第二题a小问

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.

第二题b小问

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:





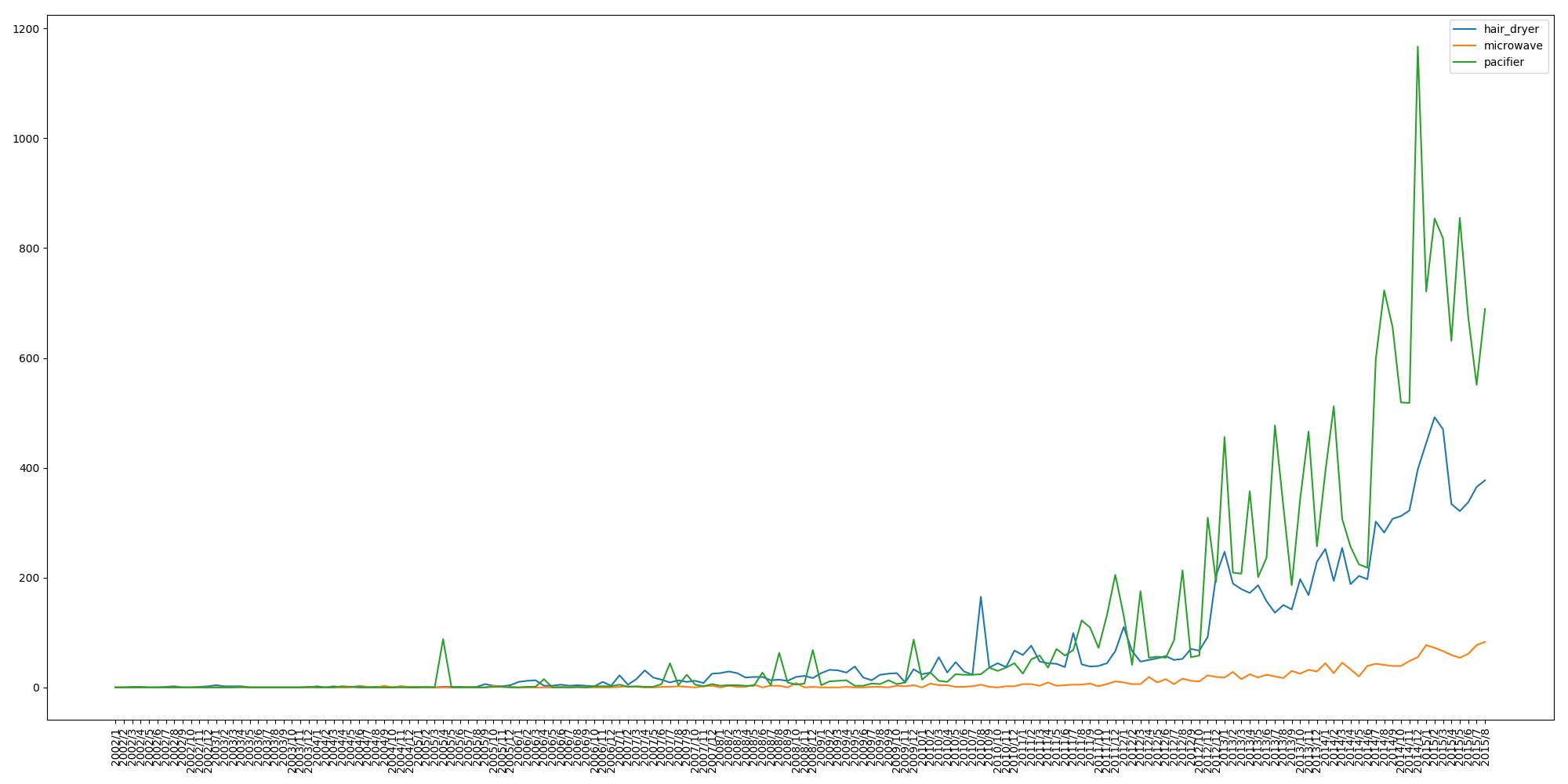
.



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:

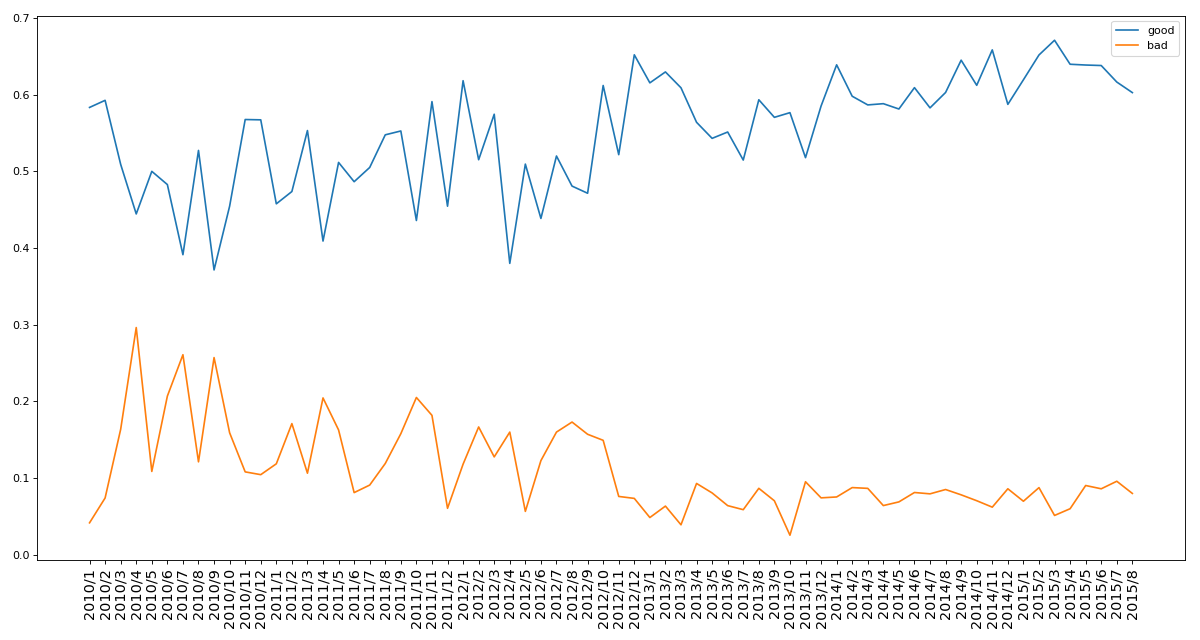


由于篇幅限制, 图像模糊, 清晰版请见附录, 通过图像可以得知2010年以前的销量非常少, 故下面分析时只取2010年1月以后的数据进行分析.

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

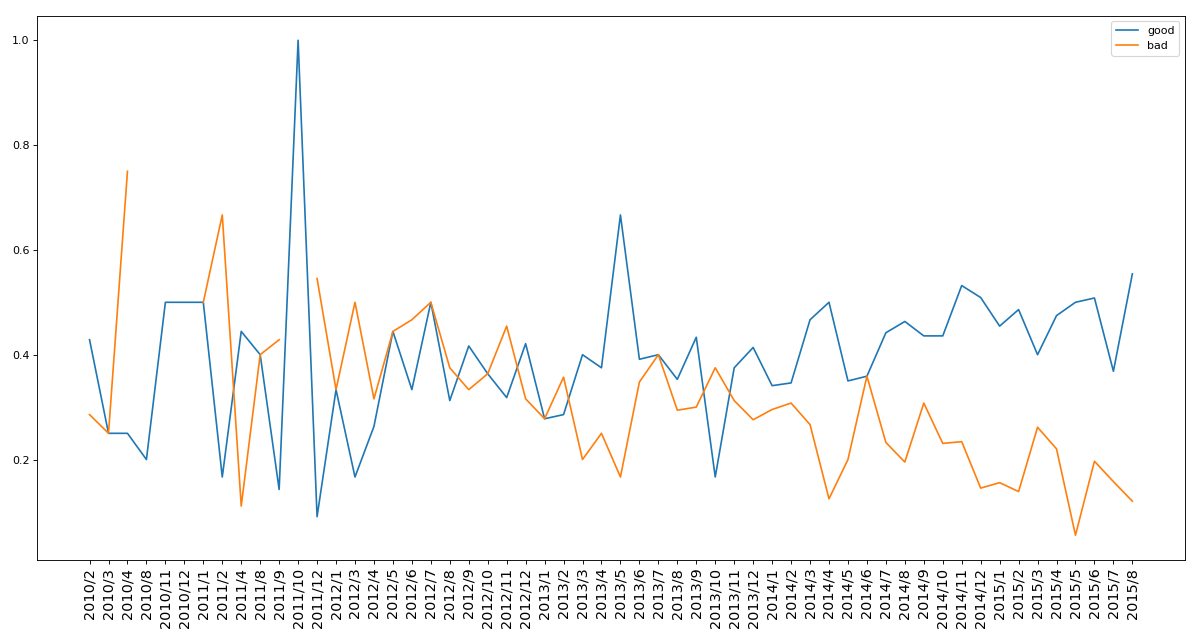
利用预处理之后的数据, 得到hair\_dryer的月份好评率和差评率如下图xxx:



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.

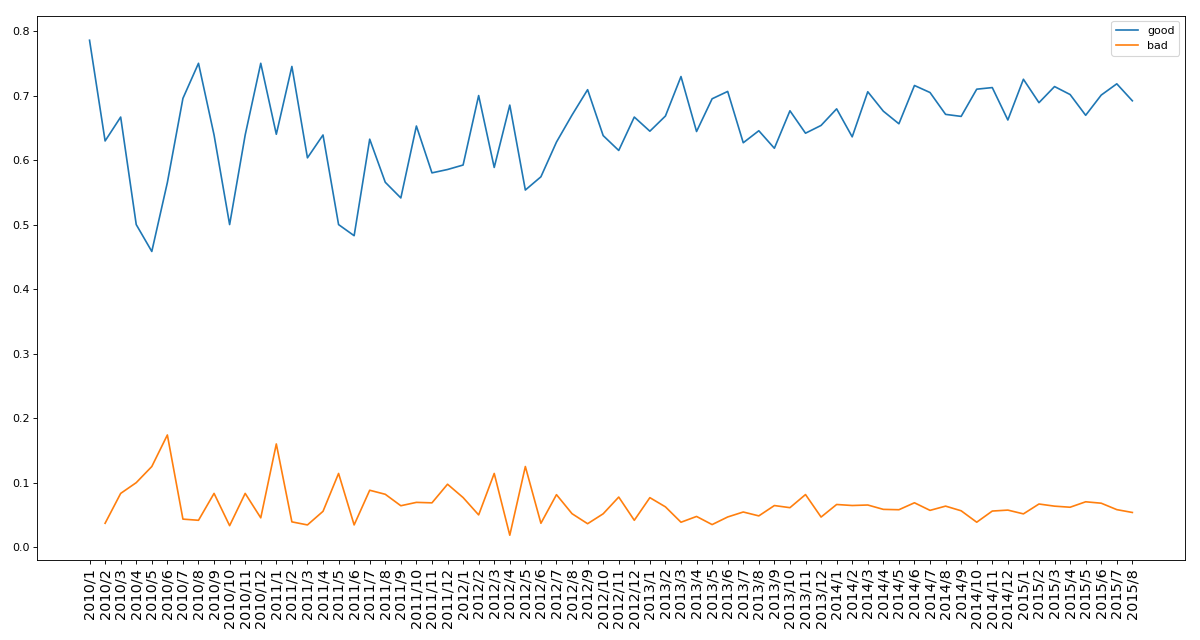
1. microwave时间模式分析

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:



附录

