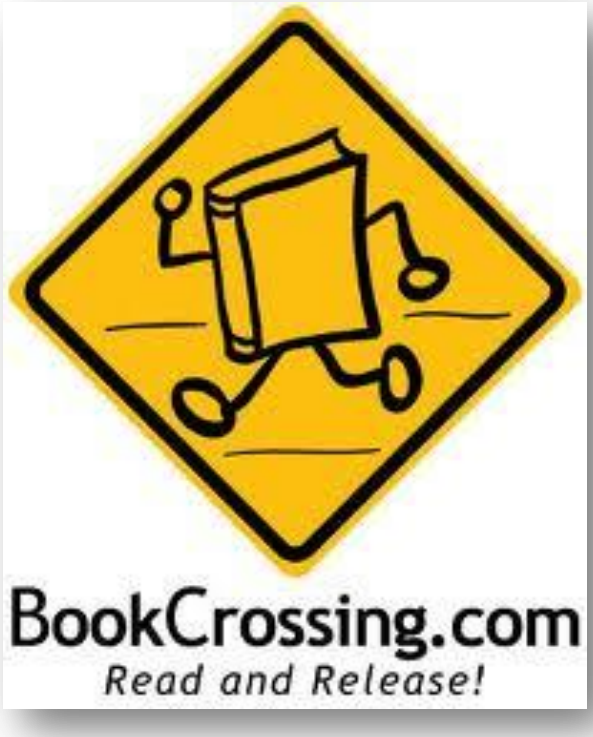
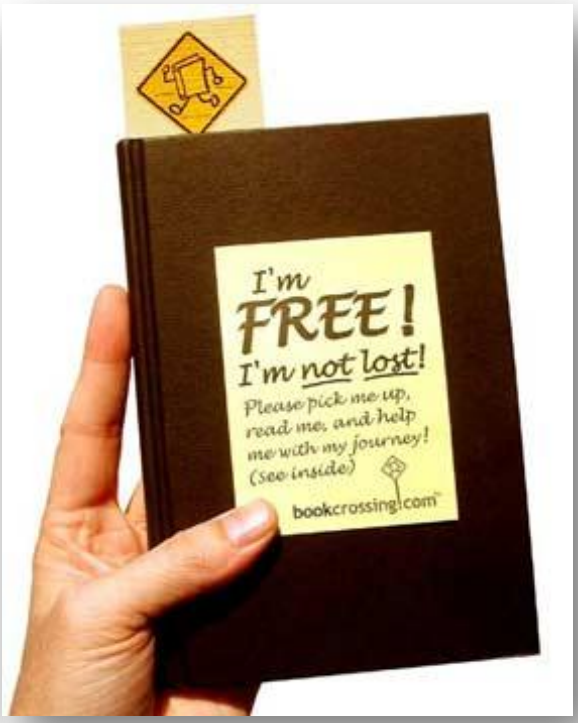
Group A3

Anurag Sharma
Shashvat Rai
Siddhartha Chatterji
Siddharth Raman Singh
Nitesh Batra
Sandip Chaudhuri





BookCrossing

Data Mining Group Project

Executive summary

Our Analysis aims at developing a recommendation system for BookCrossing.com – an online book sharing community. Using user level book ratings and basic location / demographic information – we recommend a two pronged recommendation system.

Background and Problem description:

BookCrossing is defined as "**the practice of leaving a book in a public place to be picked up and read by others, who then do likewise**." The term is derived from **bookcrossing.com**, a free online book club which began in order to encourage the practice, aiming to "make the whole world a library."

The exchanging of books may take any of a number of forms, including 'releasing' books in a public place, direct swaps with other members of the website, or "book rings" in which books travel in a set order to participants who want to read a certain book. There are currently **901,229** BookCrossers and **6,725,334** books travelling through **132** countries

Essentially each BookCrossing member

- Registers on the website, providing location, name and age.
- Picks up a book, reads it and assigns it a rating on the website
- Releases the book to the next user

BookCrossing.com has a great deal of data on book ratings across genres and geographies and basic demographic / geographic information (location and age). At present it has no recommendation systems in place (other than a naïve rule based on the number of users reviewing a particular book and average rating).

Our aim was to use BookCrossing data to suggest a recommendation system for new and existing members of the site.

Data: Sources and Challenges

Our data was obtained from a trawl of the BookCrossing website over a four week period by Carl-Nicolas Zeigler of the University of Freiburg. The raw data consisted of User Id's, ISBN Number, book ratings, location and age for 278,000 users across 10 countries. We however confined ourselves to US data since 30% of BookCrossing members are based in the United States. We also used Amazon.com metadata to obtain genre / category information on the books rated.

Tools/ programming languages used:- SAS, Spotfire, XLMiner, Perl script

The process followed in preparing our data is illustrated in Figure 1 below:

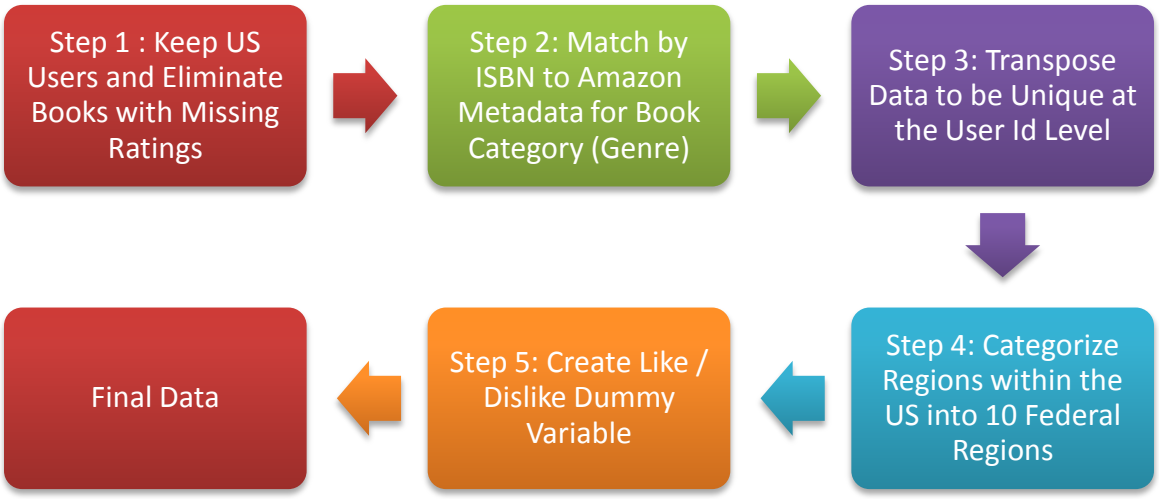


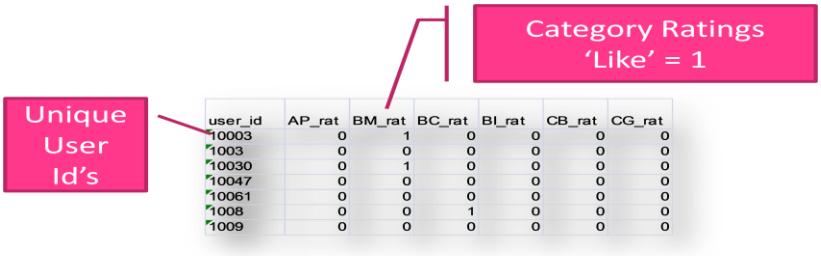
Figure 1: Data Preparation Process Overview

- Step 1: The raw data was comma delimited however some internal fields also had commas within them (for example - book titles), this step of the cleaning process was achieved by running search and replace commands using regular expressions.
- Step 2: This involved collating data for user information (user ID, state and age), books information (book ISBN and name) and ratings data (book ISBN, user ID and book rating) into a common data file. This involved the following set of operations:-
  - The data from Amazon was in the xml format and a "**parser**" was written in **Perl** to pullout primary category and Amazon sales rank for each book (refer **Appendix 1**)
  - A "left inner join" operation was then performed on the resultant tables using **book ISBN** as the primary key to collate the data for ratings data and books information; **user ID** was used as the primary key for the final merge operation with user data.
- Step 3: Transposition of the data was critical for our analysis and involved significant effort and time. We used **SAS**, as excel cannot handle transposition operations for a file with 500k rows, to generate genre-based rating columns (categorical) at individual user level for every user rating.
- Step 4: Dummies were generated for user locations after classifying US states into 10 federal regions.
- Step 5: Each user rates books on a scale of 0-10, with higher scores being better. We went with the assumption that **a person rating a book 5 or higher 'likes' the book**.

Accordingly a dummy variable was defined for like/does not like for every category and individual. Finally, we binned continuous variables age into 4 age bins.

- Step 6: The Final data contained **48,134** ratings by **9998** unique users across the US.
- Some data snapshots are presented in **Appendix 2**

The final form in which data was obtain is presented in Figure 2 below:



user_id	AP_rat	BM_rat	BC_rat	BI_rat	CB_rat	CG_rat
10003	0	1	0	0	0	0
1003	0	0	0	0	0	0
10030	0	1	0	0	0	0
10047	0	0	0	0	0	0
10061	0	0	0	0	0	0
1008	0	0	1	0	0	0
1009	0	0	0	0	0	0

Figure 2: Final Data Format (State, Age not shown)

Findings from the application of data mining models

Since our final data is entirely categorical; hence we applied the following techniques.

- **Association Rules**, given the binary matrix form of the data and using the 'like' dummies created for each category. This was to capture category affinity at an individual level without using location and age information.
- **A Naïve Bayes classification** for each category / genre, using location and binned age as predictors.
- An assumption we made to simplify the analysis is that individual title preferences are indicative of genre / category preferences.

Association Rules

Table 1 : Association Rules

Rule #	Conf. %	Antecedent (a)	Consequent (c)	Support(a)	Support(c)	Support(a U c)	Lift Ratio
1	56.99	BM_rat, LNF_rat=>	BC_rat	479	1587	273	3.586978
2	56.66	MNT_rat, NOF_rat=>	BC_rat	383	1587	217	3.565846
3	50.53	CB_rat, LNF_rat=>	BC_rat	469	1587	237	3.180366
4	73.2	BC_rat, ROM_rat=>	MNT_rat	306	3126	224	2.338924
5	90.04	HOR_rat, LNF_rat=>	MNT_rat	241	3940	217	2.282575
6	67.69	SFF_rat, ROM_rat=>	MNT_rat	588	3126	398	2.162695

BM	Biography & Memoirs
LNF	Literature and Fiction
NOF	Non Fiction
CB	Childrens Books
BC	Book Club
MNT	Mystery and Thriller
ROM	Romance
RNS	Religion and Spirituality
SFF	Science Fiction & Fantasy

- We retained association rules with a lift over benchmark greater than 2. The confidence cutoff was set to 50%
- Some combinations recommended are intuitive , for instance Horror and Mystery / Thrillers
- Others are not so intuitive, Biographies and Book Clubs or Romances and Mystery / Thrillers

Naïve Bayes Classification

We used a Naïve Bayes model using the 'like' dummy for a category as the dependent variable – and location / age as explanatory variables. Results for two categories are presented below (Table 2)

Table 2: Naïve Bayes Results for Two Categories

Mystery and Thrillers

	Classes-->			
Input Variables	1		0	
	Value	Prob	Value	Prob
State	0	0.982613573	0	0.975173421
Number_1	1	0.017386427	1	0.024826579
State	0	0.9326977	0	0.950711939
Number_2	1	0.067302299	1	0.049288061
State	0	0.913628716	0	0.939028843
Number_3	1	0.086371284	1	0.060971157
State	0	0.895120583	0	0.908360716
Number_4	1	0.104879417	1	0.091639284
State	0	0.615255188	0	0.546184739
Number_5	1	0.384744812	1	0.453815261
State	0	0.936623668	0	0.921625898
Number_6	1	0.063376332	1	0.078374102
State	0	0.94615816	0	0.950468541
Number_7	1	0.05384184	1	0.049531459
State	0	0.971396523	0	0.977120604
Number_8	1	0.028603477	1	0.022879396
State	0	0.851374089	0	0.886698308
Number_9	1	0.148625911	1	0.113301692
State	0	0.9551318	0	0.944626993
Number_10	1	0.0448682	1	0.055373007
Binned_Age	1	0.206954571	1	0.31118413
	2	0.198541784	2	0.282828283
	3	0.212563096	3	0.173420957
	4	0.38194055	4	0.23256663

Travel

	Classes-->			
Input Variables	1		0	
	Value	Prob	Value	Prob
State	0	0.985915493	0	0.976432672
Number_1	1	0.014084507	1	0.023567328
State	0	0.971830986	0	0.947326015
Number_2	1	0.028169014	1	0.052673985
State	0	0.830985915	0	0.935240205
Number_3	1	0.169014084	1	0.064759795
State	0	0.929577465	0	0.905831403
Number_4	1	0.070422535	1	0.094168597
State	0	0.52112676	0	0.558767247
Number_5	1	0.478873239	1	0.441232753
State	0	0.943661972	0	0.924161547
Number_6	1	0.056338028	1	0.075838453
State	0	0.957746479	0	0.949642461
Number_7	1	0.042253521	1	0.050357539
State	0	1	0	0.975929097
Number_8	1	0	1	0.024070903
State	0	0.873239436	0	0.880451204
Number_9	1	0.126760563	1	0.119548796
State	0	0.985915493	0	0.946218149
Number_10	1	0.014084507	1	0.053781851
Binned_Age	1	0.112676056	1	0.293886595
	2	0.338028169	2	0.267297814
	3	0.267605634	3	0.179776413
	4	0.281690141	4	0.259039178

Age		
Intervals value		#records
From	To	
13	27	8196
28	34	7343
35	45	7790
46	103	7642

- Our dependent variables are the 'Like' dummy for Mystery and Thrillers and Travel
- For Mystery and Thrillers, Age Bin 4 (Older people) have a higher probability to like the genre.
- For Travel books, Age Bin 2 (28 to 34) show a higher probability to like the genre.

- Federal regions 5 and 9 have the maximum concentration of data, no other significant regional trends were observed.
- Significant lift over random obtains from the Naïve Bayes rules (Figure 3)

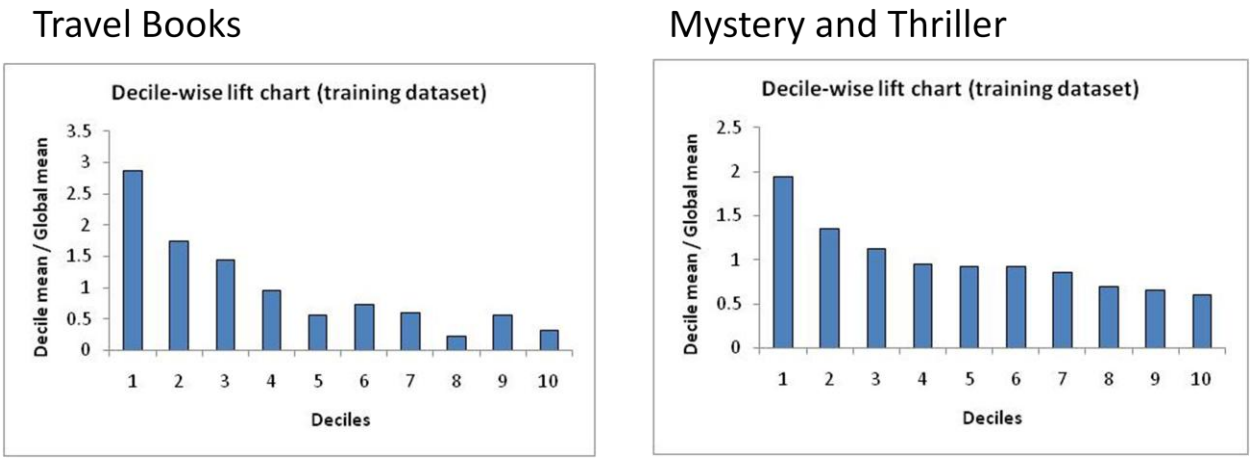


Figure 3: Lift from Naïve Bayes Rules for Two Sample Categories

Conclusion / Applications

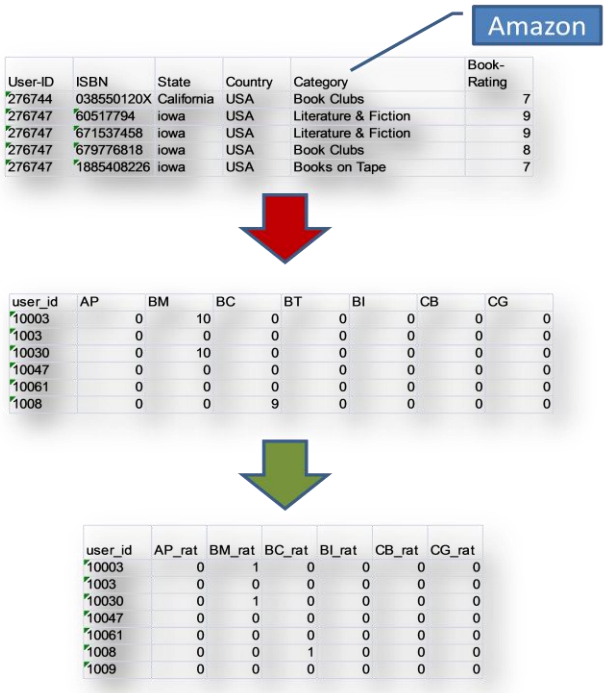
We propose a two – tier recommendation system for BookCrossing.com based on the following

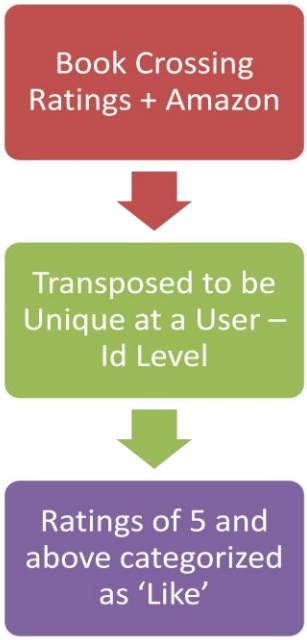
- When a user registers, use the Naïve Bayes results to propose genre recommendations based on demographics.
- Once the user reveals his genre preferences via assigned ratings, use affinity rules to propose more refined recommendations.

Future improvements

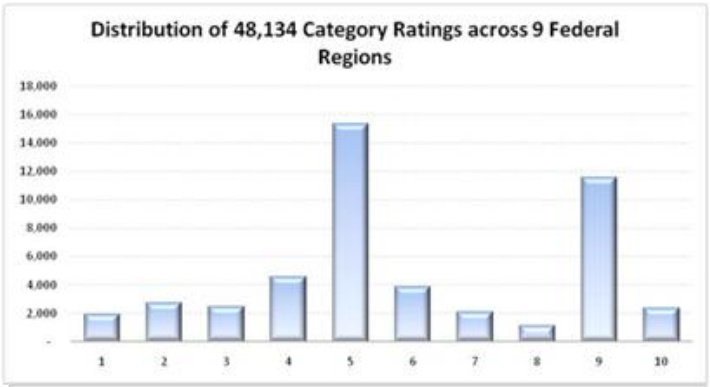
- An obvious extension of our analysis is to incorporate data from geographies other than the US.
- Further analysis can comprise of extending and increasing the level of granularity in recommendation system. For example, moving it to "book title" level, refining the age bins to gain deeper insights and introducing more levels of rating bins than the current system of like & dislike.

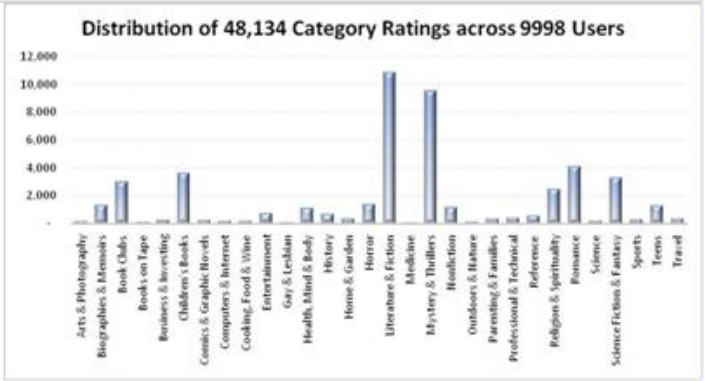
Appendix 1: Data Transformation





Appendix 2





- Federal Regions 5 and 9 dominate for book crossing ratings
- Some genres clearly dominate
- Nature of data suggested the use of association rules