BIDM Group Project

Group members:
Anurag Sharma
Nitesh Batra
Sandip Chaudhary
Siddharth Chatterji
Siddharth Singh

Data Sources

- Book-crossing: The phenomenon.
- Dataset: Sourced through University of Freiburg, Computer Science Department
  - Collected by Carl-Nicolas Zeigler by crawling book crossing community website for 4 weeks in Aug 2004
  - Raw data:
    - 278K+ users,271K+ book details
    - 1.14M+ book ratings
- Cross referenced by ISBN # to Amazon meta data: Category and sales details for 500K+ books.
- Final Data : 30 categories of books, rated by 9998 users across the United States. Minimal demographic data (age)

Data Preparation

Amazon

User-ID	ISBN	State	Country	Category	Book-Rating
276744	038550120X	California	USA	Book Clubs	7
276747	60517794	iowa	USA	Literature & Fiction	9
276747	671537458	iowa	USA	Literature & Fiction	9
276747	679776818	iowa	USA	Book Clubs	8
276747	1885408226	iowa	USA	Books on Tape	7

Book Crossing Ratings + Amazon

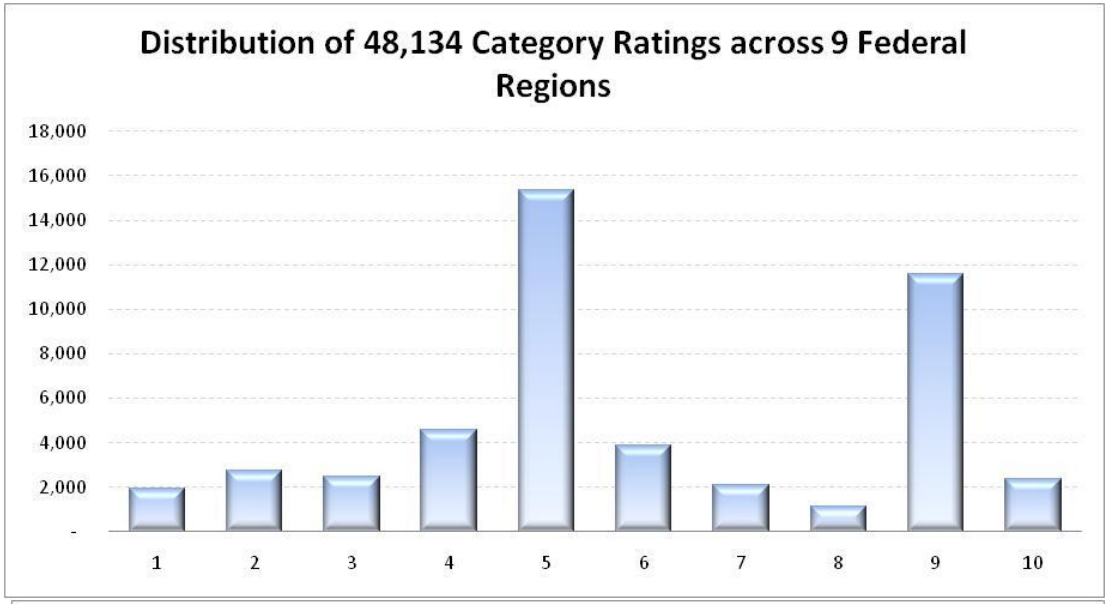
user_id	AP	BM	BC	BT	BI	CB	CG
10003	0	10	0	0	0	0	0
1003	0	0	0	0	0	0	0
10030	0	10	0	0	0	0	0
10047	0	0	0	0	0	0	0
10061	0	0	0	0	0	0	0
1008	0	0	9	0	0	0	0

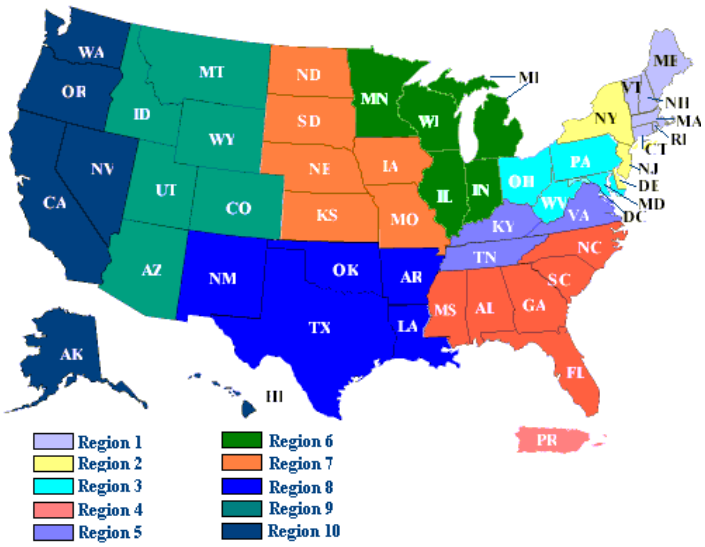
Transposed to be Unique at a User – Id Level

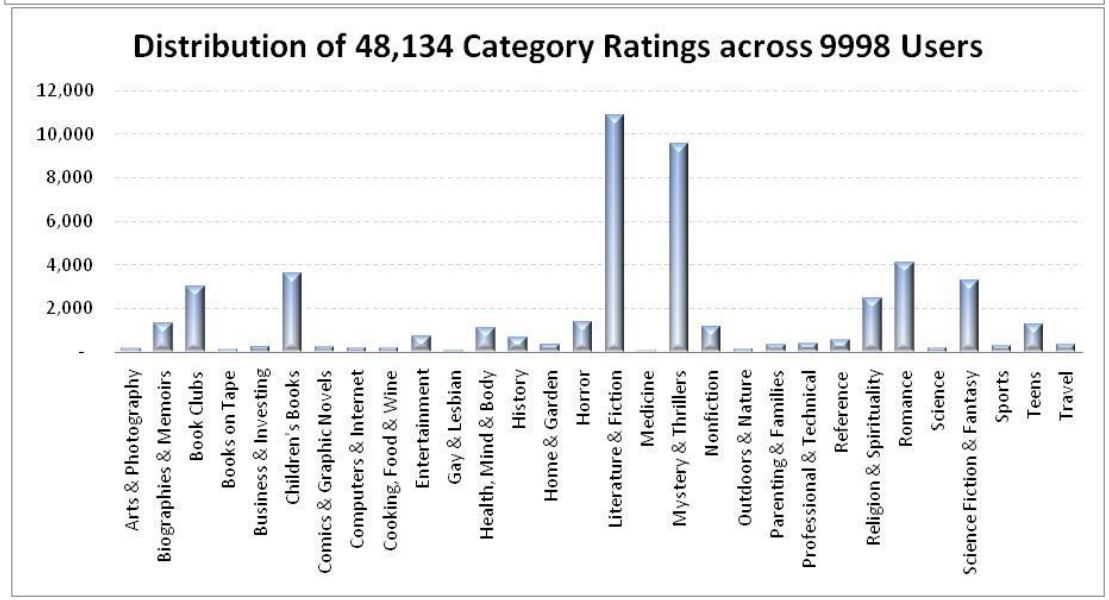
user_id	AP_rat	BM_rat	BC_rat	BI_rat	CB_rat	CG_rat
10003	0	1	0	0	0	0
1003	0	0	0	0	0	0
10030	0	1	0	0	0	0
10047	0	0	0	0	0	0
10061	0	0	0	0	0	0
1008	0	0	1	0	0	0
1009	0	0	0	0	0	0

Ratings of 5 and above categorized as ‘Like’

Data –Preliminary Look







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

Likes : Affinity 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

- Retained Rules with a Lift greater than 2
- Romances and Mystery / Thrillers
- Horror and Mystery / Thrillers
- Science Fiction and Mystery / Thrillers
- Biographies / Memoirs and Book club recommendations

Naïve Bayes

Mystery and Thrillers

	Classes-->			
Input Variables	1		0	
	Value	Prob	Value	Prob
State Number_1	0	0.982613573	0	0.975173421
	1	0.017386427	1	0.024826579
State Number_2	0	0.9326977	0	0.950711939
	1	0.067302299	1	0.049288061
State Number_3	0	0.913628716	0	0.939028843
	1	0.086371284	1	0.060971157
State Number_4	0	0.895120583	0	0.908360716
	1	0.104879417	1	0.091639284
State Number_5	0	0.615255188	0	0.546184739
	1	0.384744812	1	0.453815261
State Number_6	0	0.936623668	0	0.921625898
	1	0.063376332	1	0.078374102
State Number_7	0	0.94615816	0	0.950468541
	1	0.05384184	1	0.049531459
State Number_8	0	0.971396523	0	0.977120604
	1	0.028603477	1	0.022879396
State Number_9	0	0.851374089	0	0.886698308
	1	0.148625911	1	0.113301692
State Number_10	0	0.9551318	0	0.944626993
	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 Number_1	0	0.985915493	0	0.976432672
	1	0.014084507	1	0.023567328
State Number_2	0	0.971830986	0	0.947326015
	1	0.028169014	1	0.052673985
State Number_3	0	0.830985915	0	0.935240205
	1	0.169014084	1	0.064759795
State Number_4	0	0.929577465	0	0.905831403
	1	0.070422535	1	0.094168597
State Number_5	0	0.52112676	0	0.558767247
	1	0.478873239	1	0.441232753
State Number_6	0	0.943661972	0	0.924161547
	1	0.056338028	1	0.075838453
State Number_7	0	0.957746479	0	0.949642461
	1	0.042253521	1	0.050357539
State Number_8	0	1	0	0.975929097
	1	0	1	0.024070903
State Number_9	0	0.873239436	0	0.880451204
	1	0.126760563	1	0.119548796
State Number_10	0	0.985915493	0	0.946218149
	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

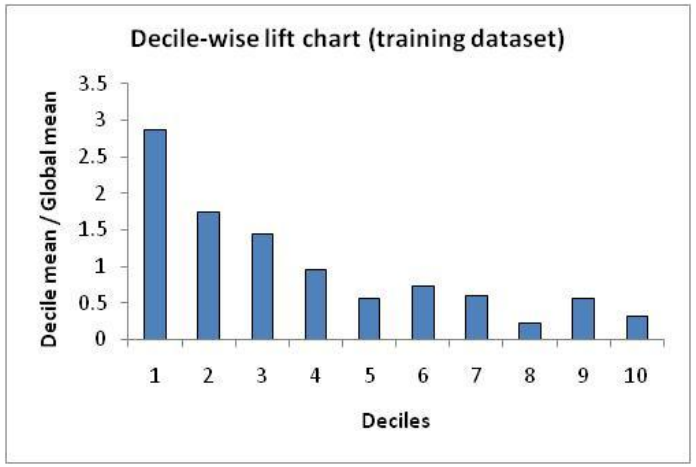
- Older People prefer the Mystery and Thriller Category
- Young Adults prefer the Travel Category
- Zones 5 and 9 have the highest concentration of data

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

Applications

- Recommendation systems for Bookcrossing.com based on Affinity Rules/ Age and State of Origin
- Naïve Bayes Rule itself provides significant lift

Travel Books



Mystery and Thriller