Smart Rating for Electronic gadgets

BUSINESS ANALYTICS USING DATA MINING

Deepanshu Saini – 61310308
Rachna Lalwani – 61310845
Saurabh Thaman – 61310113
Vikramadith Raman – 61310387
Jeevan Murthy – 61310542
Karthik Venkiteswaran - 61310809

Business Goal

- Objective: Aggregator websites of Electronic gadgets listings should be able to decide if they want to pick a particular listing from an e-commerce website and display on its website.
- The average ratings on these websites for the products go about a long way in deciding the sales.
- Business Problem: Many online aggregators of electronic gadgets especially cameras and mobile handsets realize that a Low rating for their products impacts the sales both on the online and offline channels. Negative publicity through Word of Mouth could further impact the sales of these products.

Data Mining Objective

- 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.

- The Data available for the prediction has the following details about the electronic gadgets listed on the Websites:

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

Column	Transformation
brand	Created categorical variables for each brand. Reduced categories by studying the pivot table of avRating (output) with brand categories.
freeShipping	No Change.
inStock	No Change.
avRating	Created Binned variables: Values less than 2 -> LOW otherwise HIGH
reviewCount	No Change.
listPrice	No Change.
shippingPeriod	Missing values were generated used KNN – prediction from available datapoints using sitenames, category, instock and freeshipping as inputs
siteName	Created categorical variables for each website.
category	Created categorical variables for each category.

Methods

- BENCHMARK: Naïve Bayes
  - The Naïve Bayes method is used for the prediction of the HIGH-LOW rating of the test data based on majority rule when no predictor is available for the product.
  - The performance of the Naïve Bayes is used as benchmark for comparison.

- Method Adopted: Logistic Regression
  - Logistic Regression is used for the prediction of the HIGH-LOW rating of the test data.
  - The predictors used are: Brand, Free Shipping, In Stock, Review Count, List Price, Shipping Period, Category and Site Name.

Evaluation metrics

- The Logistic Regression model generated will be used to predict the HIGH – LOW ratings of the test data.
  - The performance of the model are compared with the predictions made by the Naïve Bayes for the test data.
  - The HIGH rating is considered as success and a probability cutoff of 0.6 is used to predict the probability as success.

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

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	

Conclusion

- Logistic works much better than Naïve.
- We used Logistic Regression due to data poverty. Hence we divided data only in training and validation sets.
- K-NN would require the data to be divided amongst 3 sets and Naïve Bayes provides a biased probabilities when working with categories.
- Since we are not interested in ranking but only in prediction, we would look at only confusion matrix as an evaluation metric and not the lift chart.
- We used 0.6 as the cut-off probability so that products with high probability of having "high" ratings get selected more through the engine. It turns out that 0.6 also gives us the least error in the validation set.
- We have an engine that crawls product listings on multiple websites, predicts whether the customer is likely to rate that product "high" or "low" based on past ratings and shows listing with higher likelihood of "high" rating on the aggregator website.
- Can be used for Price Comparison websites e.g. Junglee.com, Kayak.com etc.