信:

非常荣幸有此次机会能够向您和贵公司展示我们对hair\_dryer, microwave, pacifier三个产品数据的挖掘和分析结果, 在这里我将向您全面报告我们的建模过程和我们在数据中的发现.

首先, 通过初步观测提供的三个数据集, 我们发现每个产品的marketplace都在US , 并且product\_category分别在每个数据集内部都一样, 我们认为其对分析无用, 故对其进行删除. 同样,发现product\_parent数据与product\_id数据冗余, 选择删除了product\_parent列数据. 同时, 对数据进行了其他预处理后, 进入建模阶段.

在经历自然语言处理中的情感分析过程后, 我们将非量化的评论文字映射到数值空间中, 用[-1,1]的数值范围来表示评论中蕴含的具体情感. 数值越大, 积极情感越浓厚, 数值越小, 则消极情感越浓厚. 情感分析完成后, 我们进行了特征工程, 一个好的特征工程能够让模型更好的抓住数据特征, 更好的发现数据中的内在模式, 提高模型的准确率. 通过特征工程建立了新的特征好评率rate和删除了一些非正常情况, 例如1星的评论:” 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.”, 可是这应该是5星的好评才对, 所以针对此类评星和评级完全相反的数据进行删除处理, 幸运的是, 这样的数据并不多, 三个数据集一共只有27条. 随后, 我们将数据按8:2的比例进行划分训练集和测试集, 并将评级大于等于4星的作为正例, 小于四星的作为反例, 利用机器学习算法LightGBM对每个数据集进行单独的数据建模. 最终通过测试集的验证得到了模型在三个数据集上的F1值为0.889, 0.874, 0.939, 结果表明我们的模型建立的非常成功! 不仅如此, 我们还通过模型得知了review\_body, review\_date, rate 和total\_votes对模型的重要程度最大. 贵公司应该主要关注这四类数据的变化, 尽量少或不关注评论者等其他重要程度低的数据.

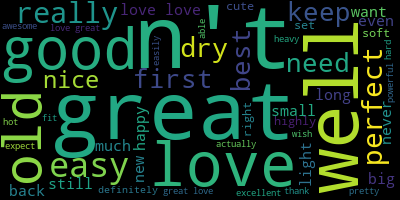
一旦贵公司打算上线销售三种产品, 可以利用上述模型对产品的销售数据进行评估, 通过模型预测结果和用户实际评分进行对比, 得到二者是否趋于一致的结果. 若一致, 则抓住此类用户的评价特点, 有针对性的进行改进产品, 以获得更好的销售. 若不一致, 则可以关注该类用户是否为虚假用户或恶意用户等等.

我们还利用了熵权法从评级和评价进行综合分析得到产品得分, 可以衡量产品的声誉等相关信息. 并且根据所给的数据分数据集讨论了基于时间的模式度量, 绘制统计图分析得出, xxxxx. 为了评价一个产品的潜在成功或失败, 我们结合了建立的机器学习模型LighGBM和熵权法分析的结果, 选取了重要程度大的与文本和评级有关的指标进行模糊综合评价建模, 最终得出可以通过计算特定分值来指示产品的潜在成功与否.

最后, 我们利用相关性分析技术得到了特定的评级会引发某种类型的评论, 即用户的评论可能会受到之前人评级的影响. 例如以下评论:” - DO NOT BUY - the 5 star reviews are fake”. 同时我们建立了特定词汇表, 对所有评论数据惊醒词频统计, 我们绘制了如下两张词云图, 词云图的单词的字体大小与其在评论中出现的频次有关. 频次越高, 字体越大. 通过词云图我们可以看到, 一星级的词云图大多为否定词汇, 例如n’t like, never, bad, 而五星级词云图则大多为great, love, well, perfect等. 虽然五星的词云图中也有n’t, 但是通过具体分析我们发现其主要表示积极情感, 例如” like n’t allow finger print”等等.



一星级词云图



五星级词云图

我们的数据挖掘结果呈现完毕了, 相信您看到之后一定会有大量的收获, 最后希望贵公司推出的产品能够大获成功!