Anomaly Detection On Business Items With Machine Learning Algorithms

Andre Pietsch
Product Manager Splunk, OTTO Hamburg Germany

Stefan Scholz
Sr Consultant, LC Systems Munich Germany
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

- About Otto IT
- About LC Systems
- Initial Situation
- Next Generation Of Data Analytics
- Results
About OTTO
History

Werner Otto
Founder of the company in 1949
CEO until 1981

Dr. Michael Otto
Chairman & CEO of Otto group until 2007
Chairman of Otto group supervisory board until today
Selection Of Catalogs
OTTO – Subsidiary Of The Otto Group

- Employees: 4,350
- Revenue 2.6 Billion € (2015/2016)
- 90% Online
- > 2.2 Million items
- 6,000 Brands

Headquarters: OTTO-Campus in Hamburg, Germany
OTTO – Subsidiary Of The Otto Group

- #1 Furniture online Germany (before IKEA)
- #2 Germany (after Amazon)
- > 678,000 Facebook fans (03/2016)
- > 28,000 Twitter followers
- 90,000 Service requests via Facebook and Twitter (2015/2016)

Headquarters: OTTO-Campus in Hamburg, Germany
Otto Group - International Success

Otto group is a globally active retail and retail-related services group with 123 major companies in more than 30 countries.

49,597 Employees (2015/2016)
Revenue 12,5 Billion EUR (2015/12)
Otto Group - Main Business Segments

- **Multichannel Retail**: E-Commerce, Catalogs and Over-The-Counter Retail.
- **Financial Services**: Receivables, Payment Services, Information and Liquidity Management.
- **Services**: Procurement, Quality Control, Transportation and Warehousing, Delivery to Private and Business Customers.
About LC Systems
Locations

Headquarter
LC Systems-Engineering AG
Postfach 40, Seestrasse 24
CH-9326 Horn

Office Basel
Reinacherstrasse 129
CH-4053 Basel

Office Berne
Schwarztorstr. 9
CH-3007 Berne

Office Germany
LC Systems GmbH
Landsberger Straße 302
D-80687 Munich

www.lcsystems.ch
www.lcsystems.de
Together On The Road To Success
(Not Conclusive)

Pharma  Finance  Automotive Industry  Internet Provider  Research and Development  Service Provider  Trading

Novartis  Credit Suisse  RBS Coutts  Volkswagen  Swisscom  CERN  DATEV  Otto

Zürcher Kantonalbank  Daimler  Telefonica  O2  LZ  Citrix  Eventim

PostFinance  Liechtensteinische Landesbank  T-Systems  Fraunhofer  1&1  Fiducia GAD

1822direkt

Julius Bär
Data Analytics
Using And Evaluating Data In The Best Possible Way

- Services
  - Consulting
  - Workshops
  - Use cases
  - Proof of Concept
  - Project definition
  - Methodically structured implementation
  - Operation
  - Managed services
  - Training
Initial Situation
The Backend – Historically Grown

- SOA organized
- online UseCases
- Batch processing
- Mainframe in the Basement
The Backend – Some Order Into Chaos

- 4-5 Environments at the same time
- 1500 Interfaces
- 63 Systems per environment
- 200 User interfaces
- 13 Billion events in 36 days
How Can I Get An Alert?

| tstats prestats=t span=1h count where index="_*" OR index="**" by _time | timechart span=1h count

Start with:
- Unqualified eventcounts

Qualify them to:
- Transactions from payment provider
- Webshop requests
- Social media contacts
How Can I Get An Alert?

| tstats prestats=t span=1h count where index="_*" OR index="*" by _time | timechart span=1h count | eval threshold=20000000

Method of „fierce inspection“
- Needs a lot of human resources
And What About All The Other Peaks?

Peaking and plunging affects your business
- Transactions from payment provider
- Webshop requests
- Social media contacts
Next Generation Of Data Analytics
"...[we are] left with only one option, [we] gonna have to science the shit out of this." (Marc Watney, SOL 71)

Data Science – The Science

Unsupervised Learning
- Clustering
  - K-Means
  - EM-Algorithmus
  - DBSCAN
  - Birch
  - Spectral Clustering

Supervised Learning
- Regression
  - Linear Regression
  - Decision Trees
  - RandomForest

Continuous

Categorical
- Association analysis
  - Apriori
  - FP-Growth

Already available in Splunk or Machine Learning Toolkit
Data Science – The Science

For beginners:
- Transform to categorical
- Use statistic functions of splunk

ML Toolkit used for thesis:
- Known algorithms
- Great variety of algorithms

Train only with a sample (70%):
- Overfitting is a risk
- Model could get too accustomed to data

Unsupervised Learning
- Clustering
  - K-Means
  - EM-Algorithmus
  - DBSCAN
  - Birch
  - Spectral Clustering

Supervised Learning
- Regression
  - Linear Regression
  - Decision Trees
  - RandomForest

Classification
- Logistic Regression
- Support Vector Machine
  - Decision Trees
  - RandomForest

+ Already available in Splunk> or Machine Learning Toolkit
### Data Science – The Science

#### Documentation of ML Toolkit:


#### Especially the commands:

- sample rate=0.7
- fit "MyModel" ...
- apply "MyModel" ...

#### Evaluate the results to rank the algorithms:

```
`confusionmatrix("behavior","predicted(behavior)")`
```

<table>
<thead>
<tr>
<th></th>
<th>Logistic Regression</th>
<th>Support Vector Machine</th>
<th>RandomForest Classifier</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accuracy</td>
<td>0.92 (92 %)</td>
<td>1.00 (100 %)</td>
<td>1.00 (100 %)</td>
</tr>
<tr>
<td>Classification Error</td>
<td>0.08 (8 %)</td>
<td>0.00 (0 %)</td>
<td>0.00 (0 %)</td>
</tr>
<tr>
<td>Precision</td>
<td>0.48 (48 %)</td>
<td>1.00 (100 %)</td>
<td>1.00 (100 %)</td>
</tr>
<tr>
<td>Recall</td>
<td>1.00 (100 %)</td>
<td>1.00 (100 %)</td>
<td>1.00 (100 %)</td>
</tr>
<tr>
<td>F-1</td>
<td>0.65 (65 %)</td>
<td>1.00 (100 %)</td>
<td>1.00 (100 %)</td>
</tr>
</tbody>
</table>

#### Results Prediction accuracy (on trained data)

<table>
<thead>
<tr>
<th></th>
<th>Logistic Regression</th>
<th>Support Vector Machine</th>
<th>RandomForest Classifier</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accuracy</td>
<td>0.95 (95 %)</td>
<td>0.93 (93 %)</td>
<td><strong>0.99 (99 %)</strong></td>
</tr>
<tr>
<td>Classification Error</td>
<td>0.05 (5 %)</td>
<td>0.07 (7 %)</td>
<td>0.01 (1 %)</td>
</tr>
<tr>
<td>Precision</td>
<td>0.59 (62 %)</td>
<td>undefinierbar</td>
<td>1.00 (100 %)</td>
</tr>
<tr>
<td>Recall</td>
<td>1.00 (100 %)</td>
<td>0.00 (0 %)</td>
<td>0.91 (91 %)</td>
</tr>
<tr>
<td>F-1</td>
<td>0.74 (74 %)</td>
<td>0.00 (0 %)</td>
<td>0.95 (95 %)</td>
</tr>
</tbody>
</table>

Results Prediction accuracy (on test data)
Data Science – The Science

Visualize the predictions to understand found anomalies.

```plaintext
| eval predictedAnomaly=if('predicted(behavior)'=="Anomaly", count, null)
| table _time, count, predictedAnomaly
```
Data Science – The Process

What is the scientific goal?
What would you do if you had all data?
What do you want to predict or estimate?

How was the data sampled?
Which data is relevant?
Are there privacy issues?

Plot the data.
Are there anomalies?
Are there patterns?

Build a model.
Fit the model.
Validate the model.

What did we learn?
Do the results make sense?
Can we tell a story?

Ask an interesting question
GET the data
EXPLORE the data
MODEL the data
Communicate and visualize the results
The Data Science Process

Translated to Splunk Technology

- What type of information is available?
- Use Case
- Business Case
- Predict or estimate
- How to present the data?
The Data Science Process

Translated to Splunk Technology

• Location of the data
  – Log Files → Universal Forwarder
  – Databases → DB Connect
  – Applications → http Collector, REST
  – Other sources → REST, HEC, SYSLOG...

• Verify the data
  – Field-Extractions
  – CIM compliance

• Permissions
  – Users / Roles
  – Indexes
The Data Science Process

Translated to Splunk Technology

- Plot the data
  - Identify patterns with the patterns tab
  - Searching for anomalies
    - correlate
    - associate
    - analyze fields
    - cluster / kmeans
    - anomalies
    - anomalous value

Ask an interesting question

GET
the data

EXPLORE
the data

MODEL
the data.

Communicate
and visualize
the results.
The Data Science Process

Translated to Splunk Technology

- Splunk certified App
  - Machine Learning Toolkit and Showcase
    - need Splunk 6.4
    - need Python for Scientific Computing
  - Splunk IT Service Intelligence
- 3rd party vendors
  - Prelert Anomaly Detective® App for Splunk®
The Data Science Process

Communicate and visualize

- Use Splunk Apps Visualization
- Create your own Views and Dashboards
- Talk to the requester
- Verify outcome
- Deploy to production

- Estimate the knowledge
- If needed restart the loop and include the new findings
Results
Data Science – The Process

Technologies that have been analyzed

- Prelert Anomaly Detective® App for Splunk®
- Machine Learning Toolkit and Showcase
  - need Splunk 6.4
  - need Python for Scientific Computing
- Splunk IT Service Intelligence
Main Findings Of The Collaboration

Companies have a huge interest in identifying anomalies, but most of them are still in the process of understanding and preparing their data.

C. Günther (LC Systems)

A simple approach to anomaly detection is to use the static methods integrated in splunk>. If you want to use ML algorithms for that, the ML Toolkit or the Anomaly Detective App (Prelert) is recommended.

P. Drieger (Splunk)

You need a baseline to use your data for anomaly detection, because you have to define what's "normal".

M. Borner (LC Systems)
Thanks To

Hamburg University of Applied Sciences - Prof. Dr. Schoeneberg, Niklas Netz

LC Systems - Christian Günther, Mika Borner

splunk> - Philipp Drieger, Holger Sesterhenn
What Now?

Related breakout sessions and activities...

- Splunk for Donuts – Intermediate – Thursday | 2:35PM
- Solve Big Problems with ML – Advanced – Thursday | 1:30PM
- Predicting Incidents with ML – Advanced – Thursday | 2:35PM
- Splunk UBA – A Data Scientist in a Box – Beginner – Tuesday
- A (VERY) Brief Introduction to ML – Beginner – Wednesday
- Demystifying ML and Anomaly Det. – Intermediate - Wednesday
THANK YOU
Backup