Building ML Solutions using MLTK

Repeatable ML Workflows

Andrew Stein | Analytical Architect
Iman Makaremi | Senior Data Scientist
September 2017 | Washington, DC
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.
Agenda

- MLTK? What?!
- Assistant: Numeric Outlier Detection
- Assistant: Predict Numeric Fields
- Assistant: Predict Categorical Fields
What?
Splunk has a Machine Learning Toolkit App!

- What is Splunkbase
- What is the App
- Where can I go to learn more
What? The Machine Learning Toolkit has Assistants!

What is an Assistant

- Predict Numeric Fields
- Predict Categorical Fields
- Detect Numeric Outliers
- Detect Categorical Outliers
- Forecast Time Series
- Cluster Numeric Events
A Successful Machine Learning Process

Data Driven Decisions

01 Acquire Data

02 Generate Insights

03 Publish Insights

04 Data Driven Decisions
Assistant: Numeric Outlier Detection
Assistant: Detect Numeric Outlier
One workflow, multiple use cases

- Site Reliability Engineer
  - Reduced Time to Action by 50%
- Design Specialist
  - 16 Million Dollars Saved
Website Accessibility

- A Collection of **Several Applications**
- **Daily Code Releases** for each Application
- Dynamic Behaviour of Internet Browsing
- Data
  - Summary Indexes
- Goal
  - Catching a problematic code release in a timely manner is important for customer satisfaction.
- Action Taken
  - Inform the release team about the problems
Mobility 3GPP Core KPI

Data
- The 3GPP Core receives transactions from each subscriber to maintain connection.
- The KPI captures the behavior of TELUS’s network.

Goal
- Monitor this dynamic KPI and alert on performance degradation.

Action Taken
- Radio engineers informed about problems as soon as something occurred.
Assistant : Detect Numeric Outliers

Step 1 : Data Driven Decisions

Define your Problem!

- I want **detect and alert** on metric(s) that deviate **significantly** from their **past behaviour**.
Assistant: Detect Numeric Outliers

Step 2: Acquire Data

- **Industrial Data**: SCADA, AMI, Meter Reads
- **Native Inputs**: TCP, UDP, Logs, Scripts, Wire, Mobile
- **Modular Inputs**: MQTT, AMQP, COAP, REST, JMS
- **HTTP Event Collector**: Token Authenticated Events
- **Technology Partnerships**: Kepware, AWS IoT, Cisco, Palo Alto

**Real Time**

**Splunk Enterprise**

**Splunk Cloud**

**External Lookups/Enrichment**

- Asset Info
- Maintenance Info
- Data Stores
Assistant : Detect Numeric Outliers
Step 3 : Generate Insights

- Place search in Detect Numeric Outliers Assistant
- Validate results using the visualizations
- Create alerts
- One unified workflow.
- Go customize to your hearts content
Assistant : Detect Numeric Outliers

Step 4 : Publish Insights

- Reuse searches and visualizations
- Built your own report
- Schedule outlier detection searches

"It's a non-linear pattern with outliers......but for some reason I'm very happy with the data."

J.B. Lacloche ©
To the Toolkit Example!
Assistant: Predict Numeric Field
Assistant: Predict Numeric Field
One workflow, multiple use cases

- Site Reliability Engineer
- Senior Design Specialist

Reduce Sales Loss

Improved Cell Tower Performance
Reduced Troubleshooting Time
Aggregated Order Behavior

- Failed orders result in
  - Unhappy Customer
  - Loss of Revenue

- Data
  - Order Counts by HTTP Response Codes

- Goal
  - Detecting deviation in relationship with the response codes

- Action Taken
  - Inform the release team about the problems
Radio Access Network Interference

- **Interference**
  - The level of noise within the frequency band in cell towers

- **Uplink Interference Impacting Factors**
  - Number of Subscribers
  - Connection Types
  - Radio Conditions

- **Data**
  - Uplink Rate/minute for Each Cell Tower
  - Number of Subscribers

- **Goal**
  - Find underperforming cells and the characteristics of their problem.

- **Action Taken**
  - Reconfigure Underperforming Cells
  - Identify non-standard devices on the Network
Assistant: Predict Numeric Field

Step 1: Data Driven Decisions

Define your Problem!

- I want to predict a numeric field given multiple other fields.

- I want to predict the future.
Assistant: Predict Numeric Field

Step 2: Acquire Data

Industrial Data
- SCADA, AMI, Meter Reads

Native Inputs
- TCP, UDP, Logs, Scripts, Wire, Mobile

Modular Inputs
- MQTT, AMQP, COAP, REST, JMS

HTTP Event Collector
- Token Authenticated Events

Technology Partnerships
- Kepware, AWS IoT, Cisco, Palo Alto

Real Time

Engineers
- Search

Data Analysts
- Alert

Security Analysts
- Visualize

Business Users
- Predict
- Develop

External Lookups/Enrichment

Asset Info

Maintenance Info

Data Stores
Assistant: Predict Numeric Field

Step 3: Generate Insights

- Preprocessing Step
- 7 Algorithms
- Model Analysis
- Customized Visualizations
- Create Alerts

Step 3:

Preprocessing Step

7 Algorithms

Model Analysis

Customized Visualizations

Create Alerts

$R^2 = 0.06$

I don't trust linear regressions when it's harder to guess the direction of the correlation from the scatter plot than to find new constellations on it.

https://xkcd.com/1725/
Assistant : Predict Numeric Field

Step 4 : Publish Insights

- Consume Predicted Field as an Alert
  - Directly from the Assistant, or
  - In another Search Bar

- Schedule Model Training
  - Batch Training
  - Partial Training

- Alert on Model Deviation
To the Toolkit Example!
Assistant: Predict Categorical Field
Assistant: Predict Categorical Field

One workflow, multiple use cases

• Wizard
• Senior Design Specialist

Work in Progress

At least 15% increase in expected sales adoption
Sales Monitoring

- Enterprise Adoption of Sales Analytics
- Data
  - Sales data
- Goal
  - Diminish Churn
  - Tailor the Customer Experience
- Action Taken
  - Align Sales Initiatives With Marketing Initiatives
Bot Detection

- Good and bad bots scrapping
- Bad bots go unnoticed
- A Liability Issue
- Data
  - Browsing Log
- Goal
  - Detect Bad Bots
- Action Taken
  - Ban or reduce access for malicious scrapping
Assistant : Predict Categorical Field
Step 1 : Data Driven Decisions

▶ Define your Problem!

• I want to **predict a categorical field** given **multiple other fields**.
  - Good or Bad Bot
  - Malware or Not
  - Churn or Not
  - Face Recognition
  - Cat or Dog or …
Assistant: Predict Categorical Field

Step 2: Acquire Data

- **Industrial Data**: SCADA, AMI, Meter Reads
- **Native Inputs**: TCP, UDP, Logs, Scripts, Wire, Mobile
- **Modular Inputs**: MQTT, AMQP, COAP, REST, JMS
- **HTTP Event Collector**: Token Authenticated Events
- **Technology Partnerships**: Kepware, AWS IoT, Cisco, Palo Alto

**Engineers**
- Search

**Data Analysts**
- Alert

**Security Analysts**
- Visualize

**Business Users**
- Predict
- Develop

**Real Time**

**External Lookups/Enrichment**

- Asset Info
- Maintenance Info
- Data Stores
Assistant : Predict Categorical Field

Step 3 : Generate Insights

- Preprocessing Step
- 6 Algorithms
- Model Analysis
- Custom Visualizations
- Create Alerts
Assistant : Predict Categorical Field

Step 4 : Publish Insights

▶ Consume Predicted Field as an Alert
  • Directly from the Assistant, or
  • In another Search Bar
▶ Schedule Model Training
  • Batch Training
  • Partial Training
▶ Alert on Model Deviation
To the Toolkit Example!
Q&A
Go See These Talks

▶ Advanced Machine Learning Using the Extensible ML API
  • Alexander Johnson & Zidong Yang

▶ Automation of Event Correlation and Clustering With Built-In Machine Learning Algorithms in Splunk IT Service Intelligence (ITSI)
  • Vineetha Bettaiah & Ross Lazerowitz

▶ Prioritizing Anomalies Using the Machine Learning Toolkit
  • Harsh Keswani

▶ Splunk Machine Learning Capabilities and Condition-Based Maintenance: Train Doors on the German Public Rail Transport System
  • Henning Brandt & Daniel Pal