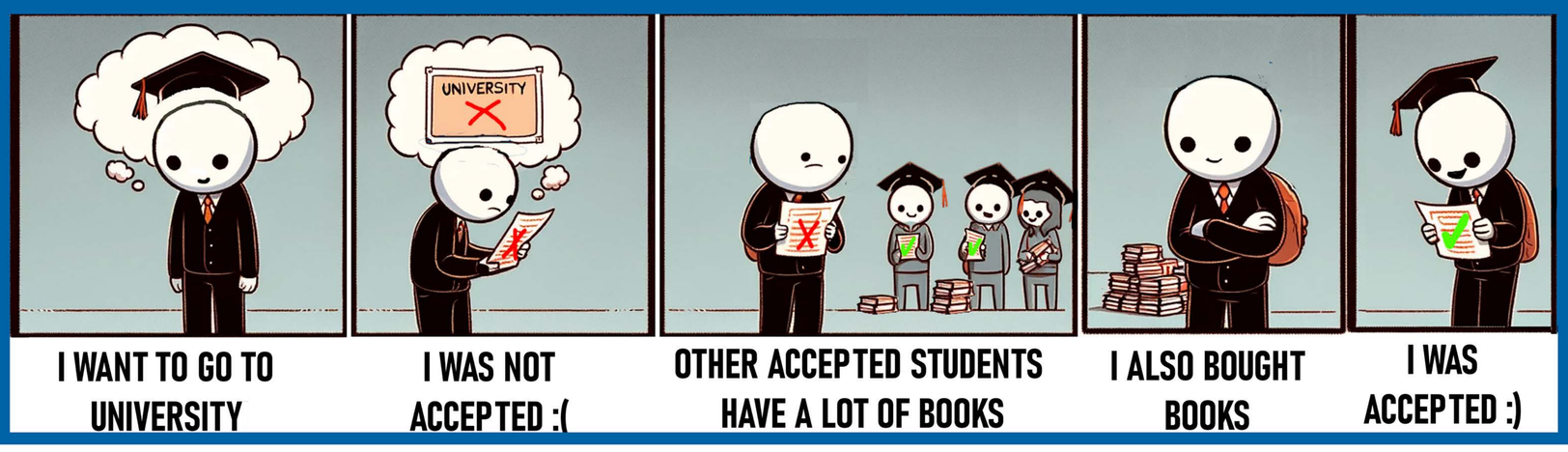
How do ML models influence us?



by: Blanka Visy
project supervisor: Ass.-Prof. Dipl.-Ing. Dr. techn. Sebastian Tschiatschek

universität wien

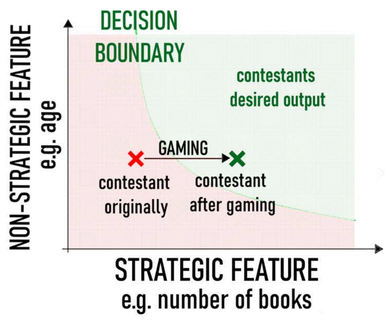
"SYSTEM WANTS CORRECT PREDICTIONS

USERS WANT POSITIVE PREDICTIONS"

-Nir Rosenfeld, 2021

1. INTRODUCTION: Strategic classification

- Setting: Classifiers make decisions about users based on the users' attributes
- Information from the classifier can be available to the users
- Gaming = manipulation of a users' attributes to modify the classifiers' decision => shift in distribution between training and deployment



How do models influence people?

How to get accurate decisions when gaming happens?

THE GAME

Players: *Jury* (Decision maker)
Contestant (users)

The game:

1. *Jury* publishes classifier $f : X \to \{-1, 1\}$
2. **User learns** of the decision of the model and the decision rule
3. Users not receiving desired decision try to alter features to get desired outcome while minimizing change costs:

$$\Delta(x) = \arg\max_{y \in X} f(y) - c(x, y)$$

new utility **cost of change**

4. If feasible and worthwhile, user makes changes; if not, they maintain current features.

Payoffs: *Jury*: accuracy on the new, shifted distribution

Contestant: utility of prediction of the classifier – costs of feature change

We consider mixed costs: $c(x, y) = q \cdot \|x - y\|_2^2 + (1 - q)\|x - y\|_1$

2. METHODOLOGY

Users: solve optimisation problem to get desired output

Jury options – what to do to avoid decreasing performance:

- **Repeated Risk Minimization** (RRM): continuously publish and optimize models
- Utilize **algorithms** to correct predictions based on transition model assumptions.

Algorithms to improve models while gaming:

One way to solve gaming: make model decisions robust to distribution shift based on assumptions to it

One strategy-robust learning algorithm by [1]:

- Input: labeled examples (can be even a black box model), description of a separable cost function
- Output: corrected labels for assumed transition model

[1]: Hardt, Moritz, et al. "Strategic classification.", 2016

3. EXPERIMENTS: Student Performance Prediction:

- Binary classification dataset to predict students' final grade on a 0-20 scale (0: below 10, 1: over 10).
- The goal of the students (contestants): get 1 as a prediction by the ML model

The features in X (bold: strategic features):

Feature(s)	Description
student info	sex, age, home adress, current health status
family info	family size, parents cohabitation status, parents education and job
studytime	weekly study time
Dalc	workday alcohol consumption
Walc	weekend alcohol consumption
…	

Costs for the users for change:

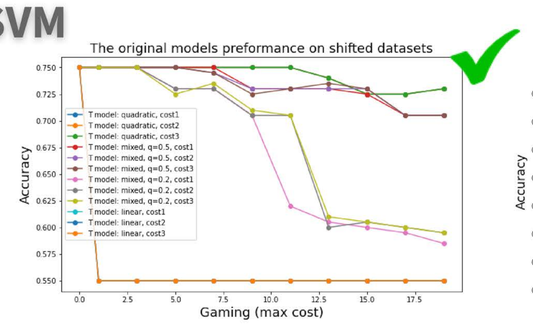
Assumed possible costs:			
Changeable features:	cost1	cost2	cost3
studytime	+1	+2	+3
Dalc	-1	-1	-1
Walc	-1	-1	-2
Meanings:	+: costly to increase -: costly to decrease		1: low cost 2: moderate cost 3: high cost

Other datasets: syntetic bank loan spam classification

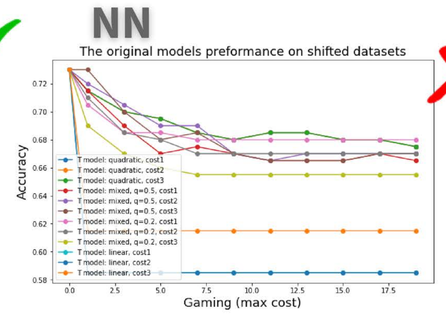
4. RESULTS: Which model should be chosen?

One-time shift

SVM

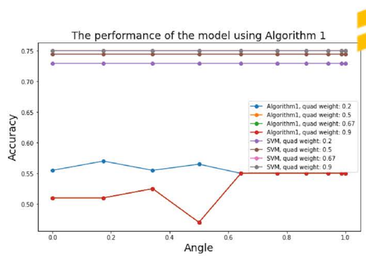


NN

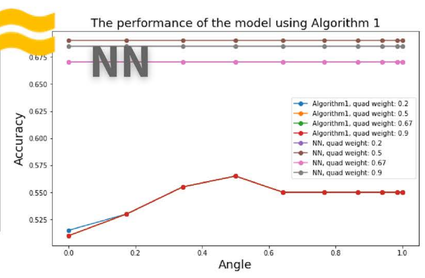


- SVM: more robust in performance
- NN: similar ACC for different costs and quadratic weights
- For small amounts of gaming, SVM performs better
- For high amounts of gaming, results are mixed

SVM

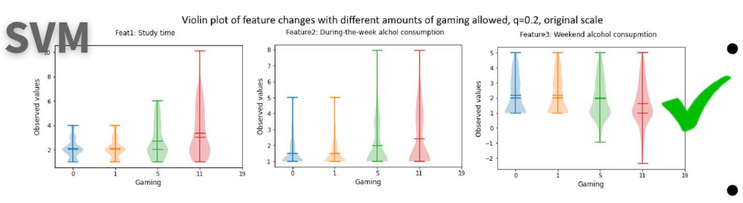


NN

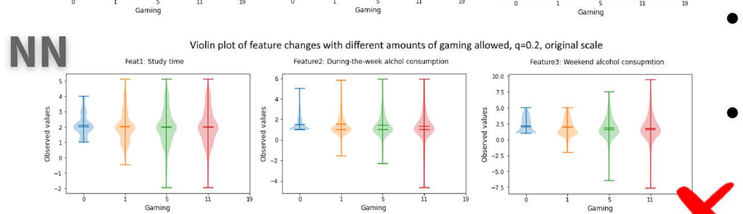


- The algortihm does not improve the performance in any of the cases
- Different quadratic weights act similarly

SVM



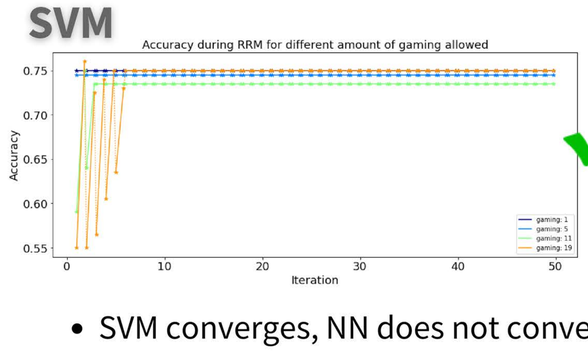
NN



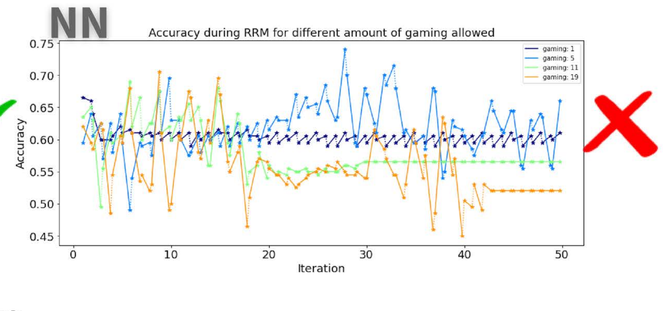
- Study time increases for SVM, the median remains constant for NN.
- Alcohol consumption during the week increases for SVM
- Weekend alcohol consumption decreases in both cases.

Repeated Risk Minimization

SVM

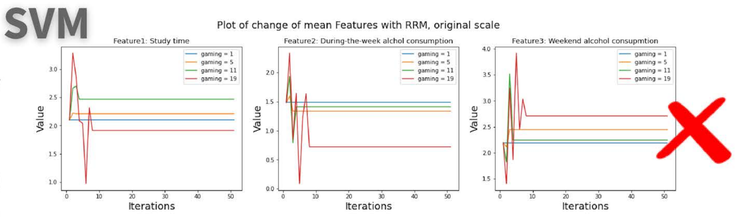


NN



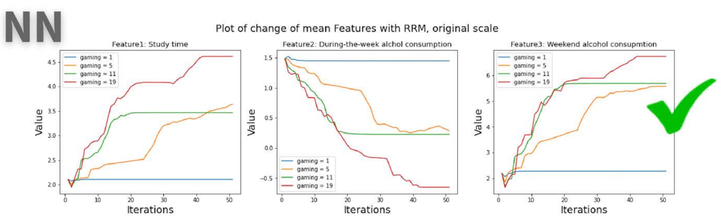
- SVM converges, NN does not converge

SVM



- Study time increases for NN, mostly increases for SVM
- Alcohol cosumption during the week decreases for both SVM and NN
- Alcohol consumption on the weekends rises in both cases

NN



5. SUMMARY

For this dataset, **SVM** is more robust, converges better in RRM and has better influence on users than **NN**

Winner model: SVM

Strategic classification: When a model's decision is significant, users often modify their features, creating a shift between the model's training and its actual use. This affects both the decision-makers and the users, so it's crucial to consider these impacts during model selection.

GAMING **PROS: it can motivate users to truly improve**
CONS: incorrect decisions can increase, this is bad for both players