SMART RATING FOR ELECTRONIC GADGETS

Data Mining Project Report

12/27/2012

Group A2

Deepanshu Saini – 61310308

Rachna Lalwani – 61310845

Saurabh Thaman – 61310113

Vikramadith Raman – 61310387

Jeevan Murthy – 61310542

Karthik Venkiteswaran - 61310809

SMART RATING FOR ELECTRONIC GADGETS

Executive Summary

Business Objective & Goal:

Aggregator websites of Electronic gadgets listings decide if they want to pick a particular listing from an e-commerce website and display on their own website. Selection of the right items to list is essential to customers selecting products for purchase, therefore leading to higher revenues.

Business Problem:

Many e-commerce websites do not collect user ratings, or do not provide this data to our website. Hence, it is difficult to determine whether to display the product or not. Simply removing all items without ratings could lead to high opportunity costs, hence, another solution is required.

Data Mining Objective:

Predict the *High - Low* rating for any new electronic gadget listing to be introduced on the aggregator website monitored by bargain.in. The same model can also be used to predict the High - Low Rating on websites that do not support the feature of Average Rating.

Analytical Method

Using Logistic regression method a model is created to predict high-low rating of the test data where high rating is considered as success measure and cut off point at 0.6 is kept is used to predict the probability of success. 10.25% prediction error observed in Naïve method is improved to prediction error of 2.91% in Logistic regression method.

Recommendation

Our model can be used when crawling product listings on multiple websites, to predict whether the customer is likely to rate that product “high” or “low” based on past ratings and this can be used to show listings with higher likelihood of “high” rating on the aggregator website.

SMART RATING FOR ELECTRONIC GADGETS

Background

Business Objective: Increase coverage of listings that can be included in our aggregator website, by including those items that do not have ratings. This has the potential to improve our sales revenues by selection of appropriate items that are likely to be purchased by customers, and also avoids the opportunity cost of simply not listing such products.

Data Mining Objective: Predict the *High - Low* Rating for any new electronic gadget listing to be introduced on the aggregator website monitored by our website. The same model can also be used to predict the High - Low Rating on websites that do not support the feature of Average Rating.

Data

Data is being collected online from the website with details about the electronic gadgets listed on the ecommerce websites. The following columns are available for the predictions:

Column	Description
brand	Brand of the Product
color	Color of the Product
freeShipping	1 = Free Shipping Available, 2 = Free Shipping Not Available, 0 = Data Unavailable
inStock	1 = Product In-Stock, 2 = Product Out-of-Stock, 0 = Data
avRating	Average rating of the product
reviewCount	No. of users who rated the product
listPrice	Price of the product on "date"
shippingPeriod	Shipping period of the product
siteName	Name of the website from which the product is sold

category	Category of the product
date	Timestamp of the product and price information (mm/dd/yyyy)
TimeNextPrice	number of days until the next available price information

Data Preparation

The data required the following transformations in order to perform the modeling.

Column	Transformation
brand	Created categorical variables for each brand. Reduced categories by studying the pivot table of avRating (output) with brand categories.
avRating	Created Binned variables: Values less than 2 -> LOW otherwise HIGH
shippingPeriod	Missing values were generated used KNN – prediction from available datapoints using sitenames, category, instock and freeshipping as inputs (see Appendix B)
siteName	Created categorical variables for each website.
category	Created categorical variables for each category.

Data Mining Methodology

Given the business objective we set out to achieve, the constraints imposed by this data set were primarily concerned with missing data *(missing ratings for almost 4000 of the 8000 odd records)*. Moreover, out of these 4000 records for which we had the ratings data, shipping period data was missing for over 2000 records. We worked our way around the problem by predicting the shipping period values using the KNN method, the amount of data available was still a constraint (see Appendix B).

As a first step, we built a Naïve rule as a benchmark. This was a simple model wherein we predicted the rating based on majority. All records were of the "High" variety, all new records were deemed to be of the same category. The percentage error in this case is around 10.25% which is the benchmark all our other models to beat.

Next, we selected the Logistical regression model in order to predict the rating of new products. The overbearing rationale behind this was the data intensive nature of K-NN.

In addition, we have also tried the Naïve Bayes model due to the small size of the training set and with the knowledge that with a smaller training set, we should prefer a high bias and low variance approach like NB I order to compensate for over fitting seen in logistical regression. Although, Naïve Bayes provides biased probabilities while working with categories but we can compare the accuracy of NB results with our logistic regression model results.

Evaluation

Let's first assess the benchmark Naïve model. This is a fairly straightforward model wherein the prediction is based on majority and the error rates are around 10.25% (see Appendix C Part 1).

Moving to the logistic regression we assess the confusion matrix, as our aim is only prediction and not ranking. Based on this requirement and a base probability cutoff of 0.6, the errors fall dramatically from 10% to 2.2% . Thought the percentage for LOW variety is still high, the same can be ascribed to the fact that we have kept a higher threshold for High variety to get better accuracy there for business reasons (see Appendix C Part 2).

Next we also evaluate the Naïve Bayes model and compare the error with that of logistic regression. The total error in this case is around 3.69% which is comparable to the logistical regression and beats the Naïve rule but logistic regression still better in predicting a high rated product. This can be interpreted as a fact that the fitting which was done in the logistical model was not over fitted to a great extent (see Appendix C Part 3).

Conclusion and Recommendations

Based on the modeling exercise carried out, we were able to come up with certain recommendations for the user of this model and some learnings which we would like to document.

Business

- The model is of utmost use to product owners who intend to launch their product via these websites. They can assess on which of the sites will their products garner a high rating and which of the sites will underrate their products.
- This information can then be combined with financial aspects (fee and logistics) of each website to assess if investing in advertising in a certain site makes sense for us in terms of RoI

Data Mining

- The relative advantages of each of the methodologies came to the fore when we started building the models. In our assessment, we recommend that we use Naïve bayes in case of a small training set, while with bigger sets, we can go for KNN of logistic regression.
- There can be some overfitting in the logistic regression model, so running multiple models for the same data set is always useful
- Data Cleaning and poverty – The most important and cumbersome part of the data mining activity is the data cleaning. This took around 65 – 70 % of our time, but once this was finished efficiently, it gives us a very clear idea of limitations within which we had to work, making rest of the assessment easier

Appendix A - Data Preparation

The 'Brand' data used bins based on the average of avRating as indicated below:

Brand	Average of avRating	Brand Bin
CANON	0	1
FUJIFILM	0	1
LEMON	0	1
NIKON	0	1
OLYMPUS	0	1
ZEN	0	1
VIDEOCON	0.833333333	1
SONY ERICSSON	2.24137931	2
SONY	2.46	2
XOLO	3	2
IBALL	3.317073171	2

Appendix B - KNN Modeling used to predict shipping period

Validation error log for different k

Value of k	Training RMS Error	Validation RMS Error
1	0.509430257	1.009858031
2	0.553386494	0.975308169
3	0.581469294	0.945325292
4	0.606034786	0.928684079
5	0.616797432	0.924115084
6	0.623277912	0.924362044
7	0.632727139	0.925207089
8	0.636751456	0.92204528
9	0.637605427	0.919661762
10	0.640076348	0.91914998

11	0.639572248	0.919019191	
12	0.640930549	0.915819837	<--- Best k
13	0.641143384	0.917192682	
14	0.641571739	0.916948353	
15	0.641571739	0.917577642	
16	0.641622179	0.917502665	
17	0.641970986	0.916978861	
18	0.642008698	0.917030285	
19	0.642159104	0.915949376	
20	0.6422549	0.916105775	

Training Data scoring - Summary Report (for k=12)

Total sum of squared errors	RMS Error	Average Error
423.1157282	0.640930549	-0.00194102

Validation Data scoring - Summary Report (for k=12)

Total sum of squared errors	RMS Error	Average Error
518.3326523	0.915819837	0.01955578

Test Data scoring - Summary Report (for k=12)

Total sum of squared errors	RMS Error	Average Error
360.8500462	0.934734682	-0.07863182

Appendix C - Confusion Matrices

1. Naive

Summary report (Naïve)			
	Predicted class		
Actual Class	High	Low	Grand Total
High	3771	0	3771
Low	431	0	431
Grand Total	4202	0	4202
	% Error	10.2570205	

2. Logistic

Validation Data scoring - Summary Report (Logistic regression)				
	Cut off Prob.Val. for Success (Updatable)			0.6
	Classification Confusion Matrix			
		Predicted Class		
	Actual Class	High	Low	
	High	1484	7	
	Low	42	148	
	Error Report			
	Class	# Cases	# Errors	% Error
	High	1491	7	0.47
	Low	190	42	22.11
	Overall	1681	49	2.91

3. Naive Bayes

Validation Data scoring - Summary Report (Naïve Bayes)				
	Cut off Prob.Val. for Success (Updatable)			0.6
	Classification Confusion Matrix			
		Predicted Class		
	Actual Class	High	Low	
	High	1470	30	
	Low	32	149	
	Error Report			
	Class	# Cases	# Errors	% Error
	High	1500	30	2.00
	Low	181	32	17.68
	Overall	1681	62	3.69