What's your ML Test Score?
A rubric for ML production systems

Eric Breck, Shanqing Cai, Eric Nielsen, Michael Salib and D. Sculley
{ebreck, cais, nielsene, msalib, dsculley}@google.com
Testing Code

- Write Code
- Verify Expected Behavior
- Running System
- Unit Tests
- Integration Tests
- Monitoring
Testing ML Code

1. Write ML Code
2. Verify Expected Behavior
3. Train Model
4. Verify Model
5. Running System
Testing ML Code

1. Write ML Code
2. Verify Expected Behavior
3. Train Model
4. Verify Model
5. Running System

- Unit Tests
- Integration Tests
- Monitoring
Testing ML Code

1. Write ML Code
2. Verify Expected Behavior
3. Train Model
4. Verify Model
5. Running System

- Unit Tests
- Integration Tests
- Monitoring
Okay, so what should we test?
An ML Test Rubric

- Tests for ML Infrastructure
- Tests for Model Development
- Tests for Features and Data
- Monitoring of Running ML Systems
A Few Examples
Test that the distribution of each feature matches your expectation.
Test that the distribution of each feature matches your expectation.
Test the relationship between offline and online metrics
Test the relationship between offline and online metrics

Model 1 - 0.95 AUC

Model 2 - 0.93 AUC
Test models via a canary process before serving in production
Test models via a canary process before serving in production

Tests for
- Features and Data
- Model Development
- ML Infrastructure

Monday

Model
Test models via a canary process before serving in production.
Test models via a canary process before serving in production

Wednesday

Model

Server
Test models via a canary process before serving in production.
Test that training and serving features compute the same values
Test that training and serving features compute the same values
Test that training and serving features compute the same values
Test that training and serving features compute the same values
Test that training and serving features compute the same values.
What’s my score?
<table>
<thead>
<tr>
<th>Score</th>
<th>Interpretation</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>Research project</td>
</tr>
<tr>
<td>1-2</td>
<td>Not totally untested</td>
</tr>
<tr>
<td>3-4</td>
<td>A first pass at basic productionization</td>
</tr>
<tr>
<td>5-6</td>
<td>Reasonably tested</td>
</tr>
<tr>
<td>7-10</td>
<td>Appropriate for mission-critical systems.</td>
</tr>
<tr>
<td>12+</td>
<td>Exceptional levels of automated testing and monitoring.</td>
</tr>
</tbody>
</table>
Questions?

Thank You!

dsculley@google.com