DATA 643 Proj 1

Dan Fanelli

June 5, 2017

Global Baseline Predictors and RMSE

Figure 1:

LastFM Users and Artists: Listen Counts

LastFM provides a data set that gives us counts of how many times a LastFM user has listened to a particular artist. This system should recommend musical artists to users.

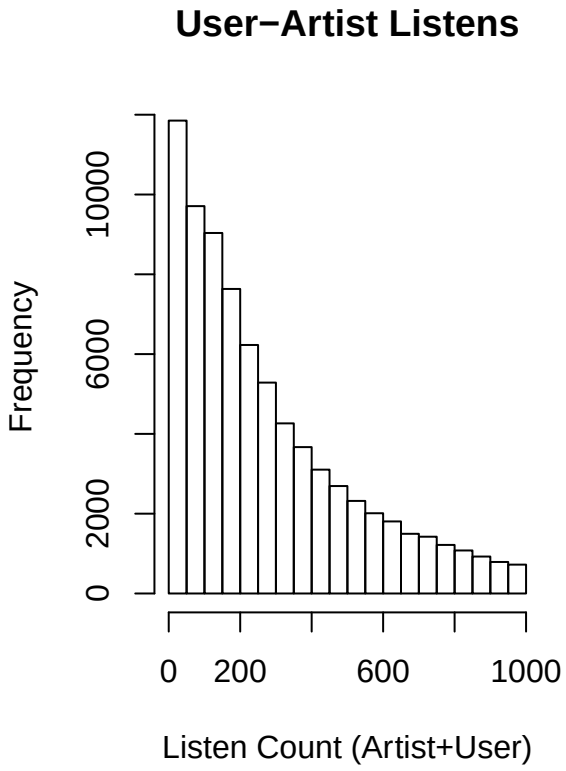
We will assume that a user's "listen count" per artist is analogous to an explicit "rating" of the artist

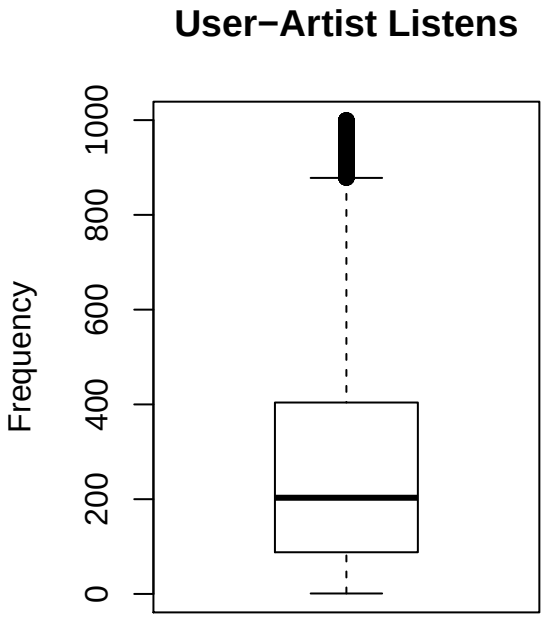
We well exclude listen counts above 1000 for the sake of visualizing the data and avoiding skew via huge listen counts. (Maybe listen counts above 5000 are "bots" or invalid in some other way)

Table 1: A Sample of the Initial Data

	userID	artistID	artist	listen_count
11131	652	233	Nine Inch Nails	3716
76341	450	7054	The Alan Parsons Project	256
490	839	45	Mindless Self Indulgence	706
9240	585	220	Red Hot Chili Peppers	571

	userID	artistID	artist	listen_count
41613	862	1034	Kerli	10623
1201	1835	56	Daft Punk	165
29633	1844	546	The Ting Tings	204
39186	1063	930	Nightwish	47
54713	1133	1941	Gnarls Barkley	70
48619	172	1409	Calvin Harris	439





The Raw average of ALL user-item combination = 277.81

Calculate the RMSE for raw average for both your training data and your test data.

RMSE for Raw Avg:

```
## RMSE (TRAIN) 2.7
```

```
## RMSE (TEST) 5.25
```

Using your training data, calculate the bias for each user and each item

USER BIAS (TRAIN)

```
##    userMean userID user_bias
## 1  99.71429      3 -178.0957
```

```
## 2 285.75000      4    7.9400
## 3 452.00000      5  174.1900
## 4  19.25000      6 -258.5600
## 5 753.00000      7  475.1900
## 6 468.66667      8  190.8567
```

ARTIST BIAS (TRAIN)

```
##   artistMean artistID artist_bias
## 1   212.0000        1   -65.81000
## 2   385.5000        6   107.69000
## 3   341.6667        7    63.85667
## 4   587.0000        8   309.19000
## 5    96.0000        9  -181.81000
## 6   141.0000       12  -136.81000
```

USER BIAS (TEST)

```
##   userMean userID  user_bias
## 1 254.0000      4  -23.81000
## 2 206.3333      5  -71.47667
## 3  17.0000      6 -260.81000
## 4 617.5000      7  339.69000
## 5 383.6667      9  105.85667
## 6 337.0000     11   59.19000
```

ARTIST BIAS (TEST)

```
##   artistMean artistID artist_bias
## 1   134.0000        2  -143.81000
## 2   218.3333        7   -59.47667
## 3   532.0000       10   254.19000
## 4    10.0000       17  -267.81000
## 5    16.0000       27  -261.81000
## 6   198.5000       30   -79.31000
```

From the raw average, and the appropriate user and item biases, calculate the baseline predictors for every user-item combination.

Table 2: Train Data Predictions

	userID	artistID	artist	listen_count	listen_count_PREDICTION
76341	450	7054	The Alan Parsons Project	256	302.1900
490	839	45	Mindless Self Indulgence	706	774.2733
9240	585	220	Red Hot Chili Peppers	571	548.9400
1201	1835	56	Daft Punk	165	218.8400
29633	1844	546	The Ting Tings	204	127.1067
39186	1063	930	Nightwish	47	0.0000

Table 3: Test Data Predictions

	userID	artistID	artist	listen_count	listen_count_PREDICTION
29319	1017	543	Nicole Scherzinger	298	174.1900
690	407	53	Air	46	0.0000
81045	729	9639	End of Green	22	0.0000
3583	941	89	Lady Gaga	459	503.4757
40790	1259	982	Foo Fighters	404	623.3567
73896	365	6048	Godspeed You! Black Emperor	19	0.0000

Calculate the RMSE for the baseline predictors for both your training data and your test data.

- Base RMSE (train) = **2.7**
- OUR RMSE (train) = **1.74**
- OUR IMPROVEMENT (train) = **35.46%**
- Base RMSE (test) = **5.25**
- OUR RMSE (test) = **3.52**
- OUR IMPROVEMENT (test) = **32.95%**

Summarize your results.

Though the final result only shows the evaluation of the full data set, but the steps leading up to it were as follows:

- When using train and test sets of **80** and **20**, respectively, the train improvement was **16.35%** and the test improvement was **10.05%**
- When using train and test sets of **800** and **200**, respectively, the train improvement was **26.72%** and the test improvement was **24.91%**
- When using train and test sets of **8000** and **2000**, respectively, the train improvement was **35.07%** and the test improvement was **31.38%**

It seems safe to say that as the data sets grow, the gains in using this strategy also increase.