Increasing Trust and Understanding in Machine Learning with Model Debugging

©Patrick Hall*

H_2O.ai

July 25, 2019

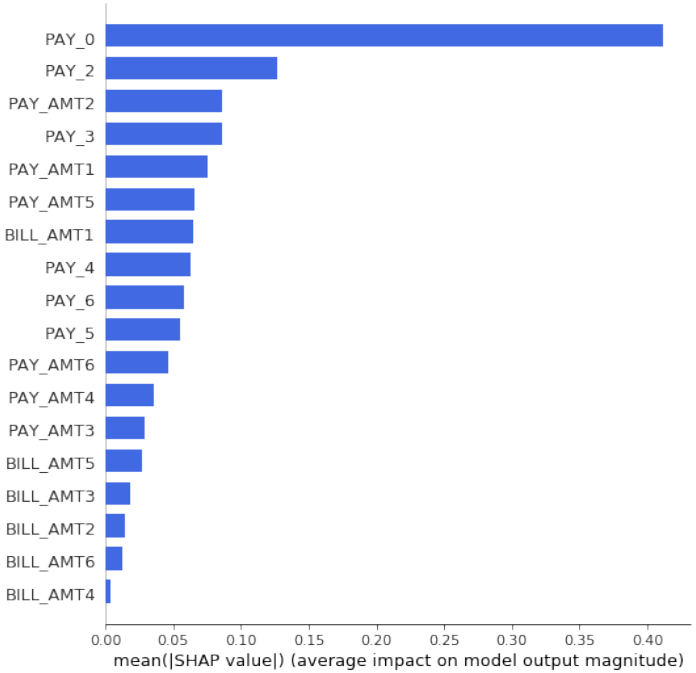
Contents

What is Model Debugging?

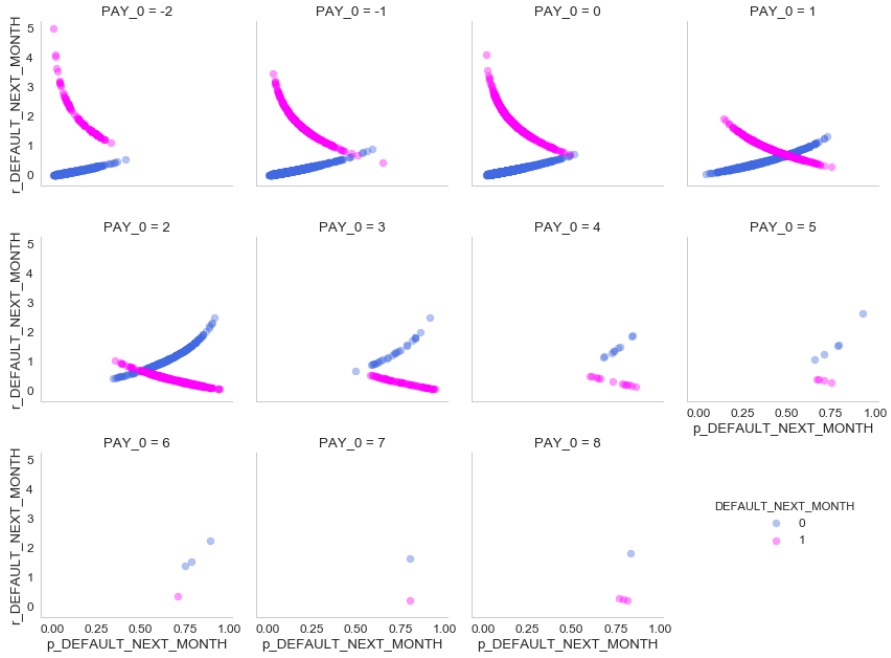
- Model debugging is an emergent discipline focused on discovering and remediating errors in the internal mechanisms and outputs of machine learning models.
- Model debugging attempts to test machine learning models like code (because the models are code).
- Model debugging promotes trust directly and enhances interpretability as a side-effect.

Why Bother With Model Debugging?

Machine learning models can be **inaccurate**.



This probability of default classifier, g_{mono}, over-emphasizes the most important feature, a customer's most recent repayment status, PAY_0.



g_{mono} also struggles to predict default for favorable statuses, $-2 \leq \texttt{PAY\_0} < 2$, and often cannot predict on-time payment when recent payments are late, $\texttt{PAY\_0} \geq 2$.

H2O.ai

Why Bother With Model Debugging?

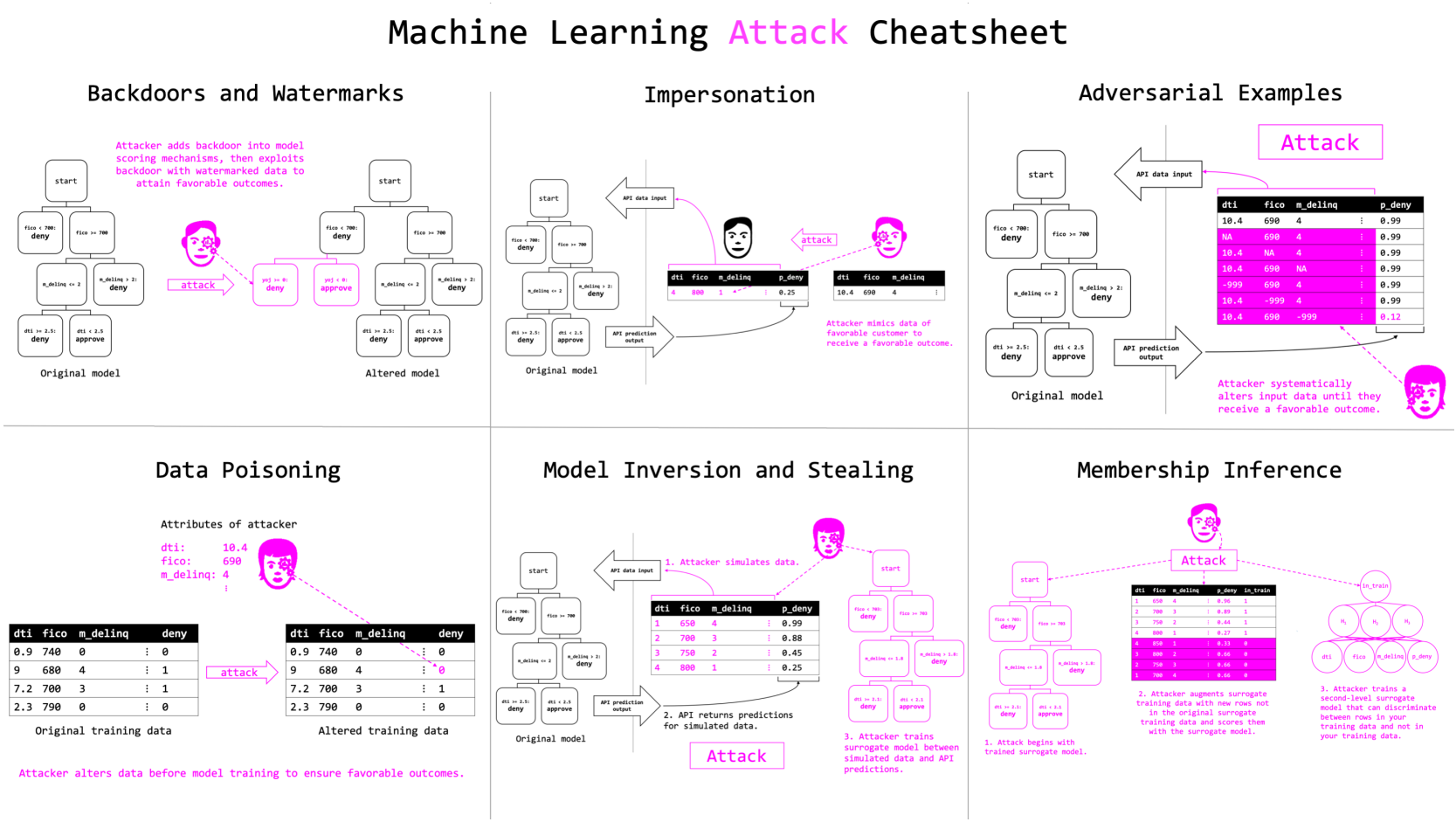
Machine learning models can exhibit **disparate impact** or other types of sociological bias.

	Adverse Impact Ratio	Accuracy Disparity	TPR Disparity	TNR Disparity	FPR Disparity	FNR Disparity
`single`	0.89	1.03	0.99	1.03	0.85	1.01
`divorced`	1.01	0.93	0.81	0.96	1.25	1.22
`other`	0.26	1.12	0.62	1.17	0	1.44

Group disparity metrics are out-of-range for g_{mono} across different marital statuses.

Why Bother With Model Debugging?

Machine learning models can have **security vulnerabilities**.



Hackers can manipulate models and steal models and data!

References

This presentation: