Big data killed the click
and many other unsung metric heroes

dstillery

Claudia Perlich
Chief Scientist
@claudia_perlich
What is a metric?

“A process of measuring something”
Are the metrics you are tracking helping you make decisions?
Metric hardships ...

The things we measure are usually not the things we care about ... but they are what we CAN measure

- Unobservable
- Long time before outcome
Better decision using metric?

Option A

Metric > Metric

Option B
Old Hero of Advertising: Click

Measure interest of consumers in the product
Measuring vs. Optimization vs. Incentives ...
Predictive Modeling & Learning from Data

If you want me to look good along some metric, would it not be great if I could predict it and act based on predictions?

Predict who will click on the ad and only show ads to those people with high probability

For now: let's try to identify females on Facebook instead …
Experiments: Predicting Facebook Gender (female)

Data:
Facebook public dataset with 200K anonymized users, their demographics and their likes

Methodology:
Logistic regression on sparse representation

60% Women
Predict female gender based on age only

Overall Accuracy: 60%

Accuracy in top 1%: 75%
Signal & Noise in predictive modeling
Distribution of predictions in case of a lot of noise ...

Baserate: 60%
Overall Accuracy: 60%
AUC: 58%

Accuracy in top 1%: 75%
Distribution of predictions in case of a lot of signal ...

Overall Accuracy: 83%

Accuracy in top 1%: 100%
Welcome Big Data: adding more and more ‘likes’

75%  86%  100%

100%  100%  100%
Quick summary so far

• ‘signal to noise’ changes the distribution of the probabilities: more signal, more spread

• High accuracy in the top 1% even if the overall accuracy is not great

• All in all this is good news, no?
Not all clicks are made equal …

The majority of clicks happen accidentally, very few lead to conversion

• Click rate is pretty much constant across campaigns

• Post click conversion is easily one to two orders of magnitude less common
Old days of the Click Metric …

Strategy 1: target women

Strategy 2: target men

intentional > accidental
Optimization days of big data and predicting clicks...

Model learns to predict the accidental much more easily than the intentional ones ...

I will only target clicks that are likely to happen accidentally but not randomly ...
Signal is mostly in the context = accidentals

Top 10 Apps by CTR

<table>
<thead>
<tr>
<th>App Category</th>
<th>CTR App / Avg CTR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Game</td>
<td>5</td>
</tr>
<tr>
<td>Game</td>
<td>5</td>
</tr>
<tr>
<td>Game</td>
<td>5</td>
</tr>
<tr>
<td>Flashlight</td>
<td>6</td>
</tr>
<tr>
<td>Social Video</td>
<td>6</td>
</tr>
<tr>
<td>Game</td>
<td>7</td>
</tr>
<tr>
<td>Game</td>
<td>7</td>
</tr>
<tr>
<td>OnlinePub</td>
<td>7</td>
</tr>
<tr>
<td>Flightlight</td>
<td>9</td>
</tr>
<tr>
<td>Game</td>
<td>8</td>
</tr>
</tbody>
</table>
Optimizing on Clicks = targeting at random …

Results across 70 Dstillery campaigns

AUC = 0.5 is random
Another victim: Post view site visit rate

Performance Index

2 weeks in 2012

2x
Bots are coming to conversion events

- ‘Cookie Stuffing’ increases the value of the ad for retargeting
- Messing up Web analytics ...
- Messes up my models because a botnet is easier to predict than a human
Bot activity has more signal – higher predictions

Humans are hard
to predict …
Eliminate labels generated by bots ....

3 more weeks in spring 2012
Silver Lining: it works both ways …
Can you target consumers in the market for upscale car and about to go to a car dealership?
Can you predict who will go to a car dealership using all their digital information?

How would you even know who goes to car dealership?
Challenges: Location data & cross walk accuracy

• Uncertainty about the dealer location
• Superman effect
• People piles ...
• Association between devices is probabilistic
How much random noise can a model absorb?

We will randomly switch the gender value for increasing percentages

```
00000d41ed774823fca142945ec915c0,0,...........,en_GB,  
00000dee02d70cf8c0d79f96b6d1c59d,0,...........,en_US,  
00000f232abfe25a80156fe069395460,0,1992,20,2,2,........,19,-5  
00000f4ba0cff946b1c0e3b051287ede,0,1993,19,2,........,en_US,310,8  
000130571654e3afaa62f49d2e4f63,0,2,2,........,en_US,193,7  
0001544469ae9b408869a463a1dd77a,1,1984,28,2,........,en_US,,-4
```
What will happen if I randomly assign labels?

Percent Women in top 1%

Percent original labels

50% original labels

25% original labels
Thank You!

@Claudia_Perlich

Chief Scientist

Thank You!