PREDICTIVE RISK MODELING AND DATA MINING

David Schwartz

Child Protective Services Modeling: 3 Counties in New York State

- OCFS identified substantiated (and recidivism for cases with prior investigations), and unsubstantiated investigations as target outcomes for the modeling team. Recidivism (or reentry) was selected because it is one OCFS's CFSR Program Improvement Plan outcomes.

- Our sample consisted of 55,934 cases of child maltreatment reported to the NYS hotline between 2001-2010 where cases were defined as either having no substantiated dispositions (76.5% of all cases) or at least one substantiated disposition (23.5% of all cases).

- **Results: The team tested several predictive models, ultimately developing strong models for predicting substantiated/recidivism cases (91% accuracy) and unsubstantiated cases (almost 90% accuracy).**

- One of the most surprising findings was the differences between counties, making county a significant and high ranking variable for predicting substantiation/recidivism.

- In the next phase of the project the team will utilize statewide data, introduce advanced nonlinear data mining algorithms to optimize the models, and develop a live shadow system for testing and analysis.

Step 1: PCA/Factor Analysis Ranks 44 Significant Measures

Rank	Field
1	MANDATED_IND_1
2	REPORT_AGE_NBR_1
3	COUNTY_COUNTY_NM_1
4	VICTIM_INTAKE_AGE_NBR_1
5	NCANDS_REPORTER_CD_1
6	N_INTAKES
7	VICTIM_ETHNIC_CD
8	SUBJECT_ETHNIC_CD_1
9	PRIORS
10	SUBJECT_INTAKE_AGE_NBR_1
11	ALLEGATION_TXT_1
12	ALLEGATION_CD_1
13	N_CLOSES
14	VIC_RECS
15	SUBJECT_HISP_LATINO_IND_1
16	SUBJECT_SEX_CD_1
17	ALLEGATION_GROUP_TXT_1
18	PERP_RECS
19	VICTIM_SEX_CD
20	INQTR_1
21	INMONTH_1
22	VICTIM_HISP_LATINO_IND
23	INDAY_1

Step 2: Several Different Algorithms Were Tested

Initial analysis of the overall sample dataset:

- 9998 total randomly selected cases, 44 features (Final predictors)

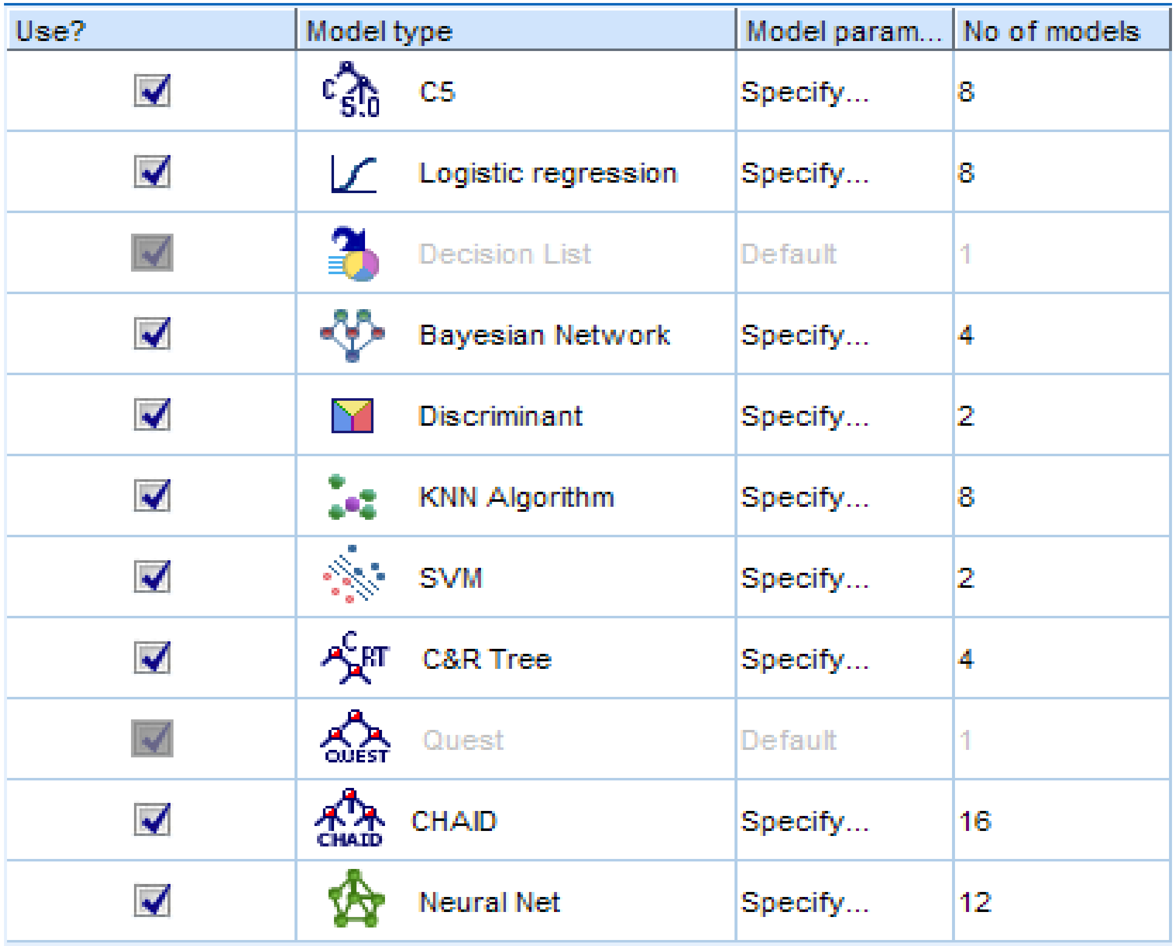
- **Unfounded Cases:** The best prediction model is strong (almost 90% accurate) at finding cases not substantiated within our predefined low range
  - Total records in range: 4404.
  - Actual records with a known value of 0 = 3926 cases. That shows 89% accuracy.

- **Indicated Cases:** The best prediction model is good (70% accurate) at finding cases substantiated within our predefined high range. Less than 25% of the total sample had a substantiated outcome.
  - Total records in range: 195.
  - Actual records with a known value of 1 = 138 cases. That shows 70% accuracy.

Our next step was to include Safety Factor data into the model

- Caretaker previously committed or allowed others to abuse or maltreat child
- Caretaker’s current alcohol abuse seriously affects his/her ability to care for child
- Caretaker’s current drug abuse seriously affects his/her ability to care for child
- Child has or is likely to experience physical or psychological harm due to domestic violence
- Caretaker’s mental illness/developmental disability impairs ability to supervise, protect or care for child
- Caretaker is violent and appears out of control
- Caretaker is unable/unwilling to meet child’s basic needs for food, clothing, shelter and/or medical care
- Caretaker is unwilling/unable to provide adequate supervision of child
- Caretaker caused serious physical harm to child or has made a plausible threat of serious
- Caretaker views/describes/acts negatively toward child and/or has extremely unrealistic expectations of child
- Child’s whereabouts are unknown, or the family is about to flee or refuse access to the child
- Caretaker caused serious physical harm to child or has make a plausible threat of serious harm
- Caretaker views/describes/acts negatively toward child and/or has extremely unrealistic expectations of child
- Child’s whereabouts are unknown, or the family is about to flee or refuse access to the child
- Current allegation or history of sexual abuse and caretaker is unable/unwilling to adequately protect child
- Physical living conditions are hazardous
- Child is afraid of or extremely uncomfortable around people living in or frequenting the home
- Child has Positive Toxicology for drugs and/or alcohol
- Child is on sleep apnea monitor
- Weapon noted in CPS report or found in the home
- Other/criminal activity (specify):

Several Models Reach 87-91% Accuracy Predicting Substantiation and Recidivism Using New Safety Factors Data



Use?	Model type	Model param...	No of models
☑	C5	Specify...	8
☑	Logistic regression	Specify...	8
☑	Decision List	Default	1
☑	Bayesian Network	Specify...	4
☑	Discriminant	Specify...	2
☑	KNN Algorithm	Specify...	8
☑	SVM	Specify...	2
☑	C&R Tree	Specify...	4
☑	Quest	Default	1
☑	CHAID	Specify...	16
☑	Neural Net	Specify...	12

Juvenile Recidivism PRM with Integrated Data in Philadelphia

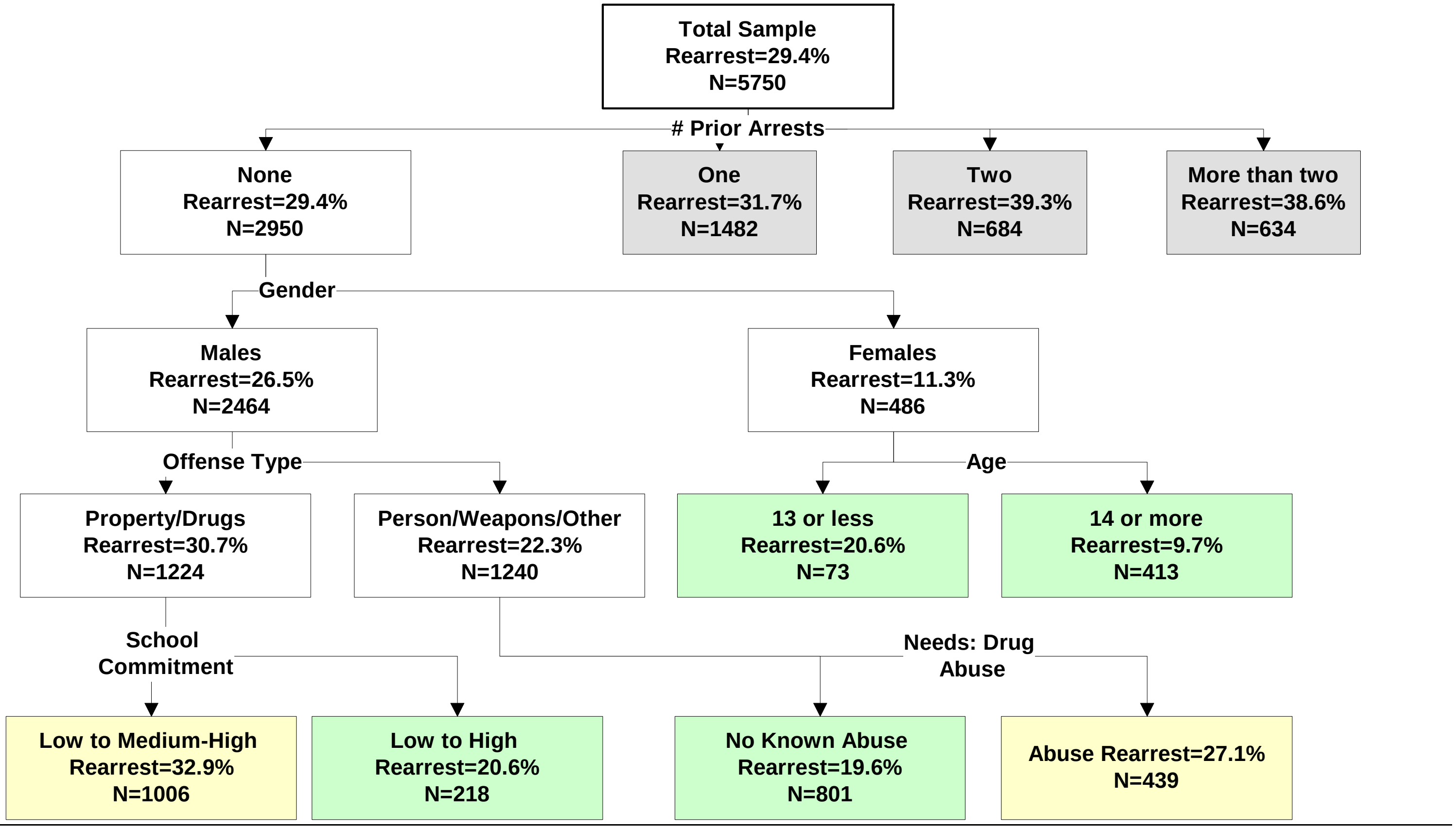
- NCCD's Wisconsin SDM Risk Instrument... carefully developed, validated, widely used.
- Configural Analysis.. customized risk assessment tool.
- Neural Networks.. more sophisticated customized analysis.. the future?
- Utilize single data source.
- Compare classifications and predictive validity.

Integrated Data Source

- Database of 40,000+ cases.
- All dispositions from Family Court where more than regular probation (program or state).
- Different data sources – court records, DHS staff assessments, self report.
- Current sample 8,239 cases, dispositions with staff/self-report data complete.

CHAID Model: Outline

Figure 2.1 Dendogram of Predictors of Re-Arrest (1 Prior exploded)



Neural Network Model

- Iterative procedure with ability to 'learn'
- Builds predictive models that can evolve and update themselves.. dynamic quality
- Widely used in other fields (e.g. medicine, finance) because highly adaptive and robust
- Research comparing neural networks with logistic/linear regression using child protective services concluded neural network produced superior prediction and classification results.

Comparing the methods

- Results show significant variation by method
- From individual rights perspective…
  - Wisconsin classifies 1215 low risk.. correct 75%
  - Configural classifies 2349 low risk.. correct 83%
  - Neural network classifies 5424 low risk.. correct 97%
- From public safety perspective…
  - Wisconsin classifies 1752 high risk.. correct 30%
  - Configural classifies 1521 high risk.. correct 43 %
  - Neural network classifies 2815 high risk.. correct 81%

Translating Results Into Practice with Practitioners

- Fancy predictive analytic models are useless unless they are integrated into practice
  - Gap Foundation Plan Ahead Model – Provider/Youth engagement was the most important predictor of outcomes.
  - Girl Scouts of Northern California and Thrive Foundation Model – Volunteer understanding of programs was the most important predictor of volunteer retention. In focus groups volunteers explained that Girl Scouts national office was forcing them to follow a new curriculum they hated.
  - OCFS Model – County decision-making examined.
  - Practitioners must be able to make meaning and leverage the analyses to inform their own practice.