①数据预处理

首先，去除与数据分析无关的字段，如"marketplace","customer\_id","review\_id","product\_id","product\_category" 生产地、顾客id、评论id、商品品类。 去掉数据集中的无关商品评价，如在婴儿奶嘴商品评价数据集中的枕头、吹风机的商品评价。 将数据标记“n”、“y”转换为“0”、“1”，将字符串数据转换成浮点数，便于下一步的统计分析和预测。

将评价标题与正文拼接起来，去除无关的标点符号，将字母统一转换为小写，用spaCy方法将词语转换成它的词根形式，便于后续的文本分析。将评分化为0-1变量。评分4 5为正面，1 2 3为负面。

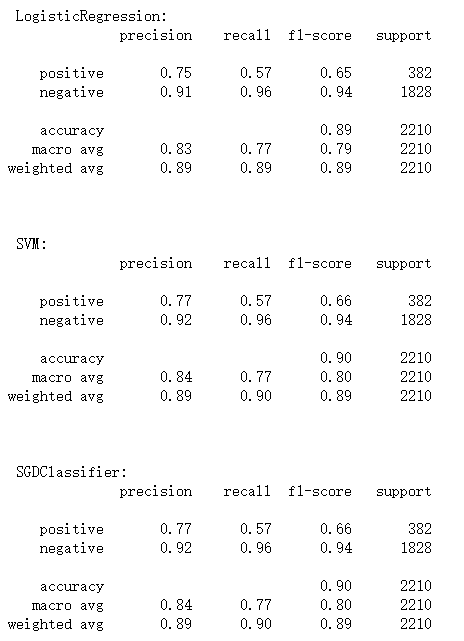
②利用情感分析算法对评论进行量化，提取评论的情感倾向作为情感值。然后我们对评论进行词频分析，找出人们关注的关键词。用Python的NLTK包里面就有训练好的情感量化模型，剔除禁用词后，使用TF-IDF词向量的方法将文本向量化。以正/负面作为标签，分别用 SGD Classifier、线性回归模型、支持向量机、伯努利朴素贝叶斯分类器和多项式朴素贝叶斯分类器进行文本情感分析。用准确率、召回率、F1度量、ROC曲线四种指标评价三种模型的拟合效果，其中SVM模型表现最优.

·准确率指标：

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Classifier | SGD Classifier | Logistic Regression | SVM | Bernoulli NB | Multinomial NB |
| Accuracy | 0.8950226244343892 | 0.8927601809954752 | 0.890497737556561 | 0.8402714932126697 | 0.8307692307692308 |

从上表可以看出， SGD Classifier、线性回归模型、支持向量机的准确率都高达89%。

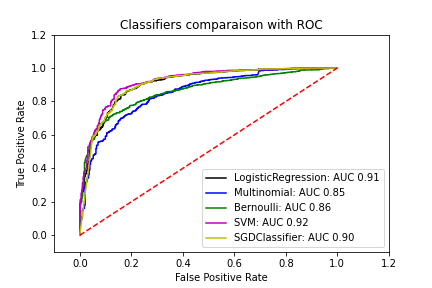
·召回率、F1度量指标：



f1-score.png

由上图可看出SGD Classifier、线性回归模型、支持向量机的召回率几乎相同、而SVM的F1度量更高，说明SVM模型表现更优。

·ROC曲线指标：AUC（Area Under Curve）被定义为ROC曲线下的面积。因为很多时候ROC曲线并不能清晰的说明哪个分类器的效果更好，所以我们使用AUC值作为模型的评价标准：AUC更大的分类器效果更好



ROC\_pacifier\_all\_withSW.png

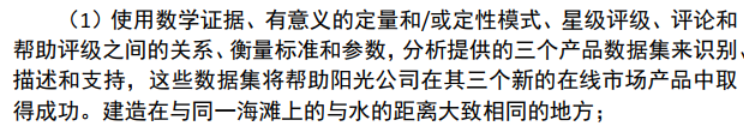
由于SVM模型表现更优

根据SVM模型得出的系数大小和显著性筛选出每一个商品最正面的词和最负面的词，从这些词提取顾客最关注的产品的特性。coef表示系数，系数为正时是正面评价，绝对值越大影响力最大。系数为负时是负面评价，绝对值越大影响力越大。对ratings和量化评论文本reviews进行SVM模型分析，发现人们的情绪与评分水平有很强的相关性。

分析：奶嘴正面：

1. 阅读说明书详细且清楚
2. 奶嘴易用、非常方便。

理一下思路：对于要求（1）



（1）需要找到奶嘴、微波炉、吹风机的特性

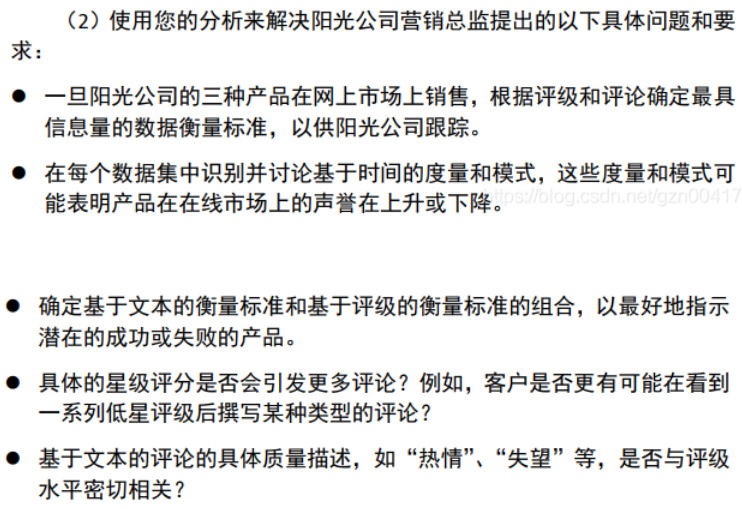
---->建立分类模型【SVM】 【rating>3为正，<=3为负，作为标签】

---->根据SVM得出的系数大小和显著性【即系数最大的词，影响力最大】

---->筛选出每一个商品最正面的词和最负面的词【系数排序，输出前30及后30】

---->分析：这些词可以反映顾客最关注的产品特性(比如目前跑出来的奶嘴有 方便易用 个性化 质量好 设计可爱 可能会卡嘴 浪费材料 heavy hard 等特性)，告诉商家要关注此类特性【写信】

对于要求（2）



①自己建立一个 f(star\_rating，reviews，vine, helpup\_votes, .....)？

②

可放在附录

**奶嘴：**coef\_:各特征的系数（重要性）。

**负面如下（30条）：**

|  | **feature** | **coef** |
| --- | --- | --- |
| **4498** | disappointed | -3.041049 |
| **7337** | hates | -2.259913 |
| **4303** | didn | -2.098272 |
| **4242** | did | -2.069764 |
| **4511** | disappointing | -2.036833 |
| **13159** | poor | -2.013776 |
| **10129** | maybe | -1.988821 |
| **851** | babies mouth | -1.985188 |
| **17287** | thought | -1.962938 |
| **16533** | stuck | -1.933167 |
| **11570** | okay | -1.849089 |
| **18674** | waste | -1.838236 |
| **9770** | loves mam | -1.809987 |
| **4723** | doesn work | -1.792152 |
| **4440** | difficult | -1.772528 |
| **14202** | red | -1.734007 |
| **11887** | opened | -1.720552 |
| **14331** | replaceable | -1.700044 |
| **12852** | pic | -1.695825 |
| **7448** | heavy | -1.689678 |
| **7265** | hard | -1.685590 |
| **9285** | logo | -1.678271 |
| **18192** | useless | -1.670727 |
| **18036** | unhappy | -1.652933 |
| **12827** | personalized | -1.645773 |
| **12895** | pictured | -1.638039 |
| **15875** | specifically | -1.564997 |
| **1036** | barely | -1.563301 |
| **5922** | firm | -1.562670 |
| **11397** | nope | -1.561830 |

**正面如下（30条）：**

| **feature** | **coef** |
| --- | --- |
| **11161** | nicei | 1.059042 |
| **7190** | handy | 1.062561 |
| **8849** | lifesaver | 1.063271 |
| **5871** | finally | 1.073029 |
| **12677** | passy | 1.076505 |
| **7944** | house | 1.081059 |
| **3827** | cute wanted | 1.105039 |
| **12105** | pacifierlove | 1.108847 |
| **19185** | worry | 1.121819 |
| **13803** | quality didn | 1.121975 |
| **6917** | grandson | 1.156259 |
| **15219** | simple | 1.183078 |
| **4066** | definitely | 1.188359 |
| **769** | awesome | 1.192225 |
| **7225** | happy | 1.203141 |
| **7576** | highly | 1.214900 |
| **19050** | wonderful | 1.228934 |
| **12492** | pacifiersi | 1.255714 |
| **6635** | glad | 1.255834 |
| **9449** | lose | 1.259590 |
| **1651** | born | 1.260755 |
| **5118** | easy | 1.298796 |
| **16975** | teething | 1.337997 |
| **15341** | sleep | 1.342016 |
| **314** | allows | 1.357767 |
| **13934** | read description | 1.468950 |
| **17793** | tried | 1.712084 |
| **1163** | best | 2.099175 |
| **9647** | loves | 2.589572 |
| **12732** | perfect | 3.042944 |