Problem Chosen	2020 MCM/ICM Summary Sheet	Team Control Number
C		2010638

Sales Strategy Recommendation Based on Commodity Data Mining

Summary

We conduct data analysis and information mining from the following four aspects: correlation between review, star rating and helpfulness rating, product brand rating, prediction of product's reputation and impact of star rating on reviews, so as to propose reliable sales strategies and suggestions for product improvement.

Firstly, after data pre-processing, We perform word segmentation and preliminary sentiment analysis on the text data through the NLTK tool, and quantify it as emotion score, ranging from [-1,1].We adopt the methods of data visualization, descriptive statistics and correlation analysis, and further construct a multivariate Logistic regression model to analyze the relationship between helpfulness rating and review length, star rating and compound. The results show that helpfulness rating has an inverted "U-type" relationship with review length and a positive "U-type" relationship with star rating.

The next step, we conduct analysis based on the rating and evaluation model. The LDA analysis model is constructed to find the topic feature of each product, based on which we can propose suggestions for improvement of products. At the same time, we summarize five indexes that affect the product sales from the topic features, namely quality, price, appearance, service and size. Then a computer search algorithm based on the text similarity is used to calculate the index score of each review. In addition, we combine the score with the analytic hierarchy process to determine the weight of each index and build a weighted brand scoring system. Finally, we cluster all product brands through systematic clustering to select potential high-quality brands and recommend them to sunshine company.

Further, we calculate the comprehensive score of review and star rating, and take it as the product's reputation, which is conducive to forecast the future reputation of three products through time series analysis. And it depicts that the three products have seasonal characteristics and will probably maintain a stable seasonal cycle in the near future. However, the peak time of reputation comprehensive score of the three products is discrepant, and further analysis illustrates that the product sales figure is larger during the peak period of reputation score, according to which sunshine company can make a sales plan.

Finally, we analyze the relationship between star rating and review. Through the establishment of distributed lag model, it is found that customers' reviews in the current period will be affected by other customers' ratings and reviews. Meanwhile, we observe the time-varying synchrony between emotional score and star rating, which is clear that reviews containing positive words result in higher star ratings, while reviews containing negative words result in lower star ratings. In other words, there is a strong correlation between star ratings and specific quality descriptors.

Key words: correlation analysis, multinomial logistic regression, natural language processing, time series analysis