Increasing Trust and Understanding in Machine Learning with Model Debugging

©Patrick Hall*

H2O.ai

July 25, 2019

Contents

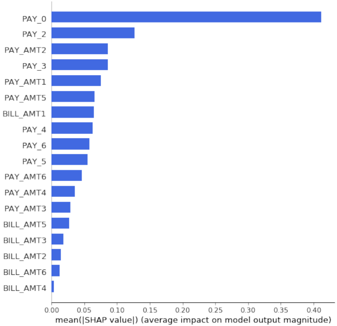
What is Model Debugging?

- 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.

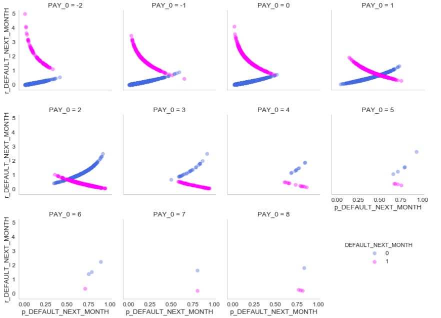
[†] See https://debug-ml-iclr2019.github.io/ for numerous model debugging approaches.

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, PAY_0 ≥ 2.

H2O.ai

Why Bother With Model Debugging?

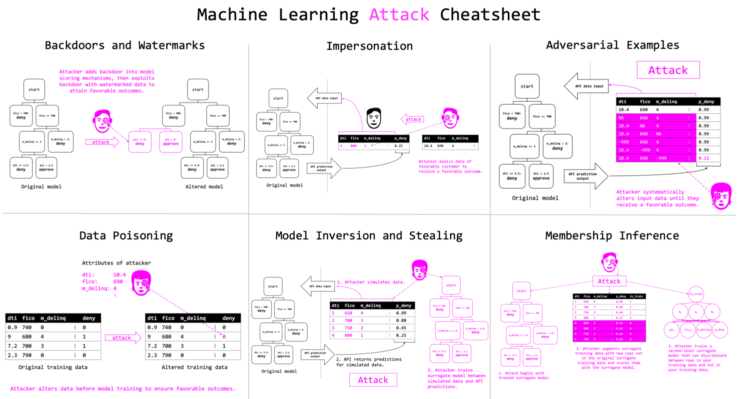
Machine learning models can perpetuate **sociological biases** [1].

	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?

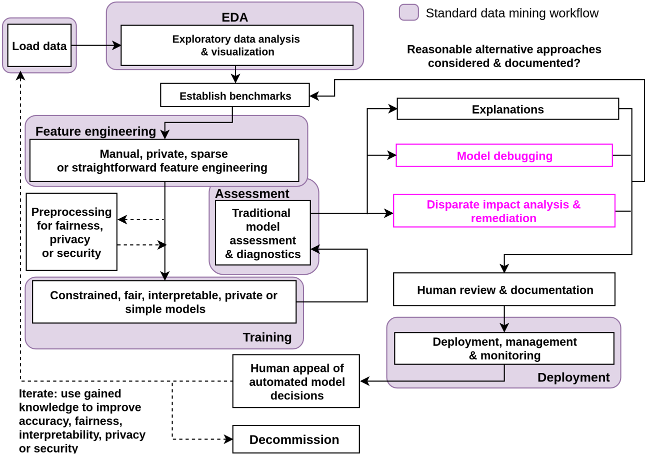
Machine learning models can have **security vulnerabilities** [2], [3], [4].



Hackers, competitors, or malicious or extorted insiders can manipulate model outcomes, steal models, and steal data!

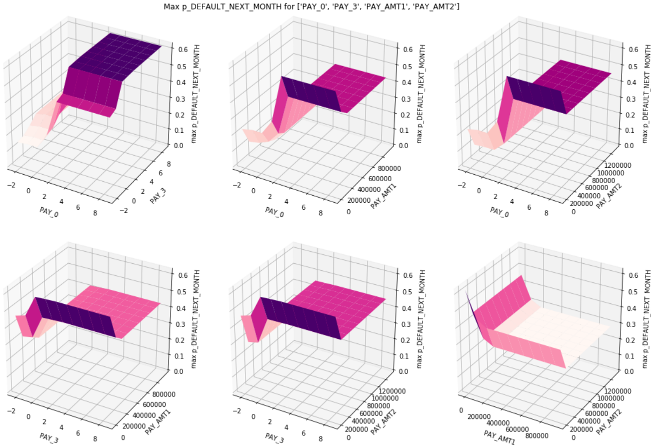
How to Debug Models?

As part of a holistic, low-risk approach to machine learning.‡



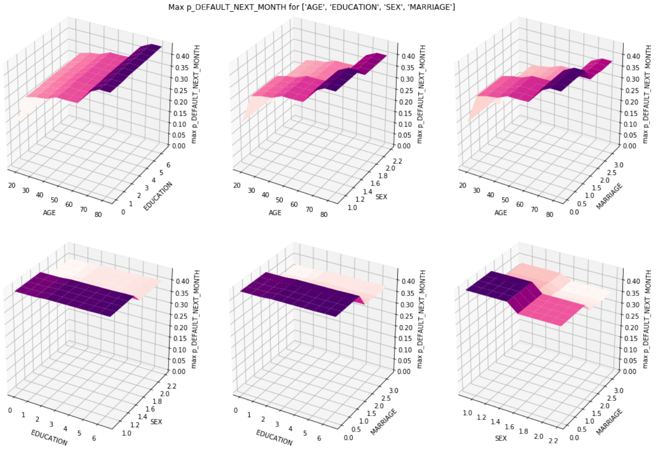
‡See https://github.com/jphall663/hc_ml for more information.

Sensitivity Analysis: Search for Adversarial Examples

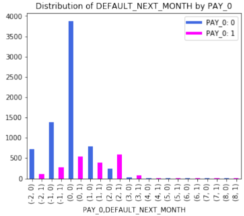


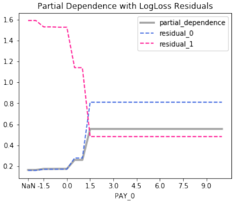
H2O.ai

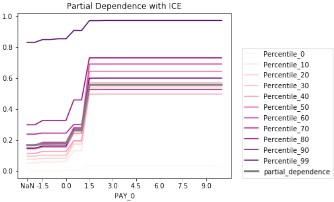
Sensitivity Analysis: Search for Adversarial Examples



Sensitivity Analysis: Partial Dependence and Individual Conditional Expectation (ICE)

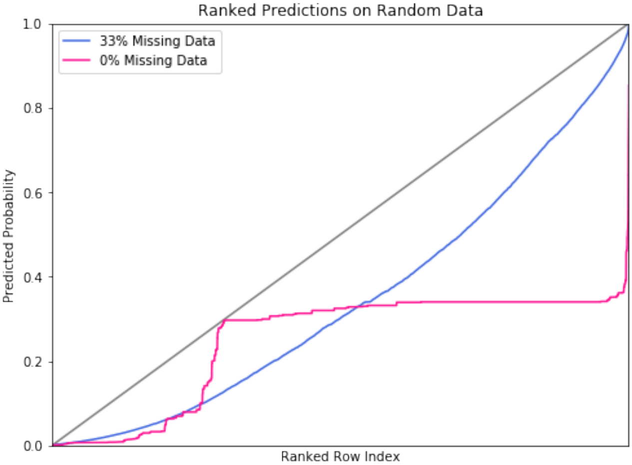






H2O.ai

Sensitivity Analysis: Random Attacks



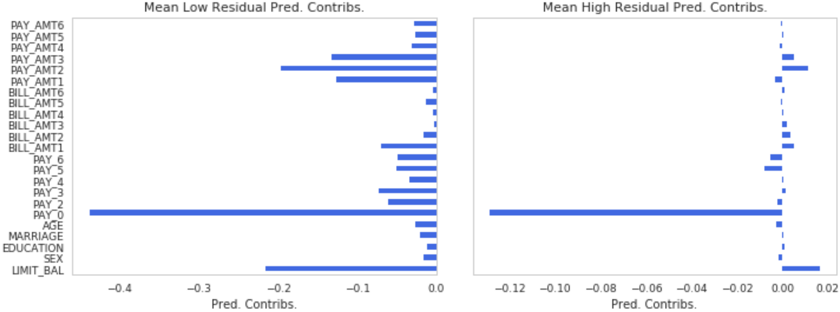
Residual Analysis: Disparate Errors

PAY_0	Prevalence	Accuracy	True Positive Rate	Precision	Specificity	Negative Predicted Value	False Positive Rate	False Discovery Rate	False Negative Rate	False Omissions Rate
-2	0.049	0.857	0.3	0.119	0.885	0.961	0.115	0.881	0.7	0.039
-1	0.117	0.805	0.383	0.267	0.861	0.913	0.139	0.733	0.617	0.087
0	0.05	0.864	0.345	0.143	0.891	0.963	0.109	0.857	0.655	0.037
1	0.822	0.457	0.368	0.93	0.871	0.229	0.129	0.07	0.632	0.771
2	1	0.709	0.709	1	0.5	0	0.5	0	0.291	1
3	1	0.748	0.748	1	0.5	0	0.5	0	0.252	1
4	1	0.571	0.571	1	0.5	0	0.5	0	0.429	1
5	1	0.444	0.444	1	0.5	0	0.5	0	0.556	1
6	1	0.25	0.25	1	0.5	0	0.5	0	0.75	1
7	1	0.5	0.5	1	0.5	0	0.5	0	0.5	1
8	1	0.75	0.75	1	0.5	0	0.5	0	0.25	1

SEX	Prevalence	Accuracy	True Positive Rate	Precision	Specificity	Negative Predicted Value	False Positive Rate	False Discovery Rate	False Negative Rate	False Omissions Rate
Male	0.3	0.773	0.513	0.655	0.884	0.809	0.116	0.345	0.487	0.191
Female	0.242	0.788	0.495	0.573	0.882	0.845	0.118	0.427	0.505	0.155

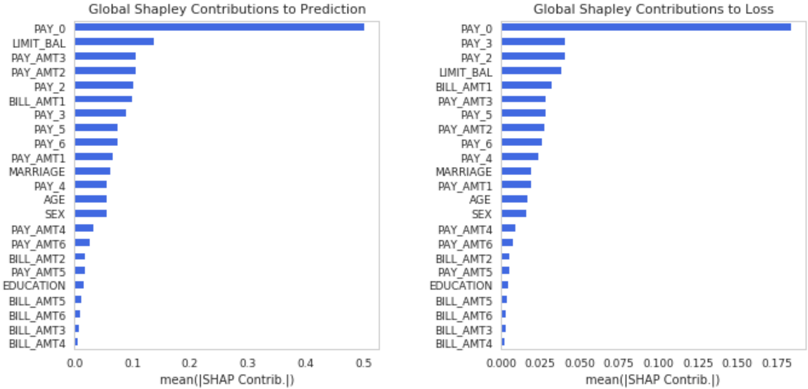
H_2O.ai

Residual Analysis: Mean Local Feature Contributions

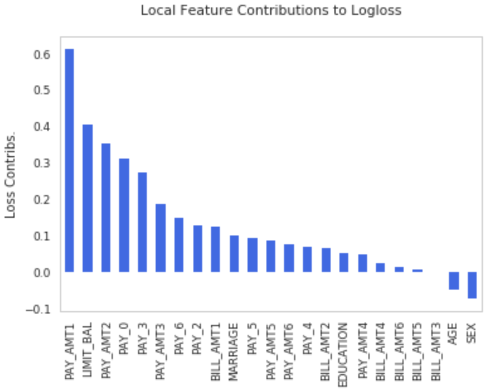


H2O.ai

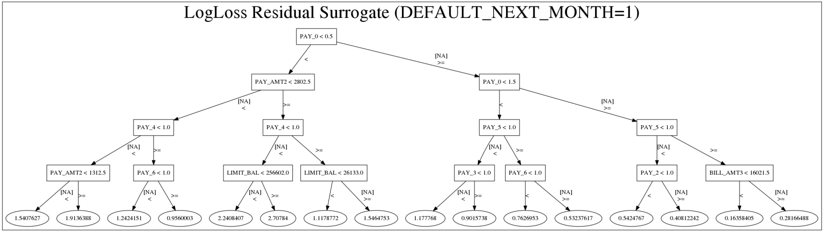
Residual Analysis: Global Importance for Predictions and Logloss



Residual Analysis: Local Feature Contributions to Logloss

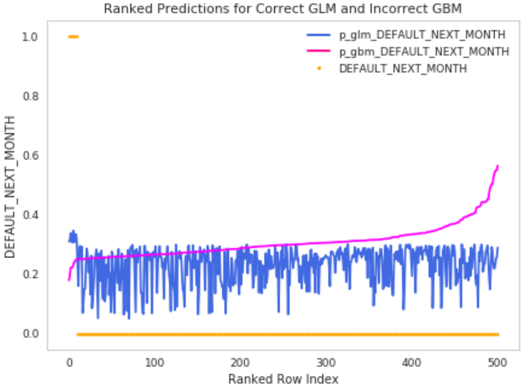


Residual Analysis: Surrogate Decision Trees



H2O.ai

Benchmark Models



References

This presentation:

References

[1] Solon Barocas, Moritz Hardt, and Arvind Narayanan. *Fairness and Machine Learning*. URL: http://www.fairmlbook.org. fairmlbook.org, 2018.

[2] Marco Barreno et al. "The Security of Machine Learning." In: *Machine Learning* 81.2 (2010). URL: https://people.eecs.berkeley.edu/~adj/publications/paper-files/SecML-MLJ2010.pdf, pp. 121–148.

[3] Reza Shokri et al. "Membership Inference Attacks Against Machine Learning Models." In: *2017 IEEE Symposium on Security and Privacy (SP)*. URL: https://arxiv.org/pdf/1610.05820.pdf. IEEE. 2017, pp. 3–18.

[4] Florian Tramèr et al. "Stealing Machine Learning Models via Prediction APIs." In: *25th {USENIX} Security Symposium ({USENIX} Security 16)*. URL: https://www.usenix.org/system/files/conference/usenixsecurity16/sec16_paper_tramer.pdf. 2016, pp. 601–618.