Using the Splunk Machine Learning Toolkit to Create Your Own Custom Models

Dr. Adam Oliner
Director of Engineering, Data Science, Splunk

Manish Sainani
Principal Product Manager, 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 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.
Who are we?

- Dr. Adam Oliner
  - Director of Engineering, Data Science & Machine Learning
  - Splunker for 2 years
  - Embarrassingly overeducated

- Manish Sainani
  - Principal Product Manager, Machine Learning
  - Splunker for 2 years
  - First ML hire at Splunk!
What are we doing here?

- Overview of Machine Learning
- The Assistants: Guided Machine Learning
  - Prepare
  - Fit
  - Validate
  - Deploy
- Examples
  - DIY Anomaly Detector
  - Customer Applications
Overview of ML at Splunk

Core Platform Search

Packaged Premium Solutions

Custom ML

Machine Learning Toolkit

splunk> Platform for Operational Intelligence
Splunk Machine Learning Toolkit

Extends Splunk platform functions and provides a guided modeling environment

**Assistants:** Guide model building, testing, & deploying for common objectives

**Showcases:** Interactive examples for typical IT, security, business, IoT use cases

**Algorithms:** 25+ standard algorithms available prepackaged with the toolkit

**SPL ML Commands:** New commands to fit, test and operationalize models

**Python for Scientific Computing Library:** 300+ open source algorithms available for use

Build custom analytics for any use case
What’s New since our 0.9 Beta Release (last year’s .conf)?

- New name and abbreviation ;-) 
- No event limits (removal of 50K limit on fitting models) 
- Configurable resource caps via mlspl.conf 
- Search head clustering support 
- Distributed / streaming apply 
- Scheduled fit 
- New algorithms (next slide) 
  - Feature engineering and selection 
  - Stochastic gradient descent (e.g.) 
  - ARIMA 
- Multi-algorithm support across Assistants 
- Scatterplot matrix viz 
- Alerting 
- Tooltips 
- In-app tours 
- Cluster Numeric Events assistant 
- Videos videos videos for each assistant across IT, Security, IoT and Business Analytics 
- ML-SPL Cheat Sheet
The Assistants: Guided Machine Learning
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)
Machine Learning Process

- Collect Data
- Clean/Transform
- Explore/Visualize
- Model
- Evaluate
- Publish/Deploy
Machine Learning Process with Splunk

1. Collect Data
2. Clean/Transform
3. Explore/Visualize
4. Model
5. Evaluate
6. Publish/Deploy

- props.conf, transforms.conf, Datamobiles
- Add-ons from Splunkbase, etc.
- Pivot, Table UI, SPL

Alerts, Dashboards, Reports

ML Toolkit
Custom Machine Learning – Success Formula

- Domain Expertise (IT, Security, ...)
  - Identify use cases
- Splunk Expertise
  - Set business/ops priorities
- Data Science Expertise
  - Drive decisions
  - Statistics / math background
  - Algorithm selection
  - Model building

Splunk ML Toolkit facilitates and simplifies via examples & guidance

SPL
- Data prep
- Operational success
Guided ML with the Assistants

- Guides you through various analytics
  - Prepare, fit, validate, and deploy
- Automatically generates all the relevant SPL
Assistants: Fit
Assistants: Validate

- R² Statistic: 0.9480
- Root Mean Squared Error (RMSE): 5.55

Graphs showing residuals line chart and residuals histogram.
Assistants: Deploy

example server power

Fit Model On A Schedule

Fit Model

Open in

Open in Search
Show SPL
Schedule Alert

Scheduled Training
Alerts

Schedule an alert

Alert me when the predicted value is not between 0 and 100

Search to refine the parameters or further explore the data.

Open in Search
Show SPL
Schedule Alert
The Assistants

1. Predict Numeric Fields
2. Predict Categorical Fields
3. Detect Numeric Outliers
4. Detect Categorical Outliers
5. Forecast Time Series
6. Cluster Numeric Events
Predict Numeric Fields

- **Algorithms**
  - LinearRegression
    - ... including Lasso, Ridge, and ElasticNet
  - KernelRidge
  - DecisionTreeRegressor
  - RandomForestRegressor
  - SGDRegressor

- **Validation**
  - Four visualizations of prediction error
  - $R^2$ and RMSE
Predict Categorical Fields

- Algorithms
  - LogisticRegression
  - DecisionTreeClassifier
  - RandomForestClassifier
  - SGDClassifier
  - SVM
  - Naïve Bayes
    - BernoulliNB and GaussianNB

- Validation
  - Precision, recall, accuracy, F1
  - Confusion matrix

Precision: 0.97
Recall: 0.97
Accuracy: 0.97
F1: 0.97

Classification Results (Confusion Matrix)

<table>
<thead>
<tr>
<th>Predicted actual</th>
<th>Predicted 2008 BMW M3</th>
<th>Predicted 2011 Ferrari 458</th>
</tr>
</thead>
<tbody>
<tr>
<td>2008 BMW M3</td>
<td>4405 (99.7%)</td>
<td>0 (0%)</td>
</tr>
<tr>
<td>2011 Ferrari 458</td>
<td>0 (0%)</td>
<td>3327 (97%)</td>
</tr>
<tr>
<td>2011 Ford Mustang GT500</td>
<td>0 (0%)</td>
<td>0 (0%)</td>
</tr>
<tr>
<td>2013 Audi R5</td>
<td>73 (1.9%)</td>
<td>54 (1.4%)</td>
</tr>
<tr>
<td>2014 Chevrolet Corvette</td>
<td>11 (0.2%)</td>
<td>45 (0.8%)</td>
</tr>
<tr>
<td>2015 Porsche GT3</td>
<td>0 (0%)</td>
<td>0 (0%)</td>
</tr>
</tbody>
</table>
Detect Numeric Outliers

- Methods
  - Standard deviation
  - Median absolute deviation
  - Interquartile range

- Validation:
Detect Categorical Outliers

- Statistical methods
- Validation:

### Data and Outliers

<table>
<thead>
<tr>
<th>customer_id</th>
<th>distance</th>
<th>price</th>
<th>product_id</th>
<th>quantity</th>
<th>shop_id</th>
<th>probable_cause</th>
<th>isOutlier</th>
</tr>
</thead>
<tbody>
<tr>
<td>u92</td>
<td>1063.275</td>
<td>62.51</td>
<td>p4188</td>
<td>2</td>
<td>s1</td>
<td>price</td>
<td>▶️ 1</td>
</tr>
<tr>
<td>u150</td>
<td>1463.661</td>
<td>28.624</td>
<td>p4184</td>
<td>1</td>
<td>s1</td>
<td>price</td>
<td>▶️ 1</td>
</tr>
<tr>
<td>u186</td>
<td>7833.517</td>
<td>83.191</td>
<td>p280</td>
<td>1</td>
<td>s1</td>
<td>price</td>
<td>▶️ 1</td>
</tr>
<tr>
<td>u196</td>
<td>4803.592</td>
<td>54.493</td>
<td>p49</td>
<td>1</td>
<td>s1</td>
<td>price</td>
<td>▶️ 1</td>
</tr>
<tr>
<td>u196</td>
<td>4803.592</td>
<td>51.306</td>
<td>p439</td>
<td>1</td>
<td>s1</td>
<td>price</td>
<td>▶️ 1</td>
</tr>
<tr>
<td>u202</td>
<td>2114.282</td>
<td>60.324</td>
<td>p28</td>
<td>1</td>
<td>s1</td>
<td>price</td>
<td>▶️ 1</td>
</tr>
<tr>
<td>u123</td>
<td>1300.591</td>
<td>21.005</td>
<td>p2042</td>
<td>123</td>
<td>s1</td>
<td>quantity</td>
<td>▶️ 1</td>
</tr>
<tr>
<td>u137</td>
<td>961.408</td>
<td>16.92</td>
<td>p4029</td>
<td>106</td>
<td>s3</td>
<td>quantity</td>
<td>▶️ 1</td>
</tr>
<tr>
<td>u231</td>
<td>583.590</td>
<td>15.836</td>
<td>p4033</td>
<td>94</td>
<td>s2</td>
<td>quantity</td>
<td>▶️ 1</td>
</tr>
<tr>
<td>u7</td>
<td>4082.522</td>
<td>0.334</td>
<td>p112</td>
<td>3</td>
<td>s1</td>
<td>✔️ 0</td>
<td></td>
</tr>
</tbody>
</table>
Forecast Time Series

- **Algorithms**
  - State-space method using Kalman filter
  - ARIMA

- **Validation**

![Graph showing time series data with metrics R² Statistic 0.8575 and Root Mean Squared Error (RMSE) 2.10]
Cluster Numeric Events

- Algorithms
  - KMeans
  - DBSCAN
  - Birch
  - SpectralClustering

- Validation
  - Scatterplot Matrix viz
Prepare
Data Gathering and Prep

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%

Source: CrowdFlower
Splunk!

- Leading platform for collecting, cleaning, and transforming data
- Interactive Field Extractor
- Datamodels
- Hundreds of add-ons from Splunkbase
- transforms.conf
- props.conf
- etc.
Feature Engineering

- TFIDF (term-frequency x inverse document-frequency)
  - Transform free-form text into numeric attributes
- StandardScaler (i.e. normalization)
- FieldSelector (i.e. choose k best features for regression/classification)
- PCA and KernelPCA
Preprocessing in the Assistants

- **Preprocess (optional)**
  - Fields to preprocess:
    - batteryVoltage
    - engineCoolantTemperature
    - engineSpeed
    - lateralGForce
    - longitudeGForce
    - speed
    - verticalGForce
  - Select method(s) to use:
    - Apply StandardScaler
    - Apply PCA
    - Apply KernelPCA
    - To reduce dimensionality to 3 fields

- **Cluster**
  - Algorithm: Birch
  - Fields to use:
    - PC_1
    - PC_2
    - PC_3
  - K (# of centroids): 6
Fit
Fit: What’s New

- No event limits
- Configurable resource caps (ml-spl.conf)
- Search head clustering support
- Scheduled fit
- New algorithms
Fit: What’s New
Validate / Apply: What’s New

- Configurable resource caps
- Search head clustering support
- Distributed / streaming apply
- Scatterplot matrix viz
Scatterplot Matrix Viz
Deploy
Deploy anywhere in Splunk!

- Scheduled training
- Alerting
- Reports and dashboards
- Augmented search results
- etc.
Deploy: What’s New

- Distributed Apply
  - Apply models to indexed data
  - Streaming
- Scheduled training
- Alerting
What’s New: Scheduled Fit
What’s New: Alerting

Schedule an alert

Alert me when the predicted value is not between 0 and 100
Example:
DIY Anomaly Detector
Let’s Build an Anomaly Detector!

• We’ll use two Assistants
  – Predict Numeric Fields
  – Detect Numeric Outliers

• Show automatically-generated intermediate SPL
Fit a Predictive Model

Create New Model

1: Enter a search
inputlookup server_power.csv

31,272 results (12/31/69 4:00:00.000 PM to 8/5/16 4:18:30.000 PM)

Job Smart Mode

2: Field to predict
ac_power

3: Fields to use for predicting
× 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-uhalted_core_cycles

4: Split for training / test: 50 / 50

5: Save the model as
example_server_power

Fit Model On A Schedule
Set up Scheduled Training

Schedule Model Training

Title
Power Model Retraining

Description
Optional

Schedule
Run every week

On
Monday at 5:00

Time Range
Run every hour
Run every day
Run every week
Run every month
Run on Cron Schedule

Schedule

44
Open Residuals in Search
Open Detect Numeric Outliers Assistant
Detect Outliers (Large Prediction Errors)
Schedule an Alert
Schedule an Alert

Alert me when the number of outliers outside both thresholds is greater than 5
Schedule an Alert

Save As Alert

Settings

Search: | inputlookup server_power.csv | apply "example_server_power" | eval residual = ac_power^2 - predicted(ac_power) | table time, residual | streamstats window=100 current=true avg("residual") as avg stddev("residual") as stddev eval

Title: AC Power Anomaly

Description: Optional

Alert type: Scheduled Real-time

Run every week: 

On: Monday at 6:00

Trigger Conditions

Trigger alert when: Number of Results

is greater than: 5

Trigger: Once For each result

Throttle: 

Trigger Actions: Add action

Save
Manage Your New Anomaly Detector
The Assistant Generated the SPL for You

**Fit a model on all your data in search**

```
| inputlookup server_power.csv
| fit LinearRegression 'ac_power' from "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" into "example_server_power"
```

// fit and save a model using the entire dataset and provided parameters

**Plot prediction errors on a line chart**

```
| inputlookup server_power.csv
| apply "example_server_power"
| eval residual = 'ac_power' - 'predicted(ac_power)'
| table _time, residual
```

// apply the model to the entire dataset to predict "ac_power"

// calculate the prediction error
The Assistant Generated the SPL for You

```
| inputlookup server_power.csv | apply "example_server_power" |
| eval residual = 'ac_power' - 'predicted(ac_power)' | table _time, residual |

| streamstats window=100 current=true avg("residual") as avg_stddev("residual") as stdev |
| eval lowerBound=(avg-stdev*4), upperBound=(avg+stdev*4) |
| eval isOutlier=if('residual' < lowerBound OR 'residual' > upperBound, 1, 0) |
| table _time, "residual", lowerBound, upperBound, isOutlier |
```

You Built an Anomaly Detector!

- You built a predictive model of AC Power
- When the prediction error from this model is an outlier compared to past errors, you generate an alert
- This predictive model automatically retrains itself on a schedule you control
- You didn’t have to type any SPL
Machine Learning Customer Success

**TELUS**
Network Optimization
Detect & Prevent Equipment Failure

**docomo**
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

*Optimizing operations and business results*
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
Detect Network Outliers

Reduced downtime + increased service availability = better customer satisfaction

ML Use Case
Monitor noise rise for 20,000+ cell towers to increase service and device availability, reduce MTTR

Technical overview
• A customized solution deployed in production based on outlier detection.
• Leverage previous month data and voting algorithms

“The ability to model complex systems and alert on deviations is where IT and security operations are headed ... Splunk Machine Learning has given us a head start...”
Reliable website updates

Proactive website monitoring leads to reduced downtime

ML Use Case
- Very frequent code and config updates (1000+ daily) can cause site issues
- Find errors in server pools, then prioritize actions and predict root cause

Technical overview
- Custom outlier detection built using ML Toolkit Outlier assistant
- Built by Splunk Architect with no Data Science background

“Splunk ML helps us rapidly improve end-user experience by ranking issue severity which helps us determine root causes faster thus reducing MTTR and improving SLA”
What Now?

http://tiny.cc/splunkmlapp

- Get the Machine Learning Toolkit from Splunkbase
- Go watch Machine Learning Videos on Splunk Youtube Channel [http://tiny.cc/splunkmlvideos](http://tiny.cc/splunkmlvideos)
- Go to Machine Learnings talks:
  - Advanced Machine Learning in SPL with the Machine Learning Toolkit by Jacob Leverich
  - Extending SPL with Custom Search Commands and the Splunk SDK for Python by Jacob Leverich
- Several Customers and Partner Talks
  - Cisco, Scianta Analytics, Asian Telco, etc.
- Early Adopter And Customer Advisory Program: [mlprogram@splunk.com](mailto:mlprogram@splunk.com)
- Product Manager: Manish Sainani [ms@splunk.com](mailto:ms@splunk.com)
- Field Expert: Andrew Stein [astein@splunk.com](mailto:astein@splunk.com)
THANK YOU