Putting Big Data & Analytics to Work!

Prof. dr. Bart Baesens

Department of Decision Sciences and Information Management, KU Leuven (Belgium)

School of Management, University of Southampton (United Kingdom)

Bart.Baesens@kuleuven.be
Twitter/Facebook/YouTube: DataMiningApps
www.dataminingapps.com
Presenter: Bart Baesens

- Studied at KU Leuven (Belgium)
  - Business Engineer in Management Informatics, 1998
  - PhD. in Applied Economic Sciences, 2003
- PhD. : Developing Intelligent Systems for Credit Scoring Using Machine Learning Techniques
- Professor at KU Leuven, Belgium
- Lecturer at the University of Southampton, UK
- Research: Big Data & Analytics, Credit Risk, Fraud, Marketing, ...
- YouTube/Facebook/Twitter: DataMiningApps
- www.dataminingapps.com
- Bart.Baesens@kuleuven.be
Example Publications
Living in a Data Flooded World!
The Analytics Process Model

1. Identify Business Problem
2. Identify Data Sources
3. Select the Data
4. Clean the Data
5. Transform the Data
6. Analyze the Data
7. Interpret, Evaluate, and Deploy the Model

- Preprocessing
- Analytics
- Post-processing
Feel the vibe!

Web Analytics

Customer Lifetime Value

Customer Segmentation

Fraud Detection

Response Modeling

Market Basket Analysis

Social Network Analytics

Churn Prediction

APPLICATIONS

Custom Tool
Example: marketing context

<table>
<thead>
<tr>
<th>Customer</th>
<th>Age</th>
<th>Recency</th>
<th>Frequency</th>
<th>Monetary</th>
<th>Churn</th>
</tr>
</thead>
<tbody>
<tr>
<td>John</td>
<td>35</td>
<td>5</td>
<td>6</td>
<td>100</td>
<td>Yes</td>
</tr>
<tr>
<td>Sophie</td>
<td>18</td>
<td>10</td>
<td>2</td>
<td>150</td>
<td>No</td>
</tr>
<tr>
<td>Victor</td>
<td>38</td>
<td>28</td>
<td>8</td>
<td>20</td>
<td>No</td>
</tr>
<tr>
<td>Laura</td>
<td>44</td>
<td>12</td>
<td>4</td>
<td>280</td>
<td>Yes</td>
</tr>
</tbody>
</table>

Analytical Software (SAS, R, Python, …)
Analytics

- Term often used interchangeably with data science, knowledge discovery, ...
- Essentially refers to extracting useful business patterns and/or mathematical decision models from a preprocessed data set
- **Predictive analytics**
  - Predict the future based on patterns learnt from past data
  - Classification (churn, response) versus regression (CLV)
- **Descriptive analytics**
  - Describe patterns in data
  - Clustering, Association rules, Sequence rules
Analytic Model requirements

- **Business relevance**
  - Solve a particular business problem

- **Statistical performance**
  - Statistical significance of model
  - Statistical prediction performance

- **Interpretability + Justifiability**
  - Very subjective (depends on decision maker), but CRUCIAL!
  - Often need to be balanced against statistical performance

- **Operational efficiency**
  - How can the analytical models be integrated with campaign management?

- **Economical cost**
  - What is the cost to gather the model inputs and evaluate the model?
  - Is it worthwhile buying external data and/or models?

- **Regulatory compliance**
  - In accordance with regulation and legislation
Post processing

• Interpretation and validation of analytical models by business experts
  • Trivial versus unexpected (interesting?) patterns

• Sensitivity analysis
  • How sensitive is the model wrt sample characteristics, assumptions and/or technique parameters?

• Deploy analytical model into business setting
  • Represent model output in a user-friendly way
  • Integrate with campaign management tools and marketing decision engines

• Model monitoring and backtesting
  • Continuously monitor model output
  • Contrast model output with observed numbers
Two Analytical Disconnects

• **Data versus Data Scientist**
  – Data: unstructured, distributed, noisy, time-evolving
  – Data Scientist: patterns in data, statistical significance, predictive power, structure the unstructured!

• **Data Scientist versus Business Expert**
  – Data Scientist: decision trees, logistic regression, random forests, area under ROC curve, top decile lift, R-squared, etc.
  – Business Expert: customers, marketing campaigns, risk mitigation, portfolios, profit, return on Investment (ROI), etc.

Visual Analytics as a mediator!
The Power of Visual Analytics

Carte Figurative des pertes successives au bataillon de l'Armée Française dans la campagne de Russie 1812-1813.

Paris, le 20 Novembre 1869.

Les nombres d'hommes présents, rejetés par les blessures et les maladies, sont, au total, plus importants que ceux de l'année précédente. Le portrait du bataillon qui combattit en Russie, le 1er août et en octobre, est celui qui a connu les pertes les plus importantes. La carte montre les pertes successives avec précision, tant sur le plan de la santé que sur le plan de la guerre.

TABLEAU GRAPHIQUE de la température en degrés du thermomètre de Réaumur au dessus du zéro.

Charles Minnard, 1869
Visual Analytics versus the Analytics Process Model

• **Data preprocessing**
  – Use Visual Analytics to find outliers, missing values, frequent/suspicious/interesting patterns, etc.
  – Visualisation unit: **Data**!

• **Model representation**
  – Use Visual Analytics to represent models in a user-friendly way
  – Visualisation unit: **Model formula**!
Visual Analytics versus the Analytics Process

**Model**

- **Model usage**
  - Use Visual Analytics to integrate models with other applications (e.g. GIS)
  - Visualisation unit: **Model interaction**!

- **Model backtesting**
  - Use Visual Analytics to monitor model performance
  - Visualisation unit: **Model performance**!
Data Preprocessing: cluster plot

http://blog.gramener.com/18/visualising-securities-correlation
Model Representation: Scorecards

\[
P(\text{Good} \mid \text{Age, Gender, Salary, ...}) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 \text{Age} + \beta_2 \text{Gender} + \beta_3 \text{Salary} + \ldots)}}
\]

<table>
<thead>
<tr>
<th>Characteristic Name</th>
<th>Attribute</th>
<th>Scorecard Points</th>
</tr>
</thead>
<tbody>
<tr>
<td>AGE 1</td>
<td>Up to 26</td>
<td>100</td>
</tr>
<tr>
<td>AGE 2</td>
<td>26 - 35</td>
<td>120</td>
</tr>
<tr>
<td>AGE 3</td>
<td>35 - 37</td>
<td>185</td>
</tr>
<tr>
<td>AGE 4</td>
<td>37+</td>
<td>225</td>
</tr>
<tr>
<td>GENDER 1</td>
<td>Male</td>
<td>90</td>
</tr>
<tr>
<td>GENDER 2</td>
<td>Female</td>
<td>180</td>
</tr>
<tr>
<td>SALARY 1</td>
<td>Up to 500</td>
<td>120</td>
</tr>
<tr>
<td>SALARY 2</td>
<td>501-1000</td>
<td>140</td>
</tr>
<tr>
<td>SALARY 3</td>
<td>1001-1500</td>
<td>160</td>
</tr>
<tr>
<td>SALARY 4</td>
<td>1501-2000</td>
<td>200</td>
</tr>
<tr>
<td>SALARY 5</td>
<td>2000+</td>
<td>240</td>
</tr>
</tbody>
</table>

Model Representation: Nomogram

Van Belle and Van Calster (2015)
Model Representation

• Bridge the gap between the analytical model and the business user
• Minimize information loss between analytical model and visual representation
• Business user engagement to foster trust
• Note: model interpretability depends upon business application
  – Credit risk versus medical diagnosis
  – Fraud detection versus fraud prevention
Model Representation: Decision Tables

RULE1: IF Avg Usage < 25 AND Internet Plan = Y AND Service Calls > 3
    THEN Churn

RULE2: IF Avg Usage < 25 AND Internet Plan = N THEN Churn

RULE3: IF Avg Usage ≥ 25 AND Internet Plan = Y THEN Not Churn

RULE4: IF Avg Usage < 25 AND Service Calls ≤ 3 THEN Not Churn

Rule Conflicts?
Rule Coverage?

Baesens, Van Vlasselaer, Verbeke, 2015.
### Model Representation: Decision Tables

<table>
<thead>
<tr>
<th>Contribution Rule(s):</th>
<th>R4</th>
<th>R1</th>
<th>R2</th>
<th>R2</th>
<th>R3</th>
<th>R3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Avg Usage</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Internet Plan</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Service Calls</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Churn</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Not Churn</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>&lt; 25</th>
<th></th>
<th>≥ 25</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Avg Usage</td>
<td>Y</td>
<td>N</td>
<td>Y</td>
<td>N</td>
</tr>
<tr>
<td>Internet Plan</td>
<td>Y</td>
<td>&gt; 3</td>
<td>N</td>
<td>&gt; 3</td>
</tr>
<tr>
<td>Service Calls</td>
<td>≤ 3</td>
<td>&gt; 3</td>
<td>≤ 3</td>
<td>&gt; 3</td>
</tr>
</tbody>
</table>

- **Conflict!** in the table indicates overlapping or conflicting rules.
- **No coverage!** indicates areas of the decision table that are not covered by any rules.
Model Usage: Geospatial plots

Model Usage: Segmentation

Google Analytics  www.dataminingapps.com
## Model Backtesting: Traffic Light Indicator Approach

<table>
<thead>
<tr>
<th></th>
<th>Baa1</th>
<th>Baa2</th>
<th>Baa3</th>
<th>Ba1</th>
<th>Ba2</th>
<th>Ba3</th>
<th>B1</th>
<th>B2</th>
<th>B3</th>
<th>Caa-C</th>
<th>Av</th>
</tr>
</thead>
<tbody>
<tr>
<td>PD</td>
<td>0.26%</td>
<td>0.17%</td>
<td>0.42%</td>
<td>0.53%</td>
<td>0.54%</td>
<td>1.36%</td>
<td>2.46%</td>
<td>5.76%</td>
<td>8.76%</td>
<td>20.89%</td>
<td>3.05%</td>
</tr>
<tr>
<td>DR</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1993</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.83%</td>
<td>0.00%</td>
<td>0.76%</td>
<td>3.24%</td>
<td>5.04%</td>
<td>11.29%</td>
<td>28.57%</td>
<td>3.24%</td>
</tr>
<tr>
<td>1994</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.59%</td>
<td>1.88%</td>
<td>3.75%</td>
<td>7.95%</td>
<td>5.13%</td>
<td>1.88%</td>
</tr>
<tr>
<td>1995</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>1.76%</td>
<td>4.35%</td>
<td>6.42%</td>
<td>4.06%</td>
<td>11.57%</td>
<td>2.51%</td>
</tr>
<tr>
<td>1996</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>1.17%</td>
<td>0.00%</td>
<td>3.28%</td>
<td>13.99%</td>
<td>0.78%</td>
</tr>
<tr>
<td>1997</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.47%</td>
<td>0.00%</td>
<td>1.54%</td>
<td>7.22%</td>
<td>14.67%</td>
<td>1.41%</td>
</tr>
<tr>
<td>1998</td>
<td>0.00%</td>
<td>0.31%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.62%</td>
<td>1.12%</td>
<td>2.11%</td>
<td>7.55%</td>
<td>5.52%</td>
<td>15.09%</td>
<td>2.83%</td>
</tr>
<tr>
<td>1999</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.34%</td>
<td>0.47%</td>
<td>0.00%</td>
<td>2.00%</td>
<td>3.28%</td>
<td>6.91%</td>
<td>9.63%</td>
<td>20.44%</td>
<td>3.35%</td>
</tr>
<tr>
<td>2000</td>
<td>0.28%</td>
<td>0.00%</td>
<td>0.97%</td>
<td>0.94%</td>
<td>0.63%</td>
<td>1.04%</td>
<td>3.24%</td>
<td>4.10%</td>
<td>10.88%</td>
<td>19.65%</td>
<td>3.01%</td>
</tr>
<tr>
<td>2001</td>
<td>0.27%</td>
<td>0.27%</td>
<td>0.00%</td>
<td>0.51%</td>
<td>1.38%</td>
<td>2.93%</td>
<td>3.19%</td>
<td>11.07%</td>
<td>16.38%</td>
<td>34.45%</td>
<td>5.48%</td>
</tr>
<tr>
<td>2002</td>
<td>1.26%</td>
<td>0.72%</td>
<td>1.78%</td>
<td>1.58%</td>
<td>1.41%</td>
<td>1.58%</td>
<td>2.00%</td>
<td>6.81%</td>
<td>6.86%</td>
<td>29.45%</td>
<td>3.70%</td>
</tr>
<tr>
<td>Av</td>
<td>0.26%</td>
<td>0.17%</td>
<td>0.42%</td>
<td>0.53%</td>
<td>0.54%</td>
<td>1.36%</td>
<td>2.46%</td>
<td>5.76%</td>
<td>8.76%</td>
<td>20.9%</td>
<td>3.05%</td>
</tr>
</tbody>
</table>

# Model Backtesting: Traffic Light Indicator Approach

<table>
<thead>
<tr>
<th>Color</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Green</td>
<td>everything is okay</td>
</tr>
<tr>
<td>Yellow</td>
<td>decreasing performance, which can be interpreted as an early warning</td>
</tr>
<tr>
<td>Orange</td>
<td>performance difference that should be closely monitored</td>
</tr>
<tr>
<td>Red</td>
<td>severe problem</td>
</tr>
</tbody>
</table>

Colors can be defined based on p-values.  
- p-value less than 0.01 = red  
- p-value between 0.01 and 0.05 = orange  
- p-value between 0.05 and 0.10 = yellow  
- p-value higher than 0.10 = green

Visualizing Temporal Patterns

- E.g. Churn Prediction in Telco

Homophily!
Conclusions

• Be aware but critical about emerging technologies (e.g. deep learning)
• Validation of patterns is key!
• Profit driven analytics (TCO and ROI)
• Visual analytics
Courses

• Analytics: Putting it all to Work (1 day)
  https://support.sas.com/edu/schedules.html?ctry=us&id=1339

• Advanced Analytics in a Big Data World (3 days)
  https://support.sas.com/edu/schedules.html?ctry=us&id=2169

• Credit Risk Modeling (3 days)
  https://support.sas.com/edu/schedules.html?ctry=us&id=2455

• Fraud Analytics using Descriptive, Predictive and Social Network Analytics (2 days)
  https://support.sas.com/edu/schedules.html?ctry=us&id=1912
More Information

**E-learning course: Advanced Analytics in a Big Data World**

https://support.sas.com/edu/schedules.html?id=2169&ctry=US

The E-learning course starts by refreshing the basic concepts of the analytics process model: data preprocessing, analytics and post processing. We then discuss decision trees and ensemble methods (random forests), neural networks, SVMs, Bayesian networks, survival analysis, social networks, monitoring and backtesting analytical models. Throughout the course, we extensively refer to our industry and research experience. Various business examples (e.g. credit scoring, churn prediction, fraud detection, customer segmentation, etc.) and small case studies are also included for further clarification. The E-learning course consists of more than 20 hours of movies, each 5 minutes on average. Quizzes are included to facilitate the understanding of the material. Upon registration, you will get an access code which gives you unlimited access to all course material (movies, quizzes, scripts, ...) during 1 year. The E-learning course focuses on the concepts and modeling methodologies and not on the SAS software. To access the course material, you only need a laptop, iPad, iPhone with a web browser. No SAS software is needed.
E-learning course: Fraud Analytics

https://support.sas.com/edu/schedules.html?ctry=us&id=1912

This new E-learning course will show how learning fraud patterns from historical data can be used to fight fraud. To be discussed is the use of descriptive analytics (using an unlabeled data set), predictive analytics (using a labeled data set) and social network learning (using a networked data set). The techniques can be applied across a wide variety of fraud applications, such as insurance fraud, credit card fraud, anti-money laundering, healthcare fraud, telecommunications fraud, click fraud, tax evasion, counterfeit, etc. The course will provide a mix of both theoretical and technical insights, as well as practical implementation details. The instructor will also extensively report on his recent research insights about the topic. Various real-life case studies and examples will be used for further clarification.
The E-learning course covers both the basic as well some more advanced ways of modeling, validating and stress testing Probability of Default (PD), Loss Given Default (LGD) and Exposure At Default (EAD) models. Throughout the course, we extensively refer to our industry and research experience. Various business examples and small case studies in both retail and corporate credit are also included for further clarification. The E-learning course consists of more than 20 hours of movies, each 5 minutes on average. Quizzes are included to facilitate the understanding of the material. Upon registration, you will get an access code which gives you unlimited access to all course material (movies, quizzes, scripts, ...) during 1 year. The course focusses on the concepts and modeling methodologies and not on the SAS software. To access the course material, you only need a laptop, iPad, iPhone with a web browser. No SAS software is needed. See https://support.sas.com/edu/schedules.html?ctry=us&id=2455 for more details.