Using The Latest Features From The Splunk Machine Learning Toolkit To Create Your Own Custom Models

Adam J. Oliner  |  Director of Engineering
Manish Sainani  |  Director of Product

September 2017  |  Washington, DC
Forward-Looking Statements

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 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.

Splunk, Splunk>, Listen to Your Data, The Engine for Machine Data, Splunk Cloud, Splunk Light and SPL are trademarks and registered trademarks of Splunk Inc. in the United States and other countries. All other brand names, product names, or trademarks belong to their respective owners. © 2017 Splunk Inc. All rights reserved.
Outline

▶ Splunk Machine Learning Toolkit
▶ Platform Extensions: ML-SPL, etc.
▶ The Assistants: Guided Machine Learning
▶ What’s New
▶ Demo
  • DIY Anomaly Detector
Splunk Machine Learning Toolkit

platform extensions and guided modeling dashboards
Machine Learning

▶ A process for generalizing from examples

▶ Examples

• A, B, … → # (regression)

• A, B, … → a (classification)

• $X_{\text{past}} \rightarrow X_{\text{future}}$ (forecasting)

• like with like (clustering)

• $|X_{\text{predicted}} - X_{\text{actual}}| >> 0$ (anomaly detection)
Data Gathering And Prep

Source: CrowdFlower

What data scientists spend the most time doing

- Building training sets: 3%
- Cleaning and organizing data: 60%
- Collecting data sets: 19%
- Mining data for patterns: 9%
- Refining algorithms: 4%
- Other: 5%
Machine Learning Process With Splunk

Alerts, Dashboards, Reports

Splunk Machine Learning Toolkit

props.conf, transforms.conf, Datamodels, Splunkbase

Pivot, Table UI, SPL

Splunk
Overview Of Machine Learning At Splunk

Splunk Premium Solutions

Rich Ecosystem of Apps & Add-Ons

Splunk IT Service Intelligence™
Splunk Enterprise Security™
Splunk User Behavior Analytics™

Mobile
Forwarders
Syslog/TCP
IoT Devices
Network Wire Data
Hadoop
Relational Databases
Mainframe Data

splunk>enterprise

splunk>cloud

Platform for Operational Intelligence

© 2017 SPLUNK INC.
Splunk Machine Learning Toolkit extends Splunk with new tools and guided modeling

- **Assistants**: Guide model building, testing, & deployment for common tasks
- **Showcase**: 25+ interactive examples from IT, security, business, and IoT
- **Algorithms**: 25+ standard algorithms plus an extensibility API
- **SPL ML Commands**: New commands to fit, test, and operationalize models
- **Python for Scientific Computing Library**: 300+ open-source algorithms
Machine Learning Customer Success

**Network Optimization**
Detect & Prevent Equipment Failure

**Security / Fraud Prevention**

**Telco**
Prevent Cell Tower Failure
Optimize Repair Operations

**Zillow**
Prioritize Website Issues
and Predict Root Cause

**Entertainment Company**
Predict Gaming Outages
Fraud Prevention

**Concanon**
Machine Learning Consulting Services

**Scianta Analytics**
Analytics App built on ML Toolkit

---

© 2017 Splunk Inc.
Machine Learning Toolkit Customer Use Cases

- Reducing customer service disruption with early identification of difficult-to-detect network incidents
- Minimizing cell tower degradation and downtime with improved issue detection sensitivity
- Speeding website problem resolution by automatically ranking actions for support engineers
- Ensuring mobile device security by detecting anomalies in ID authentication
- Predicting and averting potential gaming outage conditions with finer-grained detection
- Preventing fraud by identifying malicious accounts and suspicious activities
- Improving uptime and lowering costs by predicting/preventing cell tower failures and optimizing repair truck rolls
Platform Extensions: ML-SPL, etc.

custom search commands for machine learning
SPL, Macros, & Viz
Oh, my!

- **Commands (ML-SPL)**
  - fit
  - apply
  - summary
  - listmodels
  - deletemodel
  - sample

- **Macros**
  - regressionstatistics
  - classificationstatistics
  - classificationreport
  - confusionmatrix

- **Viz**
  - forecastviz
  - histogram
  - modvizpredict
  - splitby(1-5)

- **Viz**
  - Outliers Chart
  - Forecast Chart
  - Scatter Line Chart
  - Histogram Chart
  - Downsampled Line Chart
  - Scatterplot Matrix
ML-SPL

What is it?

- A suite of SPL commands specifically for machine learning
  - modeling
  - sampling
- Most are implemented using modules from the Python for Scientific Computing add-on for Splunk
  - scikit-learn
  - numpy
  - pandas
  - statsmodels
  - scipy
ML-SPL Commands

- Fit (i.e., train) a model from search results
  
  ... | fit <ALGORITHM> <TARGET> from <VARIABLES ...>
  <PARAMETERS> into <MODEL>

- Apply a model to obtain predictions from (new) search results
  
  ... | apply <MODEL>

- Inspect a model (e.g., display coefficients)
  
  | summary <MODEL>
ML-SPL Commands: fit

... | fit <ALGORITHM> <TARGET> from <VARIABLES> <PARAMETERS> into <MODEL>

optional

Examples:

... | fit LinearRegression system_temp from cpu_load fan_rpm into temp_model

... | fit KMeans k=10 downloads purchases posts days_active visits_per_day into user_behavior_clusters
ML-SPL: Algorithms

- 25+ algorithms OotB
  - prediction, clustering, forecasting, feature engineering
- Extensibility API for 300+ more
- Pipeline for advanced use cases

... | fit TFIDF message
| fit StandardScaler files bytes
| fit KMeans message_tfidf_* SS_* k=5
| fit PCA message_tfidf_* k=2
| ...

```
fit TFIDF message
```
```
fit StandardScaler files bytes
```
```
fit KMeans message_tfidf_* SS_* k=5
```
```
fit PCA message_tfidf_* k=2
```
```
...```
ML-SPL Commands: apply

... | apply <MODEL>

Examples:

... | apply temp_model
... | apply user_behavior_clusters
ML-SPL Commands: summary

Examples:

    ... | summary <MODEL>

    ... | summary temp_model

    ... | summary user_behavior_clusters
ML-SPL Commands

| listmodels
| deletemodel <MODEL>
Randomly sample or partition events

... | sample <PARAMETERS>

Four modes

- Ratio ... | sample 0.01
- Count ... | sample 20
- Proportional ... | sample proportional=“some_field”
- Partition ... | sample partitions=10
Plug: Machine Learning In SPL With The Machine Learning Toolkit

Tuesday, September 26th @ 2:15pm in Ballroom B

- Custom search commands!
- Tabular munging!
- Jacob Leverich!
- Shang Cai!
The Assistants

guided machine learning
Guided ML With The Assistants

- Guides you through an analytic
  - Prepare, fit, validate, and deploy
- Automatically generates all the relevant SPL

```bash
| inputlookup server_power.csv
| fit StandardScaler "total-cpu-utilization", "total-disk-accesses", "total-disk-blocks", "total-disk-utilization", "total-instructions_retired", "total-last_level_cache_references", "total-memory_bus_transactions", "total-unhalted_core_cycles" with_mean=true with_std=true into example_server_power.StandardScaler_0
| fit LinearRegression fit_intercept=true "ac_power" from "SS_*" into "example_server_power" // fit and save a model using the entire dataset and provided parameters
```
Assistants: Prepare

**Preprocessing Steps**

- **StandardScaler**
  - **Preprocess method**: StandardScaler
  - **Fields to preprocess**: business_acres, property_tax_rate, distance_to_employment_center
  - **Standardize Fields**: with respect to mean and with respect to standard deviation

- **PCA**
  - **Preprocess method**: PCA
  - **Fields to preprocess**: highway_accessibility_index, distance_to_employment_center, pupil_teacher_ratio, crime_rate
  - **K (# of Components)**: 2
Assistants: Fit
Assistants: Validate

- **Residuals Line Chart**
- **Residuals Histogram**
- **R² Statistic**: 0.9480
- **Root Mean Squared Error (RMSE)**: 5.55

<table>
<thead>
<tr>
<th>Prediction Feature</th>
<th>Linear Regression Coefficient</th>
</tr>
</thead>
<tbody>
<tr>
<td>total-cpu-utilization</td>
<td>-0.060032370862</td>
</tr>
<tr>
<td>total-disk-accesses</td>
<td>-0.00900054884969</td>
</tr>
<tr>
<td>total-disk-blocks</td>
<td>-0.00028470246115</td>
</tr>
<tr>
<td>total-disk-usage</td>
<td>0.000082299744</td>
</tr>
<tr>
<td>total-instructions-retired</td>
<td>1.6539277239e-09</td>
</tr>
</tbody>
</table>
Assistants: Deploy

![Diagram of Assistants interface with options for Fit Model on a Schedule and Schedule an alert with parameters for alert conditions.](image-url)
The Assistants

- Predict Numeric Fields
- Predict Categorical Fields
- Detect Numeric Outliers
- Detect Categorical Outliers
- Forecast Time Series
- Cluster Numeric Events
What’s New

since last .conf
What’s New
(since .conf 2016)

- Detect Numeric Outliers improvements
- Preprocessing
- Model Management
- ML-SPL extensibility API
- Spark Support (private beta)
- New algorithms:
  - ACF & PACF
  - ARIMA
- Load Existing Settings is per-user
- Downsampled Line Chart supports drilldown
Detect Numeric Outliers

data distribution viz

Data Distribution

points within threshold

standard deviations from mean

count

Open in Search  Show SPL
Preprocessing
build a pipeline of data prep

- In Predict Numeric, Predict Categorical, and Cluster Numeric assistants

Preprocessing Steps

- StandardScaler
  - Preprocess method: StandardScaler
  - Fields to preprocess: business_acres, property_tax_rate, distance_to_employment_center
  - Standardize Fields: with respect to mean, with respect to standard deviation
  - Apply

- PCA
  - Preprocess method: PCA
  - Fields to preprocess: highway_accessibility_index, distance_to_employment_center, pupil_teacher_ratio, crime_rate
  - K (# of Components): 2
  - Apply
Assign permissions to models to control access

Manage models via the UI

Experiments
  - Assistant configurations
  - May produce 1+ models
ML-SPL Extensibility API
featuring: primo documentation

- Make more algos available to fit / apply
  - 300+ in PSC
  - Custom algorithms
- Expose new or different parameters
- Docs include examples
  - Correlation Matrix
  - Agglomerative Clustering
  - Support Vector Regressor
  - Savitzky-Golay Filter
- Use in your apps / dashboards / etc.!
Plug: Advanced Machine Learning using the Extensible ML API

Wednesday, September 27th @ 4:35pm in Ballroom B

- Implementation details!
- Extensibility API!
- Alexander Johnson!
- Zidong Yang!
Spark Support
private beta open now

Use your existing Spark cluster with MLTK
• Distributed fit on massive datasets
• Apply MLlib models for supported algos

sfit / sapply

Contact mlprogram@splunk.com
• What is your use case (e.g., predicting server downtime)?
• Why do you want / need Spark (i.e., why isn't MLTK sufficient)?
Demo
Thank You

Don't forget to rate this session in the .conf2017 mobile app