.conf2015

Machine Learning and Analytics in Splunk

Adam J. Oliner

Principal Engineer, Splunk

Brad Block, Ari Brown, Jacob Leverich, Andrea Longdon, Di Lu, Iman Makaremi, Ben Moskowitz, Manish Sainani, Sergey Slepian, Sinduja Sreshta, Zidong Tang, Brooke Wenig, Jonas Yao, Fred Zhang

splunk>

Disclaimer

During the course of this presentation, we may make forward looking statements regarding future events or the expected performance of the company. We caution you that such statements reflect our current expectations and estimates based on factors currently known to us and that actual events or results could differ materially. For important factors that may cause actual results to differ from those contained in our forward-looking statements, please review our filings with the SEC. The forward-looking statements made in the this presentation are being made as of the time and date of its live presentation. If reviewed after its live presentation, this presentation may not contain current or accurate information. We do not assume any obligation to update any forward looking statements we may make.

In addition, any information about our roadmap outlines our general product direction and is subject to change at any time without notice. It is for informational purposes only and shall not, be incorporated into any contract or other commitment. Splunk undertakes no obligation either to develop the features or functionality described or to include any such feature or functionality in a future release.

Agenda

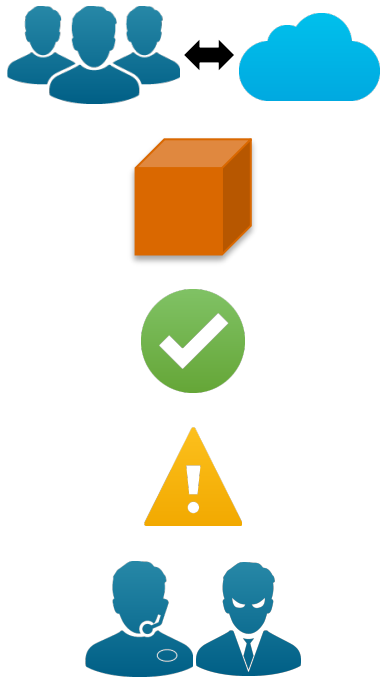
- Machine learning and statistics
- ML Toolkit and Showcase app
- Demo!
- How to acquire and use the app

ML and Statistics

- Process for generalizing from examples
  - ML: (labeled) examples ➜ model
  - Stats: sample ➜ population
- E.g., face recognition
  - Hard to write the code from scratch
  - Easy to give examples of faces and non-faces
    - Label which is which = supervised learning
    - Tell it split into two groups, no labels = unsupervised learning

Insider Threat (User Behavior Analytics)

1. Log cloud storage data transfer
2. Build a predictive model
3. Refine until predictions are accurate
4. Detect large prediction errors
5. Investigate



Capacity Planning

1. Log resource utilization (e.g., disk capacity)
2. Build a predictive model based on past values
3. Refine until predictions are accurate
4. Forecast resource saturation or demand
5. Act





Predict Customer Churn

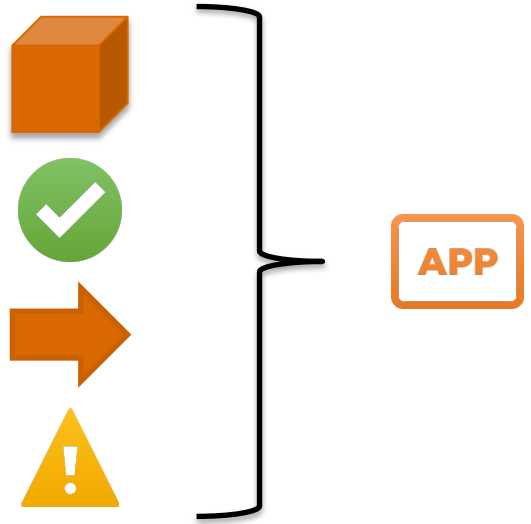
1. Build a model that predicts customer churn
2. Refine until predictions are accurate
3. Predict when customers will churn
4. Inspect the model to see what factors drive churn
5. Act





The Process

1. Clean & transform
2. **Fit a model**
3. **Refine the model**
4. **Apply to make predictions**
5. **Detect anomalies**
6. Alert
7. Act



ML Toolkit and Showcase App

splunk>

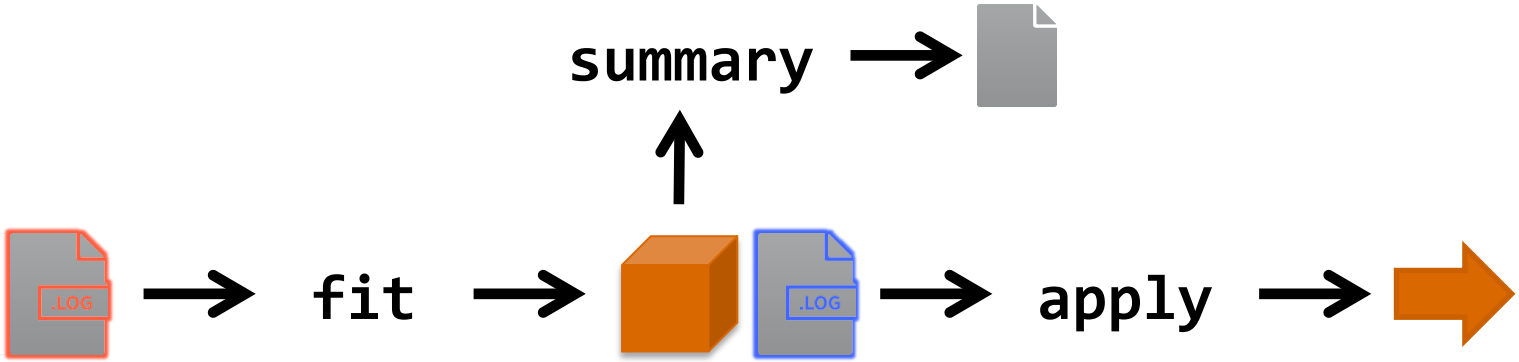
ML Toolkit and Showcase App

ML SPL

- Generic grammar
  - Follows the lead of popular ML libraries
  - Doesn't clutter SPL
- **`fit, apply, summary`**

ML SPL

- **Fit** a (persistent) model using training data
- **Apply** a model to new data to make predictions
- Inspect a **summary** of the model



ML SPL Example

- *[training data]* | **fit** LinearRegression into my_model costly_KPI from metric1 metric2 metric3
- *[test data]* | **apply** my_model as pred_kpi_value
- | **summary** my_model

Demo!

splunk>

Using the App

- Act on alerts and reports
- Use dashboards as tools
- Adapt dashboards to your needs
- Apply models
- Use ML SPL
- Build custom dashboards and visualizations
- Build custom analytics

Behind the Curtain

- Uses only public interfaces and libraries
- Distribution of the python data science ecosystem
  - scitkit-learn, pandas, numpy, scipy, and much more
  - On Splunkbase: Python for Scientific Computing
- “Just an app”
- Source code is packaged in the app

Preview Release Caveats

- Limited to 50k training events
  - Can apply model to unlimited events
- Limited to a single search head
  - No support for search head clustering
  - Does not distribute work to indexers
- Limited to an initial set of analytics
- Community-supported app
- Plus all the other caveats you'd expect of a preview release

Gimme! Gimme!

- ML Toolkit and Showcase App
  - Preview Release is Free on Splunkbase
- Dependencies
  - Splunk 6.3
  - Python for Scientific Computing

http://tiny.cc/splunkmlapp