Project 4 (Fletcher) Proposal

Jason Salazer-Adams

Overview

A product is either introduced in the market or currently in the market, and want to understand whether or not the product is trending. One way to predict whether or not a product will trend is based on the initial comments from customer reviews. If the review is similar to other reviews of previous products which have trended, then there is a likelihood this product will trend too. Understanding whether or not a product will trend will be vital in the product's forecasting process. Traditionally, the forecasting process of an item will follow a process of,

1. Generate a statistical forecast. Potentially, use history from a similar product if no history exists.
2. Sales team provides additional market insight on top of the statistical forecast.

If a product could be identified as trending based on actual user feedback, then this information could be added to forecasting process, resulting in a more accurate forecast.

I am planning to use Amazon customer review data and in particular the toy data. I may change this to clothing as it could be more applicable in that industry. In general, I think this model will lend itself well to products which can be cheaply and quickly made so inventories can be easily adjusted based on near term market feedback.

Determining product trend

The big challenge will be determining the ground truth for a trending product or not. I did some research and developed a formula based on inspiration from a blog post.

The definition of a trending product would be,

- Highly rated, i.e. more 5 star ratings compared to 1 star ratings.
- Consistently rated, i.e. little variability in the star ratings.
- Reviewed by many people.

The formula would look something like this,

(Review Proportion * avg Rating) / std Rating

The Review Proportion is the proportion of reviews based on the total reviews in the data set, likewise the average and standard deviation are calculated for the data set. If there is no variability in star rating, then the minimum variability will be used. The data set to calculate the product trend will be defined as all reviews in the past 30 days from the last known review for each product.

The star ratings are from 1-5, and I would like my trend statistic to indicate a positive and not a negative trend. I will be smoothing the star rating by raising each rating to the power of 1.5. The trend statistic will be normalized utilizing the tanh function. I found this function based on a Wikipedia page on activation functions. This will ensure all trend scores are from 0 to 1, since the original trend score is positive. The trend score will be converted to a 0-1 decision variable by making a cutoff of the top 1% of scores within the data set. Admittedly, this adds bias to the data set since the cut-off is based on data, instead of some other independent classification. However, I plan on looking at products over 1-2 years, so hopefully the score will average out.

Features

The feature set will be the first review in the 30 day range since the last review. Taking the first review means the model is attempting to predict whether or not the product will trend in the next 30 days or not. I am also planning to generate a topic/group of comments feature by utilizing unsupervised learning. I am planning to stay away from features that are inputs or highly correlated to the trend score, or else those would be leaky variables.

MVP

I created a quick and dirty MVP of 10000 random toy products between 2014-2015. Logistic Regression was able to identify 18/93 positive cases in my training data set. There appears to be some potential for building a model. The MVP notebook can be found here.