Machine Learning as an Attack Surface

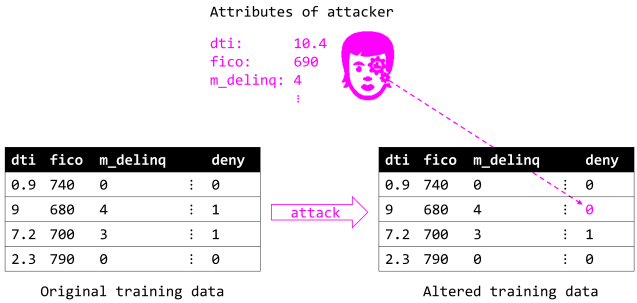
©Patrick Hall*

H2O.ai

July 14, 2019

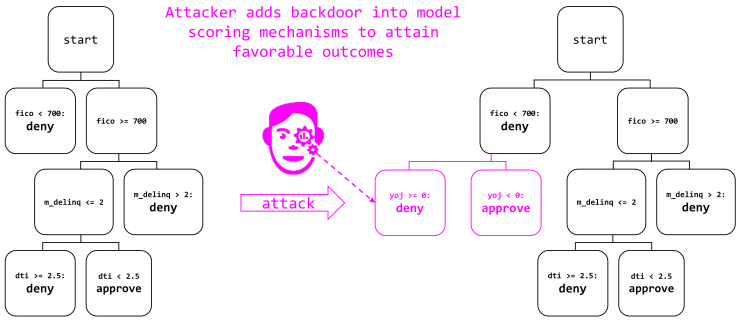
Contents

Data Poisoning Attacks

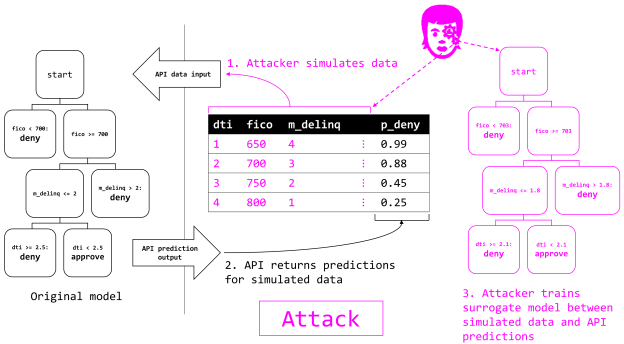


Attacker alters model training data to ensure favorable outcomes

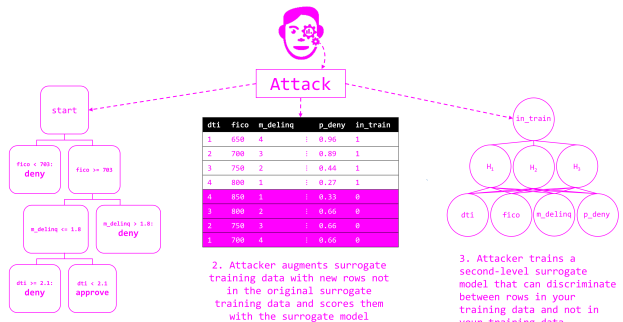
Watermark (i.e. Backdoor) Attacks



Surrogate Model Inversion Attacks



Membership Inference Attacks

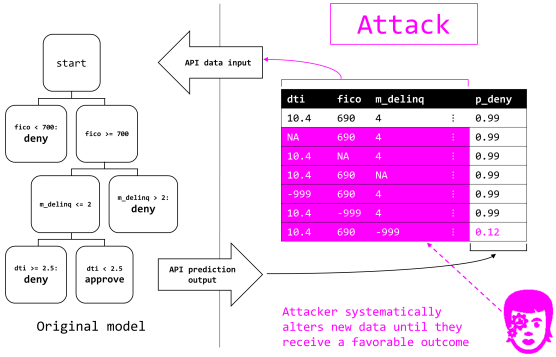


1. Attack begins with trained surrogate model

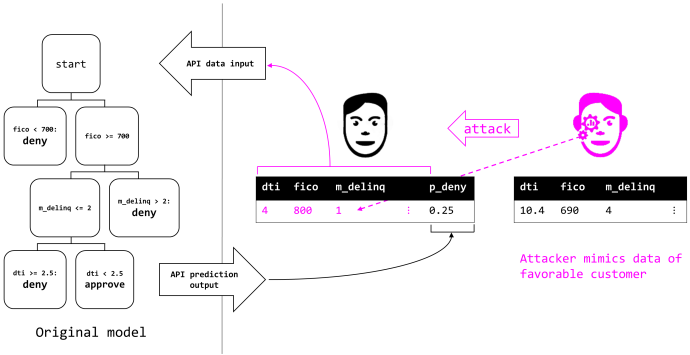
2. Attacker augments surrogate training data with new rows not in the original surrogate training data and scores them with the surrogate model

3. Attacker trains a second-level surrogate model that can discriminate between rows in your training data and not in your training data

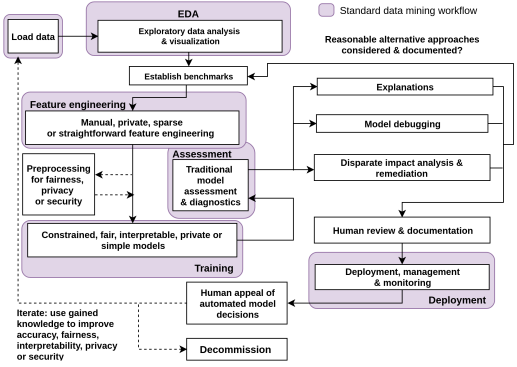
Adversarial Example Attacks



Impersonation Attacks



A Blueprint for Low-Risk Machine Learning



References

Proposals for model vulnerability and security:
https://www.oreilly.com/ideas/proposals-for-model-vulnerability-and-security

Can Your Machine Learning Model Be Hacked?!:
https://www.h2o.ai/blog/can-your-machine-learning-model-be-hacked/