东莞理工学院

粤台产业科技学院

本 科 期末 设 计

（2022 届）

毕业设计题目：亚马逊在线平台评论分析

指导教师姓名及职称：詹家榜 副教授

学生姓名：高胜良、陈科润、李灿勤

学　　号：201843302107、201843302104、201843302115

系　　别：计算机科学与技术系（跨境电商方向）

专业班级：跨境电商 1 班

起止时间：2021 年 3 月——2021 年 7 月

目录

亚马逊在线平台评论分析

摘　要： 在亚马逊的在线平台中，亚马逊为客户提供了对各类产品的购买进行评分和评价的机会，客户可以根据这些数据来协助自己的购买决策。同样公司也可以根据这些数据，来改变自身产品销售策略或是改善自身产品的功能。本研究通过对亚马逊平台上微波炉、吹风机、奶嘴相关数据的清洗与分析来帮助公司制定在亚马逊在线平台上销售这三种新产品的策略以及确定这三种产品中潜在的重要功能来达成提高产品销售率的目标。

本研究综合了 star_rating、review_body、review_headline、vine 和 purchase 数据，运用 TF-IDF 模型、朴素贝叶斯算法、逻辑回归分类、层次分析法、相关性分析和多项式拟合完成了对评论的定量并得到了一个模型，通过该模型本研究能得到一条评论对产品更客观的评分。并且根据从该模型得到的评分，本研究识别了评分与时间的关系模式，判断在某一时间段产品的声誉是否正在上升或下降。同时本研究将每个低星级的评论和处于该评论前、后一定数量的评论的星级提取出来，通过观察其前、后评论星级的相关性来探索低星级评论是否会对后面的评论产生影响或引起更多评论。

本研究的分析始于准确亚马逊销售产品的评分和评价，本研究模型和算法也相对应准确，所以本研究的结论将在此提出更加具有可信度。

关键词： 1.词频统计 2. TF-IDF 模型 3. 层次分析法 4. 相关性分析 5. 多项式拟合

一、引言

（一）、背景分析

产品所得得到最终评分是一种综合性的考量，它不仅仅是简单的客户购物软件上给的 star_rating 那么简单，还要综合分析客户说给的评论与客户类型因素。各类产品处于当今时代如果不进行改进，不具备相对应创新极可能会失败。本研究说研究正是基于产品满意度的分析与建模，并基于模型，帮助企业得到产品改进方案与决策，使产品增加成功率。

（二）、术语与定义

predictions_stars：根据本研究的 TF-IDF 模型将 review_headline 与 review_body 转化为星级再取平均的数据

score：根据最终模型得出的各成分权重计算得到的星级的数据

final_score：根据 score 分数定为（0-5）区间数据

vine：亚马逊的评价员，可认为他的评价更有价值

score_sum：一个时间节点内所有用客观评分模型求得的评分的均值

score_average：一个时间节点（一天）内，所有评论的 score 的均值

result_average：截止于一个时间节点内所有 final_score 的均值

average_before：当出现客户评分为一星时该条评论之前的 20 个 star_rating 的均值

average_after：当出现客户评分为一星时该条评论之后的 20 个 star_rating 的均值

分类器:常规任务是利用给定的类别、已知的训练数据来学习分类规则和分类器，然后对未知数据进行分类（或预测）

代价函数（cost function）:将随机事件或其有关随机变量的取值映射为非负实数以表示该随机事件的“风险”或“损失”的函数。

梯度下降法（Gradient descent）:找到一个函数的局部极小值，必须向函数上当前点对应梯度（或者是近似梯度）的反方向的规定步长距离点进行迭代搜索。

（三）、研究架构

本研究主要工作就是根据亚马逊提供的三种产品的评分与评价数据，进行预处理，词频统计，客户对商品的情感偏好隐藏在评论的字里行间，近年来出现了很多方法来挖掘客户隐藏在评论文本中对商品的情感信息，商品评论可以补充传统推荐系统中仅仅依靠打分数据进行推荐的不足。再基于词频统计本研究利用TF-IDF 模型得到 predictions_stars，之后使用层次分析法对各项影响评分客观性的因素赋予权重来得出本研究的模型，统计整理 predictions_stars，利用时间维度看 final_score 的变化，从而分析出对应产品的声誉上升与降低，并挖掘声誉降低与升高内在原因，为企业提供参考与建议。

二、算法说明

（一）、TF-IDF 模型

首先本研究先将购买信息的 review_headline 和 review_body 进行 TF-IDF 语言分析，下面本研究将解释 TF-IDF 模型的定义与实现

在一份给定的文件里，词频（term frequency，TF）指的是某一个给定的词语在该文件中出现的频率。这个数字是对词数(term count)的归一化，以防止它偏向长的文件。（无论一个词语是否重要，同一个词语在长文件里可能会比短文件有更高的词数）对于在某一特定文件里的词语 t_i 来说，它的重要性可表示为：

$$\mathrm{tf_{i,j}} = \frac{n_{i,j}}{\sum_k n_{k,j}}$$

以上式子中 $n_{i,j}$ 是该词在文件 d_j 中的出现次数，而分母则是在文件 d_j 中所有字词的出现次数之和。

逆向文件频率（inverse document frequency，IDF）是一个词语普遍重要性的度量。某一特定词语的 IDF，可以由总文件数目除以包含该词语之文件的数目，再将得到的商取对数得到：

$$\mathrm{idf_i} = \log \frac{|D|}{|\{j : t_i \in d_j\}|}$$

其中

|D|：语料库中的文件总数

$|\{j : t_i \in d_j\}|$：包含词语 t_i 的文件数目（即$n_{i,j} \neq 0$的文件数目）如果该词语不在语料库中，就会导致被除数为零，因此一般情况下使用 $1 + |\{j : t_i \in d_j\}|$

然后

$$tfidf_{i,j} = tf_{i,j} \times idf_i$$

某一特定文件内的高词语频率，以及该词语在整个文件集合中的低文件频率，可以产生出高权重的 TF-IDF。因此，TF-IDF 倾向于过滤掉常见的词语，保留重要的词语

从而本研究用代码实现 TF-IDF 进行情感词的提取。

提取好后本研究进行分类构建分类器算法，对 TF-IDF 模型处理后的文本进行机器学习和数据挖掘。本研究使用的方法有两种第一种方法是朴素贝叶斯分类，第二种是逻辑回归分类器首先本研究先将两种方法的原理及公式先展现，再将结果得出结果展现。

（二）、朴素贝叶斯算法[1]

输入:输入样本数据 Y= (x1, y1) (x2, y2) … (xi, yi)，其中 Y 为文本

$$Y_n(x) = sign(\sum_{i=1}^{M} \alpha_i y_i(x))$$

Step1:将每个样本的权值设定为: $w_i = \frac{1}{N}$, i=1, 2, 3, …, N;

Step2:for t=1, …, M, 对于朴素贝 Y_c，计算其误差: $\varepsilon_t = \sum_{i=1}^{N} w_i^{(t)} (h_c(x_i) \neq y_i)$，经过朴素贝叶斯模型得到的

分数以误差值的方式得到, $\alpha_t = \ln\{\frac{1-\varepsilon_t}{\varepsilon_t}\}$

对每个样本权值更新: $w_{t+1,i} = \frac{w_{t,i}}{z_t}\begin{cases} e^{-\alpha t} & if h_t(x_i) = y_i \\ e^{\alpha t} & if h_t(x_i) \neq y_i \end{cases}$,将系数归一化得到 Z_t, $\sum_{i=1}^{N} w_i = 1;$

$$Y_n(x) = sign\left(\sum_{i=1}^{M} \alpha_i y_i(x)\right)$$

Step3:最后得到 M 个朴素贝叶斯分类器:

（三）、逻辑回归分类

（1）寻找预测函数

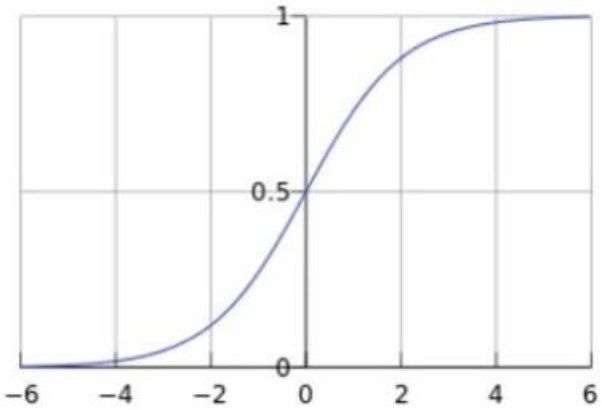
$$z = w^T x + b$$

假设有一个二分类问题，输出为 y∈{0,1}，而线性回归模型产生的预测值为（w 是参数向量）

使用一个理想的函数来帮本研究实现 z 值到 0/1 值的转化。于是到了 sigmoid 函数

$$\phi(z) = \frac{1}{1+e^{-z}}$$

图像是



把 sigmoid 计算得到的值作为预测点为类别 1 的概率。概率大于 0.5 归类为

1，小于 0.5 归类为 0。至此，得到了预测函数模型

$$\hat{y} = \begin{cases} 1 & if\ \phi(z) \geq 0.5 \\ 0 & otherwise \end{cases}$$

这里面，ϕ(z)表示的是类别取 1 的概率大小，那类别取 0 的概率大小 j 就是 1-ϕ(z)

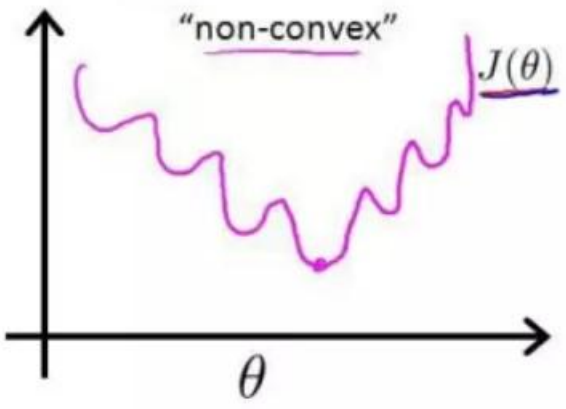
（2）构造代价函数

$$J(w) = \sum_i \frac{1}{2}(\phi(z^{(i)}) - y^{(i)})^2$$

按照线性回归的思路，如果利用误差平方和来当代价函数，得到

但是，此时的预测函数模型ϕ并不是线性的，如果把ϕ带入代价函数，得到的 J 是类似于下图的非凸函数，它有很多极值，使用梯度下降会很难找到代价函数最小的情况，所以这样构建代价函数并不合适。

所以最大似然估计可以解决这个问题。根据预测函数构造一个它的分布的概



率密度，利用已知的样本反推参数。

根据上一步得到的预测函数，可以知道概率如下：

$$P(y=1|x,w)=\varphi(z)$$

$$P(y=0|x,w)=1-\varphi(z)$$

将这 2 个式子合并，得到概率公式:

$$p(y|x;w)=\phi(z)^{y}(1-\phi(z))^{(1-y)}$$

$$L(w)=\prod_{i=1}^{n}p(y^{(i)}|x^{(i)};w)=\prod_{i=1}^{n}(\phi(z^{(i)}))^{y^{(i)}}(1-\phi(z^{(i)}))^{1-y^{(i)}}$$

由最大似然估计可知，联合概率就是：

最大似然估计的目标是找到参数 W 使得 L(w)最大，那么对 L 加一个负号，就可以得到代价函数，也就是找到参数 W 使得 -L(w)最小。为了简化计算，对 L(w)取对数

$$l(w)=lnL(w)=\sum_{i=1}^{n}y^{(i)}ln(\phi(z^{(i)}))+(1-y^{(i)})ln(1-\phi(z^{(i)}))$$

最后得到代价函数

$$J(w)=-l(w)=-\sum_{i=1}^{n}y^{(i)}ln(\phi(z^{(i)}))+(1-y^{(i)})ln(1-\phi(z^{(i)}))$$

三、对数据的量化与挖掘

先利用 Python 对 review_headline 与 review_body 进行词频分析，得到 3 个频率最高的词语，如下图所示。

	review_headline	review_body
count	11468	11470
unique	7696	11197
top	Five Stars	good
freq	1284	20

1.1
Hair dryer

	review_headline	review_body
count	1615	1615
unique	1346	1601
top	Five Stars	None available.
freq	149	4

1.2
microwave

	review_headline	review_body
count	18939	18937
unique	12611	18085
top	Five Stars	good
freq	1875	72

1.3
pacifier

可观察出大部分人在 review_headline 中给出 Five Stars，但根据 review_body 本研究能得出客户对产品更为详细的评价，如在对微波炉的评论中就有大部分客户给出”none available”,这明显是对产品感到不满所以本研究

在对产品进行评论的预测评分时要结合评论标题与评论主体。

在对评论进行词频分析之后，本研究考虑到没有购买产品的客户存在随意好评或者恶意差评的行为，所以本研究将顾客分为已购买过产品的客户和未购买过产品的客户。

本研究首先探究已购买过产品的客户和未购买过产品的客户对产品的评分情况：

star_rating	sum	Y	N
1	1032	739	293
2	639	509	130
3	999	859	140
4	2096	1804	292
5	6704	5900	804

2.1
Hair dryer

star_rating	sum	Y	N
1	402	116	285
2	112	56	56
3	134	92	42
4	300	246	51
5	667	579	85

2.2
microwave

star_rating	sum	Y	N
1	1192	468	117
2	945	440	65
3	1426	678	98
4	2716	1255	191
5	12660	6950	936

2.3
pacifier

以上对不同客户类型单独的对产品的打分统计可看出，两者对产品的评分具有差异性，本研究认为购买者对产品更加了解，所以赋予相对高的权重给购买者的评分及评论。

所以确定客户对产品的满意程度的影响因子有：star_rating、verified_purchase、vine、review_body 。

四、TF-IDF 求解参数

可以使用梯度下降法求解参数不再详细说明

通过代码的构建和运行本研究得到了使用多项式朴素贝叶斯分类器，验证客户给出的打分星级的预测准确率：0.5993723849372385（Hair dryer）

0.5717821782178217 （pacifier）

0.6237623762376238 （microwave）

使用网格搜索，找到最优超参数组合对应的逻辑回归模型，验证客户给出的打分星级的预测准确率：

0.6705020920502092（Hair dryer）

0.6237623762376238(pacifier)

0.5717821782178217(microwave)

五、层次分析法建立模型

之后,本研究用TF-IDF模型来为review_body定量,得到了根据review_body所预测的评分，下图为一部分所得的预测评分。

1.1Hair dryer 的 predictions_stars

predictions_stars
1
5
4
3
5
5
1
4
5
2
5
5
5
3
1
5
1
5
4
4
5
5
5
5
3
5
5

1.2microwave 的 prediction stars

predictions_stars
5
5
5
5
4
5
1
3
5
5
5
5
3
4
5
5
5
4
5
5
1
5
5
5
5
2

1.3pacifier 的 prediction stars

predictions_stars
5
5
5
5
5
5
5
4
5
5
5
5
4
5
5
5
4
5
5
1
5
5
5
5
1
5
3
5

为了得到各个权重，本研究将用到层次分析法。本研究根据所参考的文献与经验以及下图的标准主观地完成各个判断矩阵。

Judement value	Meaning
1	Equal Importance
3	Moderate Importance
5	Strong Importance
7	Very Importance
9	Extremely Important
2、4、6、8	The median of two judgment values
Reciprocal	The ratio of A to B is 3, Then the ratio of B to A is 1/3

PS:

1.The meaning of a_{ij} is that compared with the index j, the importance of i

2. When i = j, the two indexes are the same, so it is equally important to record it as 1, which explains that the main diagonal is 1.

3.a_{ij}> 0 and satisfy $a_{ij} * a_{ji} = 1$ (we call this matrix satisfy the positive and negative inverse matrix)

stars_ratings 和 product_review 的判断矩阵（满意度）

score	product_review	stars_ratings	result_weight
product_review	1.00		
stars_ratings		1.00	

score	product_review	stars_ratings	result_weight
product_review	1.00	3.00	0.75
stars_ratings	0.33	1.00	0.25

前面内容中本研究提到，客户可以分为 4 种类型：

1. 既是评论员又购买了产品
2. 是评论员但没有购买产品
3. 不是评论员但购买了产品
4. 既不是评论员又没有购买产品

本研究认为这四种不同客户所给出的评分与评价的参考价值不同，所以本研究依然使用层次分析法去确定四种客户类型的权重。经过统计，本研究发现既是评论员又购买了产品的客户几乎没有，所以不将此类型的客户计入统计量。

三种不同类型客户的判断矩阵（参考价值）

product_review	vine but no purchase	no vine but purchase	no vine or purchase	result_weight
vine but no purchase	1.0000			
no vine but purchase		1.0000		
no vine or purchase			1.0000	

本研究用同样的方法完成判断矩阵，并得到三种不同类型客户的权重及满意度与参考价值的权重，如下图所示：

product_review	vine but no purchase	no vine but purchase	no vine or purchase	result_weight
vine but no purchase	1.0000	0.3333	3.0000	0.2426
no vine but purchase	3.0000	1.0000	7.0000	0.6694
no vine or purchase	0.3333	0.1429	1.0000	0.0880

score	satisfaction	Reference value	result_weight
satisfaction	1.00	3.00	0.75
Reference value	0.33	1.00	0.25

本研究建立的客观评分模型为：

$$Y = x*(A*a + B*b) + 5*z*c$$

（满意度权重：x 参考价值权重：z 评分：A 评分权重：a 评论：B 评论权重：b

评论者类型权重：c）

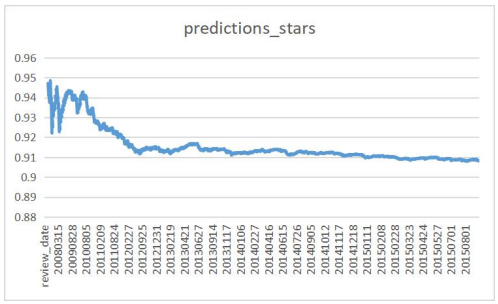
六、探究客观评分模型相关系数

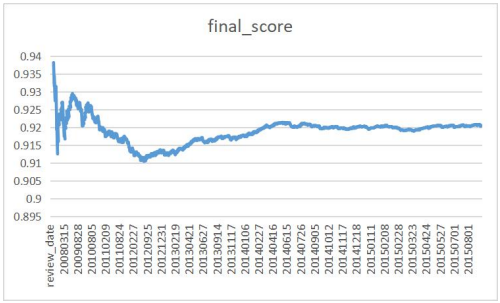
通过将各个值代入模型计算得到 score，区间为（0.86,4.58675），为了更直观的观察，本研究将 score 的区间转换为(0,5),得到对应的 final_sccore 数据集（以 hair_dryer 的数据为例）：

score
4.58675
4.39925
4.58675
4.58675
3.11000
4.58675
0.86000
3.08675
4.58675
4.58675
3.86000
4.58675
3.08675
3.83675
4.02425
4.58675
4.58675
3.83675
4.58675
4.58675
4.58675
1.58675
4.58675
4.58675
4.58675
4.58675
4.58675
2.33675

final_score
5.00
4.75
5.00
5.00
3.02
5.00
0.00
2.99
5.00
5.00
4.02
5.00
2.99
3.99
4.25
5.00
5.00
3.99
5.00
5.00
5.00
0.98
5.00
5.00
5.00
5.00
5.00
1.98

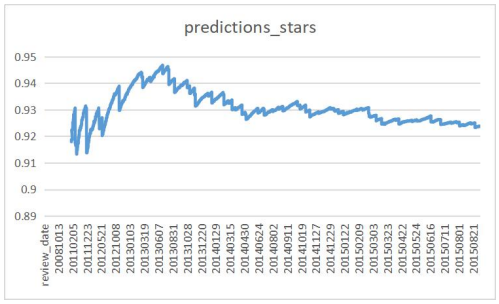
Hair_dryer:

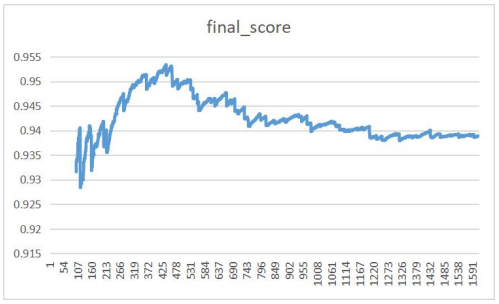




模型评分与总体的相关系数： 0.9203421752636529
评论预测与总体的相关系数： 0.9081865220631052

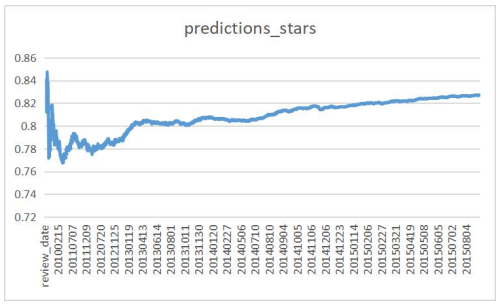
Microwave:

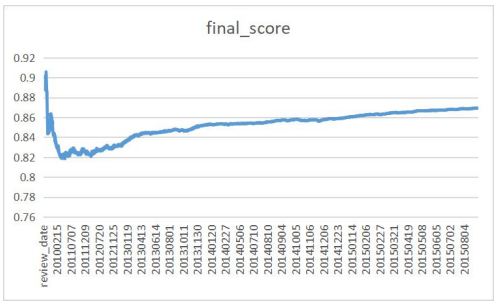




模型评分与总体的相关系数： 0.938902542770126
评论预测与总体的相关系数： 0.923864561551718

Pacifier:





模型评分与总体的相关系数： 0.8694357948896814
评论预测与总体的相关系数： 0.8272767410846057

可以看到 finals_score 数据集的相关性模型的拟合度大于用 predictions_stars 数据集所作模型。这说明本研究所做出模型更能准确客观地反映出客户的满意度，体现了模型的可信度。

七、建立时间-声誉模型

（一）、对客观评分进行数据处理与探究

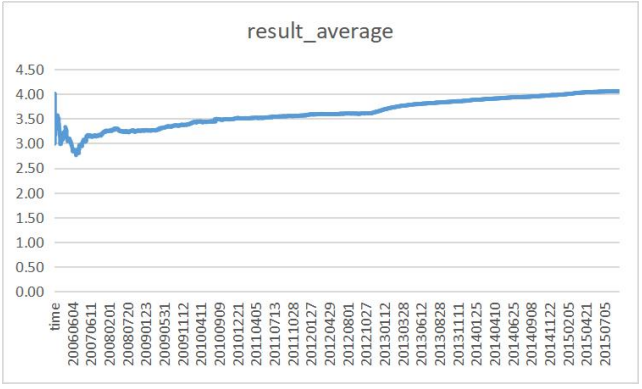
为了探究产品声誉与时间的模式本研究利用 2. 所得到数据制作的 execl 表格，并依据 excel 表格中的数据做出了各产品的评分均值变化图。下图为所用到的部分数据：

	score_sum	count	time	score_average	result_average
2	2.99	1	20020302	2.99	2.99
3	4.02	1	20020420	4.02	3.51
4	5	1	20020713	5.00	4.00
5	0	1	20020813	0.00	3.00
6	4.02	1	20020821	4.02	3.21
7	3.02	1	20021108	3.02	3.18
8	4.02	1	20021217	4.02	3.30
9	4.02	1	20021218	4.02	3.39
10	3.77	1	20030107	3.77	3.43
11	4.02	1	20030114	4.02	3.49
12	4.02	1	20030123	4.02	3.54
13	3.02	1	20030126	3.02	3.49
14	3.52	1	20030219	3.52	3.50
15	4.02	1	20030227	4.02	3.53
16	4.02	1	20030311	4.02	3.57
17	2.01	1	20030331	2.01	3.47
18	4.02	1	20030402	4.02	3.50
19	4.02	1	20030409	4.02	3.53
20	3.02	1	20040106	3.02	3.50

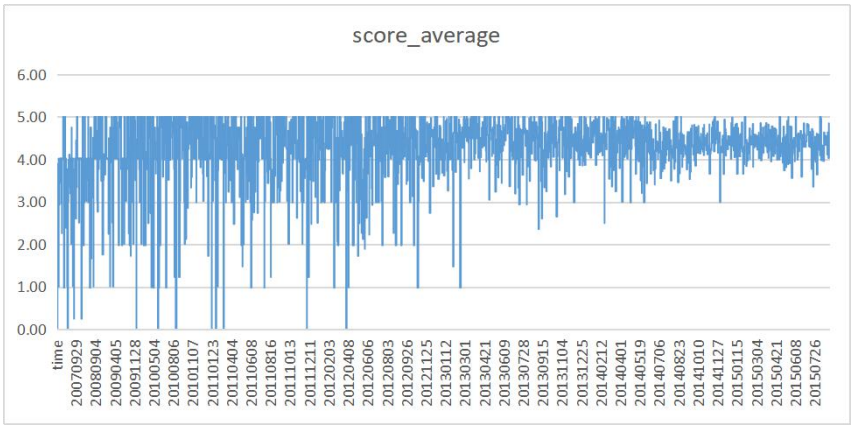
(r = result_average,s = score_average,c = count)

$$r_i = \frac{\sum_{i=1}^{n} (s_i * c_i)}{\sum_{i=1}^{n} c_i}$$

Hair dryer 截止于每个时间节点的评分均值变化图

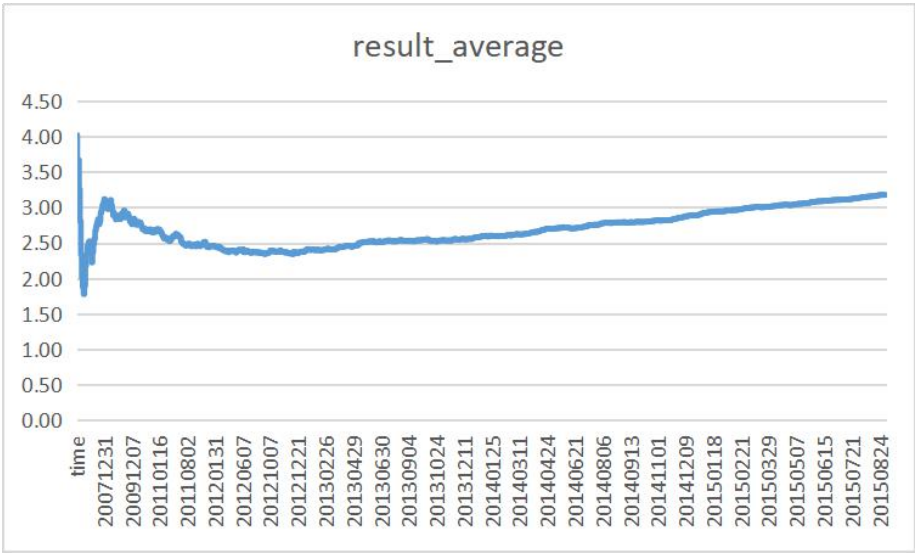


Hair dryer 的评分日均变化图

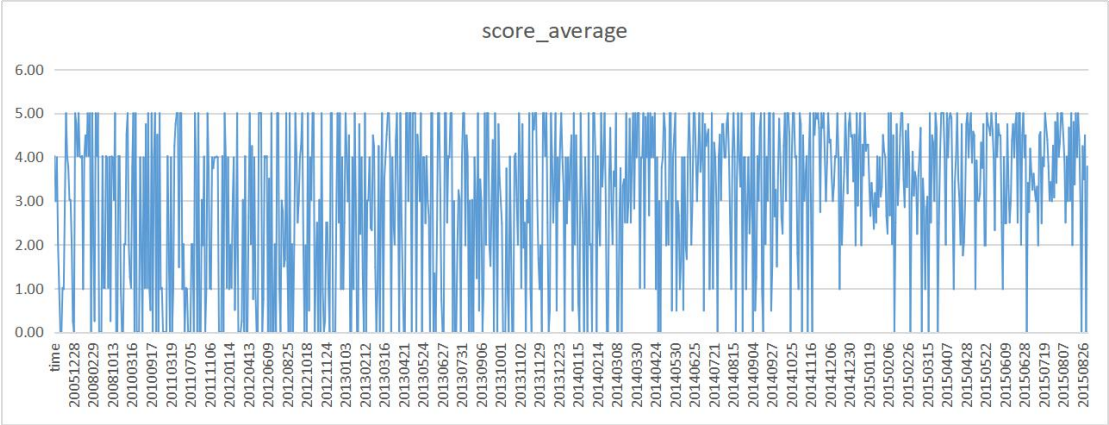


第一张图，本研究可以看出吹风机评分一开始呈现下降的趋势，这跟在此时间段产品刚推出不久并且评论数较少有很大关系，而之后评分逐步上升，最后处于一个稳定上升的状态。第二个图可以看到吹风机一开始的评分波动性很大，而随着时间推移，评分波动变小并且稳定在 3-5 分之间。综合以上两图，本研究可以很清晰看出产品声誉在不断上升。

microwave 的评分随着日期变化图

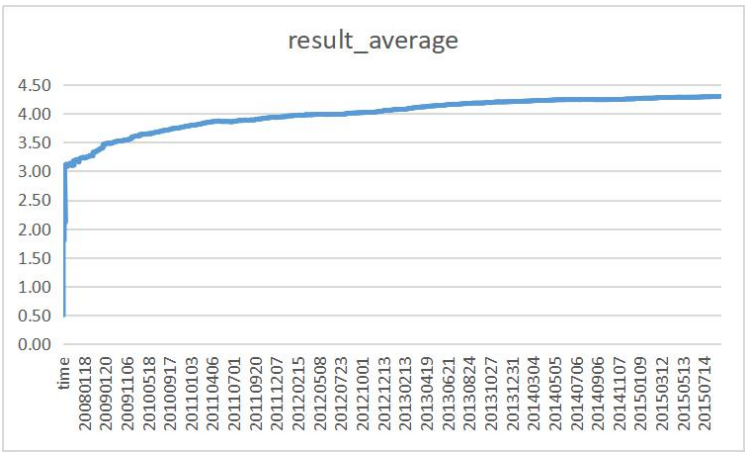


microwave 的评分日均变化图

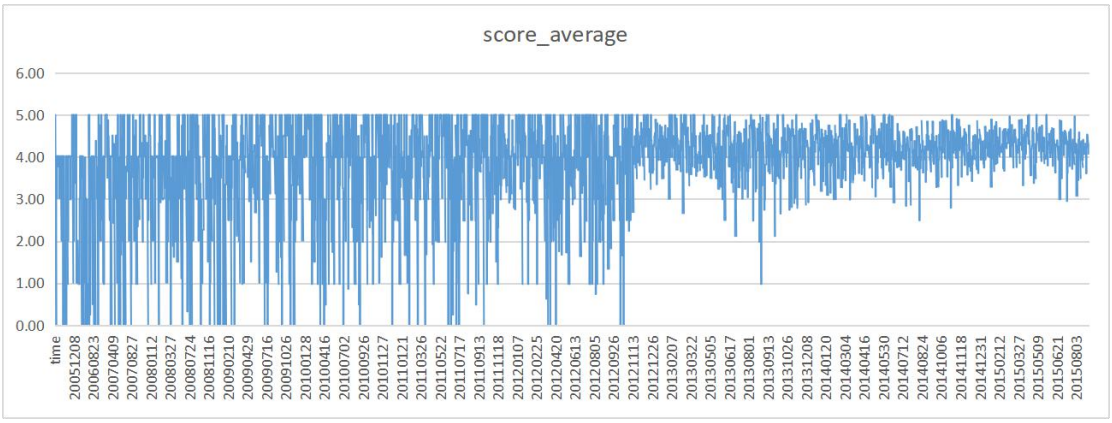


从 microwave 的 result_average 图可以看出，微波炉的声誉走向跟吹风机的声誉走向相似。虽微波炉产品的评分不及吹风机的高并且每日评分的波动性较大，但总的来说近年来微波炉产品的声誉是呈现上升趋势的。

Pacifier 的评分随着日期变化图表



Pacifier 的评分日均变化图表



结合两图可明显看出奶嘴产品的声誉呈现上升趋势

（二）、使用 R 语言时间-声誉建立预测线性模型

```
Residuals:
    Min      1Q  Median      3Q     Max
-127762   -2430     403    4076   19946

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)    2.229e+09  1.612e+07   138.3   <2e-16 ***
result_average 1.000e+00  7.297e-03   137.0   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 8784 on 2305 degrees of freedom
Multiple R-squared:  0.8907,    Adjusted R-squared:  0.8906
F-statistic: 1.878e+04 on 1 and 2305 DF,  p-value: < 2.2e-16
```

对 hairdryer 的预测模型建立

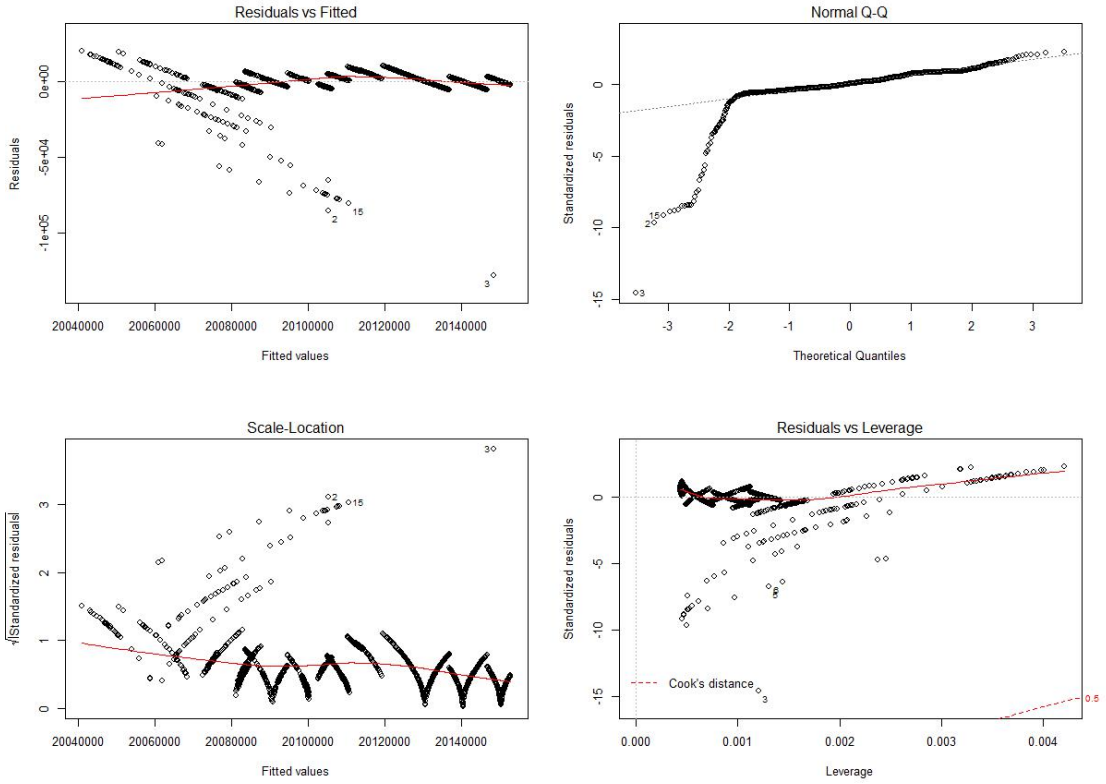
```
Residuals:
    Min      1Q  Median      3Q     Max
-130133     868    7087    9647   13508

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)    8.024e+08  6.583e+07   12.19   <2e-16 ***
result_average 3.542e-01  2.980e-02   11.88   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 19840 on 904 degrees of freedom
Multiple R-squared:  0.1351,    Adjusted R-squared:  0.1342
F-statistic: 141.2 on 1 and 904 DF,  p-value: < 2.2e-16
```

对 microwave 的预测模型建立

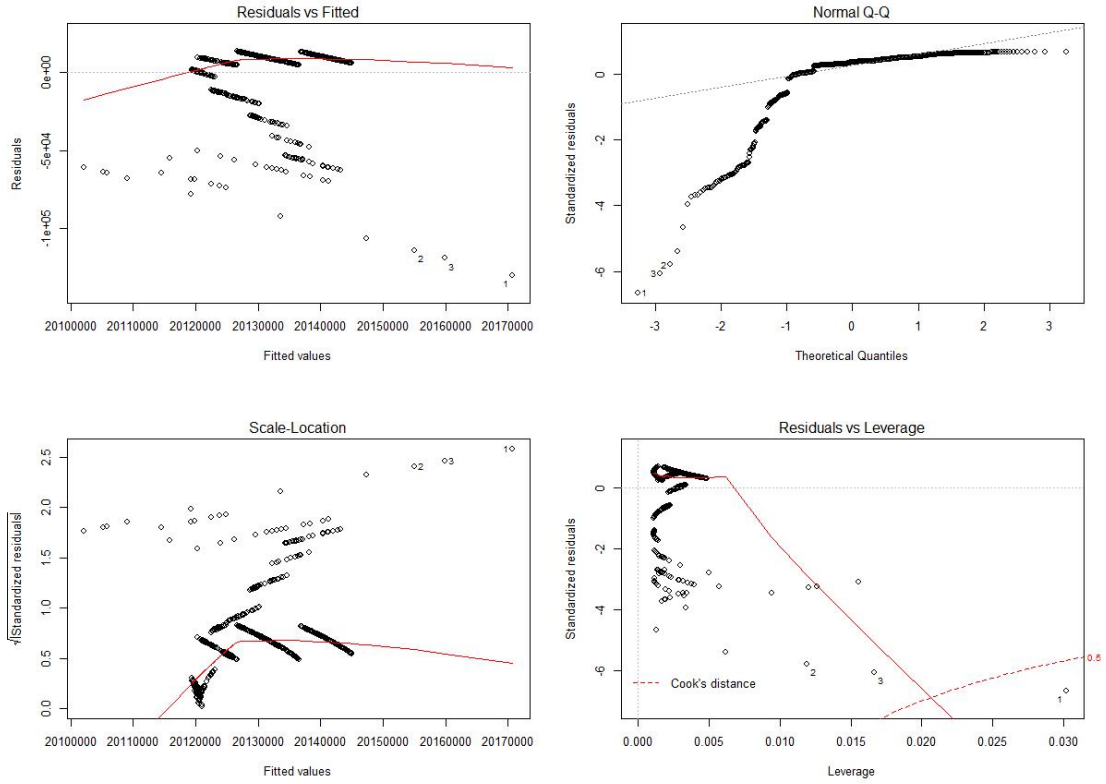
```
Residuals:
   Min     1Q Median     3Q    Max
-15736  -3761  -1261   2319 129543

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)    1.642e+09  1.185e+07   138.6   <2e-16 ***
result_average 7.344e-01  5.364e-03   136.9   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6412 on 1906 degrees of freedom
Multiple R-squared:  0.9077,    Adjusted R-squared:  0.9077
F-statistic: 1.874e+04 on 1 and 1906 DF,  p-value: < 2.2e-16
```

对 pacifier 的预测模型建立

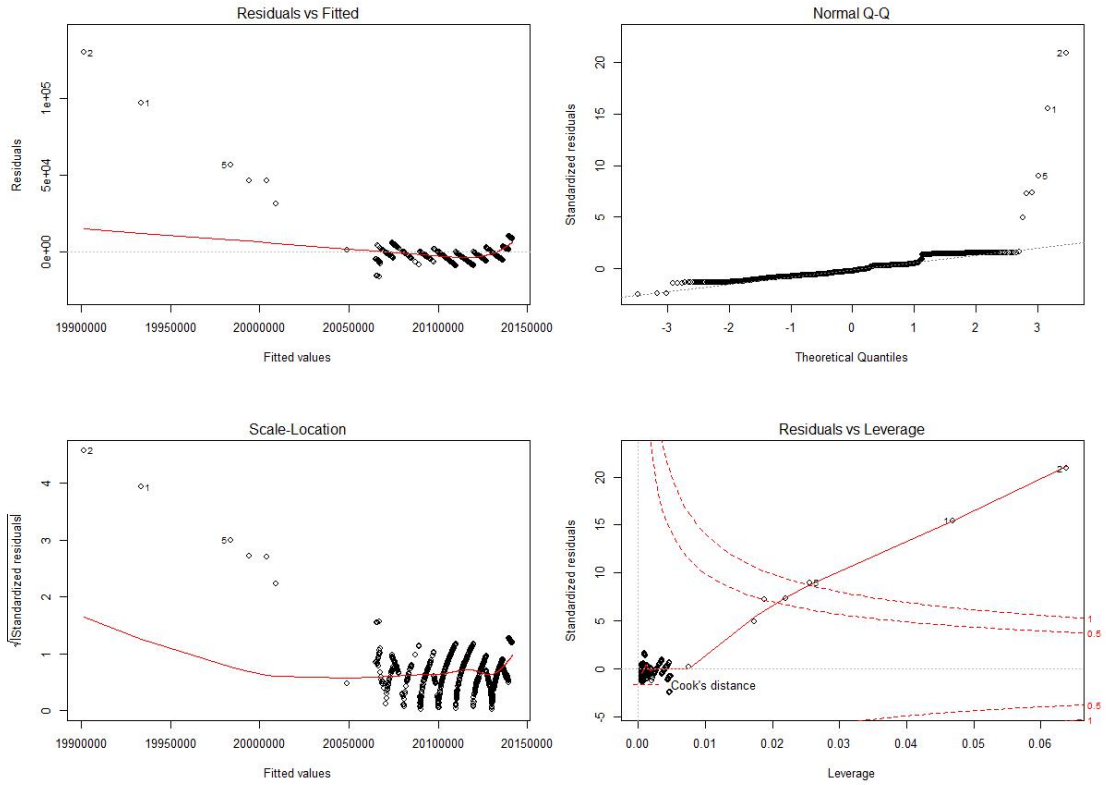
通过模型的建立本研究再通过下面 4 个图对本研究的模型进行诊断最终判断模型是否可行。



对 hairdryer 的线性回归模型诊断回归



对 microwave 的线性回归模型诊断回归



对 pacifier 的线性回归模型诊断回归

1.Residuals vs Fitted：残差与真实值之间的关系画图（残差应该是一个正态分布，与估计值无关。）

2.Normal Q-Q： 检测其残差是否是正态分布

3. Scale-Locatiob：检查等方差假设（如果方差不是一个定值那么这个模型的可靠性是大打折扣的。）

4. Residuals vs Leverage：检查数据分析项目中是否有特别极端的点（需注意，即使 R 将这些特殊的点标记了出来，也不等于他们一定需要被删除。还是要参考 Cook 距离的绝对大小。）

（三）、建立多项式模型

因为时间与声誉的关系并非一定是线性的，所以本研究通过多项式拟合与线性预测模型进行比较得出最佳模型

首先本研究通过 cross-validation 的方法在 10 次拟合以内确定最佳拟合结果下面是对这个方法的简述：

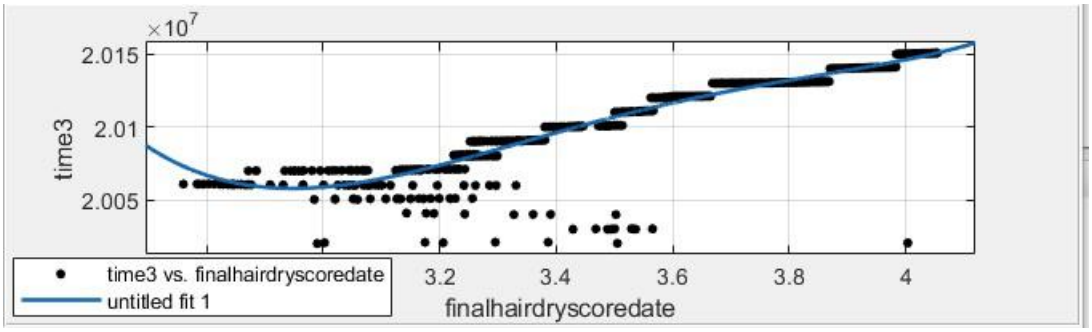
假设有 N 个样本，那么对于一个阶数 L，做 N 次拟合，每次拟合把第 i 个变量删掉，然后再计算第 i 个观测的残差的平方和，然后找到那个使得这个数值最小的 L。即：

$$\min_{L} \frac{1}{N} \sum_{i=1}^{N} \left[y_i - \sum_{l=1}^{L} x_i^l b_{-i,l} \right]^2$$

其中为删掉第 i 个观测之后的估计值。

下面就是本研究对时间-声誉进行的多项式拟合结果

Hairdryer 声誉—时间拟合结果



Linear model Poly4:

f(x) = p1*x^4 + p2*x^3 + p3*x^2 + p4*x + p5

Coefficients (with 95% confidence bounds):

p1 =	1.659e+05	(1.346e+05, 1.971e+05)
p2 =	-2.383e+06	(-2.818e+06, -1.949e+06)
p3 =	1.277e+07	(1.052e+07, 1.503e+07)
p4 =	-3.018e+07	(-3.535e+07, -2.5e+07)
p5 =	4.654e+07	(4.209e+07, 5.098e+07)

Goodness of fit:

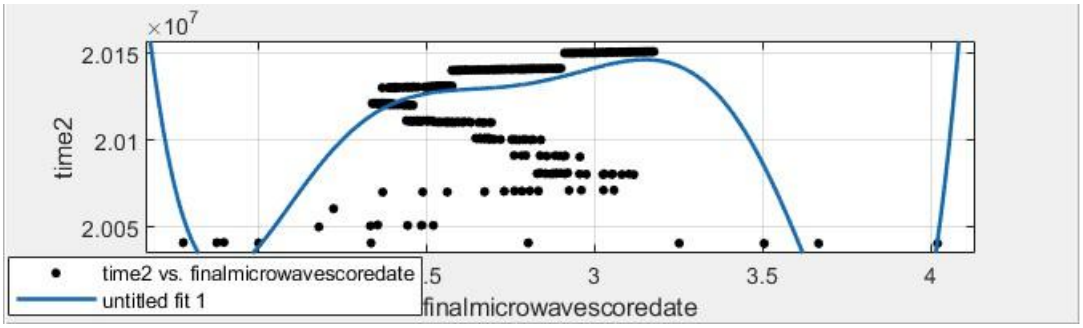
SSE: 1.551e+11

R-square: 0.9047

Adjusted R-square: 0.9045

RMSE: 8208

Microwave 声誉—时间拟合结果



Linear model Poly6:

f(x) = p1*x^6 + p2*x^5 + p3*x^4 + p4*x^3 + p5*x^2 +
p6*x + p7

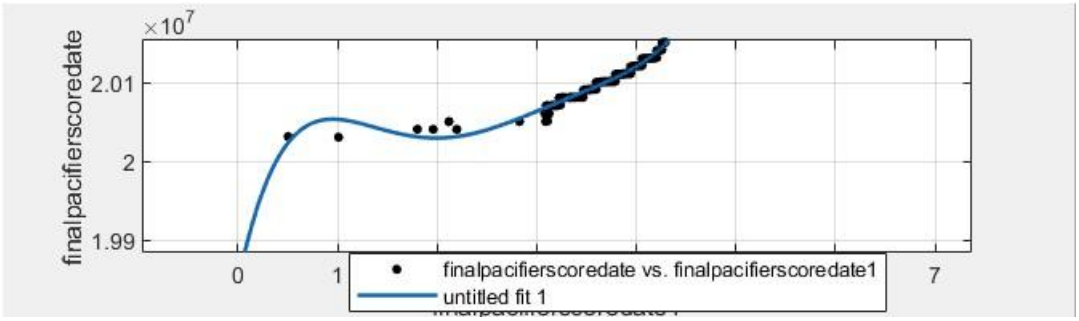
Coefficients (with 95% confidence bounds):

p1 =	4.055e+05	(3.008e+05, 5.102e+05)
p2 =	-6.898e+06	(-8.705e+06, -5.091e+06)
p3 =	4.825e+07	(3.541e+07, 6.11e+07)
p4 =	-1.777e+08	(-2.258e+08, -1.295e+08)
p5 =	3.629e+08	(2.624e+08, 4.635e+08)
p6 =	-3.898e+08	(-5.004e+08, -2.792e+08)

p7 = 1.918e+08 (1.418e+08, 2.419e+08)

Goodness of fit:
SSE: 2.883e+11
R-square: 0.2994
Adjusted R-square: 0.2947
RMSE: 1.791e+04

Pacifier 声誉—时间拟合结果



Linear model Poly5:
f(x) = p1*x^5 + p2*x^4 + p3*x^3 + p4*x^2 + p5*x + p6
Coefficients (with 95% confidence bounds):
p1 = 4218 (3670, 4766)
p2 = -5.565e+04 (-6.319e+04, -4.811e+04)
p3 = 2.766e+05 (2.374e+05, 3.157e+05)
p4 = -6.188e+05 (-7.125e+05, -5.251e+05)
p5 = 6.009e+05 (5.016e+05, 7.003e+05)
p6 = 1.985e+07 (1.981e+07, 1.988e+07)

Goodness of fit:
SSE: 1.869e+10
R-square: 0.978
Adjusted R-square: 0.9779
RMSE: 3135
（四）、模型比较

Hairdryer	linear regression	polynomial
SSE	1.779e+11	1.551e+11
R-square	0.8907	0.9047
Adjusted R-square	0.8906	0.9045
RMSE	8780.431	8208

Microwave	linear regression	polynomial

SSE	3.55919e+11	2.883e+11
R-square	0.1351	0.2994
Adjusted R-square	0.1342	0.2947
RMSE	19820.36	1.791e+04

Pacifier	linear regression	polynomial
SSE	7.835e+11	1.869e+10
R-square	0.9077	0.978
Adjusted R-square	0.9077	0.9779
RMSE	6408.308	3135

由上对比表可看出多项式的拟合模型是要比线性拟合要更好的，所以本研究可以得出最终的时间-声誉拟合结果：

Hairdryer 声誉—时间拟合结果：

f(x) = 1.659e+0.5*x^4 - 2.383e+06*x^3 + 1.277e+07*x^2 - 3.018e+07*x + 4.654e+07

Microwave 声誉—时间拟合结果：

f(x) = 4.055e+05*x^6 - 6.898e+06*x^5 + 4.825e+07*x^4 - 1.777e+08*x^3 + 3.629e+08*x^2 -3.898e+08*x + 1.918e+08

Pacifier 声誉—时间拟合结果：

f(x) = 4218*x^5- 5.565e+04*x^4 + 2.766e+05*x^3 - 6.188e+05*x^2 + 6.009e+05*x + 1.985e+07

由上面模型可知：

Hairdryer 的声誉随时间在提高，证明顾客对产品认同度不断提高对比以前成上升趋势，预计未来的销量及评价会继续升高。

Microwave 的声誉及其不稳定，可能由于购买样本量较少，但是在近期顾客对 Microwave 的评价似乎有升高的趋势，可能在近期评价会升高。

Pacifier 的声誉比其他两个产品都要好，而且销量也是在上升，在模型看来 Pacifier 的配件将维持与稳定但销量会有所上升因为 Pacifier 的良好的口碑。

八、探究极端评分对后续评分的影响

在探究完综合评分与时间的关系之后，消费者通过商品描述、价格,更主要

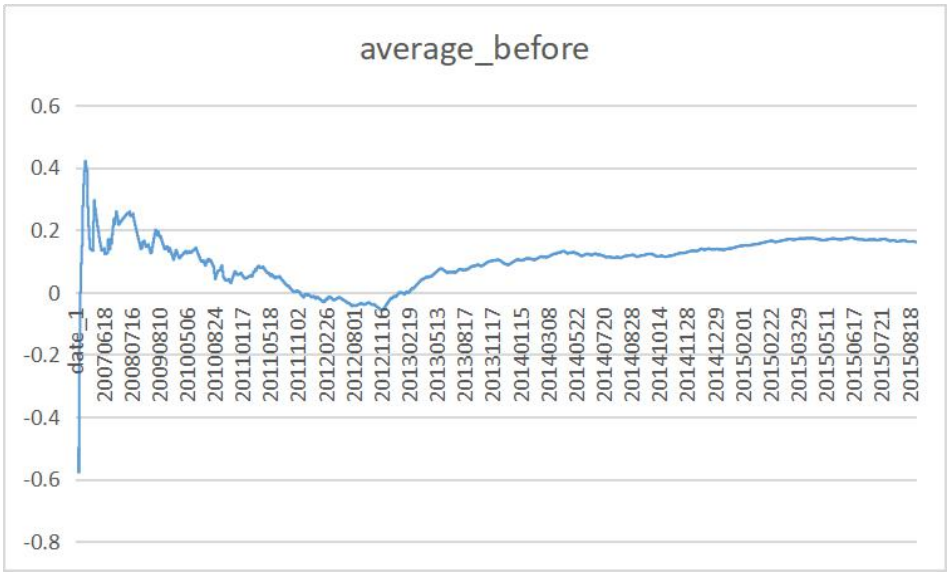
的是通过商品评论信息来考察商品的质量及其他信息[2]。为了更加深入了解一些为了深入发掘极端评论或者是打分是否会影响他人的打分或者评论，首先本研究定位到一些一星级的评论中，并且收集在此次一星打分前 20 个打分星级，之后本研究在收集该次打分之后为了更加深入了解一些为了深入发掘极端评论或者是打分是否会影响他人的打分或者评论，首先本研究定位到一些一星级的 star_rating 中，并且收集在此次一星评分前 20 个打分星级的平均即 average_before，来作为本研究判断一星评分之前的用户对产品的态度，之后本研究在收集该次一星评分之后 20 个打分星级的平均即 average_after，来作为本研究判断一星评分之后的用户对产品的态度，从而探索在一些一星评分之后用户是否有存在跟风评分一星或者说影响用户对该产品的评分的客观性。于是本研究在得出 average_before、average_after 之后利 用 Python 对这两个变量进行总体的相关性分析得到以下结果：

Hairdryer 一星评论出现前后的打分相关系数： 0.15977164786740022

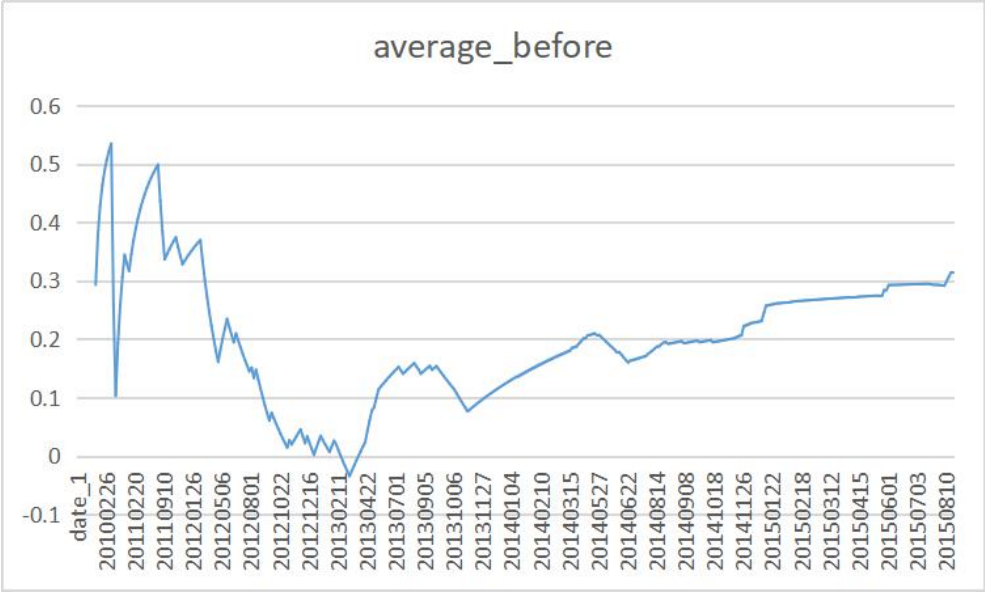
Microwave 一星评论出现前后的打分相关系数： 0.312795808934152

Pacifier 一星评论出现前后的打分相关系数： 0.15977164786740022

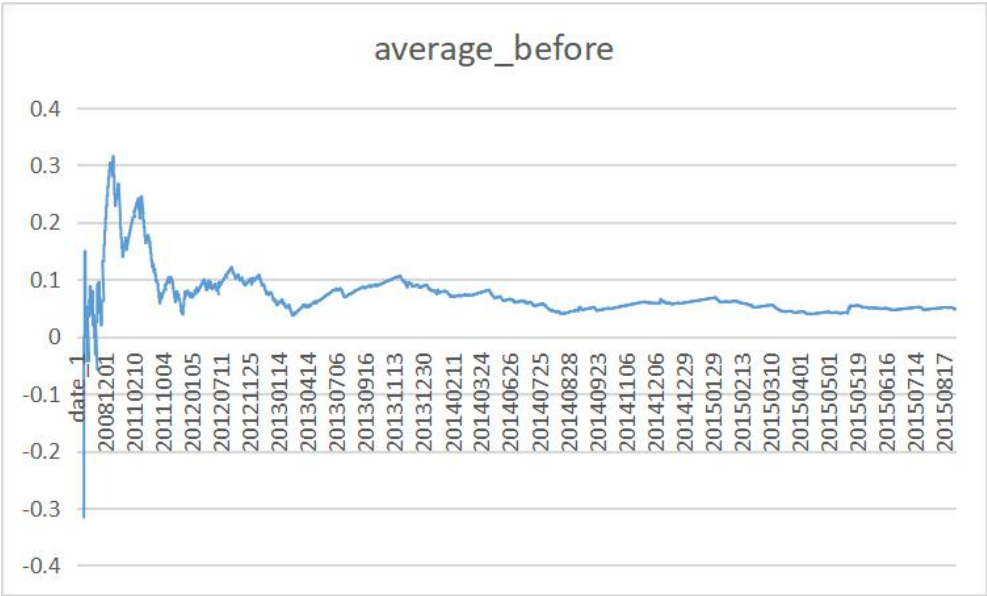
随后本研究想对这些相关性进行深入探索，所以利用 Python 对每次一星评分出现前后进行相关性分析，得到相关性变化的数据，将数据可视化得到以下结果



Hairdryer 相关性变化



Microwave 相关性变化



Pacifier 相关性变化

从以上相关性分析与挖掘可看出，在总体上来看一星评分出现前后的评分态度并没有什么太大相关性，可是在 hairdryer 和 pacifier 这两款产品早期一星评分对后续评分的影响是负相关的，所以在这两款产品早期一星评价对后续评分可能出现跟风性评分，而在后期可能由于客户对产品更加了解或者说客户对于评分更加的理智，所以对后续的评分影响并不大，冲 microwave 的相关性变化可看出这款产品的客户更加理性并没有被一星客户所影响，但是有一段时间一星评分等于后续客户的态度是有少许负面影响的，总结来说，但产品刚上市客户对产品

不了解恶意一星评分可能会对产品后续的评分以及销量产生较大的负面影响，而在产品经过市场的检验以后，人们对产品的评价也更加理性，恶意一星评分就无法构成太大影响。

九、实验讨论

随着社区互联网，特别是电子商务的快速发展，人们越来越能够便捷地在互联网上浏览和采购自己心仪的商品。随之而来的是大量商品评论和评分，这些评论和分值是用户对

所购商品和相应服务的评价，体现了商品和服务的受欢迎程度。顾客更愿意浏览和购买评价高的商品。本研究根据亚马逊提供三种产品的评分和评论，经过细致分析与推算，建立了关于对于企业有帮助的模型，本研究在分析过程中可以看到有些用户存在恶意差评，和用户随意打分问题例如评价中是“失望”之类的差评词汇，但他在进行评分时却是 5 分。本研究使用 TF-IDF 模型对评价进行预测以获得隐藏在评论中客户对该产品的真实评分 predictions_stars，近年来，评论和分值相结合的研究方法在商品推　荐系统研究中占据了越来越高的比重。评论包含了丰富的语义信息和用户对商品的偏好等情感信息，分值则是用户综合考虑各项因数之后对商品的一个整体评价。词袋模型　（Bag of Words，BoW）是一种传统的文本特征表示方法，它挖掘数据集中频繁出现的单词或者单词组，统计这些单词或者单词组在某一条文本中出现的次数，从而实现对该文本的语义表示。在此基础上本研究添加根据用户的类型不同，分为不同权重，利用层次分析法，最后所得到模型将能够精确表达出用户真实的满意度，比只依照评分与评论建立的模型拟合度更加高。本研究通过得到 final_score 并且转为 0-5 区间，便于更好的观看与分析。最后本研究去通过时间维度去看到 final_score 的变化，进而可以得到产品声誉的变化。企业可以依据此进行决策分析，在得到相应决策之后企业可以去改善自身产品，或者做出一些营销策略上的改变，这些方法都是有助于企业在经济上以及效率上的提高。因为本研究对数据预处理得足够仔细与认真，且本研究的模型准确性是较高的。利用本研究所作出的模型能够帮助企业分析产品，是一个明智的选择。

十、参考文献

[1] 吴菲,徐姗姗.Research on Algorithm of Internet Comment Tendency Analysis Based on Machine Learning[J].佳木斯大学学报(自然科学版),2019,37(01):23-26.
[2] 王禹. 电商平台购物虚假评论识别研究[D].首都经济贸易大学,2018.

十一、参考文献说明

【1】吴菲,徐姗姗.Research on Algorithm of Internet Comment Tendency Analysis Based on Machine Learning[J].佳木斯大学学报(自然科学版),2019,37(01):23-26.

本文主要研究判断网络评论信息情感倾向的方法。 针对传统 IG 算法和 TF-IDF 算法的不足,提出了一种改进的 IG 算法和 TF-IDF 算法。 针对朴素贝叶斯方法、KNN 算法和 SVM 分类算法的不足，提出了一种融合分类器。 实验表明，融合分类器取得了一定的效果，能有效提高分类精度。

【2】王禹. 电商平台购物虚假评论识别研究[D].首都经济贸易大学,2018.

本文的目的是给出一套对虚假评论进行精准、有效识别的方法及流程,并考察虚假评论的模式。主要采用数据挖掘方法实现虚假评论的识别工作,主要工作包括:获取不同电商平台的样本商品数据,对文本进行量化,通过评论时间、重复评论、评论者等级等信息进行虚假评论预识别;并采用 Logistic 回归、k 最近邻模型、SVM 模型、text-CNN 模型、fast Text 模型以及组合模型对虚假评论进行准确识别并进行验证;然后通过大量数据,考察虚假评论模式,构建虚假评论的语言模型,并从多维特征上考察虚假评论的行为属性以挖掘虚假评论在行为属性上的模式。本文创新点主要包含:1、通过数据多维特征如:重复评论、评论时间分布信息等对虚假评论进行预识别,结合预识别结果进行人工标注和后续分析;2、对传统模型算法进行了调整,另外,通过模型的分类效果对模型赋予权重,进行模型集成,提升虚假评论识别的效果;3、除此以外,本文还通过虚假评论识别结果对虚假评论信息建立语言模型,分析虚假评论的多维特征以考察虚假评论在行为属性上的模式。