Descriptive modeling in social sciences

Dragan Gamberger
Rudjer Boskovic Institute, Croatia
Motivation

FP7 project FOC “Forecasting financial crises”
http://www.focproject.net/

  defines 147 systemic banking crises in the period 1976-2011.
  (e.g. China 1998, USA 1988 and 2007)

- World bank data about countries:
  - current account balance as percentage of GDP,
  - central government debt as percentage of GDP,
  - domestic credit to private sector as percentage of GDP,
  - foreign direct investments as percentage of GDP,
  - bank capital to assets ratio
  - percentage of rural population,
  - life expectancy at birth,
  - percentage of unemployment with tertiary education,

Which properties are characteristic for countries having banking crises?

descriptive modeling
### Banking crises dataset

#### Attributes (WB data)

<table>
<thead>
<tr>
<th>Example Country</th>
<th>Example Value</th>
<th>Example Value</th>
<th>Example Value</th>
<th>Example Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>country 1</td>
<td>2.1</td>
<td>13.2</td>
<td>0.7</td>
<td>1.1</td>
</tr>
<tr>
<td>country 2</td>
<td>2.5</td>
<td>11.9</td>
<td>1.3</td>
<td>4.0</td>
</tr>
<tr>
<td>country 3</td>
<td>2.7</td>
<td>9.7</td>
<td>2.7</td>
<td>?</td>
</tr>
<tr>
<td>country 147</td>
<td>2.1</td>
<td>1.0</td>
<td>1.3</td>
<td>2.0</td>
</tr>
<tr>
<td>country 148</td>
<td>7.7</td>
<td>18.2</td>
<td>?</td>
<td>1.4</td>
</tr>
<tr>
<td>country 434</td>
<td>4.0</td>
<td>2.7</td>
<td>2.7</td>
<td>1.1</td>
</tr>
</tbody>
</table>

- **945 numerical attributes**

- **147 positive cases**
- **287 negative cases**

- **105 indicators**

- For each indicator, a period of 3 years *before* the crisis or non-crisis

#### 9 attributes for each indicator:

- \_t\_3
- \_t\_2
- \_t\_1
- \_max
- \_index\_max
- \_min
- \_index\_min
- \_average
- \_slope
Descriptive <-> Predictive modeling

Data used for:
- (automatic) classification of unclassified data

Evaluated by:
- predictive quality on unseen examples
  (objective measure)

Model

Knowledge used as:
- (novel) human knowledge

Evaluated by:
- novelty
- actionability
- interestness
  ...
  (subjective measures of human expert)
1: *Fast growing credit activity in economies with aging population*
   - Slope of credits in the period of three years before crisis > 5.8 % per year
   - Life expectancy for females in the year before the crisis > 80.2 years.

2: *High credit activity in economies with high social security*
   - Under-five mortality rate in the period of three years before crisis < 6.3 (per 1000)
   - Population ages 65 and more three years before the crisis > 14.2 % of total population.

3: *Increasing credit activity in developing economies*
   - Increasing credit activity in the period of three years before the crisis
   - Population aged 15-65 one year before the crisis < 64.3 % of total population
   - Rural population three years before the crisis < 33.7 % of total population

4: *Socioeconomic problems recognized by decreasing life expectancy*
   - Slope of life expectancy for females in the period of three years before crisis < -0.3 years per year

5: *Socioeconomic problems recognized by non-increasing quality of public health*
   - Non-increasing life expectancy for females in the period of three years before crisis
   - Under-five mortality rate in the period of three years before crisis > -0.5 (per 1000)
Subgroups

1: *Fast growing credit activity in economies with aging population*
   - slope of credits in the period of three years before crisis > 5.8 % per year
   - life expectancy for females in the year before the crisis > 80.2 years.

Supporting conditions: low mortality of children, low percentage of young population, high percentage of elderly population, high capitalization of companies.


5: *Socioeconomic problems recognized by non-increasing quality of public health*
   - non-increasing life expectancy for females in the period of three years before crisis
   - under-five mortality slope in the period of three years before crisis > -0.5 (per 1000)

Supporting conditions: high money and quasi money growth before the crisis.

A) Subgroup discovery approach does segmentation of the target set of examples and the methodology is useful when the positive class is a result of a few different models. Especially if these models have contradictory conditions.

B) Rules (including subgroup descriptions) are constructed as conjunctions of features.

*Example:*

1: Fast growing credit activity in economies with aging population
   slope of credits in the period of three years before crisis > 5.8 % per year
   life expectancy for females three years before the crisis > 80.2 years.

A feature-based view as a unifying framework for rule induction is perhaps a most distinguishing characteristic of the book!
Examples are defined by attributes

**attributes**

<table>
<thead>
<tr>
<th>NAME</th>
<th>AGE</th>
<th>SEX</th>
<th>EDU</th>
<th>PROF</th>
<th>WEIGHT</th>
<th>INCOME</th>
<th>SMOKER</th>
</tr>
</thead>
<tbody>
<tr>
<td>peter</td>
<td>30</td>
<td>male</td>
<td>low</td>
<td>worker</td>
<td>27.3</td>
<td>14000</td>
<td>yes</td>
</tr>
<tr>
<td>carl</td>
<td>55.5</td>
<td>male</td>
<td>medium</td>
<td>worker</td>
<td>90</td>
<td>20000</td>
<td>no</td>
</tr>
<tr>
<td>dora</td>
<td>?</td>
<td>female</td>
<td>high</td>
<td>teacher</td>
<td>65.2</td>
<td>1000</td>
<td>no</td>
</tr>
<tr>
<td>tanja</td>
<td>18</td>
<td>female</td>
<td>medium</td>
<td>student</td>
<td>55.1</td>
<td>0</td>
<td>no</td>
</tr>
<tr>
<td>tom</td>
<td>70</td>
<td>male</td>
<td>high</td>
<td>?</td>
<td>60</td>
<td>9000</td>
<td>yes</td>
</tr>
<tr>
<td>steve</td>
<td>35</td>
<td>male</td>
<td>medium</td>
<td>prof</td>
<td>33</td>
<td>16000</td>
<td>no</td>
</tr>
<tr>
<td>mirko</td>
<td>42.2</td>
<td>male</td>
<td>low</td>
<td>driver</td>
<td>27</td>
<td>7500</td>
<td>yes</td>
</tr>
<tr>
<td>marc</td>
<td>29</td>
<td>male</td>
<td>?</td>
<td>waiter</td>
<td>31</td>
<td>8300</td>
<td>yes</td>
</tr>
</tbody>
</table>

- **Examples**
- **Nominal** (categorical)
- **Numerical**

...
Features:

Income > 1000
Slope of credits < 5.5

For each attribute many different features may be constructed!

The first step of the rule induction process is feature construction.

Features may have only values true and false.

Features are different from binary attributes.

Features may not have unknown values.

Features may be complex in the sense that they may include information from more than one attribute or represent information from a relational database.
Why features are so important?

- There is a well-defined procedure how to construct features.

- Once the features are constructed, the rule construction process is identical regardless of the type of attributes, how features have been obtained and what is their meaning.

- Feature relevancy is well defined. It enables that irrelevant features may be immediately discarded and that only really relevant features enter the rule induction process.

- Unknown attribute values may be solved in a very systematic way in the feature construction process.

- Imprecise attribute values can be effectively handled.

- Cut-off values in the conditions of features used in rule bodies present a valuable information. They are also the basis for the transformation of subgroups into risk models.
## Handling imprecision of numerical attributes as unknown values

In the situation when $\delta=.17$ is assumed the feature based on $A_2$ is more relevant than the feature based on $A_1$ and it will be used in the rule construction process.

<table>
<thead>
<tr>
<th></th>
<th>$A_1$</th>
<th>$A_2$</th>
<th>Class</th>
<th>$A_1&lt;1.95$</th>
<th>$A_2&lt;1.95$</th>
<th>$A_1&lt;1.95$</th>
<th>$A_2&lt;1.95$</th>
</tr>
</thead>
<tbody>
<tr>
<td>ex1</td>
<td>1.60</td>
<td>1.60</td>
<td>positive</td>
<td>true</td>
<td>true</td>
<td>true</td>
<td>true</td>
</tr>
<tr>
<td>ex2</td>
<td>1.70</td>
<td>1.65</td>
<td>positive</td>
<td>true</td>
<td>true</td>
<td>true</td>
<td>true</td>
</tr>
<tr>
<td>ex3</td>
<td>1.80</td>
<td>1.70</td>
<td>positive</td>
<td>true</td>
<td>true</td>
<td>false</td>
<td>true</td>
</tr>
<tr>
<td>ex4</td>
<td>1.90</td>
<td>1.80</td>
<td>positive</td>
<td>true</td>
<td>true</td>
<td>false</td>
<td>false</td>
</tr>
<tr>
<td>ex5</td>
<td>2.00</td>
<td>2.10</td>
<td>negative</td>
<td>false</td>
<td>false</td>
<td>true</td>
<td>true</td>
</tr>
<tr>
<td>ex6</td>
<td>2.10</td>
<td>2.20</td>
<td>negative</td>
<td>false</td>
<td>false</td>
<td>true</td>
<td>false</td>
</tr>
<tr>
<td>ex7</td>
<td>2.20</td>
<td>2.25</td>
<td>negative</td>
<td>false</td>
<td>false</td>
<td>false</td>
<td>false</td>
</tr>
<tr>
<td>ex8</td>
<td>2.30</td>
<td>2.30</td>
<td>negative</td>
<td>false</td>
<td>false</td>
<td>false</td>
<td>false</td>
</tr>
</tbody>
</table>
• Select a relevant subset of supporting conditions
• For each necessary and supporting condition construct one risk factor so that:

- positive values always denote the existence of risk
- larger values always denote larger risk
- size = 0 if equal to the cut-off value
- size = 1 if equal to the mean value for the examples that are known to be members of the model.
### Subgroup -> risk model conversion

<table>
<thead>
<tr>
<th>Risk factor name</th>
<th>World Bank indicator name</th>
<th>Function</th>
<th>Cut-off</th>
<th>Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>Credit activity</td>
<td>Domestic credit to private sector (% of GDP)</td>
<td>Slope in three years period</td>
<td>5.8</td>
<td>12.0</td>
</tr>
<tr>
<td>Life expectancy</td>
<td>Life expectancy at birth, female (years)</td>
<td>Target year value</td>
<td>80.2</td>
<td>82.2</td>
</tr>
<tr>
<td>Children mortality</td>
<td>Mortality rate, under-5 (per 1,000 live births)</td>
<td>Target year value</td>
<td>8.0</td>
<td>4.8</td>
</tr>
<tr>
<td>Young population</td>
<td>Population ages 0-14 (% of total)</td>
<td>Target year value</td>
<td>21.6</td>
<td>17.4</td>
</tr>
<tr>
<td>Elderly population</td>
<td>Population ages 65 and above (% of total)</td>
<td>Value two years before the target</td>
<td>11.0</td>
<td>15.6</td>
</tr>
<tr>
<td>Capitalization of companies</td>
<td>Market capitalization of listed companies (% of GDP)</td>
<td>Maximal value in three years period</td>
<td>51.1</td>
<td>120.0</td>
</tr>
</tbody>
</table>

**PreseValue** = \( \frac{(FuncValue - CutOff)}{(Mean - CutOff)} \)
Model A – USA, Spain

USA

Banking crisis in year 2007

Credit activity
Life expectancy
Children mortality
Young population
Elderly population
Capitalization of companies

SPAIN

Banking crisis in year 2008

Credit activity
Life expectancy
Children mortality
Young population
Elderly population
Capitalization of companies
Model A – Japan, Vietnam

### JAPAN
**No banking crisis 2000-2010**

- Credit activity
- Life expectancy
- Children mortality
- Young population
- Elderly population
- Capitalization of companies

### VIETNAM
**No banking crisis 2000-2010**

- Credit activity
- Life expectancy
- Children mortality
- Young population
- Elderly population
- Capitalization of companies
Model A – Sweden, Finland

**SWEDEN**

Banking crisis in year 2008

- Credit activity
- Life expectancy
- Children mortality
- Young population
- Elderly population
- Capitalization of companies

**FINLAND**

No banking crisis 2000-2010

- Credit activity
- Life expectancy
- Children mortality
- Young population
- Elderly population
- Capitalization of companies
Model A - Cyprus

No banking crisis 2000-2010

- Credit activity
- Life expectancy
- Children mortality
- Young population
- Elderly population
- Capitalization of companies
<table>
<thead>
<tr>
<th>Risk factor name</th>
<th>World Bank indicator name</th>
<th>Function</th>
<th>Cut-off</th>
<th>Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>Life expectancy</td>
<td>Life expectancy at birth, female (years)</td>
<td>Difference between maximal value one or two years before the target year and the target year value</td>
<td>0.0</td>
<td>0.7</td>
</tr>
<tr>
<td>Children mortality</td>
<td>Mortality rate, under-5 (per 1,000 live births)</td>
<td>Slope in three years period</td>
<td>-0.5</td>
<td>0.5</td>
</tr>
<tr>
<td>Money growth</td>
<td>Money and quasi money growth (annual %)</td>
<td>Value in the year before the target year</td>
<td>5.2</td>
<td>28.5</td>
</tr>
</tbody>
</table>
Model B – Bulgaria, Italy

BULGARIA

Banking crisis in year 1996

ITALY

Banking crisis in year 2008
Model B – Sierra Leone, Portugal

**SIERRA LEONE**

- Banking crisis in year 1990

**PORTUGAL**

- Banking crisis in year 2008
1: **Fast growing credit activity in economies with aging population**
   - slope of credits in the period of three years before crisis > 5.8 % per year
   - life expectancy for females in the year before the crisis > 80.2 years.

Supporting conditions: low mortality of children, low percentage of young population, high percentage of elderly population, high capitalization of companies.


5: **Socioeconomic problems recognized by non-increasing quality of public health**
   - non-increasing life expectancy for females in the period of three years before crisis
   - under-five mortality rate in the period of three years before crisis > -0.5 (per 1000)

Supporting conditions: high money and quasi money growth before the crisis.

World Bank governance indicators

Differences in p-ranks for years 2007 and year 2004 for six governance indicators for two groups of EU countries

<table>
<thead>
<tr>
<th>Country</th>
<th>Control of corruption</th>
<th>Rule of law</th>
<th>Government effectiveness</th>
<th>Voice and accountability</th>
<th>Political stability</th>
<th>Regulatory quality</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Belgium</td>
<td>-4.32</td>
<td>0.00</td>
<td>-1.91</td>
<td>1.92</td>
<td>0.00</td>
<td>3.03</td>
<td>-1.28</td>
</tr>
<tr>
<td>Greece</td>
<td>-6.66</td>
<td>-5.74</td>
<td>-4.75</td>
<td>-8.65</td>
<td>0.48</td>
<td>0.22</td>
<td>-25.10</td>
</tr>
<tr>
<td>Hungary</td>
<td>-2.80</td>
<td>-1.44</td>
<td>-0.87</td>
<td>-4.33</td>
<td>-2.88</td>
<td>2.10</td>
<td>-10.21</td>
</tr>
<tr>
<td>Italy</td>
<td>-6.16</td>
<td>-6.22</td>
<td>-12.02</td>
<td>-1.44</td>
<td>6.25</td>
<td>-2.24</td>
<td>-21.82</td>
</tr>
<tr>
<td>Portugal</td>
<td>-4.78</td>
<td>-5.26</td>
<td>-6.23</td>
<td>-1.92</td>
<td>-5.77</td>
<td>-2.28</td>
<td>-26.25</td>
</tr>
<tr>
<td>Spain</td>
<td>-7.71</td>
<td>-0.96</td>
<td>-8.20</td>
<td>-4.81</td>
<td>-12.98</td>
<td>-0.85</td>
<td>-35.51</td>
</tr>
<tr>
<td>Austria</td>
<td>-0.96</td>
<td>3.35</td>
<td>3.43</td>
<td>2.88</td>
<td>11.06</td>
<td>3.47</td>
<td>23.23</td>
</tr>
<tr>
<td>Denmark</td>
<td>0.49</td>
<td>1.44</td>
<td>-0.49</td>
<td>-2.40</td>
<td>3.85</td>
<td>1.97</td>
<td>4.85</td>
</tr>
<tr>
<td>France</td>
<td>1.99</td>
<td>-1.91</td>
<td>-3.36</td>
<td>0.00</td>
<td>5.77</td>
<td>1.11</td>
<td>3.59</td>
</tr>
<tr>
<td>Germany</td>
<td>-0.94</td>
<td>0.00</td>
<td>3.45</td>
<td>0.00</td>
<td>14.90</td>
<td>3.00</td>
<td>20.41</td>
</tr>
<tr>
<td>Netherlands</td>
<td>1.48</td>
<td>0.00</td>
<td>-3.39</td>
<td>1.92</td>
<td>-7.69</td>
<td>-0.47</td>
<td>-8.14</td>
</tr>
</tbody>
</table>

Level of statistical significance

- 99.9%
- 97%
- 96%
- Non-sig.
- Non-sig.
- Non-sig.
- 99%

“Good governance plays a significant role in determining the extent to which a country is likely to have a crisis.”

The result demonstrates that Model B that has been the basis for selecting a subset of 6 countries is reasonable!!

Model B is based on socioeconomic problems recognized by non-increasing quality of public health.

Now we have:
- Good governance problems
- Socioeconomic problems
- Banking crises
### Difference in p-ranks for governance indicators in year 2011 and year 2008

<table>
<thead>
<tr>
<th>Total for 6 indicators</th>
<th>Total for 3 most relevant indicators</th>
<th>Control of corruption indicator</th>
</tr>
</thead>
<tbody>
<tr>
<td>Greece: -39.49</td>
<td>Greece: -19.57</td>
<td>Italy: -5.76</td>
</tr>
<tr>
<td>Malta: -29.58</td>
<td>Malta: -11.35</td>
<td>Cyprus: -5.33</td>
</tr>
<tr>
<td>Portugal: -25.13</td>
<td>Hungary: -7.59</td>
<td>Austria: -5.09</td>
</tr>
<tr>
<td>Ireland: -19.08</td>
<td>Cyprus: -6.62</td>
<td>Malta: -3.84</td>
</tr>
</tbody>
</table>
Conclusions

- Data preparation is important
- Subgroup discovery is useful for different tasks
- Subgroups may be transformed into risks models
- Comparative analysis of examples included into different subgroups may result by interesting novel knowledge
Thank you for your attention!

Questions?