Predicting radio listening habits

Gandalf Hernandez
STAT 4240 – Data Mining
Lecturer Christopher Volinsky
Why?

• Big music fan
• Fun to know how people listen to music and to what kind
• Work for a streaming music company
  • That has a radio component
• Idea
  • Predict if a user will hit the skip button on a song in the radio
  • Use as proxy for if the user liked the song
How?

• Predict a song skip by gathering data on
  • Age, sex, country, hour of the day
  • popularity of artist, popularity of song, rank of song within the artist, year the song was released
  • Difference in age between song and user
  • Which radio station they are in
  • If the user has listened to the artist and/or song before
  • If the user pressed next on the last song
  • How much the user skips on average
Data

- Have access to all the data! This should be easy!
  - Weighted random sample of 2,000 users

- Hadoop
  - User history
  - Listens, skips & country
  - Brute force – 40 machines
  - Analyzed 7000 GB of data
  - Map-reduce

- PostgreSQL
  - Two of them
  - Album, artist, song info
  - Age, sex
  - SQL interface

- Some Python scripting later and a couple of hours of waiting
  - ~ 450,000 radio listens over two weeks

Tuesday, December 6, 2011
Radio stations

Relative radio station popularity and skip ratios

Skip ratio

popularity
skip-ratio

multi
pop
hip hop
dance
country
80s
Indie
rock
mb
alternative
electronic
latin
classical
reggae
heavy metal
jazz
singer songwriter
soul
60s
70s
blues
trance
hardcore
emo
punk
death metal
black metal
folk
Sex and country

Skip ratio by sex

Skip ratio by country

Tuesday, December 6, 2011
Age groups

Skip ratio by age

<table>
<thead>
<tr>
<th>Age Group</th>
<th>Skip Ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>10-19</td>
<td>1.0</td>
</tr>
<tr>
<td>20-29</td>
<td>0.7</td>
</tr>
<tr>
<td>30-39</td>
<td>0.6</td>
</tr>
<tr>
<td>40-49</td>
<td>0.5</td>
</tr>
<tr>
<td>50-59</td>
<td>0.4</td>
</tr>
<tr>
<td>60-69</td>
<td>0.3</td>
</tr>
<tr>
<td>70-79</td>
<td>0.2</td>
</tr>
<tr>
<td>80-89</td>
<td>0.1</td>
</tr>
<tr>
<td>90-99</td>
<td>0.0</td>
</tr>
</tbody>
</table>
Hourly variance

- Could also look at weekday vs. weekend
Popularity

• Small effect of
  • Overall artist popularity and overall song popularity
    • Not totally unexpected
  • Song rank within artist
    • Big hits have minor effect?
Recording age effects

• 1950s least skipped
• 1970s–1980s most skipped
  • May be because shows up in many channels and more variety
  • 1990s onward also show high skip ratios

• Age of listener – age of recording
  • Recordings younger or same age as listener have higher skip ratios
Previous listening

Heard artist before?

Heard song before?
Past behavior

Skipped last song?

Skipped last

Listened to last
Analysis

- Logistic regression with everything
  - Simple 90% training, 10% test
- Decision tree
  - 10% training, 10% test (rpart didn’t like 420K rows)
- SVM
  - Trained on around 5% of data, 10%
- Ensemble – simple average
- Specificity – how many skips did I get right?
- Sensitivity – how many listens did I get right?

<table>
<thead>
<tr>
<th></th>
<th>Specificity</th>
<th>Sensitivity</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Log regression</td>
<td>0.73</td>
<td>0.96</td>
<td>0.90</td>
</tr>
<tr>
<td>Decision tree</td>
<td>0.87</td>
<td>0.94</td>
<td>0.92</td>
</tr>
<tr>
<td>SVM</td>
<td>0.95</td>
<td>0.84</td>
<td>0.92</td>
</tr>
<tr>
<td>Ensemble</td>
<td>0.94</td>
<td>0.89</td>
<td>0.93</td>
</tr>
</tbody>
</table>
Analysis

Combined ROC plot

sensitivity vs. 1-specificity

- logit
- tree
- svm
- ensemble

Tuesday, December 6, 2011
Preliminaries

• Characteristics can help predict skips pretty accurately
• Past behavior is an excellent predictor of the future
• Data I would really like to have
  • Artist, song similarities

• Still to do
  • More advanced ensemble
  • Possibly neutral network, random forest
  • Better model fitting (cross validation of logistic regression, tree sizing)
  • Etc., etc. 😊

• Thank you very much – Questions?