Rating music: Accounting for rating preferences

Manish Gupte

Purdue University

(mmgupte@purdue.edu)

This paper investigates how consumers' likes and dislikes influence their decision to rate a song. Yahoo!'s survey dataset which asked users how likely they were to rate based on their preference for the song was used for the analysis. Histograms reveal, online ratings are more likely when people like the song or when they hate the song. An "ordinary" song receiving rating is less likely. A model is then presented which assumes people derive utility when they give higher ratings for a song they like and derive utility when they "punish" a song they hate by giving a low rating. Also, some agents do not derive utility by "punishing". Consumers have cost to rate and there is a random draw of utility based on the song and consumer. An interpretation based on how such a utility function can account for consumer trying to influence the songs they are recommended as a result of their rating is also provided, though some aspects still need a behavioral framework. This structural model is estimated using a likelihood function. It is then suggested that these results can be used to construct a selection equation to existing recommender systems. Two experiments to increase likelihood of rating and also better study motivations to rate are suggested.

Rating music: Accounting for rating preferences

1. Introduction

This paper analyses the probability of a song receiving rating based on how much the consumer likes or dislikes the song. Yahoo! in a survey of its consumers asked them how likely they were to rate a song based on how much they liked or disliked the song. Histograms of the probability of rating reveals consumers are more likely to rate if the like the song or they hate the song. A behavioral model is proposed in which consumers view the rating as a judgment they are making on the song. In this process they derive utility by giving a high rating if they like the song. They also derive utility when they "punish" a song by giving a low rating if they hate the song. This punishment is different from spitefulness and can be viewed as being out of fairness concerns. de Quervain et al (2004) shown that people derive utility from punishment if they think it is a fair thing to do. There is also a set of consumers who do not receive utility by punishing. Also, if consumers do not strongly feel about the song they derive no utility from rating it. There is a cost of making a decision and this is part of the utility function. The likelihood of a song receiving a rating depends on the magnitude of utility consumers get by doing so. If the utility they derive is high a rating would be given more often.

A structural approach which estimates marginal utilities, cost of rating and proportion of people who do not derive utility by giving a low rating is then described. A method is then proposed which uses these estimates to correct for selection bias in ratings.

Harper, Li and Konstan (2005) perform a similar structural estimation for movie ratings. In their setup the number of movies a person rates depends on the marginal benefit and marginal cost of rating. Unlike their setup, my paper looks at the probability of particular song receiving ratings based on how much the user liked or disliked it. One of the benefits for rating in Harper, Li and Konstan's setup is getting a better movie recommended by the system. In my setup, the utility from giving a good rating for a song liked by a consumer can also be viewed as utility she is likely to receive if she was recommended a similar song in the future. Also, the punishment by rating a song low can be viewed as avoiding receiving a similar song again. However, some people do not rate when they hear a song they hate, and this observation can be explained in behavioral terms. Also, it is found in the data that songs which consumers do not feel strongly about are less likely to be rated. If the consumer rated these songs then the likelihood of receiving better songs is increased. Thus, there is a benefit to rating but the consumers do not rate. An experiment based on reframing the question from sounding like asking agents to judge songs to sound like asking agents if they would like to hear a “similar” song again is proposed in section 7 to tease out the behavioral and “rational” aspects of this question. An extension which gives monetary rewards to consumers for ratings song is then proposed. Such a reward mechanism would first have be tested using an experiment to find out the utilities from rating a song relative to monetary rewards.

Recommender systems have many different aspects as described by Schafer, Frankowski, Herlocker and Sen (2007). They write, one of the limitations of collecting ratings is they face the public goods problem. Also, Adomavicius and Tuzhilin (2005) describe classifications of such systems. My paper is focused on the probability of receiving a rating based on how strongly consumers feel about a song.

Section 2 describes the dataset and provides initial analysis, section 3 describes the model, section 4 describes the estimation method, and section 5 has results. Section 6 suggests a method to use the results and section 7 suggests future experiments.

2. Data description and initial analysis

The dataset used for this analysis is Yahoo! Music ratings for user selected and randomly selected songs, version 1.0. In a survey given to Yahoo!’s online music consumers, people were asked to say how likely they were to rate a song when they hated it, didn’t like it, were neutral about it, liked it and loved it. The responses were never (1), very infrequently (2), infrequently (3), often (4) and very often (5). Users were also asked if their preference affected their ratings. A subset of this data which uses observations where users said their preferences matter is used for the estimation. Users were also asked how frequently they rate songs and possible responses were never, a few songs since I signed up, a few songs per month, a few songs per week and a few songs per day. Histograms of the likelihood of a song receiving a rating based on likes and dislikes of the user are now presented. The likelihood of rating increases from 1 to 5, with 1 being never and 5 being very often.

Figure 1: Likelihood of rating when agents love the song

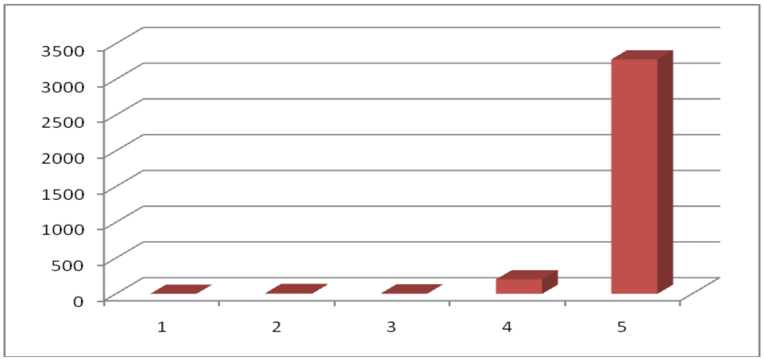


Figure 2: Likelihood of rating when agents are neutral about the song

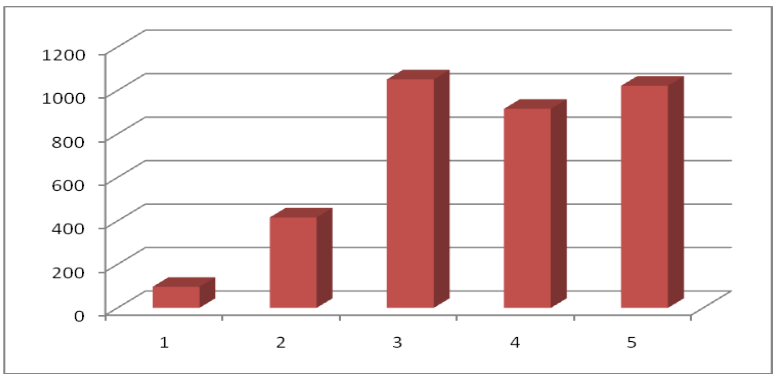
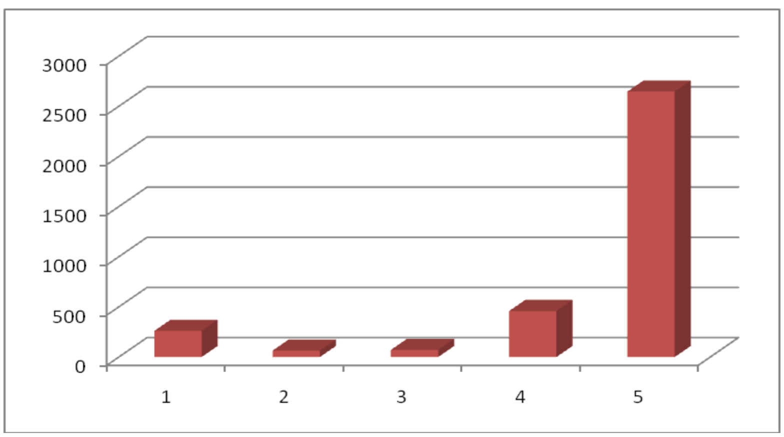


Figure 3: Likelihood of rating when agents hate the song



The histograms suggest ratings are more likely when people feel strongly about the song. Also, getting rating is less likely when one hates the song as opposed to when one loves it.

3. Model setup

A behavioral model is now presented to explain these observations. It reasonable to believe that when one is listening and ratings songs he or she is likely to "go by the heart" as opposed to "ask what I get by rating". It is possible that people rate songs to get better recommendations from the system and this interpretation of the model can also be done. Agents get a utility $U_{is} = \alpha r_s - c(1+\beta f) + \xi_{is}$ where r_s is the rating for song *s*, $c(1+\beta f)$ is the cost to rating a song and ξ_{is} is the error when $r_s = 4,5$. The error can be viewed as draws of variation in cost of rating. The cost of the rating also depends on how frequently a person rates music, *f*. People who like rating, rate more often and are hence have lower cost.

Thus, agents get utility by giving higher ratings if they like the song. This can also be interpreted as agents get positive utility if they are recommended a similar song in the future. It is assumed that the rating and how much the user likes the song are perfectly correlated, that is agents announce the true perceived rating.

The utility when agents are neutral about the song is: $U_{is} = -c(1+\beta f) + \xi_{is}$, $r_s = 3$. In this case agents do not get utility by giving a rating or they are indifferent between hearing a similar song again. Note there can be a benefit of receiving a better song if users let the system know they were neutral about the song. But, histograms show a reduction in probability of rating in such a case and hence this utility form, which is based on behavioral reasons, is assumed. Future work, as described in section 7, would consider this aspect in more detail.

The utility when agents do not like the song is: $U_{is} = \gamma[5 - r_s] - c(1+\beta f) + \xi_{is}$, $r_s = 1,2$. Thus, agents receive utility if the punish a song with a low rating. It can also be interpreted as the utility agents get by avoiding a song like the one they hated in the future. Histograms show agents do not always like to give low ratings. In this model I assume there are some of users who do not get utility by giving a lower rating. The proportion of such users is *g*. This is another behavioral aspect of the model. ξ_{is} is assumed to have a standard normal distribution. The estimation procedure is now described.

4. Empirical analysis

The estimation procedure is like an ordered probit. If *U* is the mean utility an agent gets by rating, then agent are likely to rate "very often" (which is the highest level) if $U+\xi_{is}\geq L1 \Rightarrow \xi_{is}\geq L1-U$. Agents are likely to rate "often" if $L1\geq U+\xi_{is}\geq L2 \Rightarrow \xi_{is}\geq L2-U, \xi_{is}\leq L1-U$. *L2 < L1.* A similar set of equations or levels hold as the probability of receiving a rating falls. The number of levels is 4 as 5 possible levels of likelihood of getting a rating are possible. The probability of receiving rating at a given frequency is given by integrating the error over the range. Thus, the probability of receiving a rating which gives utility U "often" is $\int_{L2-U}^{L1-U}\frac{1}{(2\pi)^{0.5}}e^{-t^2}dt$. When lower ratings are given two thing a happen. Consumers who like to punish get utility U_h and those who do not like to punish get utility U_l . $U_h > U_l$. The likelihood of getting a low rating often then becomes $(1-g)\int_{L2-U_h}^{L1-U_h}\frac{1}{(2\pi)^{0.5}}e^{-t^2}dt + g\int_{L2-U_l}^{L1-U_l}\frac{1}{(2\pi)^{0.5}}e^{-t^2}dt$.

The likelihood of receiving ratings with observed frequencies can then be written as the product of individual probabilities given a set of parameters. As this combined probability should be 1, given that this is observed data, the log of this probability is maximized to obtain the best parameters. The maximization needs constraints such that *L2 < L1, L3<L2, L4<L3.* A program written in Matab was used to estimate this model[1]. The findings are presented in the next section.

5. Main findings

The following estimates, with standard errors in brackets, were obtained by maximizing the likelihood function[2].

---

[1] Program available on request.
[2] Matlab's fmincon was used and negative of likelihood function was minimized.

$$
\begin{aligned}
&\alpha = 0.235(0.002)\\
&\gamma = 0.316(0.002)\\
&c = 1.112(0.003)\\
&\beta = -0.119(0.001)\\
&g = 0.275(0.003)\\
&L1 = -0.080(0.001)\\
&L2 = -0.626(0.002)\\
&L3 = -0.852(0.002)\\
&L4 = -1.377(0.002)
\end{aligned}
$$

As suggested by the histograms agents derive utility when they rate higher for songs they like and “punish” songs which they do not like. Also, 27.5% of people do not receive utility from punishing. Some of these results can also be viewed as agents trying to influence the recommender system to give songs which they like.

6. Suggestions to improve recommender systems

A recommender system needs to account for this selection bias. One way to incorporate it in would be to use a Heckman selection equation in regressions which predict ratings based on song and consumer characteristics. In cases when song or consumer characteristics are not available, like in the ydata-ymusic-rating-study-v1_0-train.txt dataset a selection equation can improve the raw probability estimate. An analysis of the improvement is planned in the future. Two policy implications also come out from this analysis which, are described in the following section.

7. Suggested experiments and their implications

In the real world, people rate music because of behavioral reasons or to try and influence the recommender system. If the people are not asked to “rate music”, which gives a feel of judging the song, but to say “how satisfied they would be if a similar song was recommended to them” consumers’ rating choices will be driven by trying to get the best for themselves. This change in framing the question can be used to better understand how people rate music. If ratings were driven by behavioral reasons then the 27.5% of consumers who refrained from giving a negative rating would do so. Also, agents would derive utility by letting the system know that they were neutral about the song they heard.

This will increase likelihood of receiving ratings. An experiment on these lines, which can possibly be run on the field, is suggested.

Consumers can also be given monetary incentives to rate music. This can be in the form of a lottery they get to play if they rated music. Thus, the utility function will now have a monetary part and hence the likelihood of rating will go up. The structural model presented in this paper can be modified to account for such a lottery and re-estimated to obtain the costs and benefits in monetary terms. Thus, the free-rider problem associated with recommender systems can be addressed. Clearly, higher the lottery the more likely ratings will be obtained but given benefits and costs of implementing such a system an optimal level of monetary reward can be obtained. This is another experiment which is suggested.

8. Discussion

A behavioral model was estimated to obtain estimates of how likely consumers are likely to rate a song. Extensions to this work include the suggested experiments. The estimation technique can also be improved to account for unobserved heterogeneity given variation in consumers' demographics. A simulation based method like Nevo's can be used for this purpose.

9. Acknowledgments

I would like to thank Yahoo! for providing the data which enabled this analysis. The dataset, Yahoo! Music Ratings for User Selected and Randomly Selected Songs, version 1.0, was obtained through the Yahoo! Webscope program for use solely under the terms of a signed Yahoo! Data Sharing Agreement. Further information on it can be obtained at: http://research.yahoo.com/Academic_Relations .

The following is a note from Yahoo!: "This data may be used only for academic research purposes and may not be used for any commercial purposes, by any commercial entity, or by any party not under a signed Data Sharing Agreement. The data may not be reproduced in whole or in part, may not be posted on the web, on internal networks, or in networked data stores, and may not be archived offsite. The data must be returned to Yahoo! at the end of the research project or in three years, whichever comes first.

This dataset was produced from Yahoo!'s records and has been reviewed by an internal board to assure that no personally identifiable information is revealed. You may not perform any analysis, reverse engineering or processing of the data or any correlation with other data sources that could be used to determine or infer personally identifiable information."

10. References

1. Adomavicius Gediminas, Tuzhilin Alexander, 2005, Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions, IEEE Transactions On Knowledge and Data Engineering, Vol. 17, No. 6.

2. de Quervain DJF, Fischbacher U, Treyer V, Schelthammer M, Schnyder U, Buck A, Fehr E ,2004, The neural basis of altruistic punishment, Science, Vol. 305, Issue 5688

3. Harper Maxwell, Li Xin, Chen Yan, Konstan Joseph, 2005, An economic model of user rating in an online recommender system, Lecture Notes in Computer Science, Springer-Verlag.

4. Schafer Ben, Frankowski Dan, Herlocker Jon, Sen Shilad, 2007, Collaborative filtering recommender systems, Lecture Notes in Computer Science, Springer-Verlag.

5. Nevo Aviv, A Practitioner's Guide to Estimation of Random-Coefficients Logit Models of Demand, 2000, Journal of Economics & Management Strategy, Volume 9, Number 4, Winter, 513–548.