**基于情感分析的商品评价与推荐模型**

**Abstract**

With the rapid development of the information age, the Internet and modern information technology are increasingly developed. The network has become the most convenient platform for data sharing, which also marks the arrival of a new era - the era of big data. In the era of big data, everyone is the creator and user of network data. Through big data, we can easily describe everyone's behavior habits and reflect the laws of things. E-commerce is the product of big data era. Founded in 1995 in Seattle, Washington, Amazon is the largest online e-commerce company in the United States. In the online marketplace it created, Amazon provides customers with an opportunity to rate and review purchases. In order to use these data to gain insights into the markets in which they participate, the timing of that participation, and the potential success of product design feature choices，it’s very meaningful to build a model to analyze user comments and ratings.

为了帮助阳光公司提供的hair dryer三个产品在亚马逊商城的销售取得成功，本文利用机器学习中的xgboost模型获得附件所给数据集中的各个指标（）的权重，并加入产品的好评率这一特征量预测上述三个产品的星级。

当阳光公司的三种产品上市时，为便于阳光公司根据评级和评论追踪产品销售情况，本文将为该公司提供一种信息数据测量方法。该方法首先利用？？？对评论进行情感分析，当评级与情感一致时，商品评价就是好的；当评级与评论中包含的情感不一致时，分析其不一致的原因。然后本文利用主成分分析（PCA）确定评级与评论的系数，得到评级的系数为\*\*，评论的系数是\*\*，利用评级\*a+评论\*b得到最终的评价结果。

在产品上市后，需利用数据找到产品的声誉随着时间的变化。因此需要对时间和评级建模，本文选择对评级建模可参考第一部分将评级与评论合二为后一处理；对时间进行分段处理，在处理后建立“时间-评级”模型，把产品的评价转化为数值，描绘评价随时间变化的曲线，对曲线的拐点处进行分析。由此得出产品的声誉变化。

为辨别推出的产品是成功还是失败，本文对评价内容进行文本摘要提取，基于Review公式得出分值，再利用主成分分析计算出摘要情感和分值的权重，进行组合得出产品是成功还是失败。

本文在前面时间模式的基础下，分析一段时间内是否存在较为集中的好评或差评，由此得出客户是否容易被带动情绪，探寻特定的评级是否会引发更多的评论。

本文挖掘文本摘要和评级之间的关系，建立每个星级的词云可视化结果，并进行聚类，找到文本评论的特定质量描述符与评级水平的相关程度。

**Content**

1. **Introduction**
   1. **Problem Formulation**

随着信息技术的持续大爆发，数据量也正在呈爆炸性增长的趋势，如何在大数据时代充分利用数据提取出有用的信息，这是一个非常具有意义的课题。亚马逊公司作为美国最大的电子商务公司，对用户的行为数据十分重视，该公司在其网络购物商城中为消费者提供了评分和评论的机会，本文将利用这些数据建立模型提取信息，为阳光公司新推出的三个产品上市及售后保驾护航。

该题目设立的问题可以分为两个部分，其中第一部分仅包含一个问题，第二部分包涵6个问题，具体问题如下所示：

第一部分

a.分析吹风机、微波炉、剃须刀这三种产品的数据集，定量地测定评定星级、评论和Helpfulness Rating之间的数学关系，帮助阳光公司的三个新的上市产品获取成功。

第二部分

a.提供一种可以根据评级和评论数据提取信息的方法，当推出的三个产品在网络上销售后，阳光公司可以利用这种方法追踪产品销售情况。

b.分别讨论着三个产品的数据基于时间的度量和模式，根据数据随时间的变化，分析商品声誉的变化。

c.确定一种基于文本的度量方法以及基于星级的度量方法组合，探寻产品是成功还是失败。

d.探寻客户是否具有从众心理，客户的情绪是否会随着他人的评价而被调动。

e.判断文本的特征词与评级水平是否密切相关

f.向阳光公司的市场总监写一封信推荐我们团队的测评结果并讲清楚理由。

* 1. **Model Goals**

Sunshine公司的吹风机、微波炉和剃须刀这三种产品即将上市。本文建立模型期望达到的目标是，合理利用数据提取信息，帮助sunshine公司能够根据产品的评论和评级追踪产品的销售和声誉的变化情况，以及判断产品评级和评论对消费者的影响。总体来说，我们团队建模的目的是为sunshine公司制定出正确的销售策略做出贡献。

1. **Assumptions and Justification**
2. **Symbolic Explanation**
3. **Methodology**
4. **Modeling and Analysis**

**5.1** **Data Preprocessing**

The title presents three annexes, which respectively provide data on the three categories of hair dryer, dryer and pacifier: 11470, 1615 and 18939. Also, the three data sets provided contain product user ratings and reviews extracted from the Amazon Customer Reviews Dataset thru Amazon Simple Storage Service(Amazon S3). Meanwhile, in each attachment, each row of data contains information of 15 categories of the commodity, and each category represents the specific meaning as shown in the following table\*.

Table \*:

|  |  |
| --- | --- |
| marketplace (string) | 2 letter country code of the marketplace where the review was written. |
| customer\_id (string) | Random identifier that can be used to aggregate reviews written by a single author. |
| review\_id (string) | The unique ID of the review. |
| product\_id (string) | The unique Product ID the review pertains to. |
| product\_parent (string) | Random identifier that can be used to aggregate reviews for the same product. |
| product\_title (string) | Title of the product. |
| product\_category (string) | The major consumer category for the product. |
| star\_rating (int) | The 1-5 star rating of the review. |
| helpful\_votes (int) | Number of helpful votes. |
| total\_votes (int) | Number of total votes the review received. |
| vine (string) | Customers are invited to become Amazon Vine Voices based on the trust that they have earned in the Amazon community for writing accurate and insightful reviews. Amazon provides Amazon Vine members with free copies of products that have been submitted to the program by vendors. Amazon doesn't influence the opinions of Amazon Vine members, nor do they modify or edit reviews. |
| verified\_purchase (string) | A “Y” indicates Amazon verified that the person writing the review purchased the product at Amazon and didn't receive the product at a deep discount. |
| review\_headline (string) | The title of the review. |
| review\_body (string) | The review text. |
| review\_date (bigint) | The date the review was written. |

For a large number of commodity evaluation information data, in order to facilitate the subsequent establishment of mathematical model to solve the problem, the data is preprocessed according to the following steps:

Step 1: After checking all the data, it is found that only four rows of data have missing values, which has little impact on the overall data volume, so the four rows of data are selected for deletion.

Step 2:After careful observation of the data, it can be found that the values of all sample data in the two categories of market place and product category are the same, indicating that these two categories are irrelevant variables in this study, so they are deleted; at the same time, in the data in the column of product family, except for the two outliers in the pacifier, all other data, as long as the product family ID is the same, The same is true for product family, so only one field is reserved.

Step 3:Because product ID and product title have the same meaning and data redundancy, delete product title.

Step 4:Add two columns of data, year and month, according to the data of review date.

5.2 LightGBM Model

We use TextBlob model which is used in Sentiment Analysis to quantify the reviews. The result of TextBlob Sentiment Analysis is returned as a tuple in the form of(polarity, subjectivity).

…(结果(polarity, subjectivity))

The score of bias is a floating point number with a range of [-1.0, 1.0]. Positive Numbers mean positive and negative Numbers mean negative.Alfred is a floating point number with a range of [0.0, 1.0], where 0.0 is objective and 1.0 is subjective.

We use Light gradient accelerator (Light) Gradient Boosting Machine (LightGBM) which is widely used in data mining tasks to get the weights to build model. We take star\_rating as the output and the other features as the input to train the model.

In order to obtain more data correlation information, we used feature engineering to process the data set, and then put it into the model for training.Feature engineering often plays an important role in the field of data mining. Different feature engineering can obtain different feature sets. A good feature engineering can reveal more relevant information in the data set, thus making the model more accurate.

5.2.1 Model Establishmen

LightGBM是对GBDT的高效实现，主要包括两个本分，一个是GOSS，另一个是EFB。

GOSS算法(算法1)排除了大部分具有小梯度的数据，只使用剩余的数据进行信息增益估计，LightGBM研究[1]表示:具有较大梯度的样本在计算信息增益的时候扮演着更加重要的角色，GOSS可以通过更加小规模的数据来获得非常精准的信息增益计算。

EFB算法(算法2和算法3)通过将互斥的特征捆绑在一起，来减少特征数目。互斥特征意味着它们几乎很少同时出现在非零值，并且LightGBM也表明:找到最优化特征捆绑是NP问题，但是贪心算法能够获得非常好的近似概率

算法1 GOSS

输入:训练数据 I,迭代次数d,大梯度数据采样率a,小梯度数据采样率b,损失函数loss,弱学习器L。

models←{ }，fact← (1 -a)/b

1) topN ← a × len(I) ,randN← b × len(I)

2) for i=1 to d do.

3) preds ← models. predict(1)

4) g ← loss(I ,preds) ,w←{1,1,⋯}

5) sorted ← GetSortedIndices( abs (g) )

6) topSet ← sorted [ 1: topN]

7) randSet ← RandomPick( sorted [ topN:len(I)] ,randN)

8) usedSet ← topSet + randSet

9) w[ randSet] ×= fact //给小梯度数据分配权重参数

10) newModel←L( I[ usedSet], - g[ usedSet] ,w[ usedSet])

11) models. append( newModel)

算法2 Greedy Bundlig。

输入：特征集”,最大冲突计数K;

输出：捆绑集bundles.

构造图G

1) searchOrder ←G.sortByDegree()

2) bundles←{},bundlesConflict←{}

3) for i in searchOrder do

4) needNew ← True ，

5) for j =1 to len (bundles) do

6) cnt ← ConflictCnt( bundles[j] , F[i])

7) if cnt + bundlesConflict[i] ≤ K then

8) bundles[j].add(F[i]),needNew ←False

9) break

10) if needNew then

11) 将F[i]作为新的捆绑bundle加入到bundles中

算法3 Merge Exclusive Features。

输入：数据的数目numData,一束关于互斥特征的捆绑特征F

输出：新的直方图newBin,直方图区间binRanges。

1) binRanges ← |0| , totalBin ← 0

2) for f in F do

3) totalBin += f. numBin

4) binRanges. append(totalBin)

5) newBin←new Bin( numData)

6) for i =1 to numData do

7) newBin[i]←0

8) for j =1 to len(F) do

9) if F[j].bin[i] ≠0 then

10) newBin[i]←F[j]. bin[i] + binRanges[j]

[1] KE G, MENG Q, FINLEY T, et al. LightCBM: a highly efficient gradient boosting decision tree [C]// Proceedings of the 2017 Annual Conference on Neural Information Processing Systems. New York: Curran Associates Inc., 2017: 3146 -3154.

特征工程

将通过分析数据集中的各个特征在LightGBM中重要性表现来进行特征工程操作.首先，将预处理后的数据集放入到LightGBM训练的结果如表1，表2，表3

|  |  |  |  |
| --- | --- | --- | --- |
| 特征名称 | hair\_dryer权值 | microwave权值 | pacifier权值 |
| customer\_id | 0 | 0 | 0 |
| review\_id | 0 | 0 | 0 |
| product\_id | 0.0365 | 0 | 0.0095 |
| helpful\_votes | 0.0412 | 0.0389 | 0.0444 |
| total\_votes | 0.1380 | 0.1229 | 0.1396 |
| vine | 0 | 0 | 0.0047 |
| verified\_purchase | 0.0285 | 0.0286 | 0.0253 |
| review\_body | 0.2238 | 0.3299 | 0.2857 |
| review\_date | 0.2 | 0.2131 | 0.2015 |
| year | 0.0031 | 0 | 0 |
| month | 0.0539 | 0.0819 | 0.0555 |
| rate | 0.2746 | 0.1844 | 0.2333 |

表1.特征值权值

从三类产品在LightGBM模型中学习情况可以看到，有部分的特征权值几乎为0，这些对于模型来说，可能是多余的,综合三种产品情况，删除customer\_id，review\_id，product\_id，year，vine。再次使用模型进行训练，得到表4，表5，表6

|  |  |
| --- | --- |
| 特征名称 | 权值 |
| helpful\_votes | 0.0365 |
| total\_votes | 0.1412 |
| verified\_purchase | 0.0253 |
| review\_body | 0.2253 |
| review\_date | 0.2063 |
| month | 0.0523 |
| rate | 0.3126 |

表4. hair\_dryer的处理后特征值权值

|  |  |
| --- | --- |
| 特征名称 | 权值 |
| helpful\_votes | 0.0389 |
| total\_votes | 0.1229 |
| verified\_purchase | 0.0286 |
| review\_body | 0.3299 |
| review\_date | 0.2131 |
| month | 0.0819 |
| rate | 0.1844 |

表5. microwave的处理后特征值权值

|  |  |
| --- | --- |
| 特征名称 | 权值 |
| helpful\_votes | 0.0428 |
| total\_votes | 0.1428 |
| verified\_purchase | 0.0301 |
| review\_body | 0.2936 |
| review\_date | 0.2 |
| month | 0.0571 |
| rate | 0.2333 |

表6. pacifie的处理后特征值权值

删除后,模型的特征权值变化不大，并且F1值提升不明显。然后我们观察数据发现，有些评论存在一定错误性，如给好评，却打了一星，给差评，却给了五星。

|  |  |  |
| --- | --- | --- |
| review\_id | Start\_rating | review\_body |
| R134FUK2D9TQU6 | 1 | I have used the dryer several times and it works great. I had questions which were answered promptly by other customers which was helpful in making my decision. Definitely recommend. |
| R1HI3QGXJQ2RUT | 5 | We owned these from the store and they are exactly the same. Too bad my grandson decided he was done with pacifiers one week later |
| ⋯ | ⋯ | ⋯ |

我们删除完那些”一星好评”和“五星差评”后，三类产品F1值分别从0.8766,0.8044,0.9032提升至0.8921,0.8369,0.9351.

5.3（a）

一旦阳光公司产品销售, 利用第一题的模型进行预测用户评级. 若预测评级与用户实际评级相符且评级较好, 则通过第一问得出的重要性来分析产品的成功之处, 将成功的方面做到底, 保证产品口碑不下滑. 若评级较差, 同样根据重要性来分析产品在哪些方面做的不够, 接下来应该着重提升产品劣势方面. 实现逆转. 若与预测评级与用户实际评级不符, 则应分析这些用户是否为虚假用户或恶意用户, 从而有针对性的处理该类用户, 保证正常的销售市场.

5.4（b）

针对提供的三个数据集, 本文利用之前的情感分析结果, 结合评级数据, 利用信息熵确定二者权重, 最后得到每个评价的得分, 利用该得分来分析时间模式.

信息熵有三个性质:

1. 单调性, 发生概率越高的事件, 其携带的信息量越低;
2. 非负性: 信息熵可以看做一种广度量, 非负性是一种合理的必然;
3. 累加性: 即多随机事件同时发生存在的总不确定性的量度是可以表示为各事件不确定性的量度的和，这也是广度量的一种体现。

假设用户给商品评级和评价是两个事件和, 期望他们独立, 则有他们同时发生的概率为



由累加性可知:



满足两个变量乘积函数值等于两个变量函数值的和, 应该为对数函数, 再考虑到再考虑到概率都是小于等于1的，取对数之后小于0，考虑到信息熵的第二条性质，所以需要在前边加上负号。最后参考信息论之父克劳德·香农给出的信息熵定义有信息熵公式为:



将情感分析后的数据作为评价数据, 将star\_rating作为评级数据, 最终计算产品综合得分score如下公式



emotionScore为对review\_body情感分析的分值, starRating为用户的评级. a和b是利用信息熵公式计算二者的信息熵权值, 对三个数据集分别计算权值得出表xxx:

|  |  |  |
| --- | --- | --- |
|  | a | b |
| hair\_dryer | 0.1381 | 0.8619 |
| microwave | 0.0693 | 0.9307 |
| pacifier | 0.169 | 0.831 |

计算每个产品的score来分析每个数据集的基于时间的度量模式. 本文通过按月份统计每个月的整体销量, 好评率, 差评率等等数据进行可视化分析. 其中单月份整体销量, 好评率, 差评率计算公式分别如下:





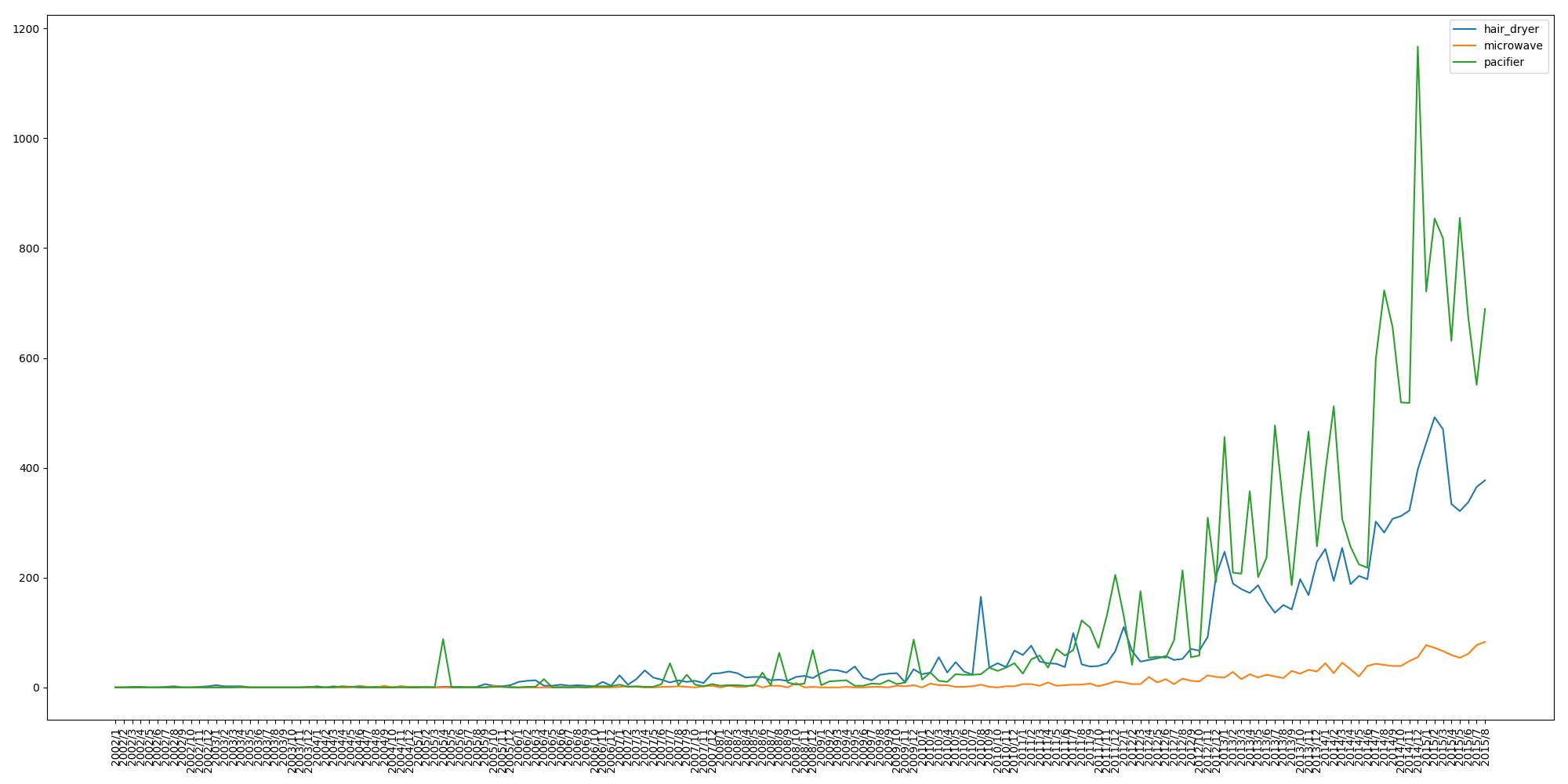
.



其中代表一行数据,即数据中的review\_date, 为真的每条数据利用公式计算的得分.

1. 整体销量分析和数据选取

利用预处理的数据, 按月份统计三个产品的销量图如下xxx:

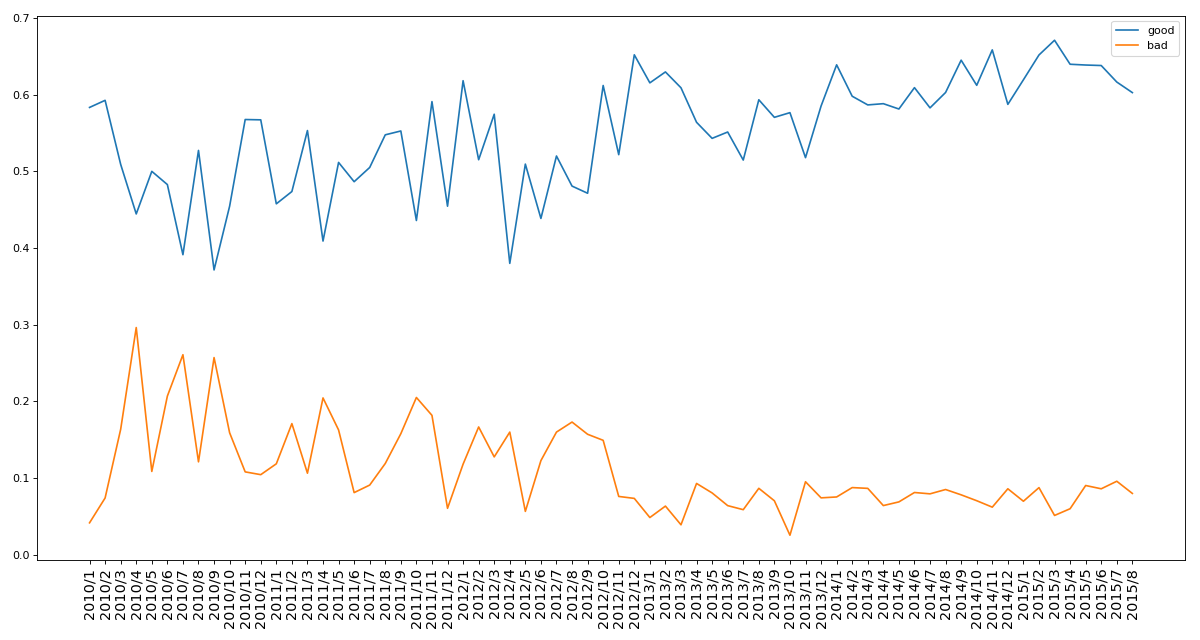


由于篇幅限制, 图像模糊, 清晰版请见附录, 通过图像可以得知2010年以前的销量非常少, 故下面分析时只取2010年1月以后的数据进行分析.

同时, 可以看出三个产品的总体销量在逐年上升, 微波炉销量相对较低, 尿不湿销量较高. 原因可能是微波炉属于耐用性产品, 而尿不湿则是一次性非耐用型产品, 吹风机销量居中. 这跟产品的特性和用途相关.

1. hair\_dryer时间模式分析

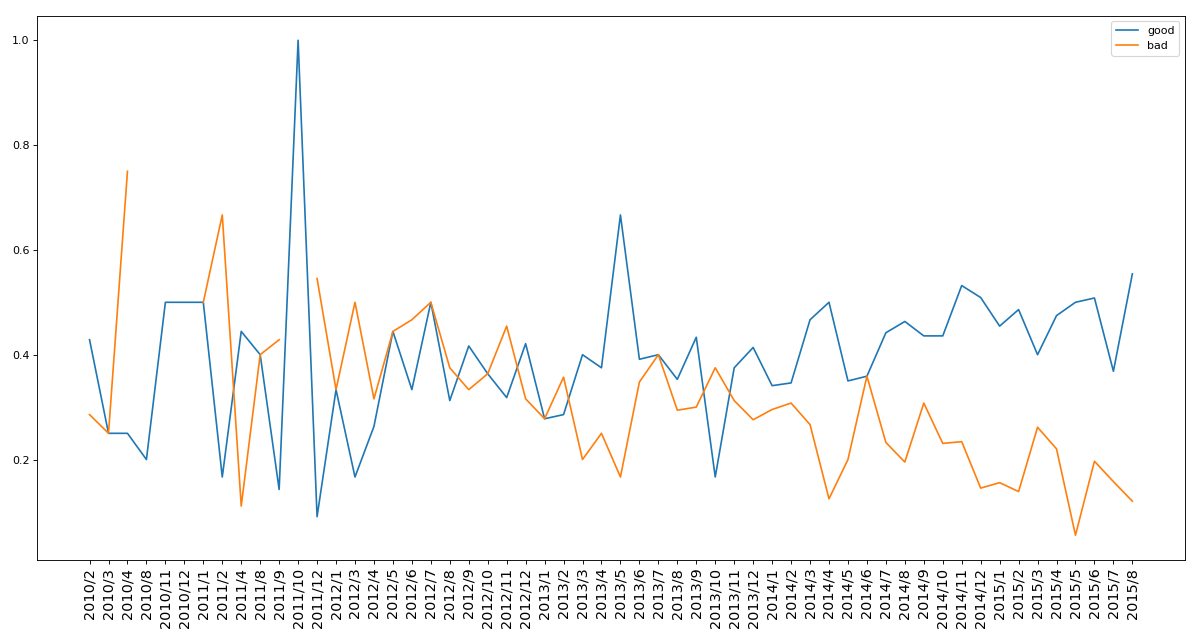
利用预处理之后的数据, 得到hair\_dryer的月份好评率和差评率如下图xxx:



从图中我们可以分析出从2010年1月以后, 吹风机的好评率总体上升, 差评率总体下降. 并且好评率始终远大于差评率. 并且在2013年1月份之前, 整个产品的差评率较高, 产品质量不客观, 但是2013年以后, 产品的好评率基本居高不下,说明吹风机在经历市场洗礼后逐渐质量口碑等逐渐上升, 人们使用起来更顺心. 但是仍会出现较轻微的波动.

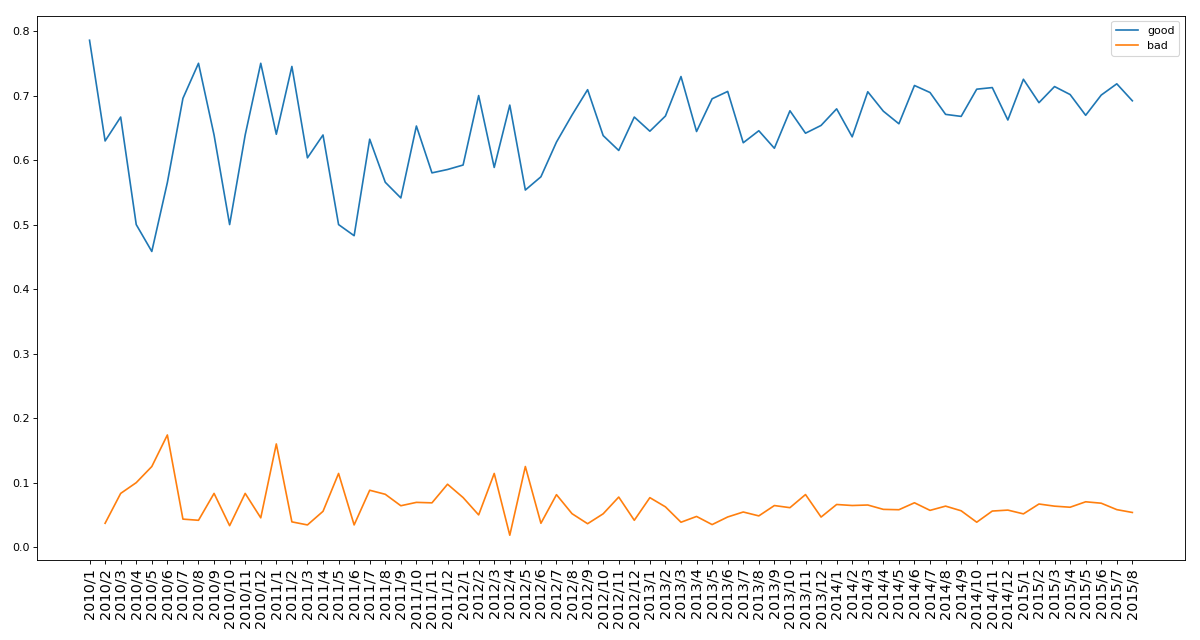
1. microwave时间模式分析

利用预处理之后的数据, 得到microwave的月份好评率和差评率如下图xxx:



1. pacifier时间模式分析

利用预处理之后的数据, 得到pacifier的月份好评率和差评率如下图xxx:



5.4（c）

要求确定一个基于文本的度量方法和基于评级的度量方法的组合, 来指示一个产品的成功或失败. 本文在第一题中建立了lgb模型并得出了相应指标对产品好坏的影响. 同时在上一小问中, 借助熵权法利用评级和评论的情感分值得到了产品的score, 上面两个解决方案都参考了评论和评级, 也即文本和评级, 故此处结合两个解决方案来最好的指示潜在的成功或失败的产品.

结合问题一的特征权重, 进一步筛选出有关文本的和有关评级的特征: review\_body, review\_date, rate, star\_rating. 同时结合上一小问的score. 建立模糊综合评价模型对产品的潜在成功与否进行评价. 模糊综合评价模型的基本步骤如下:

1. 选取特征
2. 确定评价矩阵
3. 检验评价矩阵, 计算权重
4. 计算得分, 得出结果

第一步选取特征已经完成. 接下来根据相关资料和数据确定评价矩阵A如下:



.通过对举着A进行一致性检验得出CR= 0.018<0.1 故矩阵通过一致性检验. 最后计算矩阵的特征向量并归一化后得到相应特征的权重如下:

|  |  |
| --- | --- |
| 特征 | 权重 |
| review\_body | 0.158 |
| review\_date | 0.098 |
| rate | 0.259 |
| star\_rating | 0.170 |
| score | 0.315 |

公司可以利用该权重值来评价产品的成功与否, 数值越大, 则产品越成功

5.5（d）

5.6（e）

第二问的e小问:

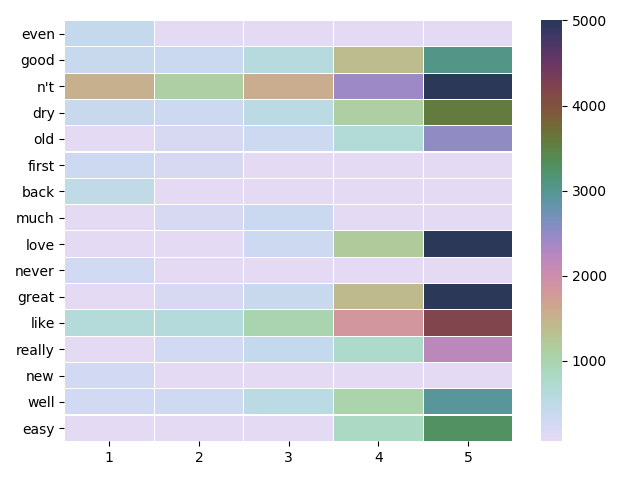
题目要求根据文本的评论的特定质量描述符, 和评级相关联, 挖掘其中是否有深层次信息. 为此本文通过对review\_body的具体内容进行挖掘, 提取其相关描述信息, 并通过统计和绘图等方式进行了挖掘.

具体步骤如下:

1. 建立特定的质量描述符词典. 通过查阅资料和相关数据, 本文建立了用户情感-评论-程度词典, 词典中包含英文的各种描述情感、评论词汇、程度词汇等等, 具体词汇表见附录.
2. 将吹风机, 微波炉, 尿不湿数据合并, 同时提取review\_body内容. 按用户评级分类, 分别从review\_body中统计出现在词典中的词语, 通过词频统计方法得到各个词语出现的频次, 将结果按频次降序排列后取出前20个词语. 得到如下表格xxx:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | 1 | 2 | 3 | 4 | 5 |
| old | 272 | 225 | 353 | 677 | 2512 |
| never | 298 | 0 | 0 | 0 | 0 |
| long | 241 | 186 | 327 | 563 | 1392 |
| new | 286 | 0 | 0 | 0 | 0 |
| good | 392 | 368 | 591 | 1402 | 3058 |
| much | 228 | 234 | 370 | 606 | 2079 |
| back | 460 | 212 | 0 | 0 | 0 |
| bad | 222 | 0 | 0 | 0 | 0 |
| high | 0 | 0 | 232 | 0 | 0 |
| however | 0 | 0 | 265 | 0 | 0 |
| need | 230 | 157 | 252 | 566 | 1645 |
| receive | 222 | 0 | 0 | 0 | 0 |
| keep | 0 | 196 | 261 | 479 | 1695 |
| cute | 0 | 0 | 0 | 0 | 1479 |
| perfect | 0 | 0 | 0 | 0 | 1600 |
| star | 0 | 0 | 0 | 440 | 0 |
| hot | 0 | 187 | 0 | 0 | 0 |
| easy | 0 | 0 | 0 | 832 | 3261 |
| dry | 386 | 326 | 535 | 1108 | 3578 |
| n't | 1509 | 1102 | 1594 | 2445 | 5932 |
| really | 237 | 268 | 417 | 784 | 2209 |
| want | 255 | 157 | 274 | 437 | 0 |
| love | 0 | 159 | 359 | 1184 | 8867 |
| great | 221 | 237 | 388 | 1405 | 5797 |
| nice | 0 | 0 | 0 | 558 | 1405 |
| small | 0 | 165 | 279 | 549 | 0 |
| like | 642 | 602 | 984 | 1851 | 4186 |
| best | 0 | 0 | 0 | 0 | 1448 |
| even | 401 | 208 | 233 | 0 | 0 |
| first | 355 | 223 | 0 | 0 | 1363 |
| still | 228 | 0 | 255 | 506 | 0 |
| set | 0 | 163 | 272 | 607 | 1378 |
| well | 285 | 302 | 531 | 1019 | 2932 |

利用表格数据绘制热力图(图xxx)进行直观分析.



图中横坐标为星级, 纵坐标为单词, 从图和表中可以看出, 评分为一星的评价出现的单词最多的是” n’t ”是个否定词, 出现了1509次, 接下来的是never, like等; 而评分为5星的出现最多的单词为love(8867次), 其次为great, like等等, 该单词在一星或二星中几乎未出现. 这表明特定的描述词汇与评级是密切相关联的. 同时, 我们也注意到了, “ n’t ”在各个星级中都出现了多次. 通过具体的review\_body数据分析, 在一星级中其多与其他积极词汇同时出现用以表达否定情感, 例如” n’t like ” 等等. 而在五星数据中其常形成许多积极情感的短语, 例如” like n’t allow finger print”等等.

通过上述分析, 我们总结出如下的星级常用词汇表xxx:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 1 | 2 | 3 | 4 | 5 |
| not like | not good | not like | like | love |
| never | not like | not good | well | great |
| back | old | much | great | easy |
| even | not well | old | not | like |
| bad | dry | however | need | well |

5.7（f）

6.Conclusion

7.Evaluating the Model

7.1Sensitivity analysis

8.Reference