Solve Big Problems with ML

Julian Andre
Staff Sales Engineer, East Strategics
Splunk, Inc.

Dr. Tom LaGatta
Staff Sales Engineer, East Strategics
Splunk, Inc.
Abstract

- Sometimes problem-solving feels like fighting fires with no relief. Leverage machine learning to help solve the problem of problem solving. We will introduce general ML concepts & workflows, and guide you through the long slog of exploratory data analysis to figure out what relates to what. Then we'll walk you through how to develop a systematic architecture to leverage ML models and improve your team's problem-solving capabilities. We'll talk about big data architectures, how to fit models on historical data and apply them in real time. We will close with a demonstration of ML capabilities in Splunk.
Why do we need ML?
Why is this so challenging using traditional methods?

- **DATA IS STILL IN MOTION**, still in a **BUSINESS PROCESS**.
- Enrich real-time **MACHINE DATA** with structured **HISTORICAL DATA**
- Make decisions **IN REAL TIME** using **ALL THE DATA**
- Combine **LEADING** and **LAGGING INDICATORS** (KPIs)
Machine Learning Customer Success

**Network Optimization**
Detect & Prevent Equipment Failure

**Security / Fraud Prevention**

**Prevent Cell Tower Failure**
Optimize Repair Operations

**Prioritize Website Issues**
and Predict Root Cause

**Predict Gaming Outages**
Fraud Prevention

**Machine Learning Consulting Services**

**Analytics App built on ML Toolkit**

*Optimizing operations and business results*
ML Toolkit Customer Use Cases

- Reduce customer service disruption with early identification of difficult-to-detect network incidents
- Minimize cell tower degradation and downtime with improved issue detection sensitivity
- Speed up website problem resolution by automatically ranking actions for support engineers
- Ensure mobile device security by detecting anomalies in ID authentication
- Predict and avert potential gaming outage conditions with finer-grained detection
- Prevent fraud by identifying malicious accounts and suspicious activities
- Improve uptime and lower costs by predicting/preventing cell tower failures and optimizing repair truck rolls

TELUS
Zillow
docomo
Entertainment Company
Telco
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
ML Use Cases
IT Ops: Predictive Maintenance

**Problem:** Network outages and truck rolls cause big time & money expense

**Solution:** Build predictive model to forecast outage scenarios, act pre-emptively & learn

1. Get resource usage data (CPU, latency, outage reports)
2. Explore data & build KPIs
3. Fit, apply & validate models on past / real-time data
4. Predict and act. Identify resource spikes, create alerts
5. Surface incidents to IT Ops, who INVESTIGATES & ACTS
Security: Find Insider Threats

**Problem:** Security breaches cause big time & money expense

**Solution:** Build predictive model to forecast threat scenarios, act pre-emptively & learn

1. Get security data (data transfers, authentication, incidents)
2. Explore data & build KPIs
3. Fit, apply & validate models on past / real-time data
4. Predict and act. Identify anomalous behaviors, create alerts
5. Surface incidents to Security Ops, who INVESTIGATES & ACTS
Business Analytics: Predict Customer Churn

**Problem:** Customer churn causes big time & money expense

**Solution:** Build predictive model to forecast possible churn, act pre-emptively & learn

1. Get customer data (set-top boxes, web logs, transaction history)
2. Explore data & build KPIs
3. Fit, apply & validate models on past / real-time data
4. Predict and act. Identify churning customers, create alerts
5. Surface incidents to Business Ops, who INVESTIGATES & ACTS
Summary: The ML Process

Problem: <Stuff in the world> causes big time & money expense
Solution: Build predictive model to forecast <possible incidents>, act pre-emptively & learn

1. Get all relevant data to problem
2. Explore data & build KPIs
3. Fit, apply & validate models on past / real-time data
4. Predict and act. Identify notable events, create alerts
5. Surface incidents to X Ops, who INVESTIGATES & ACTS
ML with Splunk
ML 101: What is it?

- Machine Learning (ML) is a process for generalizing from examples
  - Examples = example or “training” data
  - Generalizing = build “statistical models” to capture correlations
  - Process = ML is never done, you must keep validating & refitting models

- Simple ML workflow:
  - Explore data
  - FIT models based on data
  - APPLY models in production
  - Keep validating models

“All models are wrong, but some are useful.”
- George Box
Building ML Apps

• An ML application is an app which uses ML to solve a business problem
• An algorithm is just one piece of a larger solution
• Example: Outage Forecasting app, with workflows, analytics & alerts
  – Personas: deliver insights to IT Ops
  – Data: all IT-relevant data (incl. tickets)
  – Analytics: compute KPIs from raw data $\leftarrow 80\%$ of work here
  – ML: correlate outages with traffic, latency, resource usage, etc.

• Keep in mind:
  – Who is this solution designed for? Does this solve their problem?
  – What data is needed? What KPIs do we have to monitor? Who builds KPIs?
  – How do we fit/apply models as part of the app? Who validates models?
Machine Learning and Advanced Analytics at Splunk

Packaged Machine Learning
Easy to use ML integrated into standard day-to-day operations

Custom Machine Learning
Predictive analytics tailored for a customer’s specific environment and target use cases

From platform to packaged premium solutions
Machine Learning in Splunk ITSI

**Adaptive Thresholding:**
- Learn baselines & dynamic thresholds
- Alert & act on deviations
- Manage for 1000s of KPIs & entities
- Stdev/Avg, Quartile/Median, Range

**Anomaly Detection:**
- Find “hiccups” in expected patterns
- Catches deviations beyond thresholds
- Uses advanced proprietary algorithm
Splunk User Behavior Analytics (UBA)

• Understand normal & anomalous behaviors for ALL users
• UBA detects Advanced Cyberattacks and Malicious Insider Threats
• Lots of ML under the hood:
  – Behavior Baselining & Modeling
  – Anomaly Detection (30+ models)
  – Advanced Threat Detection
• E.g., Data Exfil Threat:
  – “Saw this strange login & data transfer for user mpittman at 3am in China…”
  – Surface threat to SOC Analysts
Splunk Machine Learning Toolkit

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

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

**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*
Building ML Apps
1. Where’s the Data & Who Needs it?

- Prioritize & solve the big problems:
  - Cell tower or critical infrastructure failing
  - Hard-to-find, high-risk behaviors

- Use ALL data to help solve problems:
  - E.g., can’t identify app crashes without app data
  - Enrich machine data with tickets, app data, DB, etc.

- Find the stakeholders:
  - Who owns these problems?
  - Who will invest in you to build a solution?

- Solutions not science projects:
  - If it’s mission-critical, treat it as such (Dev -> QA -> Prod)
  - Prototype: build simple MVPs, show value, iterate
2. Explore Data & Prototype in Splunk

- Data Science is 80% Data Exploration – Build KPIs!!
- Is the data in Splunk?
  - Munge it in Splunk
  - ML prototype in Splunk
  - Model analysis/validation: Splunk + other tools
  - Operationalize in Splunk
- Data not in Splunk? Why not?
  - 1000+ Splunk apps & add-ons
  - Get DB data using DB Connect
  - Get Hadoop data using Hadoop Connect
  - Get NoSQL data using Splunk apps/add-ons
3. Fit, Apply & Validate Models

- **ML SPL** – New grammar for doing ML in Splunk
- **fit** – fit models based on training data
  - \([\text{training data}] \mid \text{fit LinearRegression costly_KPI from feature1 feature2 feature3 into my_model}\)
- **apply** – apply models on testing and production data
  - \([\text{testing/production data}] \mid \text{apply my_model}\)
- **Validate Your Model** (The Hard Part)
  - Why hard? Because statistics is hard! Also: model error ≠ real world risk.
  - Analyze residuals, mean-square error, goodness of fit, cross-validate, etc.
  - Take Splunk’s Analytics & Data Science Education course
LOTS of new algorithms in ML Toolkit v2.0

- ARIMA
- SGDClassifier
- SGDRegressor
- DecisionTreeClassifier
- DecisionTreeRegressor
- AdaBoostRegressor
- BernoulliNB
- Birch
- DBSCAN
- ElasticNet
- FieldSelector
- GaussianNB
- KMeans

- KernelPCA
- KernelRidge
- Lasso
- LinearRegression
- LogisticRegression
- OneClassSVM
- PCA
- RandomForestClassifier
- RandomForestRegressor
- Ridge
- SVM
- SpectralClustering
- TFIDF
- StandardScaler
4. Predict & Act

- Forecast KPIs & predict notable events
  - When will my system have a critical error?
  - In which service or process?
  - What’s the probable root cause?

- How will people act on predictions?
  - Is this a Sev 1/2/3 event? Who responds?
  - Deliver via Notable Events or dashboard?
  - Human response or automated response?

- How do you improve the models?
  - Iterate, add more data, extract more features
  - Keep track of true/false positives
5. Operationalize Your Models

- Operationalizing closes the loop of the ML Process:
  1. Get data
  2. Explore data & fit models
  3. Apply & validate models
  4. Forecast KPIs & events
  5. Surface incidents to Ops team

- When you deliver the outcome, keep track of the response
  - Human-generated response (detailed journal logs, etc)
  - Machine-generated response (workflow actions, etc)
  - External knowledge (closed tickets data, DB records, etc)

- Then operationalize: feed back Ops analysis to data inputs, repeat

- Lots of hard work & stats, but lots of value will come out.
Show me the ML!
Example ML Architectures

• Example 1: Build models on Enterprise Security alerts
  – Data comes from: Splunk + ES indexes (index=notable, index=risk)
  – Fit workflow: fit models based on user/entity behavior
  – Apply workflow: apply model scores as part of correlation search
  – Who validates: SOC content developers
  – Action/Outcome: Deliver alerts to SOC analysts, reduce false positives & alert volume

• Example 2: Build models across clickstream + transaction data
  – Data comes from: Splunk + DB/Hadoop/NoSQL
  – Fit workflow: fit models based on customer behavior & actions
  – Apply workflow: apply model scores as part of regular jobs
  – Who validates: Business analysts + Splunk power users
  – Action/Outcome: Target qualified marketing leads, reduce customer churn
Example 1: Cluster IPs based on Security Alerts

<table>
<thead>
<tr>
<th>src</th>
<th>Account Deleted</th>
<th>Brute Force Access</th>
<th>Excessive Failed Logins</th>
<th>Excessive Hosts</th>
<th>Excessive Email</th>
<th>Unroutable Activity</th>
<th>Vulnerability Scanner Detected (by events)</th>
<th>Vulnerability Scanner Detected (by targets)</th>
<th>Watchlisted Event Observed</th>
<th>Total</th>
<th>cluster</th>
</tr>
</thead>
<tbody>
<tr>
<td>10.141.2.170</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>61</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>61</td>
<td>2</td>
</tr>
<tr>
<td>10.1.21.153</td>
<td>0</td>
<td>7</td>
<td>20</td>
<td>0</td>
<td>7</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>0</td>
<td>37</td>
<td>4</td>
</tr>
<tr>
<td>10.10.41.200</td>
<td>0</td>
<td>7</td>
<td>19</td>
<td>0</td>
<td>7</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>0</td>
<td>36</td>
<td>4</td>
</tr>
<tr>
<td>10.11.36.20</td>
<td>0</td>
<td>7</td>
<td>19</td>
<td>0</td>
<td>7</td>
<td>0</td>
<td>0</td>
<td>3</td>
<td>0</td>
<td>36</td>
<td>4</td>
</tr>
<tr>
<td>10.1.21.67</td>
<td>0</td>
<td>7</td>
<td>18</td>
<td>0</td>
<td>7</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>0</td>
<td>35</td>
<td>4</td>
</tr>
<tr>
<td>10.116.240.105</td>
<td>0</td>
<td>7</td>
<td>18</td>
<td>0</td>
<td>7</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>0</td>
<td>35</td>
<td>4</td>
</tr>
<tr>
<td>10.11.36.7</td>
<td>0</td>
<td>7</td>
<td>14</td>
<td>0</td>
<td>7</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>28</td>
<td>1</td>
</tr>
</tbody>
</table>
Example 2: Fit Regression Model on Sales Data

```java
index=pos TYPE=DONUT REGISTER=5
| timechart count, max(date_hour) as max_date_hour span=1h
| reverse | streamstats window=2 sum(count) as rollingforecast3day | reverse
| eval target =(rollingforecast3day-count) | fields - rollingforecast3day
| fit LinearRegression "target" from "max_date_hour" into retailmodel DONUT_sample
| eval residual='predicted(target)'-target | fields - target max_date_hour | rename count AS "Actual Sales", predicted(target) AS "Model", residual AS "Model Accuracy"
```

429,252 events (7/26/16 12:00:00.000 AM to 7/29/16 12:00:00.000 AM) No Event Sampling

Events Patterns Statistics (72) Visualization

*Column Chart*  *Format*
Example 2: Apply Regression Model on Sales Data

```
| timechart count, max(date_hour) as max_date_hour span=1h
| apply retailmodel_DONUT
| eval residual=count-'predicted(target)' | fields - max_date_hour,target | rename count AS "Actual Volume", predicted(target) AS "Predicted Volume", residual AS "Residual"
```

148,229 events (7/28/16 4:00:00.000 PM to 7/29/16 4:37:16.000 PM) No Event Sampling

- Column Chart

Bars represent:
- Actual Volume
- Predicted Volume
- Residual

X-axis: Time
- 4:00 PM Thu Jul 28 2016
- 8:00 PM
- 12:00 AM Fri Jul 29
- 4:00 AM
- 8:00 AM
- 12:00 PM

Y-axis: Count
Next Steps with Splunk ML

• **Reach out to your Tech Team! We can help architect ML solutions.**
  - ITSI: surface anomalous alerts & outliers, better root-cause analysis
    -- Free ITSI Cloud Sandbox! [http://splunk.force.com/SplunkCloud?prdType=ITSI](http://splunk.force.com/SplunkCloud?prdType=ITSI)
  - UBA: track anomalous behaviors, surface live threats
  - ML Toolkit for building your own ML solutions
    -- Completely free! [http://tiny.cc/splunkmlapp](http://tiny.cc/splunkmlapp)
• **Other cool ML talks:**
  -- When Recommendation Systems Go Bad
  -- Hidden Biases in Machine Learning and Big Data
• **Join the ML Early Adopter Program!**
  -- mlprogram@splunk.com