Data Science at eHarmony:

A generalized framework for personalization

by Jon Morra, Ph.D.
EHARMONY - WHO WE ARE

SOFTWARE ENGINEERS
DATA SCIENTISTS
PSYCHOLOGISTS
PRODUCT SPECIALISTS
MARKETING
CUSTOMER CARE
WHAT IS DATA SCIENCE?

Data Scientist (n.):

Person who is better at statistics than any software engineer and better at software engineering than any statistician.

- Josh Wills
ENGINEER VS SCIENTIST

Engineering Concerns

- Speed
- Scalability
- Reporting

Scientist Concerns

- Accuracy
- Precision
- Recall
- Reporting
CAN BOTH BE HAPPY TOGETHER?
Model\([A, B]\] \equiv f(A) \Rightarrow B
Aloha
A scala-based feature generation and modeling framework

Get Started »

http://eharmony.github.io/aloha
Training Data [A] → F(A) => L

VW | H2O | R | Sklearn...

Framework Specific Model [L, C]

Aloha Model [A, B]

Finalizer [C, B]

Output [B]

Specific Example [A]

Aloha Model [A, B]
ALOHA FEATURE SPECIFICATION

```json
{
  "imports": [
    "com.eharmony.aloha.feature.BasicFunctions._",
    "scala.math._"
  ],
  "features": [
    {
      "name": "male_education",
      "spec": "${male.education}"  
    },
    {
      "name": "female_education",
      "spec": "${female.education}"  
    },
    {
      "name": "height_distance",
      "spec": "${male.height} - ${female.height}"  
    }
  ],
  "label": "if(${comm_7d_c:-false} && ${comm_7d_u:-false}) 1 else -1"
}
```
HYPER PARAMETER OPTIMIZATION

GLM Parameters
• Learning rate
• L1 Regularization
• L2 Regularization
• Interactions

GBM Parameters
• Learning rate
• Number of trees
• Max tree depth
• Number of bins per split

Current Tools
• Hyperopt
• Auto-Weka
• Spearmint
HYPER PARAMETER OPTIMIZATION

https://github.com/eHarmony/spotz
MODEL TRAINER

- MatchProto, Communication?
  - Model[MatchProto, Double]
- MatchProto, What If Comm?
  - Model[MatchProto, Double]
- [UserProto, Match Goal], Comm?
  - Model[UserProto, Integer]
- JobDyadProto, Click?
  - Model[JobDyadProto, Double]
- UserProto, Fraud?
  - Model[UserProto, Boolean]
MatchProto, Communication?  Model[MatchProto, Double]
MatchProto, What If Comm?  Model[MatchProto, Double]
[UserProto, Match Goal], Comm?  Model[UserProto, Integer]
JobDyadProto, Click?  Model[JobDyadProto, Double]
UserProto, Fraud?  Model[UserProto, Boolean]
MATCHING

1. COMPATIBILITY MATCHING
2. AFFINITY MATCHING
3. MATCH DISTRIBUTION
One way communications for matches delivered to females

- Day 0
- Day 1
- Day 2
- Day 3
- Day 4
- Day 5
- Day 6

Dates:
- 5/1/16
- 5/2/16
- 5/3/16
- 5/4/16
- 5/5/16
- 5/6/16
- 5/7/16
- 5/8/16
- 5/9/16
- 5/10/16
- 5/11/16
- 5/12/16
- 5/13/16
- 5/14/16

% Matches that result in communication

Sunday
AFFINITY MEASUREMENTS

One-way communications for matches delivered to females

Date

% Matches that result in communication

DAY_6_COMM_PCT
DAY_5_COMM_PCT
DAY_4_COMM_PCT
DAY_3_COMM_PCT
DAY_2_COMM_PCT
DAY_1_COMM_PCT
DAY_0_COMM_PCT
AFFINITY ASIDES

Old Affinity: Comm
New Affinity: (Comm || View) && !Close

\[ P(\text{Comm}) \approx 0.0175 \]
\[ P(\text{Comm} | \text{Match}) \]
\[ P(\text{Affinity}) \approx 0.175 \]
One-way affinity for matches delivered to females

DAY_6_AFFINITY_PCT
DAY_5_AFFINITY_PCT
DAY_4_AFFINITY_PCT
DAY_3_AFFINITY_PCT
DAY_2_AFFINITY_PCT
DAY_1_AFFINITY_PCT
DAY_0_AFFINITY_PCT
AFFINITY MEASUREMENTS

One way communications for matches delivered to females

% Matches that result in communication

Date


DAY_6_COMM_PCT
DAY_5_COMM_PCT
DAY_4_COMM_PCT
DAY_3_COMM_PCT
DAY_2_COMM_PCT
DAY_1_COMM_PCT
DAY_0_COMM_PCT
AFFINITY MEASUREMENTS

One way communications for matches delivered to females

Date

% Matches that result in communication

DAY_6_COMM_PCT
DAY_5_COMM_PCT
DAY_4_COMM_PCT
DAY_3_COMM_PCT
DAY_2_COMM_PCT
DAY_1_COMM_PCT
DAY_0_COMM_PCT

0.00% 0.50% 1.00% 1.50% 2.00% 2.50%
Implicit vs. Explicit

- Many have acknowledged that implicit feedback is more useful
- Is implicit feedback really always more useful?
- If so, why?

Xavier Amatriain, VP Engineering at Quora
10 more lessons learned from building Machine Learning systems
MatchProto, Communication? \(\rightarrow\) Model[MatchProto, Double]
MatchProto, What If Comm? \(\rightarrow\) Model[MatchProto, Double]
[UserProto, Match Goal], Comm? \(\rightarrow\) Model[UserProto, Integer]
JobDyadProto, Click? \(\rightarrow\) Model[JobDyadProto, Double]
UserProto, Fraud? \(\rightarrow\) Model[UserProto, Boolean]
MATCH DISTRIBUTION - PROBLEM

(L_1,0) → (1,-A_1,1) → (1,-A_1,2) → (L_N+1,0)

(L_2,0) → (1,-A_2,M) → (1,-A_N,2) → (L_N+2,0)

(L_N,0) → (1,-A_N,M) → (1,-A_N,2) → (L_N+M,0)
MATCH DISTRIBUTION - SOLUTION

(L_1, 0) -> (1, -A_{1,1}) -> (L_{N+1}, 0)

(L_2, 0) -> (1, -A_{N,2})

(L_N, 0) -> (1, -A_{N,2})

(L_{N+2}, 0) -> (L_{N+M}, 0)
MATCH GOAL: An amount of matches is assigned

8 MATCHES  6 MATCHES  10 MATCHES

Can we do better?
HISTORY - CONTEXTUAL BANDITS

set of contexts available at time $t$

$$D_t \subset \mathbb{R}^d$$

one context is presented at time $t$

$$x_t \in D_t$$

action is chosen and reward is observed

$$r_t = f(x_t) + \epsilon_t$$

best action/context at time $t$

$$x_t^* = \arg\max_{x \in D_t} f(x)$$

MATCH GOALS MODELS — RESULTS

**Outbound Communication Before Test**

- Control Comms
- Treatment Comms

**Outbound Communication During Test**

- Control Comms
- Treatment Comms

<table>
<thead>
<tr>
<th>Date</th>
<th>Control Comms</th>
<th>Treatment Comms</th>
</tr>
</thead>
<tbody>
<tr>
<td>1/1/15</td>
<td>0.27535</td>
<td>0.23222</td>
</tr>
<tr>
<td>1/4/15</td>
<td>0.34</td>
<td>0.29</td>
</tr>
<tr>
<td>1/7/15</td>
<td>0.295</td>
<td>0.34</td>
</tr>
<tr>
<td>1/10/15</td>
<td>0.34</td>
<td>0.34</td>
</tr>
<tr>
<td>1/13/15</td>
<td>0.24</td>
<td>0.24</td>
</tr>
</tbody>
</table>

**T Score**

<table>
<thead>
<tr>
<th>T Score</th>
<th>P value @ 0.01</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.21901</td>
<td>0.82665</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Date</th>
<th>Control Comms</th>
<th>Treatment Comms</th>
</tr>
</thead>
<tbody>
<tr>
<td>1/24/15</td>
<td>0.23222</td>
<td>0.23782</td>
</tr>
<tr>
<td>1/27/15</td>
<td>0.24</td>
<td>0.29</td>
</tr>
<tr>
<td>1/30/15</td>
<td>0.34</td>
<td>0.34</td>
</tr>
<tr>
<td>2/2/15</td>
<td>0.34</td>
<td>0.34</td>
</tr>
<tr>
<td>2/5/15</td>
<td>0.25</td>
<td>0.24</td>
</tr>
<tr>
<td>2/8/15</td>
<td>0.24</td>
<td>0.24</td>
</tr>
</tbody>
</table>

**T Score**

<table>
<thead>
<tr>
<th>T Score</th>
<th>P value @ 0.01</th>
</tr>
</thead>
<tbody>
<tr>
<td>6.94362</td>
<td>&lt; 1e-5</td>
</tr>
</tbody>
</table>
MATCH GOAL MODEL — RESULTS

**PDF of Match Goal**

- Treatment PDF
- Control PDF

**Comm vs Match Goal**

- Treatment Comms
- Control Comms
• Best Affinity Distribution Function
• Communication Recommendations
• Meet Dave
• Empathetic
• High Income
• Tall
• Dave is a great catch

• Meet Steve
• Selfish
• Medium Income
• Short
• Steve is a mediocre catch

Does Dave deserve a better experience than Steve?
ASIDE 2

- “Are you a passionate person?”
- “What’s your philosophy on travel?”
- “How would you assess your verbal intimacy skills?”

Problem: Which question should the system suggest to maximize the P(resp.)
Solution: Contextual bandit learning!
Wrinkle: The set of questions can change over time due to business reasons
Contextual Bandits
- Specify actions ahead of time
- Build N regressors under the hood

Contextual Bandits with Action Dependent Features
- Featurize Actions
- Learn 1 model which is dependent on features of the actions not the actions themselves
- Actions can change over time

“What’s your philosophy on travel?”

“I love to travel! I’m so glad you asked that question, it’s like you used machine learning to predict what I’m most likely to respond to!”

https://github.com/JohnLangford/vowpal_wabbit
Model Trainer

Spark

H2O

Aloha

MatchProto, Communication?  \rightarrow Model[MatchProto, Double]
MatchProto, What If Comm?  \rightarrow Model[MatchProto, Double]
[UserProto, Match Goal], Comm?  \rightarrow Model[UserProto, Integer]
JobDyadProto, Click?  \rightarrow Model[JobDyadProto, Double]
UserProto, Fraud?  \rightarrow Model[UserProto, Boolean]
ELEVATED CAREERS

https://www.elevatedcareers.com
THE BOTTOM LINE

E-HARMONY COUPLES ARE HAPPIER*

<table>
<thead>
<tr>
<th></th>
<th>E-HARMONY</th>
<th>OTHER DATING SITES</th>
<th>OTHER ONLINE</th>
<th>ALL OTHER</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>WE'RE IN LOVE</em></td>
<td>6.5</td>
<td>6.2</td>
<td>6.0</td>
<td>6.0</td>
</tr>
<tr>
<td><em>WE TRUST EACH OTHER</em></td>
<td>6.4</td>
<td>6.1</td>
<td>6.1</td>
<td>6.1</td>
</tr>
<tr>
<td><em>WE ARE HAPPY</em></td>
<td>6.4</td>
<td>6.2</td>
<td>6.1</td>
<td>6.1</td>
</tr>
<tr>
<td><em>WE ARE COMPATIBLE</em></td>
<td>6.3</td>
<td>6.0</td>
<td>6.1</td>
<td>6.0</td>
</tr>
<tr>
<td><em>WE HAVE CHEMISTRY</em></td>
<td>6.3</td>
<td>6.0</td>
<td>6.0</td>
<td>6.1</td>
</tr>
</tbody>
</table>


19,131 marriages between 2005 & 2012 were surveyed.
An average of 575 people get married every day through eHarmony.
love begins here