A Case-Based-Reasoning Analysis of the COMPAS Dataset

Wijnand van Woerkom,[a] Davide Grossi,[b,c,d] Henry Prakken,[a] Bart Verheij[b]

a	Department of Information and Computing Sciences	Utrecht University
b	Bernoulli Institute for Mathematics, CS and AI	University of Groningen
c	Amsterdam Center for Law and Economics	University of Groningen
d	Institute for Logic, Language and Computation	University of Amsterdam

Overview

Previous work

- Horty developed model of a fortiori case-based reasoning in law[1]
- Used for XAI,[2] classification,[3] legal analysis (see: last presentation)
- Extended Horty's model (previous JURIX)[4]

Current work

- Applied this extension to COMPAS dataset[5]
- Analyzed its decision consistency
- Found an **error** in the dataset

---

[1] Horty, "Reasoning with Dimensions and Magnitudes" (2019).

[2] Prakken and Ratsma, "A Top-Level Model of Case-Based Argumentation for Explanation: Formalisation and Experiments" (2022).

[3] Odekerken and Bex, "Towards Transparent Human-in-the-Loop Classification of Fraudulent Web Shops" (2020).

[4] Van Woerkom et al., "Hierarchical *a Fortiori* Reasoning with Dimensions" (2023).

[5] Angwin et al., "Machine Bias" (2016).

Overview

Previous work

- Horty developed model of a fortiori case-based reasoning in law[1]
- Used for XAI,[2] classification,[3] legal analysis (see: last presentation)
- Extended Horty's model (previous JURIX)[4]

Current work

- Applied this extension to COMPAS dataset[5]
- Analyzed its decision consistency
- Found an **error** in the dataset

---

[1] Horty, "Reasoning with Dimensions and Magnitudes" (2019).

[2] Prakken and Ratsma, "A Top-Level Model of Case-Based Argumentation for Explanation: Formalisation and Experiments" (2022).

[3] Odekerken and Bex, "Towards Transparent Human-in-the-Loop Classification of Fraudulent Web Shops" (2020).

[4] Van Woerkom et al., "Hierarchical *a Fortiori* Reasoning with Dimensions" (2023).

[5] Angwin et al., "Machine Bias" (2016).

COMPAS

The COMPAS program

- COMPAS: Controversial recidivism risk prediction tool, widely used in US
- Produces three risk scores: **GRecid**, **VRecid**, **FTA**,
- combines these in a recommended supervision level **SLevel**,
- . . . etc

ProPublica

- High profile accusation by ProPublica of racial bias[6]
- Dataset was made available:
  - Risk scores from Broward county, Florida, between 2013–2014
  - Matched with defendant data on age, prior offenses, criminal history, name(!), etc. . .
- Accusations were refuted – dataset remains relevant

[6] Angwin et al., "Machine Bias" (2016).

COMPAS

The COMPAS program

- COMPAS: Controversial recidivism risk prediction tool, widely used in US
- Produces three risk scores: **GRecid**, **VRecid**, **FTA**,
- combines these in a recommended supervision level **SLevel**,
- . . . etc

ProPublica

- High profile accusation by ProPublica of racial bias[6]
- Dataset was made available:
  - Risk scores from Broward county, Florida, between 2013–2014
  - Matched with defendant data on age, prior offenses, criminal history, name(!), etc. . .
- Accusations were refuted – dataset remains relevant

[6] Angwin et al., "Machine Bias" (2016).

COMPAS

The COMPAS program

- COMPAS: Controversial recidivism risk prediction tool, widely used in US
- Produces three risk scores: **GRecid**, **VRecid**, **FTA**,
- combines these in a recommended supervision level **SLevel**,
- . . . etc

ProPublica

- High profile accusation by ProPublica of racial bias[6]
- Dataset was made available:
  - Risk scores from Broward county, Florida, between 2013–2014
  - Matched with defendant data on age, prior offenses, criminal history, name(!), etc. . .
- Accusations were refuted – dataset remains relevant

[6] Angwin et al., "Machine Bias" (2016).

Raw vs. decile risk scores

- COMPAS outputs **raw** risk scores
- Example row from the dataset:

FTA		GRecid		VRecid	
Raw	Decile	**Raw**	Decile	**Raw**	Decile
21	3	0.14	7	−0.95	9

- Difficult to interpret – converted to **decile** scores
- COMPAS manual specifies the full table of conversions, e.g. for **VRecid**:

D1	D2	D3	D4	D5	D6	D7	D8	D9	D10
−2.90	−2.50	−2.20	−2.00	−1.70	−1.50	−1.20	−1.00	−0.60	1.9

Raw vs. decile risk scores

- COMPAS outputs **raw** risk scores
- Example row from the dataset:

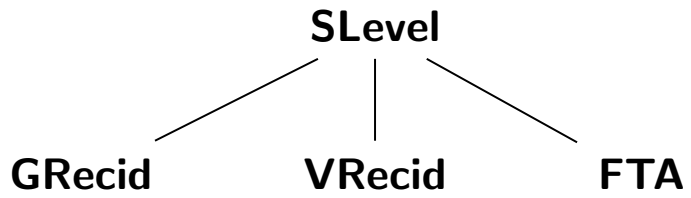
FTA		GRecid		VRecid	
Raw	**Decile**	**Raw**	**Decile**	**Raw**	**Decile**
21	3	0.14	7	−0.95	9

- Difficult to interpret – converted to **decile** scores
- COMPAS manual specifies the full table of conversions, e.g. for **VRecid**:

D1	D2	D3	D4	D5	D6	D7	D8	D9	D10
−2.90	−2.50	−2.20	−2.00	−1.70	−1.50	−1.20	−1.00	−0.60	1.9

Recommended supervision level

- Risk scores are input to supervision level recommendation:

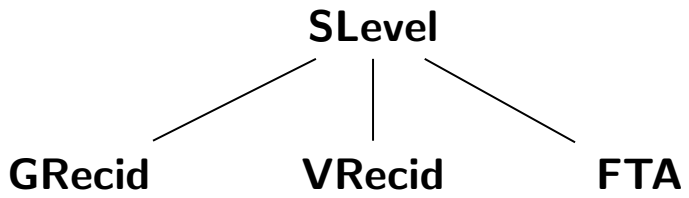


- Summary statistic in 1–4 range based "primarily" on risk scores
- Question: Are these risk scores **consistently** assigned?
- Hypothetical example (with deciles):

GRecid	VRecid	FTA	SLevel
5	2	3	2
8	9	7	3
6	4	6	2
7	5	5	?

Recommended supervision level

- Risk scores are input to supervision level recommendation:

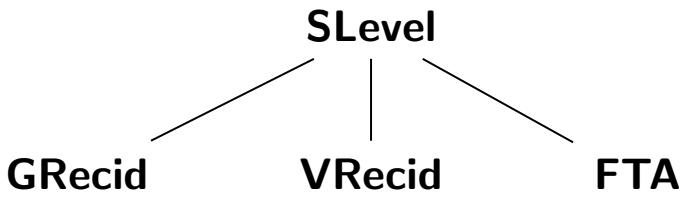


- Summary statistic in 1–4 range based "primarily" on risk scores
- Question: Are these risk scores **consistently** assigned?
- Hypothetical example (with deciles):

GRecid	VRecid	FTA	SLevel
5	2	3	2
8	9	7	3
6	4	6	2
7	5	5	?

Recommended supervision level

- Risk scores are input to supervision level recommendation:

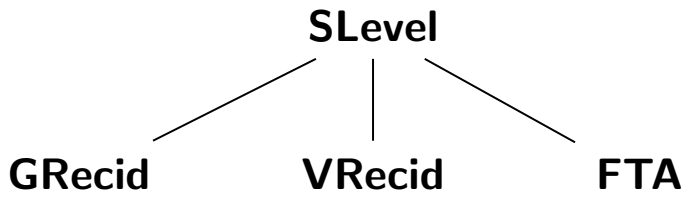


- Summary statistic in 1–4 range based "primarily" on risk scores
- Question: Are these risk scores **consistently** assigned?
- Hypothetical example (with deciles):

GRecid	VRecid	FTA	SLevel
5	2	3	2
8	9	7	3
6	4	6	2
7	5	5	$2 \leq$?

Recommended supervision level

- Risk scores are input to supervision level recommendation:

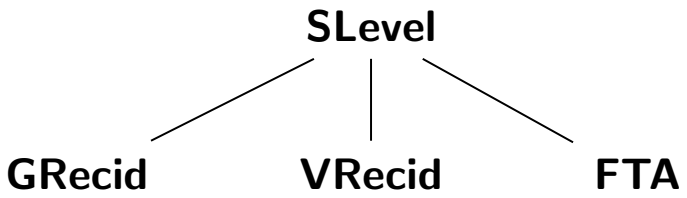


- Summary statistic in 1–4 range based "primarily" on risk scores
- Question: Are these risk scores **consistently** assigned?
- Hypothetical example (with deciles):

GRecid	VRecid	FTA	SLevel
5	2	3	2
8	9	7	3
6	4	6	2
7	5	5	$2 \leq$?

Recommended supervision level

- Risk scores are input to supervision level recommendation:

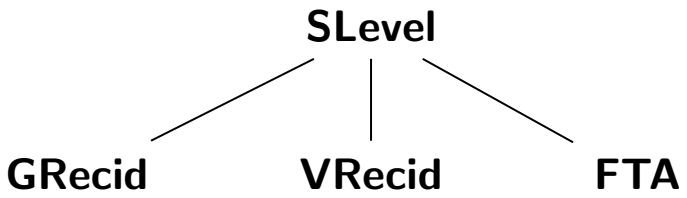


- Summary statistic in 1–4 range based "primarily" on risk scores
- Question: Are these risk scores **consistently** assigned?
- Hypothetical example (with deciles):

GRecid	VRecid	FTA	SLevel
5	2	3	2
8	**9**	**7**	3
6	4	6	2
7	**5**	**5**	$2 \leq ? \leq 3$

Recommended supervision level

- Risk scores are input to supervision level recommendation:

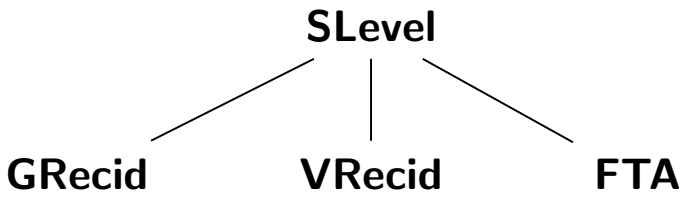


- Summary statistic in 1–4 range based "primarily" on risk scores
- Question: Are these risk scores **consistently** assigned?
- Hypothetical example (with deciles):

GRecid	VRecid	FTA	SLevel
5	2	3	2
8	9	7	3
6	**4**	**6**	2
7	**5**	**5**	$2 \leq ? \leq 3$

Recommended supervision level

- Risk scores are input to supervision level recommendation:

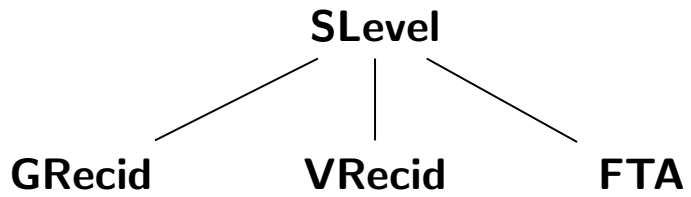


- Summary statistic in 1–4 range based "primarily" on risk scores
- Question: Are these risk scores **consistently** assigned?
- Hypothetical example (with deciles):

GRecid	VRecid	FTA	SLevel
5	2	3	2
8	9	7	3
6	**4**	**6**	2
7	**5**	**5**	$2 \leq ? \leq 3$

Recommended supervision level

- Risk scores are input to supervision level recommendation:

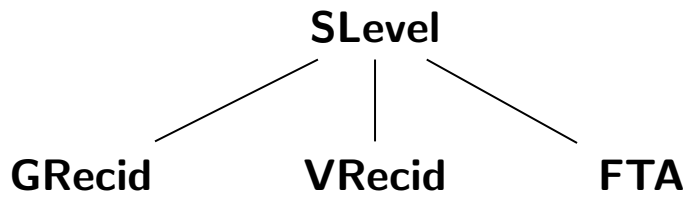


- Summary statistic in 1–4 range based "primarily" on risk scores
- Question: Are these risk scores **consistently** assigned?
- Hypothetical example (with deciles):

GRecid	VRecid	FTA	SLevel
5	2	3	2
8	9	7	3
6	4	6	2
7	5	5	4

Recommended supervision level

- Risk scores are input to supervision level recommendation:



- Summary statistic in 1–4 range based "primarily" on risk scores
- Question: Are these risk scores **consistently** assigned?
- Hypothetical example (with deciles):

GRecid	VRecid	FTA	SLevel
5	2	3	2
8	9	7	**3**
6	4	6	2
7	5	5	**4**

SLevel consistency: Results

- Calculated consistency of COMPAS **SLevel** scores
- Dataset contains around 12,000 rows
- Outcome – with raw inputs: 84%, with decile inputs: 100%
- . . . but this is **impossible**

Example rows

Table: Two rows from the COMPAS dataset

FTA		GRecid		VRecid		SLevel
Raw	**Decile**	**Raw**	**Decile**	**Raw**	**Decile**	
21	3	0.14	7	−0.95	9	3
19	3	0.11	8	−1.21	8	4

Example rows

Table: Two rows from the COMPAS dataset

FTA		GRecid		VRecid		SLevel
Raw	Decile	Raw	Decile	Raw	Decile	
21	3	0.14	7	−0.95	9	3
19	3	0.11	8	−1.21	8	4

Example rows

Table: Two rows from the COMPAS dataset

FTA		GRecid		VRecid		SLevel
Raw	**Decile**	**Raw**	**Decile**	**Raw**	**Decile**	
21	**3**	0.14	**7**	−0.95	**9**	3
19	**3**	0.11	**8**	−1.21	**8**	4

Example rows

Table: Two rows from the COMPAS dataset

FTA		GRecid		VRecid		SLevel
Raw	**Decile**	**Raw**	**Decile**	**Raw**	**Decile**	
21	3	**0.14**	**7**	−0.95	9	3
19	3	**0.11**	**8**	−1.21	8	4

Discrepancy

Cause is unclear, raises questions:

- Is one of the scores correct or are both incorrect?
- Scores swapped between people?
- Were multiple norm groups used?

Concluding

Our work

- Implemented the case-based reasoning model
- Used it to study COMPAS dataset

Takeaways

- Consistency as measure of precedent adherence
- COMPAS "conversion" error: Garbage in, garbage out

Future work

- Theoretically: Investigate formal properties of consistency
- Practically: Study other datasets, e.g. bail decisions

Thanks!

Concluding

Our work

- Implemented the case-based reasoning model
- Used it to study COMPAS dataset

Takeaways

- Consistency as measure of precedent adherence
- COMPAS "conversion" error: Garbage in, garbage out

Future work

- Theoretically: Investigate formal properties of consistency
- Practically: Study other datasets, e.g. bail decisions

Thanks!

Concluding

Our work

- Implemented the case-based reasoning model
- Used it to study COMPAS dataset

Takeaways

- Consistency as measure of precedent adherence
- COMPAS "conversion" error: Garbage in, garbage out

Future work

- Theoretically: Investigate formal properties of consistency
- Practically: Study other datasets, e.g. bail decisions

Thanks!

Concluding

Our work

- Implemented the case-based reasoning model
- Used it to study COMPAS dataset

Takeaways

- Consistency as measure of precedent adherence
- COMPAS "conversion" error: Garbage in, garbage out

Future work

- Theoretically: Investigate formal properties of consistency
- Practically: Study other datasets, e.g. bail decisions

Thanks!