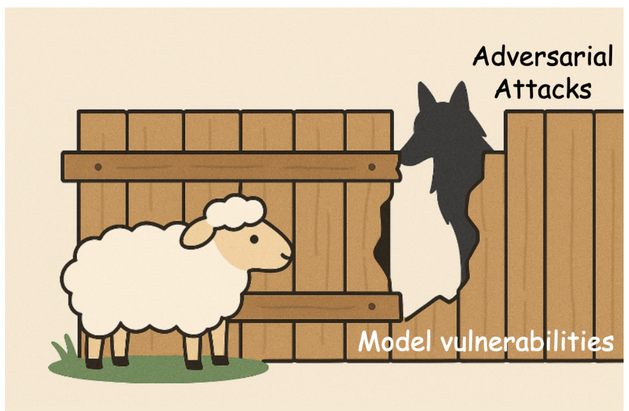
Building Robust Predictive Systems on Tabular Data

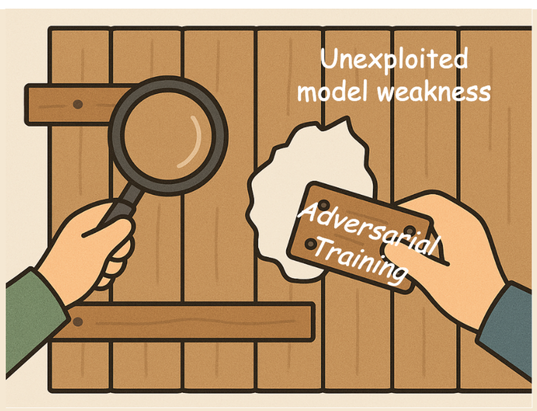
Zhipeng He[1] (*PhD Candidate*), A/Prof. Chun Ouyang[1] (*Principal Supervisor*), Prof. Alistair Barros[1] (*Associate Supervisor*), A/Prof. Catarina Moreira[2] (*External Supervisor*)
[1]*School of Information Systems, Queensland University of Technology*
[2]*Data Science Institute, University of Technology Sydney*

亡羊补牢: Mend the Fence After or Before the Sheep Are Lost

Adversarial Machine Learning as "Fence Repair": In adversarial machine learning, this ancient proverb offers a fitting analogy: *a model gets fooled, we investigate why, and then strengthen it against potential attacks.* This reactive cycle—discovering vulnerabilities through attacks, then repairing them—resembles fixing a fence only after the sheep have got lost.



Attacking the model: misclassification via model vulnerabilities.



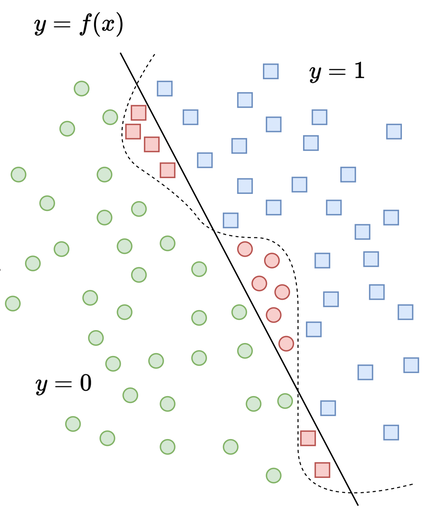
Inspecting the model: revealing and repairing model weaknesses in advance.

Mend the fence BEFORE the sheep are lost: Rather than relying on external attacks to expose model vulnerabilities after failure, we aim to discover weaknesses in advance by designing new adversarial attack algorithms that can reveal weak spots in the model before they cause failure.

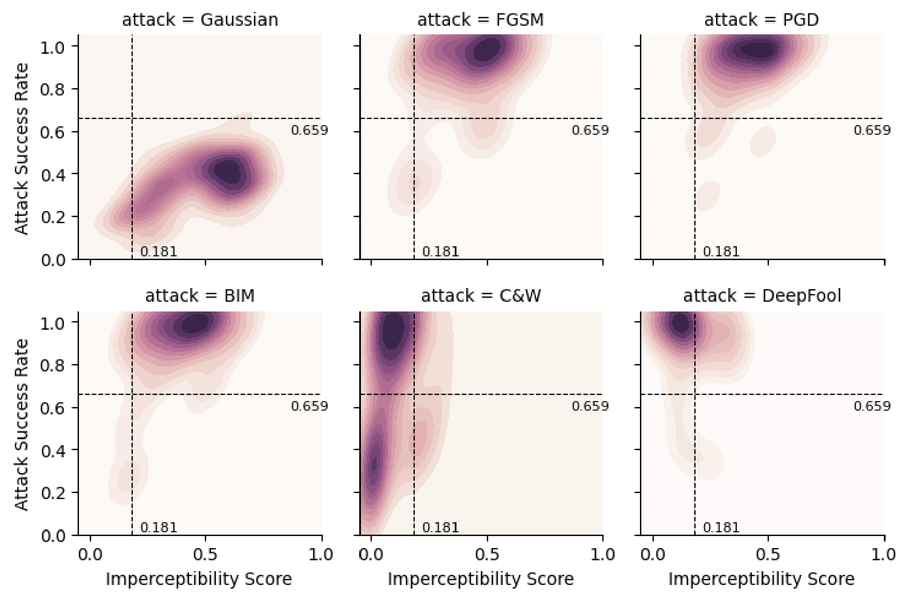
Research Gaps & Questions

Gaps: **a.** The imperceptibility of adversarial attacks on tabular data requires approaching different concepts compared to those for images. **b.** Current adversarial attacks lack imperceptibility metrics tailored for tabular data. **c.** No benchmark evaluates existing attacks under tabular imperceptibility criteria. **d.** Existing attacks are not designed for tabular imperceptibility properties.

RQ1: *What properties can be used to define the imperceptibility of adversarial attacks on tabular data?*
RQ2: *Which attacks can generate adversarial examples that are both effective and imperceptible?*
RQ3: *How can new adversarial attacks on tabular data be designed to generate both effective and imperceptible adversarial examples?*



Benchmarking Adversarial Attacks on Tabular Data [2]

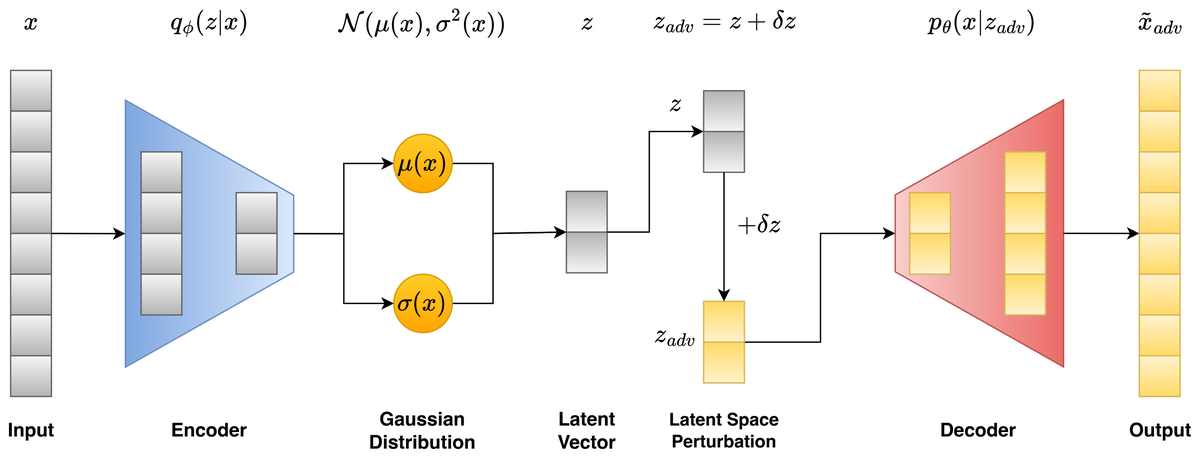


Insight 1: *DeepFool* can generate both effective and imperceptible adversarial examples.

Insight 2: Any attack can perturb numerical features; but only *PGD* can change categorical features on all models.

Insight 3: Unbounded attacks (*DeepFool* and *C&W*) generally make less changes and more likely generate in-distribution attack examples than bounded attacks (*FGSM*, *PGD* & *BIM*).

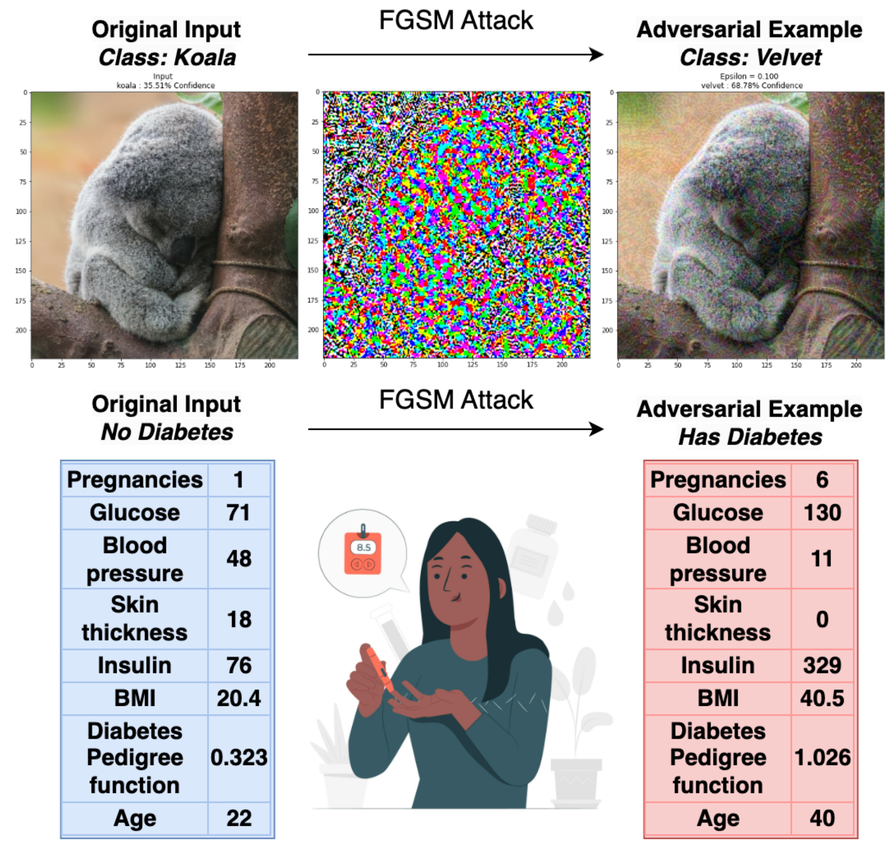
Crafting Imperceptible On-Manifold Tabular Attacks [3]



Highlights: **a.** Novel VAE-based framework generates imperceptible adversarial attacks for tabular data. **b.** Latent space approach unifies mixed types of features into coherent representation. **c.** Propose In-Distribution Success Rate to assess the deviation of adversarial examples. **d.** Our VAE attack achieves overall best performance across diverse datasets and models.

What is Adversarial Examples

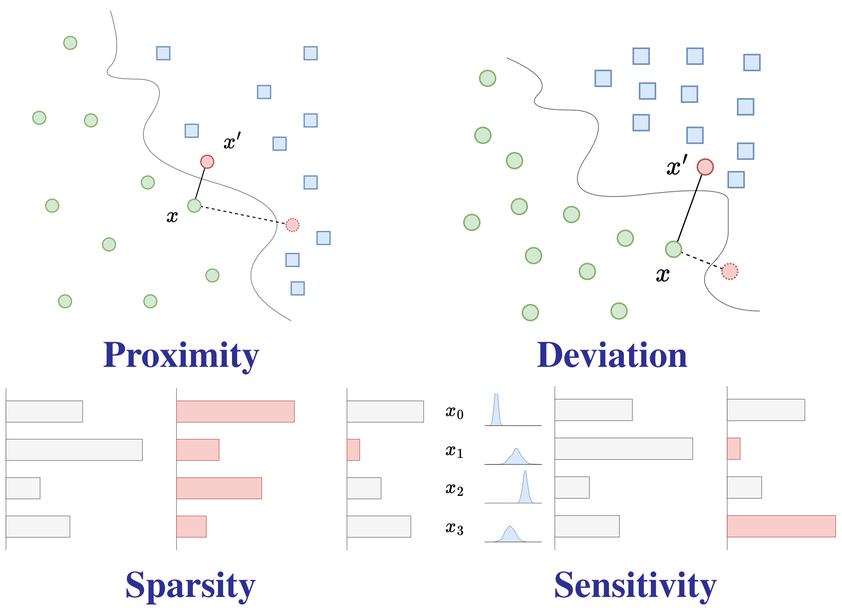
Adversarial examples are inputs that appear normal but are intentionally modified in subtle ways to fool a machine learning model. These changes are often imperceptible to humans, yet they can cause the model to make incorrect predictions.



Pregnancies	1
Glucose	71
Blood pressure	48
Skin thickness	18
Insulin	76
BMI	20.4
Diabetes Pedigree function	0.323
Age	22

Pregnancies	6
Glucose	130
Blood pressure	11
Skin thickness	0
Insulin	329
BMI	40.5
Diabetes Pedigree function	1.026
Age	40

Investigating Tabular Imperceptibility [1]



Feasibility — *Feature Interdependencies* — *Immutability*

Age	Education	Marital Status	Occupation	Relationship	Race	...	Income
43	Masters	Married-civ-spouse	Exec-managerial	Husband	White	...	>50k
28	HS-grad	Never-married	Craft-repair	Other-relative	Asian	...	<50k

Feasibility, Immutability & Feature Interdependence

References

[1] Zhipeng He et al. "Investigating imperceptibility of adversarial attacks on tabular data: An empirical analysis". In: *Intelligent Systems with Applications* 25 (2025), p. 200461.
[2] Zhipeng He et al. "TabAttackBench: A Benchmark for Adversarial Attacks on Tabular Data". In: *arXiv preprint arXiv:2505.21027* (2025).
[3] Zhipeng He et al. "Crafting Imperceptible On-Manifold Adversarial Attacks for Tabular Data". In: *arXiv preprint arXiv:2507.10998* (2025).

About Us

Homepage **GitHub** **XAMI Lab**