Monkeys & Math

How MailChimp Catches Bad Guys

John Foreman, Data Scientist
MailChimp.com
Who am I?

OR guy

Former consultant

MailChimp data scientist
Email Marketing
Conventional Wisdom
Data
Let’s Go Freemium!

3 million users
1.5 billion email addresses
35 billion emails sent
Get a Goal

Choose tools and design that decrease complexity and cost/schedule variance
Our Goal: Compliance
Cheaper, Faster, Better
Our Goal: Compliance
Cheaper, Faster, Better

❓ ✔
Permission and Hygiene

Translated into Two Measurable Response Variables:

Public Correlation
Hard Bounce Rate
Hard Bounces must be predicted

Need an abundance of features / weak learners:

- Earthlink email addresses are 10x more likely to hard bounce than Gmail
- Users whose list location does not match their billing or login locations have a 1% higher bounce rate
Lists

Content

Users
Category
- User
- List
- Campaign

Data Collection
- Acct metadata
- Address history
- Public Lists
- Campaign history
- Content

Intermediate Modeling
- Bounce Model
- Content Model

Final modeling set
- Acct Metadata
- List scores
- Prev. camp. perf.
- Content scores

Final Model

Predictions
Design to eliminate risk/complexity

<table>
<thead>
<tr>
<th>Function</th>
<th>Tool</th>
</tr>
</thead>
<tbody>
<tr>
<td>Slow Storage</td>
<td>Sharded PostgreSQL</td>
</tr>
<tr>
<td>Fast Storage</td>
<td>Redis</td>
</tr>
<tr>
<td>Modeling Language</td>
<td>R</td>
</tr>
<tr>
<td>Machine Learning Model</td>
<td>Random Forest</td>
</tr>
<tr>
<td>API + Background Processes</td>
<td>Python</td>
</tr>
</tbody>
</table>
Purchased Correlation vs Spam Button

Direct Complaint Rate vs Purchased List Correlation

- Direct Complaint Rate range: 0% to 1.5%
- Purchased List Correlation range: 0% to 100%
Predicted vs. Actual Hard Bounce Rate for New Users

\[ y = 1.0933x - 0.0055 \]

\[ R^2 = 0.93608 \]
Now that the Infrastructure’s in Place

Ad Hoc Analysis – Deep List Investigation

Send Time Optimization

Automatic Segmentation and Content Optimization
Now that the Infrastructure’s in Place
Now that the Infrastructure’s in Place
Why care about the graph?

Interest graphing

Content Optimization

A/B testing
Thanks!
@John4man