VISUAL ANALYTICS

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. : 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

**Analytics in a Big Data World**

_Bart Baesens_

**Fraud Analytics**

_Dart Baesens, Veronique Van Vlaenderen, Wouter Verbeeke_

**Credit Risk Analytics**

_Bart Baesens, Daniel Roesch, Harald Scheule_

---

*Is Your Company Ready for HR Analytics?*

HR analytics is the next big change in human resources management.

activevoice.us
Overview

• Big Data & Analytics: setting the stage
• Power and premise of Visual Analytics
• Visual Analytics and the Analytics process model
  – Data preprocessing
  – Model representation
  – Model usage
  – Model backtesting
• Software
• Guidelines
• Conclusions
Living in a Data Flooded World!

- Web/email
- Call center
- Corporate data
- Survey
- Partners
- Customers
- Analytics
The Analytics Process Model

Identify Business Problem
Identify Data Sources
Select the Data
Clean the Data
Transform the Data
Analyze the Data
Interpret, Evaluate, and Deploy the Model

Baesens, 2015.
Feel the vibe!

APPLICATIONS
- Fraud Detection
- Market Basket Analysis
- Social Network Analytics
- Customer Lifetime Value
- Customer Segmentation
- Churn Prediction
- Response Modeling
- Web Analytics

Churn Prediction
Customer Segmentation
Customer Lifetime Value
Web Analytics
Response Modeling
Market Basket Analysis
Social Network Analytics
Fraud Detection
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

DIAGRAM OF THE CAUSES OF MORTALITY
IN THE ARMY IN THE EAST

The areas of the blue, red, & black wedges are each measured from
the centre as the common vertex.
The blue wedges measured from the centre of the circle represent area
for area the deaths from Preventible or Mitigable Zymotic Diseases, the
red wedges measured from the centre the deaths from wounds, & the
black wedges measured from the centre the deaths from all other causes.
The black line across the red triangle in Nov˚ 1854 marks the boundary
of the deaths from all other causes during the month.
In October 1854, & April 1855, the black area coincides with the red,
in January & February 1856, the blue coincides with the black.
The entire areas may be compared by following the blue, the red & the
black lines enclosing them.

Coxcomb diagram
Florence Nightingale, 1858
The Power of Visual Analytics

Charles Minnard, 1869
The Power of Visual Analytics

London cholera map
John Snow, 1854

London Tube map
Harry Beck, 1931
Visuals versus Statistics: Anscombe’s Quartet

\[
y = 3.00 + 0.500x
\]

Anscombe, 1973
Visual Analytics: The Premise

• Reduce cognitive overload by having users interact with data and/or analytical models using visual tools

• “the science of analytical reasoning facilitated by interactive visual interfaces“ (Thomas and Cook, 2005)

• Help data scientists + business users to explore and better understand data + models

• "A picture is worth a thousand words"
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!
The Analytics Process Model

- Identify Business Problem
- Identify Data Sources
- Select the Data
- Clean the Data
- Transform the Data
- Analyze the Data
- Interpret, Evaluate, and Deploy the Model

Preprocessing Analytics Post-processing

Baesens, 2015.
Data Preprocessing: Statistical plots

Histogram
Radar plot
Mosaic plot
Bubble plot
Contour plot
Scatter plot
Bar chart
Pie chart
Violin plot
Box plot

Aimed at Exploratory Data Analysis!
Data Preprocessing: OLAP

- **Dicing**

- **Slicing**

- **Roll-Up**

- **Drill-Down**

OLAP techniques include:

- **Slicing**: Selecting a portion of the data cube.
- **Dicing**: Selecting a subset of dimensions to form a smaller cube.
- **Roll-Up**: Summarizing data by aggregating lower levels.
- **Drill-Down**: Detailing data by moving to lower levels.

The diagram illustrates these concepts with a 3D cube representing sales data across regions, products, and quarters.
Data Preprocessing: Correlation matrix

First steps towards predictive modeling!
Data Preprocessing: cluster plot

http://blog.gramener.com/18/visualising-securities-correlation
Data Preprocessing: Unstructured Data

http://blogs.sas.com/content/sascom/2014/11/05/what-a-sentiment-word-cloud-revealed-about-apple-pay/
Data Preprocessing: Unstructured Data

http://journals.uic.edu/ojs/index.php/fm/article/view/941/863
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
   - Interpret, Evaluate, and Deploy the Model

Preprocessing Analytics Post-processing

Baesens, 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>Avg Usage ≤ 25</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Avg Usage ≥ 25</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>Avg Usage ≤ 25</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Avg Usage ≥ 25</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>Service Calls ≤ 3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Service Calls &gt; 3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Service Calls ≤ 3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Service Calls &gt; 3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Service Calls ≤ 3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Service Calls &gt; 3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ServiceCalls ≤ 3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ServiceCalls &gt; 3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ServiceCalls ≤ 3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ServiceCalls &gt; 3</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>
Model Representation: Scorecards

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

Van Belle and Van Calster (2015)
Model Representation: Nomogram

Van Belle and Van Calster (2015)
Model Representation: Sankey plot

Customer Journey Analytics!

www.sas.com
Model Representation: Time Series
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

Baesens, 2015.
Model Usage: Treemap
Model Usage: Geospatial plots

Model Usage: Segmentation

Google Analytics

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

<table>
<thead>
<tr>
<th>PD</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>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></td>
<td>3.05%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>DR</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>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>0.00%</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>
</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

Visualing Temporal Patterns

• E.g. Churn Prediction in Telco
Virtual Reality

- Aim is to create an immersive environment for the user
- E.g. Twitter Sentiment on MIT Campus using geo-tagged Tweets (Moran, 2014)
Visual Clutter
Visual Analytics: Guidelines

- The Visual clutter trap
  - From “information overload” to “visual overload”
  - Humans can only distinguish around 8 colors in 1 visual
- Invoke business user analytical curiosity
- Interactivity
- Consistency
- Avoid scrollbars using range sizing
- Naming, Naming, Naming!
  - e.g. axes, legends, units, currencies, coding schemes, etc.
Software

- SAS Visual Analytics (SAS)
- JMP (SAS)
- Tableau (Tableau)
- QlikView (Qlik)
- Spotfire (Tibco)
- i2 Analyst Notebook (IBM)
- Microsoft BI stack (Excel, PowerPivot, SQL Server)
Conclusions

• Visual analytics permeates the entire analytics process!
• Visual analytics catalyzes
  – model discovery
  – model interpretation
  – model monitoring
• Stay focussed; avoid the visual clutter trap!
References

Follow-up SAI Events

**PROCESS MINING**
- Workshop
- Brussels
- 20-03-2017 (13:30 - 20:30)

**TEXT ANALYTICS**
- Workshop
- Brussels
- 29-05-2017 (13:30 - 20:30)
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
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 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.
More Information

**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.