SONY

Adversarial Attacks for Tabular Data

Application to Fraud Detection and Imbalanced Data

BRU SSELS LAB

F. Cartella, O. Anunciação, Y. Funabiki, D. Yamaguchi, T. Akishita, O. Elshocht
Sony Corporation

SafeAI 2021, 8th of February 2021

Adversarial attacks on imbalanced tabular data

- Adversarial attacks have been increasingly investigated for image classification tasks

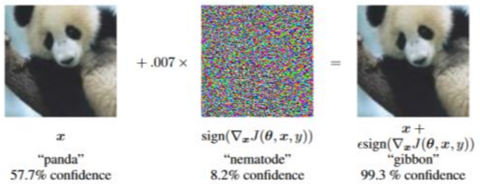
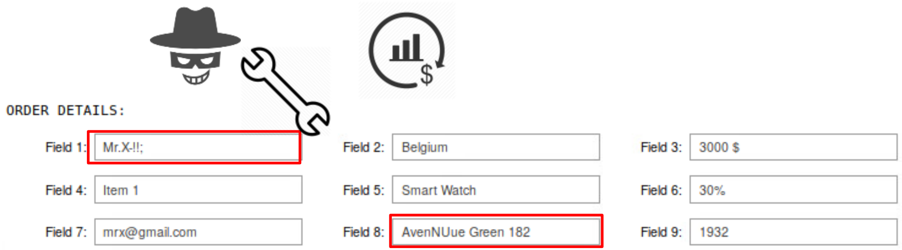


*Image from "Goodfellow, I. J.; Shlens, J.; and Szegedy, C. 2014 - Explaining and harnessing adversarial examples"



"Adversarial attacks are deliberate and imperceptible manipulations of input data made by attackers with the goal of modifying, to their advantage, the output of an AI system"

- Limited literature and tools on Adversarial attacks and defenses for AI models based on tabular data
  - AI applications based on tabular data are subject to unseen threats and attacks
  - Models designed for transactional data analysis need to be trained considering robustness and security issues

- This work focuses on adapting adversarial attacks to be effective on imbalanced tabular data (i.e., fraud detection use cases)

Main contribution and results

Three adversarial attack algorithms considered (ZOO, HopSkipJump and Boundary attacks)

	Image Classification	Fraud Detection	Solution
Class balance and bias in model	Relatively balanced data Relatively unbiased model	Highly imbalanced data Highly biased models where a properly tuned threshold is needed to take decision	Introduction of decision threshold within the attack algorithms and introduction of a novel loss function for ZOO algorithm
Data types and values range	Uniform data type and value range (i.e., integer between 0 and 255)	Heterogeneous and unconstrainted information (i.e., email addresses, amounts, ...)	Constrained perturbations to obtain realistic final values
Editability	An attacker can modify independently any of the pixel of an image	Some fields are not directly editable by attackers	Added editability constraints to the features that cannot be modified
Imperceptibility	Related to human visual perception	Related to changes made to features that are commonly checked by human operators (in case of manual inspection)	Introduction of a custom norm to drive the algorithm optimization process in obtaining adversarial examples that pass unnoticed the fraud check

- Experiment based on the German Credit Dataset (Dua and Graff 2017) for risk evaluation of loan applications
- Adversarial example considered successful when a modified risky loan application (considered as "fraud") is accepted because it is classified as safe by the model

	Boundary	HopSkipJump	ZOO
Success Rate	100%	100%	100%
Unrealistic values	0	0	0
Perturbed fields checked by humans	228 (-64%)	418 (-16%)	153 (-16%)
Perturbed non-editable fields	0	0	0

- Attack transferability tested on a real production fraud detection system
- Success rate: 13.6% of fraudulent adversarial examples accepted as not frauds