MTAT.03.287 Seminar on Business Intelligence

BI Tips and Practices

Siim Karus
siim.karus@ut.ee
Autumn 2016

https://courses.cs.ut.ee/2016/bis/
Topics

- Data Acquisition & Integration
- Data analysis
- Data Presentation
- Decision Support
DATA ACQUISITION & INTEGRATION
Transformations on data

- Even distribution for most classification tasks
- -1 .. 1 or 0 .. 1 min-max normalised for Perceptron networks (NB! Target feature)

ETL – Extract Transform Load

ETL – Extract Transform Load

• Make sure these tasks are separated/independent
  – Do not attempt to transform during extraction
• Make it incremental for operational use
  – Switching transform and load leads to loss of efficiency (but improves flexibility for exploratory studying)

Transformations on data
• Even distribution for most classification tasks
• -1 .. 1 or 0 .. 1 min-max normalised for Perceptron networks (NB! Target feature)
ETL – Extract Transform Load

- Design for scalability and maintenance
- Have fault tolerance built into
  - This also means monitoring and logging have to be set up
  - Store data on the original data source and tools (and versions) used for ETL
  - Try to quantify progress (especially with long-running transforms)

Transformations on data
- Even distribution for most classification tasks
- -1 .. 1 or 0 .. 1 min-max normalised for Perceptron networks (NB! Target feature)
Transformations on data

- Even distribution for most classification tasks
- -1 .. 1 or 0 .. 1 min-max normalised for Perceptron networks (NB! Target feature)
Data Acquisition

- There is never too much data
  - Even seemingly irrelevant data can prove to be relevant
  - Some information is implicit (yearly cycle, daily cycle, holidays, …)
Keep in Mind

- Data warehouse is not an analytical solution (warehouses have different design aims)
- External data and analysis might not be as trustworthy as internal data
- Big data usually means just more data not complex data
DATA ANALYSIS
Data Transform and Analysis

- Consider the advantages of views, materialised views, and computed columns vs a transform
- There is never too much data
  - It might help to extract possibly meaningful data or transforms as separate features
  - If you have more than 11 features, there will be dependencies among them – use it to your advantage

>7 features – ~80% interdependencies
>4 features – ~50% interdependence
Data Transform and Analysis

- Data can have errors – be sure you do not overfit your models!
  - Cross-validation and random subsampling are poor estimates for over-fitting
- Unevenly distributed can be handled by
  - Excluding over-represented data
  - Boosting undersampled data
  - Use of scoring equaliser

Exclude over-represented data
Boost undersampled data
Use scoring equaliser
Increase over-fitting probability!
Cross-validation

Much data
- overfitting
- performance overestimation

Little data
- underfitting
- performance unestimation
Classification Validation

1. Lift or (Cumulative) Profit chart
2. ROC chart
3. Accuracy
4. Any other single measure

• Do AUC and f-measure even relate to your aims?
Metrics to Use for Tweaking

- Bias
- MAE
- Error density plot
- Error vs attribute XY-plot
- Idea: find attribute values or value ranges where error or bias is abnormally large
Keep in Mind

• Analysts can do simple (descriptive) analysis on self-service basis
• Business users are not trained on using complex BI tools (nor do they have time for that)
• Analysis-focused mindset can be more valuable than specific technical know-how
DATA VISUALISATION
Design tips

- http://germoroney.wordpress.com/2012/11/05/is-a-good-dashboard-design-any-good/
Stakeholders

• Owners/top management
  – Yellow context
  – KPI-s

• Specialists/middle management
  – Blue (calm) tones
  – Clean interfaces

• Workers/lower management
  – Red (hot) tones
  – More details, colourful
Report and Dashboard Design

- Do not overlay report items
- Try to fit a dashboard on a single page
Dashboard Design Tips

- Include drill down or drill through actions
- Use standards and conventions
- Provide context
- Avoid excess precision and detail
- Provide select all (or multiple) for filters (and verify the results)
- HSL Hue ring can be used to programmatically generate palettes (Light for shading for colour-blind accessibility)
Report Delivery Tips

• Use scheduled delivery for long-running reports
  – Others can be provided on-demand (and cached)
Plots

- Scatter plot – correlations and outliers
- Distribution (Density) plot – deviations
- Bar chart – values, trend
  - Avoid 3D bar charts due to decreased legibility
- Gauges – values vs. goals/ranges
Keep in Mind

- It is often easier to empower tools that the users are familiar with than to train them to use a new service or application.
Links

- http://www.bi-bestpractices.com/
Statistics

Who's the REAL uninhabitable planet??

<table>
<thead>
<tr>
<th>Planet</th>
<th>Total Deaths</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>0</td>
</tr>
<tr>
<td>Sun</td>
<td>107,602,707,791</td>
</tr>
</tbody>
</table>

EVEN THE SUN IS SAFER THAN EARTH

Know The Facts