Model-based Discrimination Analysis: A Position Paper

Qusai Ramadan¹, Shayan Ahmadian¹, Daniel Strüber¹, Jan Jürjens¹,² and Steffen Staab¹,³

¹ University of Koblenz-Landau, Koblenz, Germany
² Fraunhofer-Institute for Software and Systems Engineering ISST, Dortmund, Germany
³ University of Southampton, UK

Tuesday 29th May 2018, Sweden, Gothenburg
Basics

- Decision-making software may **lead to undesirable discrimination**:
  - exploiting sensitive data (e.g., race)
  - learning correlations between a set of data.

1. **Black box approaches**
   - Require the software to be implemented.

2. **White box approaches**

How to discover potential discrimination during the software design phase?
(i.e., before having a faulty implementation)
Motivating Example 1

- Bank offers three services
  - Zero-Fee Money Transfer (for international merchants)
  - Vacancies Announcement (for domestic persons educated in accounting).
  - Apply for a loan.

- **Policy:** The bank disallows discriminating between the loans applicants based on their citizenship.

Business Analyst

Input $X_i$: <needed amount of money, purpose>
Motivating Example 1

Potential Input $X_i$: <needed amount of money, purpose>
Roadmap: Model-based Analysis

A Domain Expert and a Business Analyst

Business Analyst

High-level overview of a model-based discrimination analysis framework.
Information Flow Analysis

Control flow | Action | Data Store

Loan Request | receive loan request | customer::DataBase

Get characteristics | zero_Fee | vac_Ann | high_Income

Do decision | <<critical>> {nationality}

Data object

Critical Data object | Direct Data objects

high_Income

result

zero_Fee

vac_Ann

Data object

Information flow
Database Schema

- Information about whether a data object is derived or not can be represented in the database schema.

![Diagram of database schema with a focus on derived property and an explanation](image)

- But it does not tell how it is derived.
Information Flow Analysis

Activity diagram describing how the value of the zero_Fee data object is derived.

Enable the zero_Fee Money Exchange Service

Get all customers

<<centralBuffer>>
ListofCustomers::DataBase

<<Data Store>>
customer::DataBase

{weight=all}

Get Job and Citizenship Info

customer_ID:int

result: String

If merchant and international?

[true]

Update the Zero_Fee Exchange

end

[false]

Critical Data object

Direct Data objects

high_income

result

total_Fee

vac_Ann

Indirect Data objects

high_income

income > 3k euro

nationality=international

merchant

accounting

nationality=national

Data object

Information flow

There is indirect leakage of the citizenship data to the result data object.
Roadmap: Model-based Analysis

High-level overview of a model-based discrimination analysis framework.
Motivating Example 2

- What about dependencies with the **gender**?

Business Analyst

\[
P(\text{Female} \mid \neg(">3k" \text{ €} \cap \text{international} \cap \text{merchant}) \cap \text{accounting}) = 66.67\
\]

(i.e., given a national customer with educational background in accounting and the other data objects are not true)

- The nationality and the education background in this context can act as a proxy for the gender.
Roadmap: Model-based Analysis

1. Annotate the model
2. Model Analysis
3. Aggregate Data
4. Statistical Analysis

Critical Data objects
- high_income
- zero_Fee
- merchant
- vac_Ann
- nationality=national

Direct Data objects
- income > 3k euro
- nationality=international

Indirect Data objects

<table>
<thead>
<tr>
<th>P(Female)</th>
<th>&quot;&gt;3k&quot; €</th>
<th>nationality</th>
<th>merchant</th>
<th>accounting</th>
</tr>
</thead>
<tbody>
<tr>
<td>50.00%</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>66.67%</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>0.00%</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>100.00%</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>100.00%</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>0.00%</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>
Conclusion

A Domain Expert and a Business Analyst

High-level overview of a model-based discrimination analysis framework.
Challenges

- Information about the derived data are distributed in multiples diagrams.

- How to measure the discrimination by proxy? (e.g., information gain)
Conclusion

High-level overview of a model-based discrimination analysis framework.
Backup slides
Protected Characteristics

- legally protected characteristics (e.g., age, gender, race, ...).

- But not limited to those listed by the laws and regulations.

- Example: a bank may disallow discriminating between the loans applicants based on their citizenship.
Initial Statistical Analysis

\[
\begin{align*}
P(\text{international} \mid \text{zero_Fee} \cap \text{vac_Ann} \cap \text{high_Income})& \\
\text{P(\text{international})} & | \text{zero_Fee} & | \text{vac_Ann} & | \text{high_Income} \\
66.67\% & 0 & 0 & 0 \\
100.00\% & 1 & 0 & 0 \\
0.00\% & 1 & 0 & 0 \\
100.00\% & 1 & 0 & 0 \\
0.00\% & 0 & 1 & 0 \\
0.00\% & 0 & 1 & 1 \\
100.00\% & 0 & 0 & 1
\end{align*}
\]

- a societal fact (e.g., a taxi driver in Saudi Arabia).
- They could be derived from processing the citizenship information.

Table 1: Personal Data

<table>
<thead>
<tr>
<th>customer_ID</th>
<th>gender</th>
<th>Income</th>
<th>high_Income</th>
<th>nationality</th>
<th>merchant</th>
<th>accounting</th>
</tr>
</thead>
<tbody>
<tr>
<td>BA01</td>
<td>0</td>
<td>3000</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>BA02</td>
<td>0</td>
<td>4500</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>BA03</td>
<td>0</td>
<td>2500</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>BA04</td>
<td>0</td>
<td>3200</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>BA05</td>
<td>0</td>
<td>2900</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>BA06</td>
<td>1</td>
<td>5000</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>BA07</td>
<td>1</td>
<td>2450</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>BA08</td>
<td>1</td>
<td>3600</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>BA09</td>
<td>1</td>
<td>3100</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>BA10</td>
<td>1</td>
<td>1800</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>

Table 2. Services Data

<table>
<thead>
<tr>
<th>customer_ID</th>
<th>zero_Fee</th>
<th>vac_Ann</th>
</tr>
</thead>
<tbody>
<tr>
<td>BA01</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>BA02</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>BA03</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>BA04</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>BA05</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>BA06</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>BA07</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>BA08</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>BA09</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>BA10</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>