Trends in Data Science

Massive Amount of Data

Analytical Skill Gap

“Demand for deep analytical talent in the US could be 50 to 60% greater than its projected supply by 2018”

McKinsey Global Institute

Ever Faster Decision Cycle

Gartner
So how does Automated Analytics help?

<table>
<thead>
<tr>
<th>You are a Data Scientist</th>
<th>You are an Analyst</th>
<th>You are a Company</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Automate the recurring tasks and <strong>save time</strong></td>
<td>• <strong>Get access</strong> to the world of Predictive Analytics / Machine Learning</td>
<td></td>
</tr>
<tr>
<td>• Get inspiration on which direction to investigate manually</td>
<td>• <strong>Deliver new benefits</strong> by providing Predictive Models in addition to Business Intelligence</td>
<td></td>
</tr>
<tr>
<td>• Help <strong>structure your data sets</strong> for manual approach</td>
<td>• <strong>Build on</strong> existing analytical skillset</td>
<td></td>
</tr>
<tr>
<td>• <strong>Deploy</strong> models into production with ease</td>
<td>• <strong>Find a new career path</strong></td>
<td></td>
</tr>
<tr>
<td>• Have <strong>additional functionality</strong> in your portfolio to tackle day to day challenges</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Support Productivity**

**Enable Users**

**Scale**

• Benefit from Predictive Insight where needed in business processes
• Scale the use of predictive models without manual bottlenecks
• Accelerate your path to a digital business
But I am a Data Scientist, and I am efficient «by hand»

A logistic regression only takes a few lines of code in MLlib.

```scala
import org.apache.spark.mllib.classification.LogisticRegressionModel,
  LogisticRegression
import org.apache.spark.mllib.evaluation.BinaryClassificationMetrics
import org.apache.spark.mllib.util.MLlibUtils

// Load training data in LIBSVM format.
val data = MLlibUtils.loadSVMFile(sc, "data/mllib/sample_libsvm_data.txt")

// Split data into training (80%) and test (20%).
val splits = data.randomSplit(Array(0.8, 0.2), seed = 1L)
val training = splits(0).cache()
val test = splits(1)

// Run training algorithm to build the model.
val numIterations = 10
val model = SVMWithSGD.train(training, numIterations)

// Clear the default threshold.
model.clearThreshold()

// Compute raw scores on the test set.
val scoreAndLabels = test.map { point =>
  val score = model.predict(point.features)
  (score, point.label)
}

// Get evaluation metrics.
val metrics = new BinaryClassificationMetrics(scoreAndLabels)
val auROC = metrics.areaUnderROC()
println("Area under ROC = \" + auROC

// Save and load model.
model.save(sc, "target/logisticRegressionModel")
val loadedModel = SVMWithSGD.load(sc, "target/logisticRegressionModel")
```

However, most projects are more complex

The Cross Industry Standard Process for Data Mining (CRISP-DM)

The previous code only creates 1 model. The remaining aspects are not addressed yet.

Automated Predictive Analytics
The Cross Industry Standard Process for Data Mining (CRISP-DM)

- Mass-produce such best-performing models
- Monitor these models on their predictive quality
- Retrain if needed
- Calculate new scores and write back or into business applications


Explorative / Agile BI frontend
- Derive new variables in graphical interface that describe the subject
- Handle missing values and outliers
- Create robust groups
- Calculate many different models

- Evaluate models on unseen data and select the best-performing
- Interpret model and discuss insight with the business department
Automated Analytics

How?
The Principles

- The technology used in the Automated Mode of SAP Predictive Analytics is an implementation of the theory of statistical learning from Vladimir Vapnik. SAP obtained this technology with the acquisition of a company called KXEN in 2013.

- Some principles are key:
  - No hypothesis whatsoever, no testing of them
  - No required distribution of the predictors
  - Ability to handle large number of predictors
  - No assumption on relationships between predictors
  - The user has control of the process

- The process is 2 steps:
  - Preparation of the data for further processing / encoding
  - Algorithmics

- It relies on Structured Risk Minimization (SRM) which is implemented in the encoding but also in all steps of model building. The algorithmics is Ridge Regression.
Automated Predictive Analytics
The Cross Industry Standard Process for Data Mining (CRISP-DM)

- Mass-produce such best-performing models
- Monitor these models on their predictive quality
- Retrain if needed
- Calculate new scores and write back or into business applications

Explorative / Agile BI frontend
- Derive new variables in graphical interface that describe the subject
- Handle missing values and outliers
- Create robust groups

Calculate many different models
- Evaluate models on unseen data and select the best-performing
- Interpret model and discuss insight with the business department

Data Preparation
Turning raw data into wide descriptive datasets

Creating a semantic layer. The structure does not have to be persisted.

Tables  Joins  Aggregates
With understanding of time
Data Preparation

Turning raw data into wide descriptive datasets

Creating a semantic layer. The structure does not have to be persistet.

- Name
- Age
- Martial status
- Account Balance today
- Average Account Balance -1 Quarter
- Average Account Balance -2 Quarters
- Average Account Balance -3 Quarters
- Differences in Avg Account Balance in Euro
- Differences in Avg Account Balance in %
- Average Account Balance -1 Year
- Average Account Balance -2 Years
- Average Account Balance -3 Years
- Differences in Avg Account Balance in Euro
- Differences in Avg Account Balance in %
- Maximum Account Balance -1 Quarter
- Maximum Account Balance -2 Quarters
- Maximum Account Balance -3 Quarters
- Differences in Max Account Balance in Euro
- Differences in Max Account Balance in %
- Maximum Account Balance -1 Year
- Maximum Account Balance -2 Years
- Maximum Account Balance -3 Years
- Differences in Max Account Balance in Euro
- Differences in Max Account Balance in %
- …
- …
- … and thousands of further columns…

Wide descriptive datasets
Big Data is not just big
Wide, or deep, or both
Why Big Data for Predictive?
Lift with Simple Aggregates

20 Variables
• Demographics / Account Information
• Simple Aggregates (e.g. Account Balance, Total Usage)
Why Big Data for Predictive?
Lift with Complex Aggregates

100 Variables
- Pivoting Transactions (e.g. Calls by Type)
- Time-Sensitive Aggregates (e.g. Calls by Week)
Why Big Data for Predictive?
Lift with Social Network Analysis

200 Variables
• Social Network Analysis (e.g. Calls in First Circle)
• Community Detection (e.g. Community Churn Rate)
Data Preparation

Encoding the columns, Nominal and Ordinal columns

Example: Let’s consider a Variable V1 with 4 categories A, B, C and D and some missing values.

<table>
<thead>
<tr>
<th>Category / Level</th>
<th>Percent of target variable in Estimation</th>
<th>Percent of target variable in Validation</th>
<th>Assigned value in encoded dataset</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>0.1</td>
<td>0.1</td>
<td>A</td>
</tr>
<tr>
<td>B</td>
<td>0.2</td>
<td>0.2</td>
<td>B</td>
</tr>
<tr>
<td>C</td>
<td>0.15</td>
<td>0.3</td>
<td>KxOther</td>
</tr>
<tr>
<td>D</td>
<td>0.1</td>
<td>0.1</td>
<td>D</td>
</tr>
<tr>
<td>E</td>
<td>0.35</td>
<td>0.15</td>
<td>KxOther</td>
</tr>
<tr>
<td>NULL</td>
<td>0.2</td>
<td>0.2</td>
<td>KxMissing</td>
</tr>
</tbody>
</table>

Categories with low frequency (outliers) are put together in a noise category called KxOther. It contains as well categories that are not robust i.e. that don’t have the same target rate between Estimation and Validation (tested with a Chi Square Test of Independence).
Data Preparation
Binning to obtain robust groups

- Grouping can help to increase robustness. Categories are grouped depending on the target encoding.

<table>
<thead>
<tr>
<th>Category / Level</th>
<th>Percent of target variable in Estimation</th>
<th>Percent of target variable in Validation</th>
<th>Assigned value in encoded dataset</th>
<th>Grouping</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>0.1</td>
<td>0.1</td>
<td>A</td>
<td>A;D</td>
</tr>
<tr>
<td>B</td>
<td>0.2</td>
<td>0.2</td>
<td>B</td>
<td>B;KxMissing</td>
</tr>
<tr>
<td>C</td>
<td>0.15</td>
<td>0.3</td>
<td>KxOther</td>
<td>KxOther</td>
</tr>
<tr>
<td>D</td>
<td>0.1</td>
<td>0.1</td>
<td>D</td>
<td>A;D</td>
</tr>
<tr>
<td>E</td>
<td>0.35</td>
<td>0.15</td>
<td>KxOther</td>
<td>KxOther</td>
</tr>
<tr>
<td>NULL</td>
<td>0.2</td>
<td>0.2</td>
<td>KxMissing</td>
<td>B;KxMissing</td>
</tr>
</tbody>
</table>

From the encoding we can expect that A and D could be regrouped as well as B and NULL (as they have similar . This is done iteratively:

- by calculating $K_I+K_R$ for the non-regrouped categories and the regrouped ones
- If $K_I+K_R$ doesn’t decrease (with a tolerance), the group is kept
- Further grouping is tried to the point where $K_I + K_R$ decreases
Automated Predictive Analytics
The Cross Industry Standard Process for Data Mining (CRISP-DM)

- Mass-produce such best-performing models
- Monitor these models on their predictive quality
- Retrain if needed
- Calculate new scores and write back or into business applications

Modeling Ridge Regression

The Ridge Regression penalizes the size of the coefficients by minimizing this extended term:

$$
\left( \sum_{i=1}^{n} (y_i - x_i^T \beta)^2 \right) + \lambda \sum_{j=1}^{p} \beta_j^2
$$

where $\lambda$: Ridge Parameter, $p$: number of parameters.

The coefficients that minimize that error are estimated with:

$$
\hat{\beta} = (X^T X + \lambda I)^{-1} X^T y
$$

Modeling
Selecting the best model

- By playing with $\lambda$, more or less constraint is applied on the coefficients of the regression.
  - If a lot of constraint is applied, the Training error ($\varepsilon_t$) is high but the Generalization error ($\varepsilon_g$) is low
  - Inversely, if little constraint is applied, the Training error ($\varepsilon_t$) is low but the Generalization ($\varepsilon_g$) is high
Automated Predictive Analytics
The Cross Industry Standard Process for Data Mining (CRISP-DM)

- Mass-produce such best-performing models
- Monitor these models on their predictive quality
- Retrain if needed
- Calculate new scores and write back or into business applications

Explorative / Agile BI frontend
- Derive new variables in graphical interface that describe the subject
- Handle missing values and outliers
- Create robust groups
- Calculate many different models

- Evaluate models on unseen data and select the best-performing
- Interpret model and discuss insight with the business department

Closed Loop
Automatically Retrain and Apply Models

- Maintain large number of models
- Automatically retrain models when needed
- Automatically apply models and persist scores to source systems or business applications
Automated Predictive Analytics
The Cross Industry Standard Process for Data Mining (CRISP-DM)

- Mass-produce such best-performing models
- Monitor these models on their predictive quality
- Retrain if needed
- Calculate new scores and write back or into business applications

**Explorative / Agile BI frontend**
- Derive new variables in graphical interface that describe the subject
- Handle missing values and outliers
- Create robust groups
- Calculate many different models

- Evaluate models on unseen data and select the best-performing
- Interpret model and discuss insight with the business department

Big Data in Hadoop

Features
- Commodity Hardware ($1500/ TB)
- Open Source Stack ( No Licensing fee)
- Elastic scaling
- scales linearly with # of nodes
- Easy to add 1000s of (cheap) nodes
- Code executes close to the data
Hadoop Perspective for 2016

Adoption interest for Spark has topped in Hadoop eco-system

Big data workloads in production jumped by nearly 30% from 2014 to 2015

2016’s #1 trend: Apache Spark will move from talking point into deployment

55% users want to leverage Hadoop for Business users and Advanced use cases

Source: http://www.syncsort.com/
Modeling for Big Data

Traditional Tiered Architecture vs. Native Spark Modeling

- Full dataset brought to application for processing
- Limited Performance, Scalability

- Data processing beside data
- Performance and scalability built-in

**Limited Data Processing on a single server**

- Predictive Analytics
- Automated tools

**1000s of Nodes designed for cost effective Data Processing**

- No Data Transfer
- JSON
- Stats

- SAP Predictive Analytics - Automated

**FULL Data Transfer**

- Spark
- Database

Scales dynamically
Native Spark Modeling - Architecture
Performance and Scalability
With and Without Native Spark Modeling

Summary

- 14 times faster for 15K var dataset
- 10 times faster for 2K var dataset
- Native Spark Modelling performance is better with bigger and wider datasets
- Scalability = quadratic $O(n^2)$ of matrix operations
Summary

More about Automated Analytics and Big Data

>> Tutorials and blog
>> Trial version download
Thank you

Contact information:

Priti Mulchandani
Product Manager for Big Data Analytics
p.mulchandani@sap.com

Andreas Forster
Global Center of Excellence
andreas.forster@sap.com