

基于商品数据挖掘的销售策略建议

概要

我们从以下四个方面进行数据分析和信息挖掘：评论，星级和有用性评分，产品品牌评分，产品声誉预测以及星级对评论的影响，从而提出可靠的销售策略和产品改进建议。

首先，经过数据预处理后，我们通过NLTK工具对文本数据进行分词和初步情感分析，并将其量化为情感评分，范围为 $[-1, 1]$ 。我们采用数据可视化，描述性的方法统计和相关分析，并进一步构建多元Logistic回归模型，以分析帮助等级与评论时长，星级和化合物之间的关系。结果表明，帮助等级与评论时长呈倒“U型”关系，与星级之间呈正“U型”关系。

下一步，我们将基于评级和评估模型进行分析。通过建立LDA分析模型来找到每种产品的主题特征，在此基础上我们可以提出改进产品的建议。同时，我们从主题特征中总结了影响产品销售的五个指标，即质量，价格，外观，服务和尺寸。然后使用基于文本相似度的计算机搜索算法来计算每个评论的索引分数。此外，我们将分数与层次分析法相结合，以确定每个指标的权重，并建立加权的品牌评分系统。最后，我们通过系统的聚类将所有产品品牌聚类，以选择潜在的高质量品牌，并将其推荐给阳光公司。

此外，我们计算评论和星级的综合评分，并将其作为产品的声誉，这有助于通过时间序列分析预测三种产品的未来声誉。它描述了这三种产品具有季节性特征，并可能在不久的将来保持稳定的季节性周期。但是，这三种产品的声誉综合得分的高峰时间是不一致的，进一步的分析表明，在声誉得分的高峰期产品销售数字更大，阳光公司可以据此制定销售计划。

最后，我们分析了星级和评论之间的关系。通过建立分布式滞后模型，发现当期客户的评论将受到其他客户的评分和评论的影响。同时，我们观察到情绪得分与星级之间的时变同步，很明显，包含正面词语的评论会导致较高的星级评分，而包含负面词语的评论会导致较低的星级评分。换句话说，星级和特定的质量指标之间有很强的相关性。

关键字：相关分析，多项式逻辑回归，自然语言处理，时间序列分析

