KNIME in Action:

GDPR, Taking a Proactive Approach

Phil Winters
“the world’s strongest and most far-reaching law aimed at strengthening citizens' fundamental rights in the digital age”
“facilitate business by coming up with one set of rules”

• All organizations processing EU data subjects’ personal data
  – For EU organizations: “A lot more of the same”
  – But now applies to ALL organizations worldwide, not just in the EU
  – No distinction between BtoC and BtoB

• Compliance is mandatory
  – Very large fines based on worldwide income for non-compliance
  – The law takes effect 25 May 2018

GDPR Terms:

- Data Subject
- Controller, Processor
- Personal Data
  - Specific, explicit & legitimate
  - Limited to what is necessary
- Permission
- Automated Decision Making / Profiling

- Person, individual, consumer in the EU
- You are liable
- Personal Data
  - No more simply buying data
  - No more simply “because you can”
- Not just what, but HOW.
Automated Decision Making / Profiling

Predictive Analytics! That’s our job!
GDPR Terms:

- Data Subject
- Controller, Processor
- Personal Data
  - Specific, explicit & legitimate
  - Limited to what is necessary
- Permission
- Automated Decision Making / Profiling
- Special Categories of Data
- Anonymization/pseudo anonymization
- Transparent
- Protects

- Person, individual, consumer in the EU
- You are liable
- Personal Data
  - No more simply buying data
  - No more simply “because you can”
- Not just what, but HOW.
- Machine learning: our Job!
  - Discrimination / “Pseudo discrimination”
  - Representative but not traceable
  - Making models “explainable”
  - Document but use it to teach/share!
Article 22: Automated Individual Profiling & 4 techniques that will help
Article 9: Processing of special categories of personal data.
Article 9: Special Categories of Data

• Discrimination (and “pseudo discrimination”)
  – racial or ethnic origin
  – political opinions
  – religious or philosophical beliefs
  – trade union membership
  – genetic data
  – biometric data
  – health
  – sex life or sexual orientation
Flag Discriminatory & identify Pseudo-discriminatory

Use your favorite technique to determine how closely related other fields are at predicting identified discriminatory fields.

Here a simple linear correlation is used for demonstration purposes.
<table>
<thead>
<tr>
<th>Label</th>
<th>Excludes</th>
<th>Includes</th>
</tr>
</thead>
<tbody>
<tr>
<td>age</td>
<td>36, 31, 34, 23, 35</td>
<td></td>
</tr>
<tr>
<td>fnlwgt</td>
<td>123611, 164190, 203488, 148995, 121124</td>
<td></td>
</tr>
<tr>
<td>education</td>
<td>HS-grad, Some-college, Bachelors, Masters, Assoc- voc</td>
<td></td>
</tr>
<tr>
<td>education_num</td>
<td>9, 10, 13, 14, 11</td>
<td></td>
</tr>
<tr>
<td>marital_status</td>
<td>Married-civ-spouse, Never-married, Divorced, Separated, Widowed</td>
<td></td>
</tr>
<tr>
<td>occupation</td>
<td>Prof-specialty, Craft-repair, Exec-managerial, Adm-clerical, Sales</td>
<td></td>
</tr>
<tr>
<td>relationship</td>
<td>Husband, Not-in-family, Own-child, Unmarried, Wife</td>
<td></td>
</tr>
<tr>
<td>race</td>
<td>White, Black, Asian-Pac-Islander, Amer-Indian Eskimo, Other</td>
<td></td>
</tr>
<tr>
<td>sex</td>
<td>Male, Female</td>
<td></td>
</tr>
<tr>
<td>capital_gain</td>
<td>0, 15024, 7888, 7298, 99999</td>
<td></td>
</tr>
<tr>
<td>capital_loss</td>
<td>0, 1902, 1977, 1887, 1485</td>
<td></td>
</tr>
<tr>
<td>hours-per-week</td>
<td>40, 50, 60, 35</td>
<td></td>
</tr>
<tr>
<td>native-country</td>
<td>United-States, Mexico, ?, Philippines, Germany</td>
<td></td>
</tr>
<tr>
<td>Income</td>
<td>&lt;=50K, &gt;50K</td>
<td></td>
</tr>
<tr>
<td>Union_Member</td>
<td>Unknown, Non-Union, Union</td>
<td></td>
</tr>
<tr>
<td>Zip Code</td>
<td>?, 38131</td>
<td></td>
</tr>
</tbody>
</table>
Identify discriminatory

Label

- Age: 36, 31, 34, 23, 35
- Hours-per-week: 123011, 164190, 203488, 148996, 121124
- Education: HS-grad, Some-college, Bachelors, Masters, Assoc-voc
- Education-num: 9, 10, 13, 14, 11
- Marital-status: Married-civ-spouse, Never-married, Divorced, Separated, Widowed
- Occupation: Prof-specialty, Craft-repair, Exec-managerial, Adm-clerical, Sales
- Relationship: Husband, Not-in-family, Own-child, Unmarried, Wife
- Sex: Male, Female
- Capital-gain: 0, 15024, 7688, 7298, 99999
- Capital-loss: 0, 1962, 1977, 1887, 1485
- Hours-per-week: 40, 50, 45, 60, 35
- Native-country: United-States, Mexico, ?, Philippines, Germany
- Income: <=50K, >50K
- Zip Code: ?, 38131

Race: White, Black, Asian-Pac-Islander, Amer-Indian-Eskimo, Other
Union-Member: Unknown, Non-Union, Union, ?
& identify Pseudo-discriminatory

Use your favorite technique to determine how closely related other fields are to predicting identified discriminatory fields.

Here a simple linear correlation is used for demonstration purposes.

Dialog - 0:10 - identify pseudo discrim

Correlation Measure Threshold

Change

OK  Apply  Cancel
Use your favorite technique to determine how closely related other fields are to identifying discriminatory fields.

Here a simple linear correlation is used for demonstration purposes.
Flag Discriminatory & identify Pseudo-discriminatory

Use your favorite technique to determine how closely related other fields are to predicting identified discriminatory fields.

Here a simple linear correlation is used for demonstration purposes.
Preamble 26: Anonymous Data
Anonymize

• Generalization
• Perturbation
• Randomization
• Masking
• Vertex/Edge Clustering
• K-anonymity
• I-diversity
• .....
An Anonymize Data Process

Choose your own methods to anonymize Data

Create Distance Matrix → Find k-nearest Neighbors → Random Value Selection

Choose your own methods to test the quality of the anonymisation

Quality Test

Choose your own methods to test the identification of an original record

Original Check

Table Reader

Missing Value

Table Reader

Adults Table → remove rows with missing values
Test the quality of the anonymization model
Check “reverse anonymization”
Article 12: Right to Transparency
Transparency and explaining models....
Article 15: Right of Access to Information
Table Reader

Use your favorite tool or package to consolidate the information and make it understandable.

You may have a version for:
- The Legal Team
- The Data Protection Officer
- The individual who's data is being used...

THIS is probably the most important concept of all. We capture the decisions we have made and what we have done, but we EXPLAIN it as well.

Create Descriptive Documentation

Table Reader

Decision Criteria and stats

context information

Table Reader

Quality Information

de-anonymize Information

© 2018 KNIME AG. All Rights Reserved.
<table>
<thead>
<tr>
<th>Model Name</th>
<th>Income Category Prediction</th>
</tr>
</thead>
<tbody>
<tr>
<td>Created By</td>
<td>Phil</td>
</tr>
<tr>
<td>Last Updated By</td>
<td>Phil</td>
</tr>
<tr>
<td>Data Source</td>
<td>knime://LOCAL/GDPR/adults_zipcode_union.table</td>
</tr>
<tr>
<td>Workflow</td>
<td>knime://LOCAL/Random_Forest_Prediction</td>
</tr>
<tr>
<td>Tests Completed</td>
<td>Discrimination_Check</td>
</tr>
<tr>
<td></td>
<td>Anonymization</td>
</tr>
<tr>
<td></td>
<td>Model Explanation</td>
</tr>
<tr>
<td></td>
<td>Documentation</td>
</tr>
</tbody>
</table>

![Diagram of the workflow](image-url)
<table>
<thead>
<tr>
<th>Field Name</th>
<th>GDPR Special Category Type</th>
<th>Pseudodiscriminatory Relationship Description</th>
<th>Action Taken</th>
</tr>
</thead>
<tbody>
<tr>
<td>Zip Code</td>
<td>Pseudo Discriminatory</td>
<td>Correlates to &quot;race&quot; with a value of 0.53</td>
<td>Dropped</td>
</tr>
<tr>
<td>race</td>
<td>Discriminatory</td>
<td>Correlates to &quot;Zip Code&quot; with a value of 0.53</td>
<td>Dropped</td>
</tr>
<tr>
<td>Union_Member</td>
<td>Discriminatory</td>
<td>None</td>
<td>Dropped</td>
</tr>
<tr>
<td>age</td>
<td>Discriminatory</td>
<td></td>
<td>Dropped</td>
</tr>
<tr>
<td>capital_gain</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>capital_loss</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>education</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>education_num</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>fnlwgt</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>hours-per-week</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>income</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>marital_status</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>native-country</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>occupation</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>relationship</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>sex</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>workclass</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Recommendation

The following information is used in automatic profiling for the Income Prediction Model (in order of importance):

- capital_gain
- capital_loss
- education
- relationship
Article 22: Automated Individual Profiling & 4 techniques that will help
Using KNIME to take a proactive approach to GDPR

• This is not “new data science”
• It is packaging/documenting and presenting

• It can be totally automated

• KNIME Server
  – Sub workflows
  – REST micro services

  – Part of the KNIME Model Factory
Next Steps

- Presentation available immediately
- Whitepaper imminent.
  - Workflows on the example server

- Take a proactive approach and talk to your Data Protection Officer!
The KNIME® trademark and logo and OPEN FOR INNOVATION® trademark are used by KNIME AG under license from KNIME GmbH, and are registered in the United States. KNIME® is also registered in Germany.