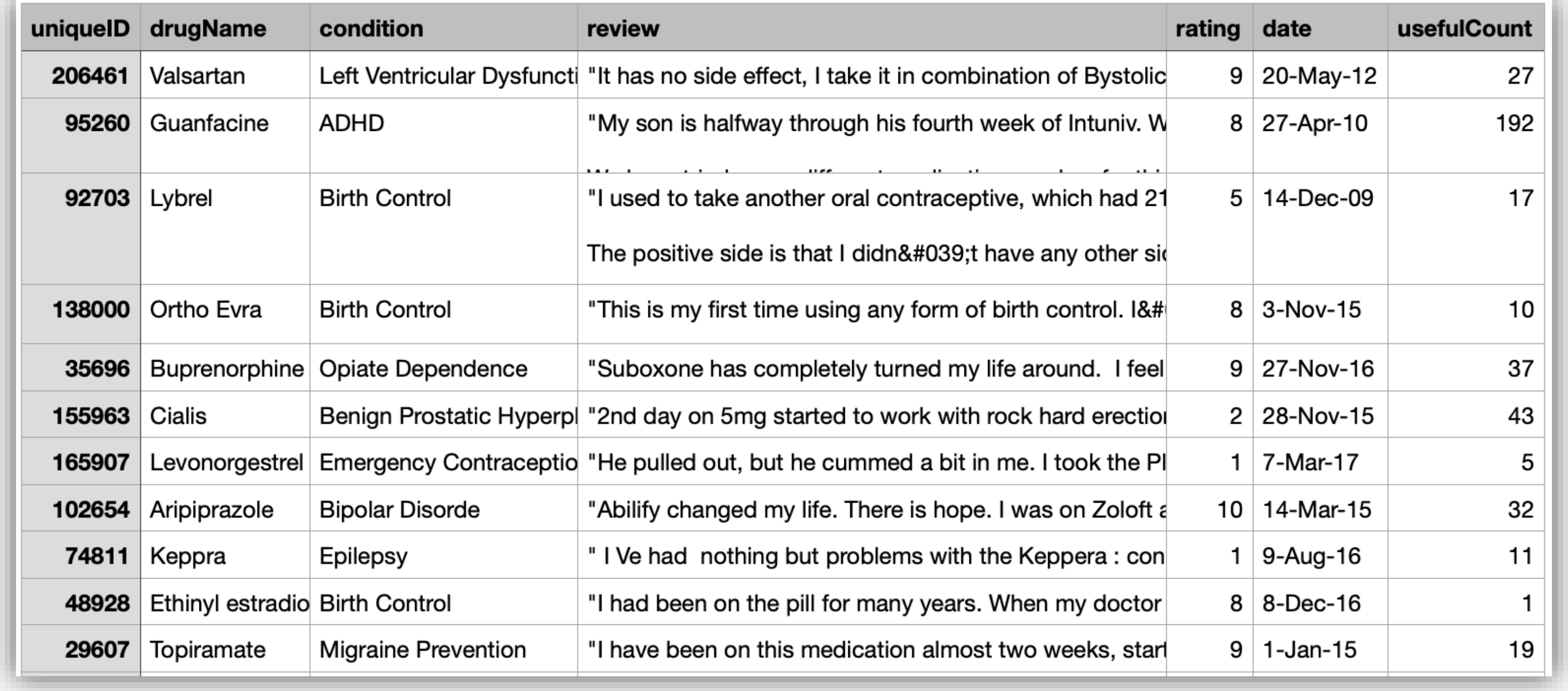
本次研究的数据集信息如下：

数据集大小

- 训练集：161,000
- 测试集：53,800

内容信息

- 药品名称(categorial)
- 对应病症(categorial)
- 患者评价(text)
- 患者打分(numerical)：10星打分制
- 评价日期(date)
- “点赞数”：认为该评价有用的用户数量



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 si	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

工作计划

- 数据预处理：刻画数据结构，讨论模型方法
- 初步模型训练与算法实现
- 进阶模型探索：多分类模型

01 数据刻画与预处理

01 数据量控制

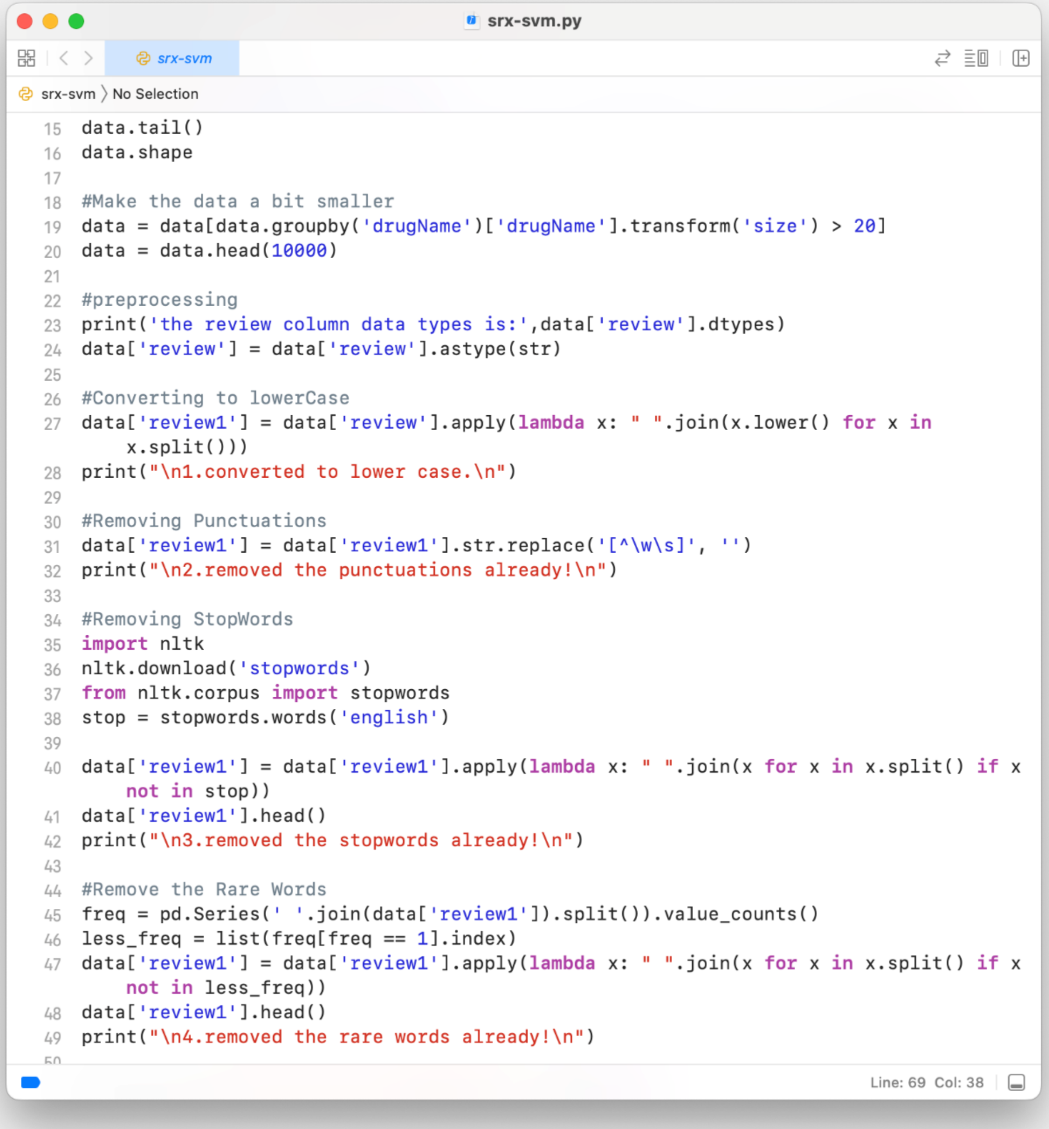
- ◆ 删除缺失值，并随机选取了10000个数据作为研究对象
- ◆ 删除无关列（uniqueID, condition,date,usefulCount列）
- ◆ 删除少于20个评价的药物，保证评价具有代表性

02 文本信息的处理

- ◆ 统一格式：删除标点符号、大写字母变为小写字母
- ◆ 删除Stop Words（如"the","a","in"等词语）
- ◆ 删除出现频率过少的词
- ◆ 提取词干：

SnowballStemmer：删除相似单词

PorterStemmer：删除单词中常见的形态词尾和固定词尾

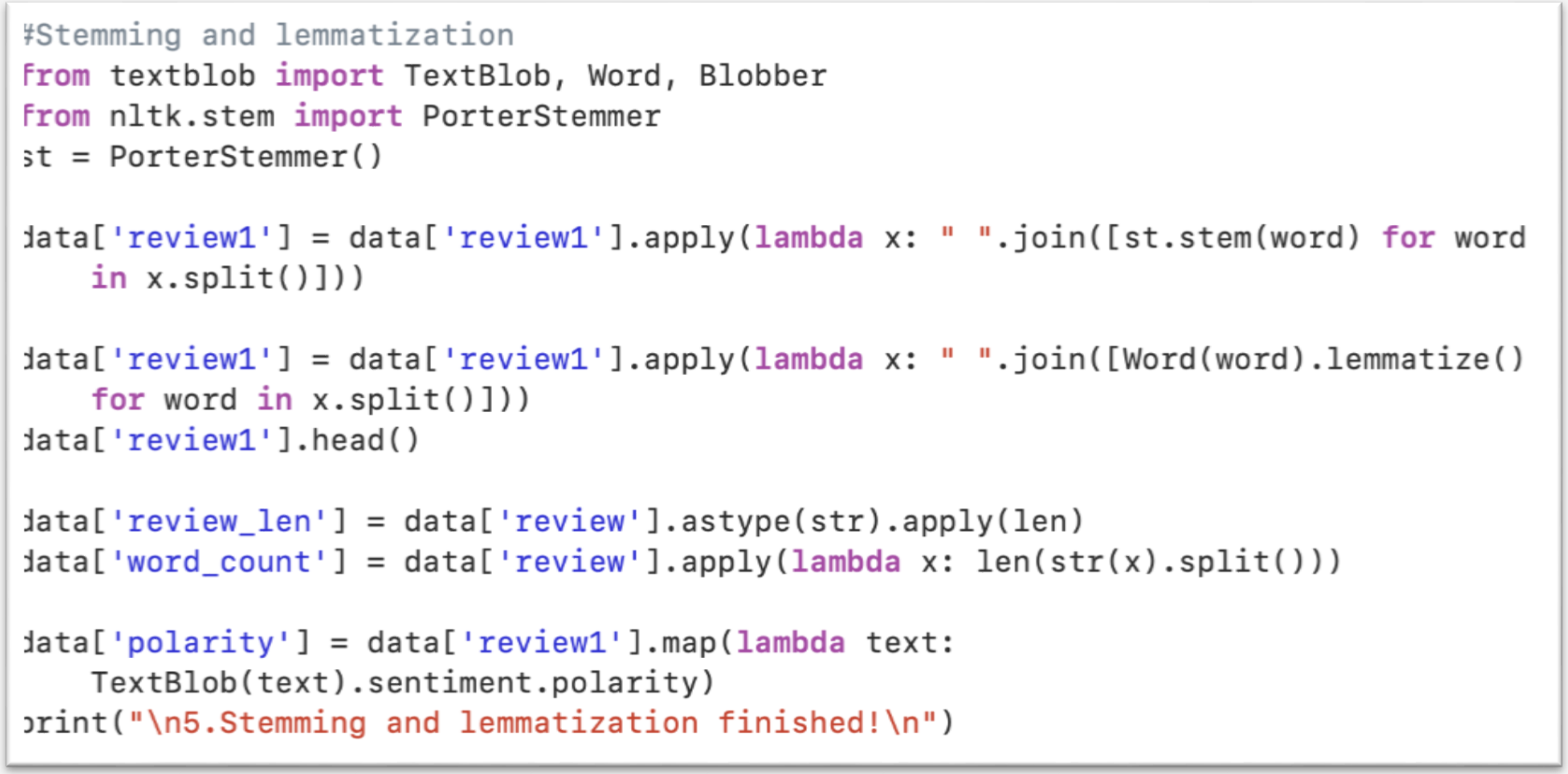


```python
data.tail()
data.shape

#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 数据刻画与预处理

03 情感极性

- 加入特征——情感极性（polarity）
- 情感极性（polarity）：取值范围为-1 ~ 1，其中-1代表消极情绪，0代表中性，1代表积极情绪。



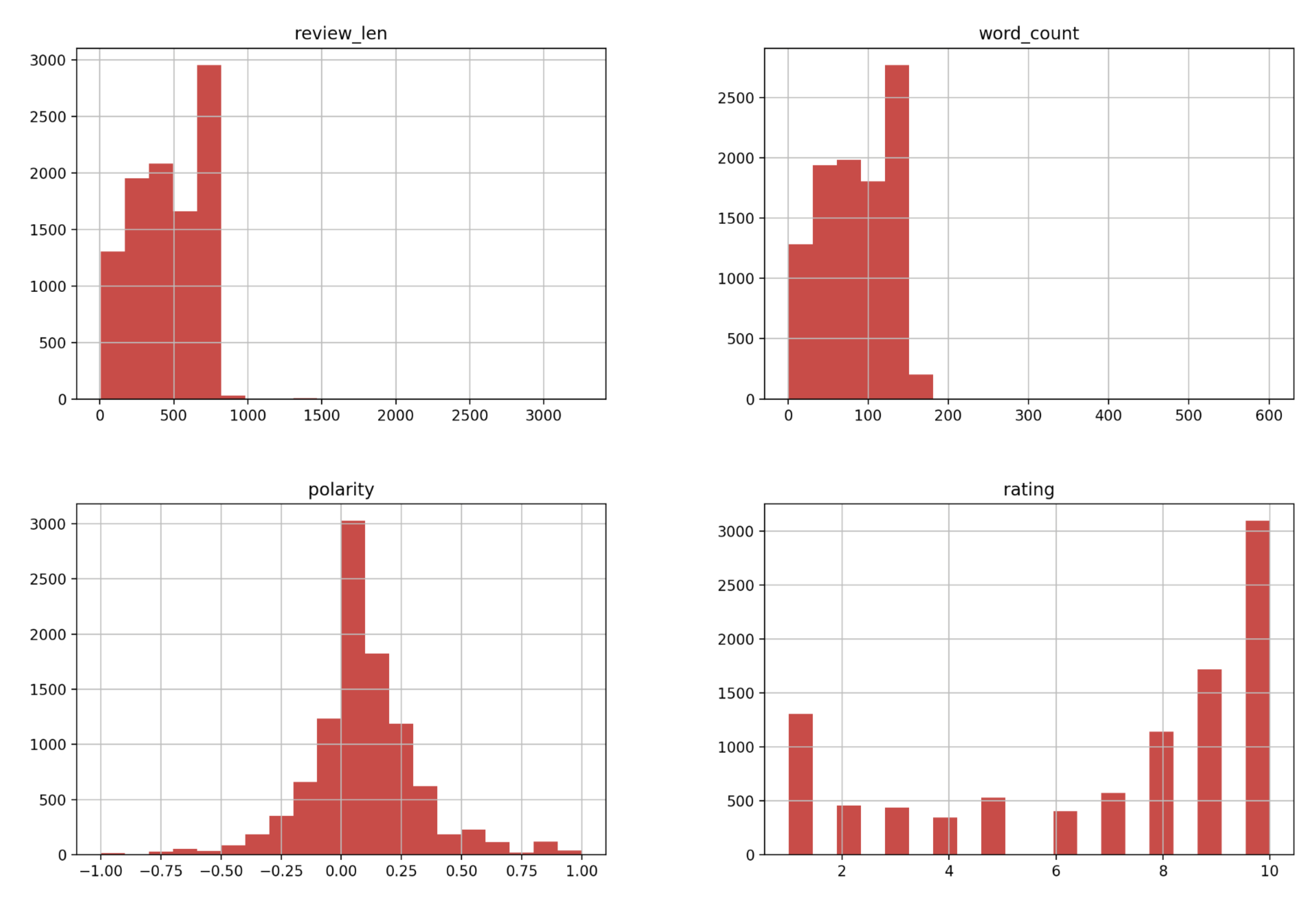
```
#Stemming and lemmatization
from textblob import TextBlob, Word, Blobber
from nltk.stem import PorterStemmer
st = PorterStemmer()

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 数据刻画与预处理

数据特征

统计得到review的

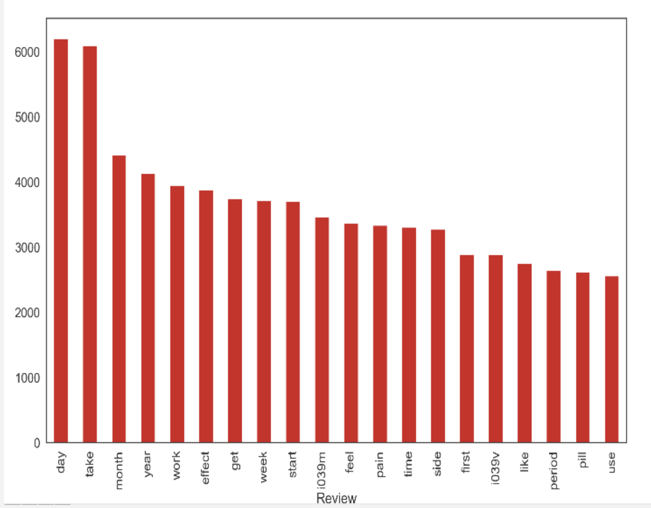
- 长度
- 词汇数量
- 极性分布
- 打分情况



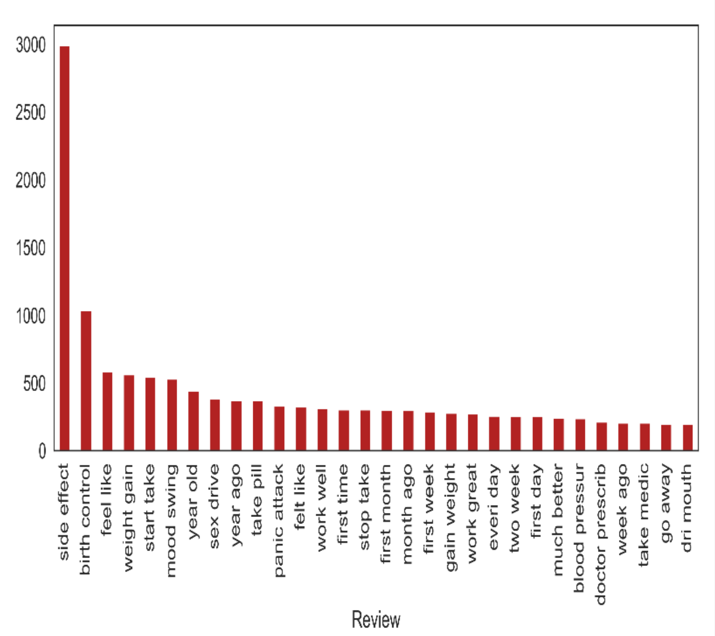
01 数据刻画与预处理

词频分析

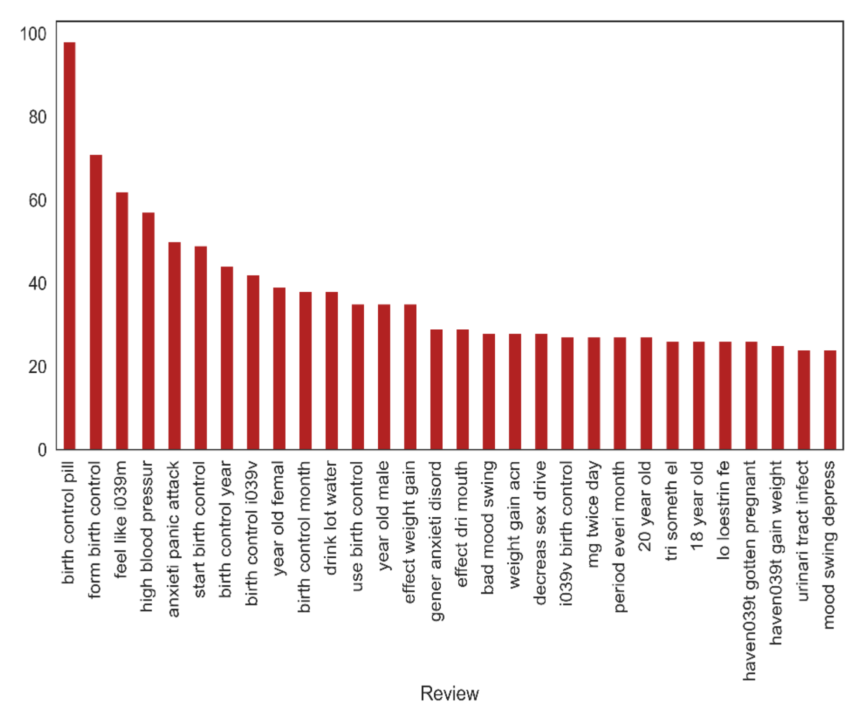
- 1-words词频分析



- 2-words词频分析



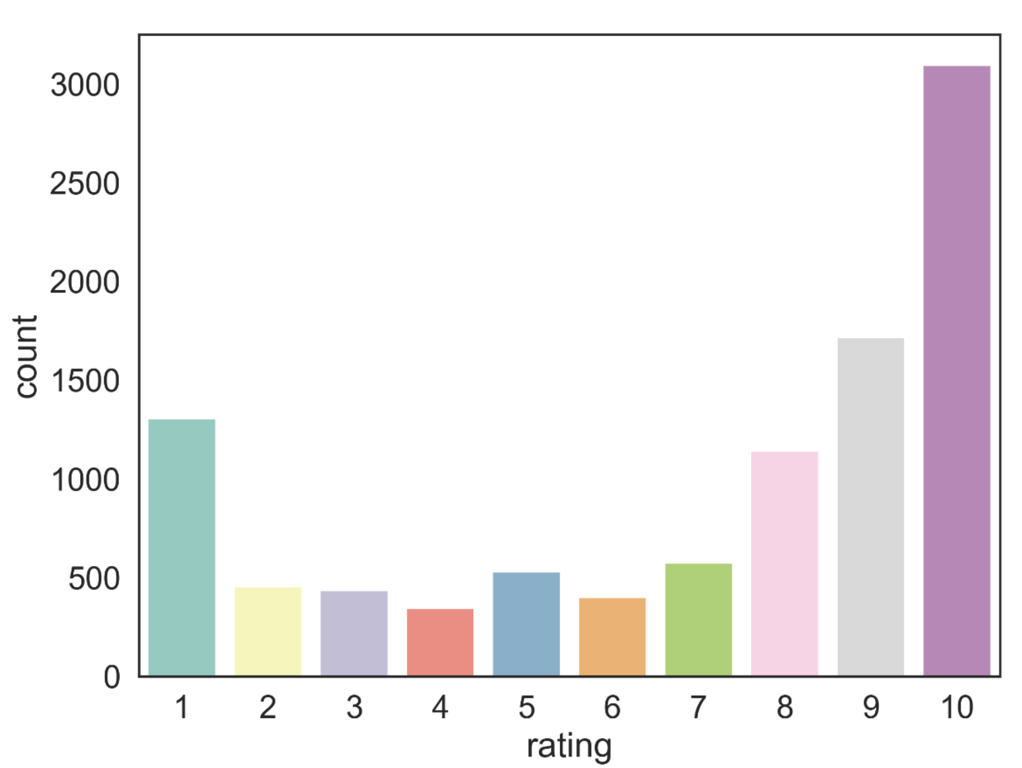
- 3-words词频分析

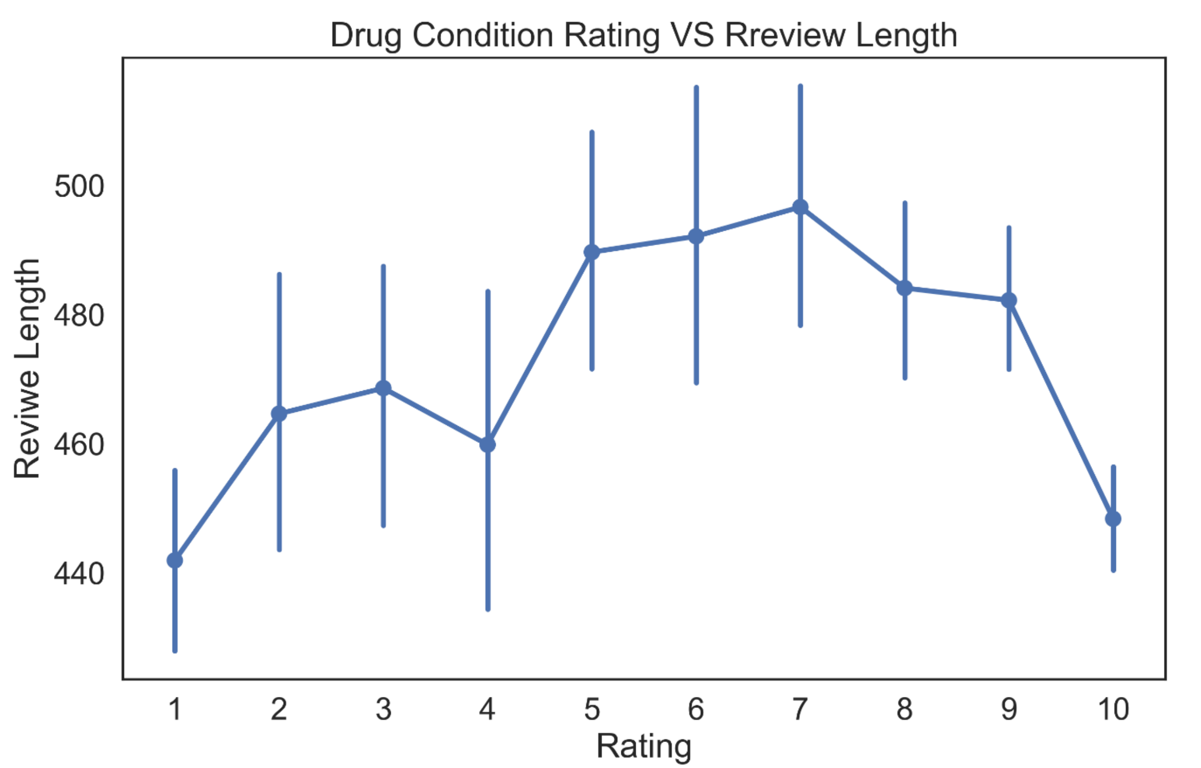


01 数据刻画与预处理

打分计数 & 评论长度与患者打分间关系

- ◆ 数据集的大多数评论的评分都是10，极端打分的数量较多
- ◆ 10分review的长度偏短，review在5～9分区间内较长。

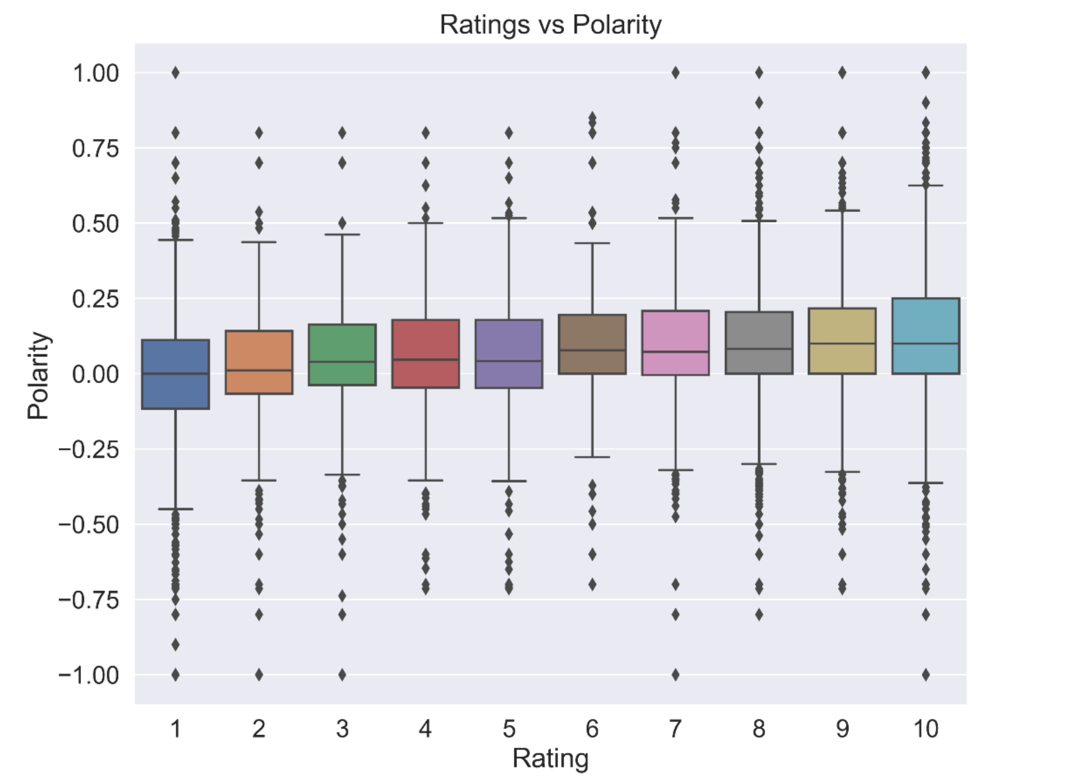


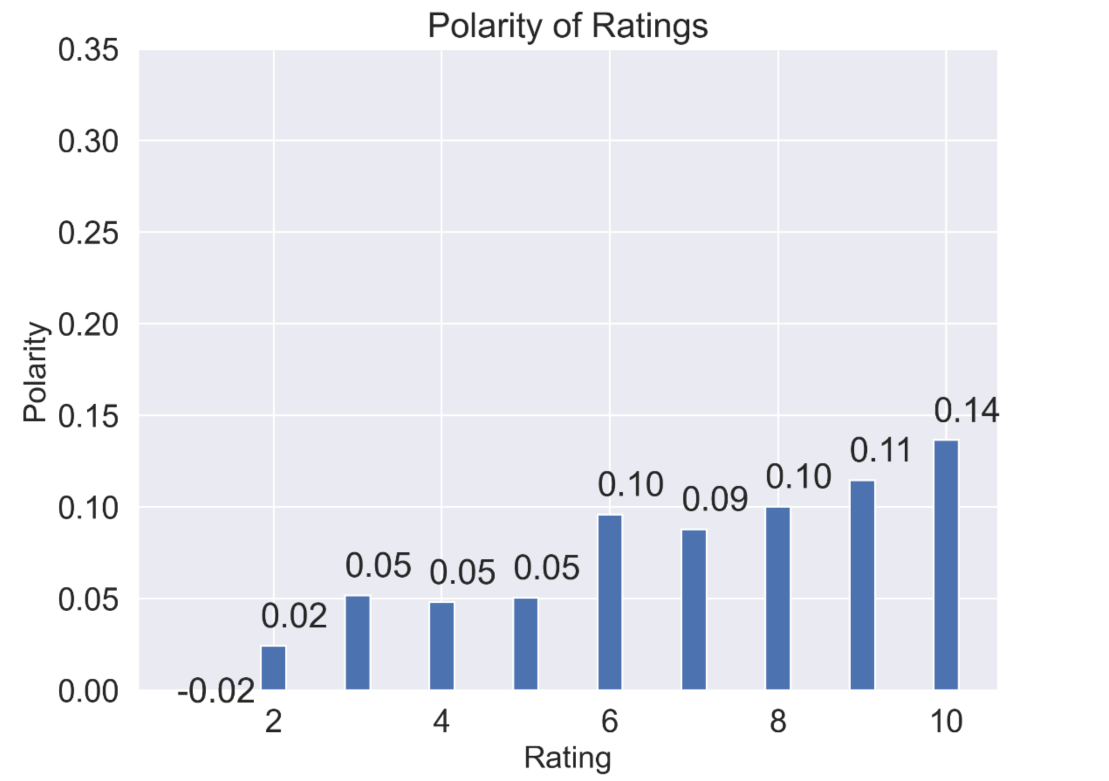


01 数据刻画与预处理

打分(rating)和情感极性(polarity)关系

- ◆ 平均极性会随着打分的提高而上升，但是在1分评价中异常值较多
- ◆ rating>3时，为积极评价； rating<3时，为消极评价-->以rating=3为分界进行初步分类。





目录

CONTENTS

01 问题解析 数据预处理

02 模型初探

03 模型改进

04 总结展望

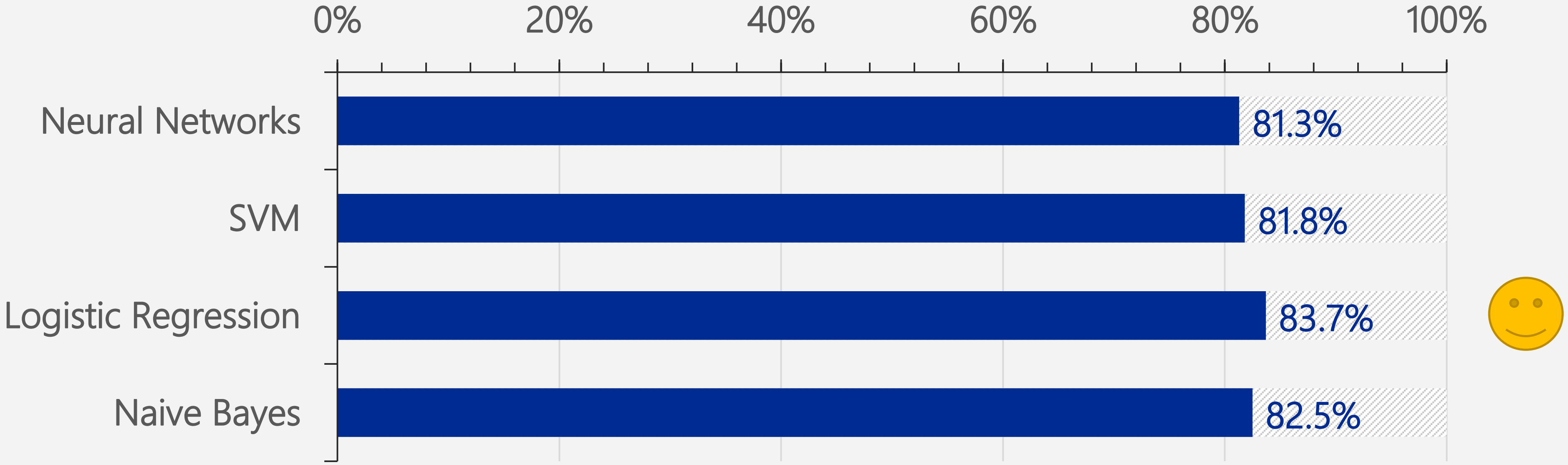
02 初步模型训练

处理思路

- ◆ 文本向量化
- ◆ 数据集划分(train/total=7173/10000)
- ◆ 以Positively Rated(0-1)作为y值，向量化文本作为x值进行模型训练

所选模型

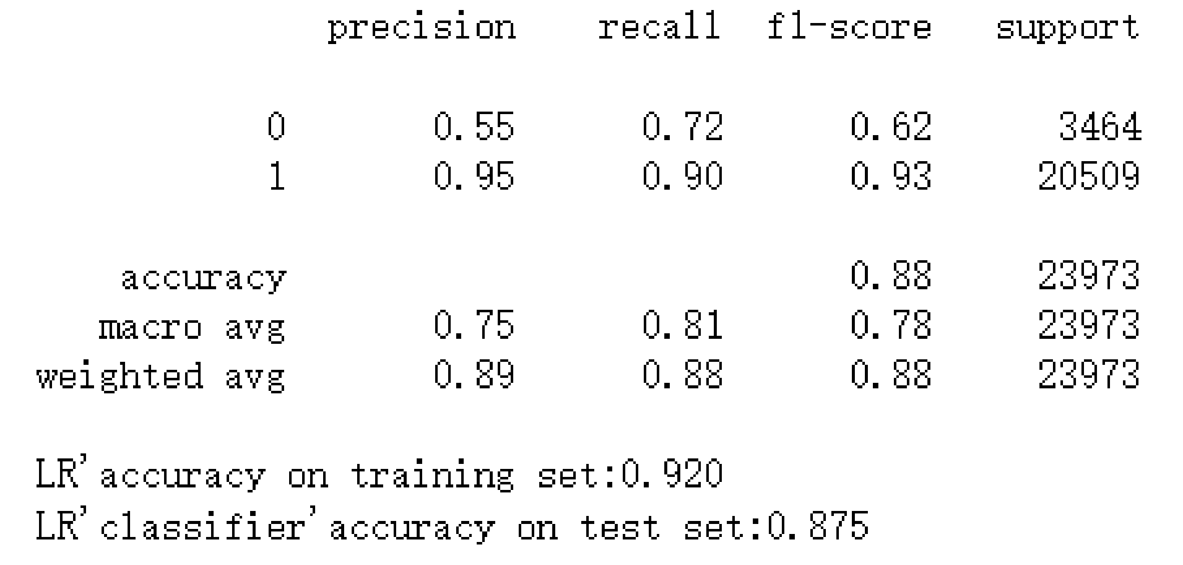
- ◆ Neural Networks
- ◆ SVM
- ◆ Logistic Regression
- ◆ Naive Bayes



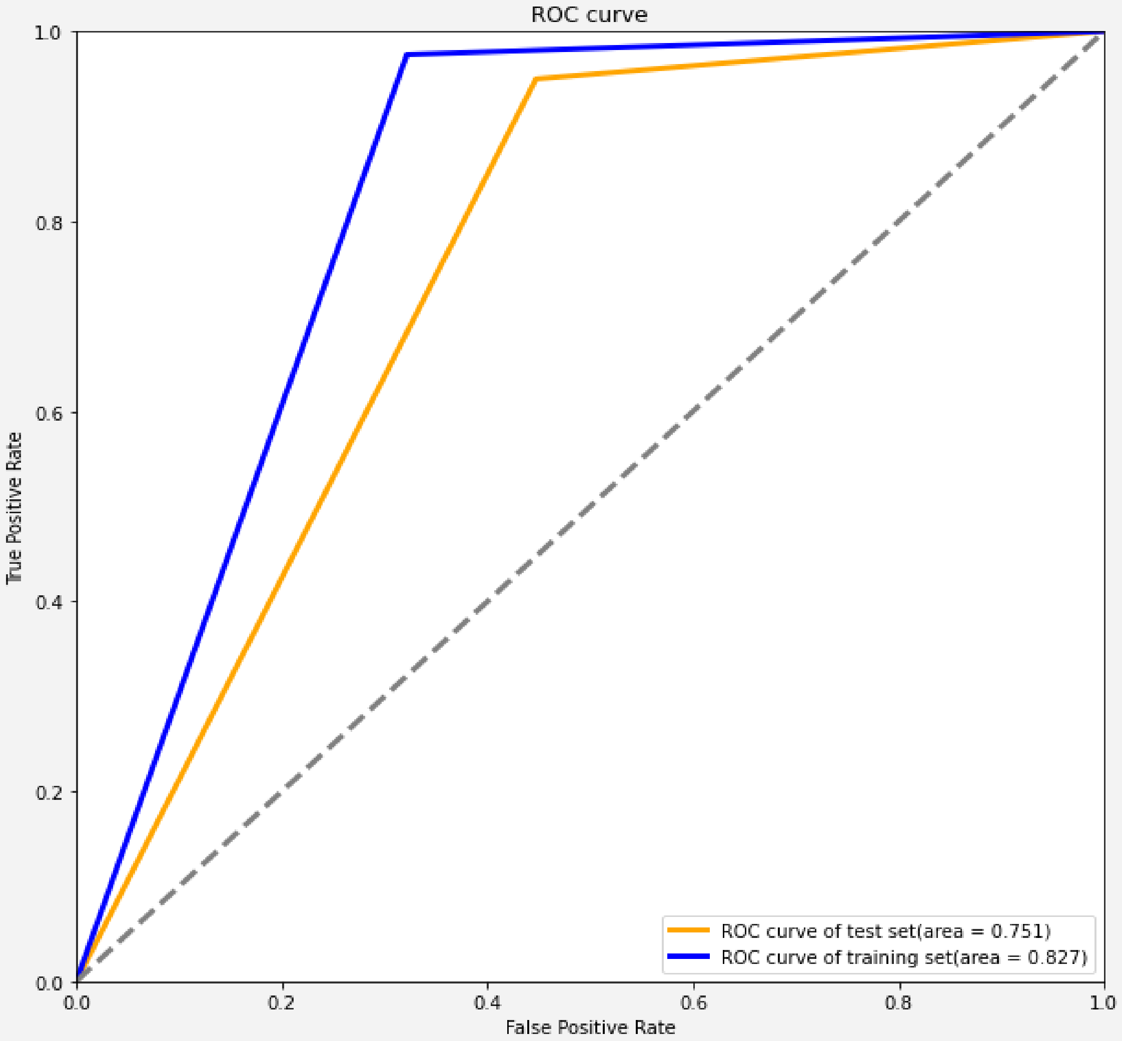
02 初步模型训练————逻辑回归模型的改进

过拟合的解决

- L2正则化
- 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
```



目录

CONTENTS

01
问题解析
数据预处理

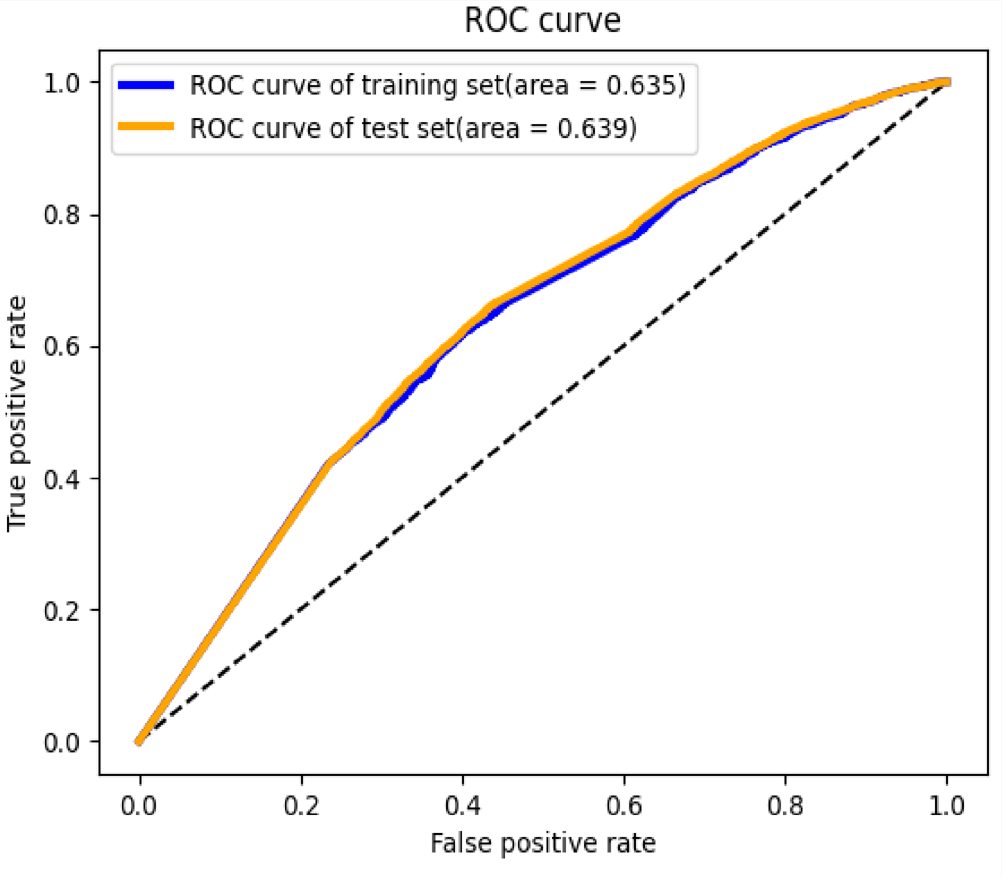
02
模型初探

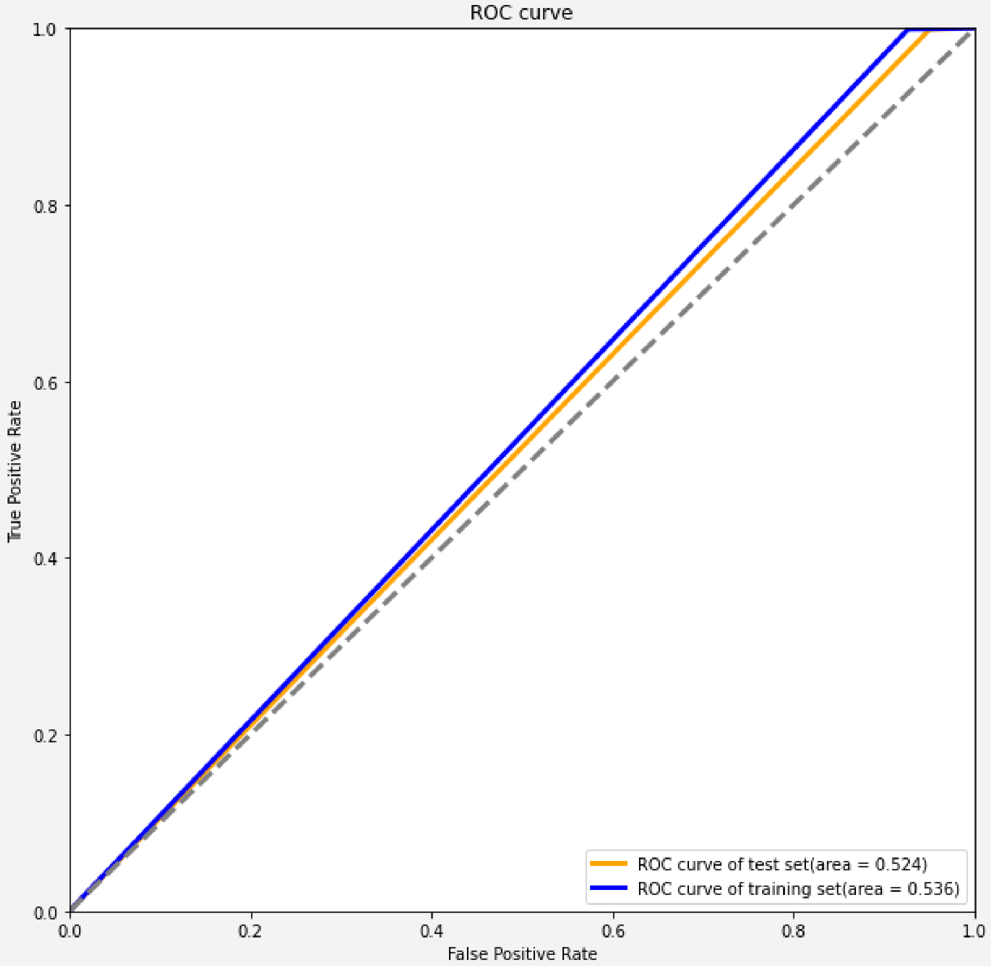
03
模型改进

04
总结展望

03 模型改进

二分类 ROC curve





✓ Neural Networks

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

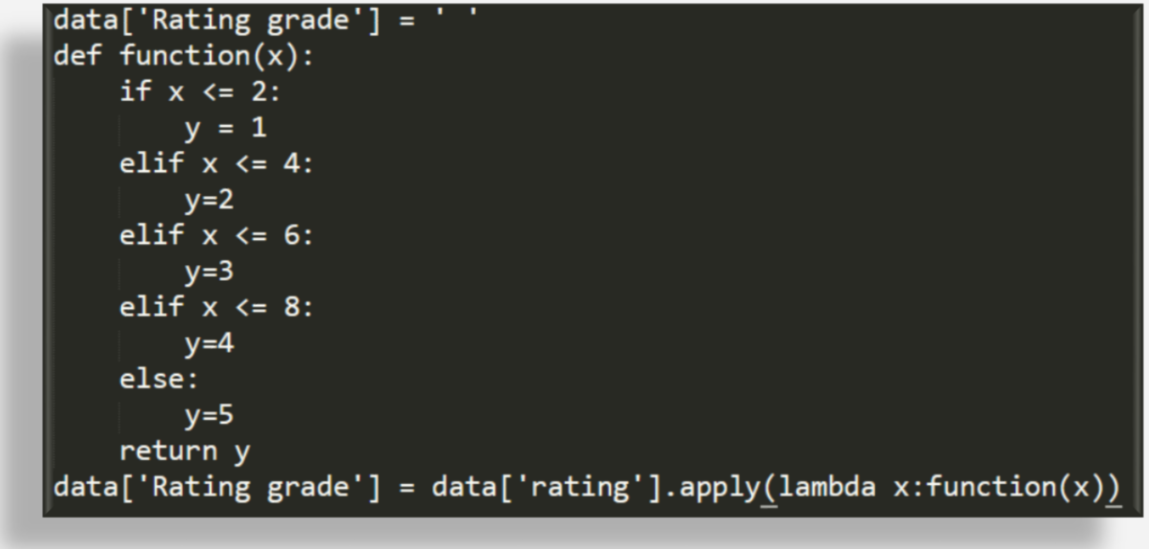
✓ Naïve Bayes

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

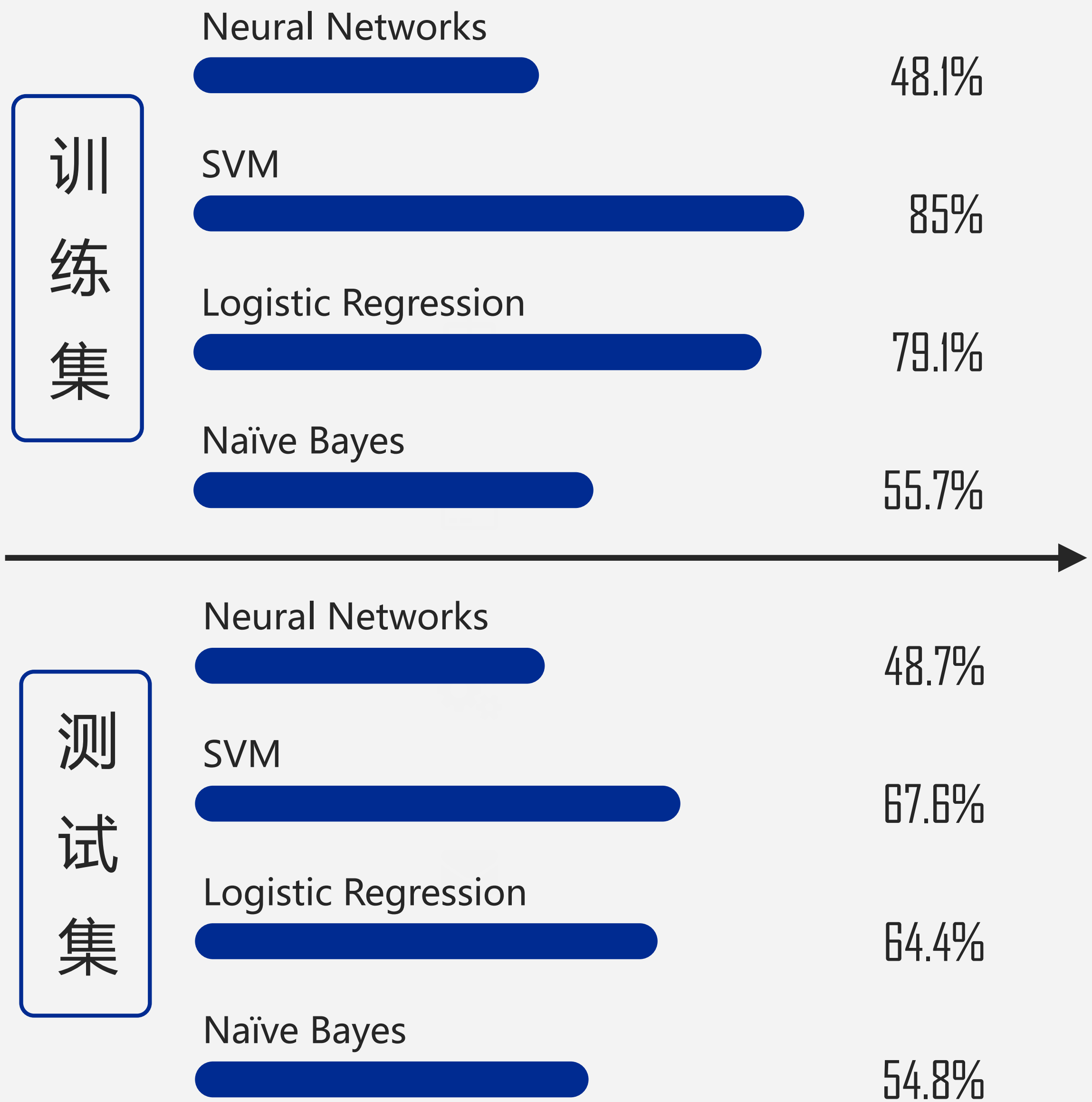
03 模型改进

将二分类转化为五分类

- 重新定义评级规则，进行五星级打分。
- 将原来的1-2分记为1星，3-4分记为2星，5-6分记为3星，7-8分记为4星，9-10分记为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))
```



目录

CONTENTS

01

问题解析
数据预处理

02

模型初探

03

模型改进

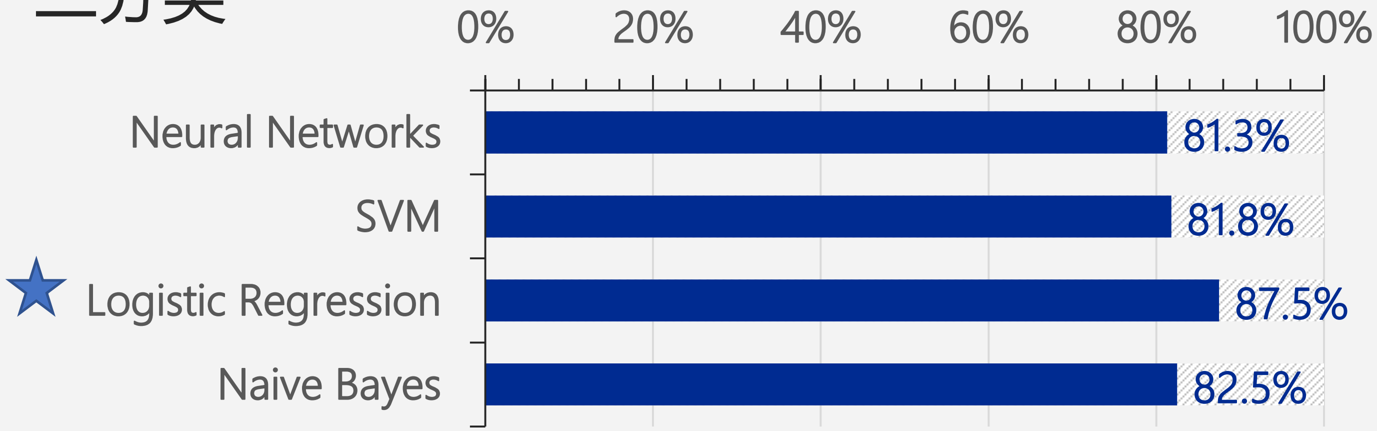
04

总结展望

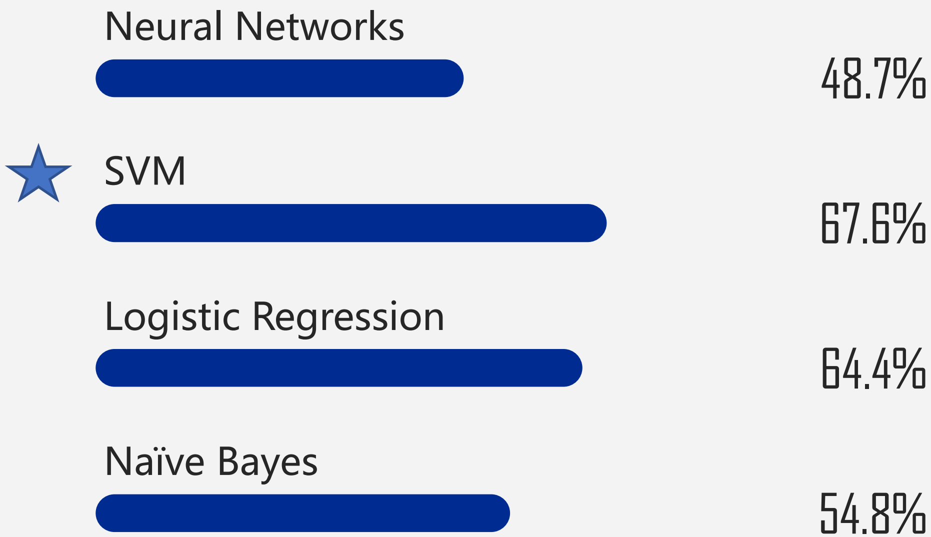
04 总结与展望

- ✓ 原始数据的特征分析
- ✓ 数据预处理
- ✓ 实现四个模型的二分类
- ✓ 绘制ROC曲线
- ✓ 实现四个模型的多分类
- ✓ 不同模型的比较分析

二分类



五分类



04 总结与展望

对数据中的其他特征进行训练，例如建立起患者评价与“点赞数”之间的联系，研究通过使用者的主观评价对网友“点赞数”分类的结果；或者在模型中加入多个特征进行训练，探究增加预测准确率的可能性。

进一步优化现有模型，增加其准确率和稳定性，并以此为基础建立一个完善的药品预测系统，以实现通过使用者的主观评价对药品进行评估，同时也能向医师提供一个临床决策的支持工具，进而针对药物的有效性、安全性等进行研究。另外这也能让保险公司与药厂在制造上有所帮助。

Thank You !