Active Editors

- An editor with 5+ edits in a main namespace
- 29% of those edited did 5+ edits (2013)
- Across wp projects, 87% of the monthly content was generated by active editors (March 2013)
Goal

- Predict active editor survival
  - Maintain the system for those predicted to survive
  - Apply interventions on those predicted to not survive
Data and Prediction Model

- (userId, edit date, number of edits, registration date)
- 12,500,726 editor records for enwp
- February 2001 to May 2014
- Sampled 10% of the data
  - 80% for training
  - 20% for testing
- Classification trees [GUIDE]
Variables for $n$-month Survival Model

- Outcome: 1, if an active editor in month $m$ is active in month $m+n$; 0, otherwise.

- Independent Variables (11)
  - Number of edits (in the registration month, current month, since registration)
  - Commitment
    - maximum length of time in months during which the editor stayed active
    - number of months between current activity month and the last time the editor was active
    - proportion of months since registration in which the editor was active
    - Number of months during which the editor stayed active
  - Registration and vintage (registration day and month, number of months since registration)
  - First time editor
ROC Curves for 1-year Survival
## Variable Importance Ranking

<table>
<thead>
<tr>
<th>Variables</th>
<th>30-day</th>
<th>1-year</th>
</tr>
</thead>
<tbody>
<tr>
<td>total number of edits since registration</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>proportion of months active</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>number of months active</td>
<td>4</td>
<td>3</td>
</tr>
<tr>
<td>first time active editor</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>maximum continuity</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>number of months since registration</td>
<td>6</td>
<td>6</td>
</tr>
<tr>
<td>number of edits in registration month</td>
<td>7</td>
<td>7</td>
</tr>
<tr>
<td>number of edits in current month</td>
<td>8</td>
<td>8</td>
</tr>
<tr>
<td>number of months since last time active</td>
<td>9</td>
<td>9</td>
</tr>
<tr>
<td>registration day</td>
<td>10</td>
<td>10</td>
</tr>
<tr>
<td>registration month</td>
<td>11</td>
<td>11</td>
</tr>
</tbody>
</table>
Discussion

• Few independent variables, good results

• Next steps
  – Add more independent variables
  – Expand to other projects
  – Other prediction algorithms for performance comparison