

- 
- 
- 

- 
- 
- 
- 

# Increasing Trust and Understanding in Machine Learning with Model Debugging

© Patrick Hall\*

H<sub>2</sub>O.ai

July 26, 2019

---

\* This material is shared under a [CC By 4.0 license](#) which allows for editing and redistribution, even for commercial purposes. However, any derivative work should attribute the author and H2O.ai.

- 
- 
- 

- 
- 
- 
- 

# Contents

## What?

### Why?

Inaccuracy

Sociological Biases

Security Vulnerabilities

### How?

Holistic, Low-Risk Approach

Sensitivity Analysis

Residual Analysis

Benchmark Models

Error Remediation

# 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.<sup>†</sup>
- 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.

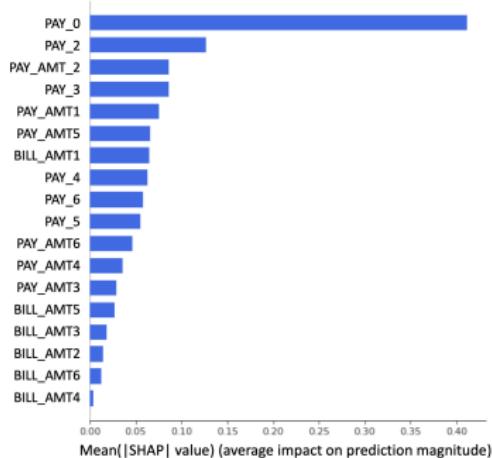
---

<sup>†</sup> See <https://debug-ml-iclr2019.github.io/> for numerous model debugging approaches.

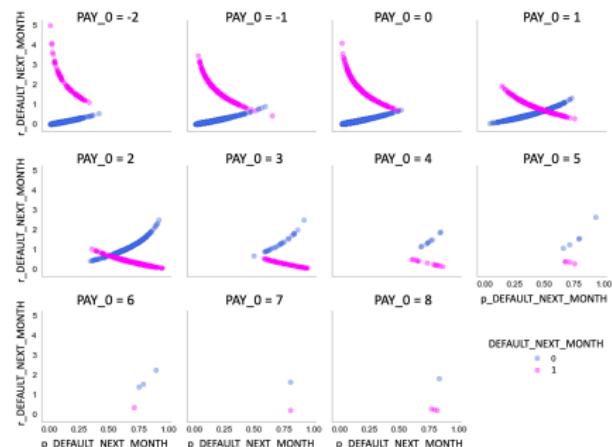


# Why Debug?

Machine learning models can be **inaccurate**.



This probability of default classifier,  $g_{\text{mono}}$ , over-emphasizes the most important feature, a customer's most recent repayment status,  $\text{PAY}_0$ .



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



# Why Debug?

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

|                 | Adverse Impact Disparity | Accuracy Disparity | True Positive Rate Disparity | Precision Disparity | Specificity Disparity |
|-----------------|--------------------------|--------------------|------------------------------|---------------------|-----------------------|
| <b>single</b>   | 0.885                    | 1.029              | 0.988                        | 1.008               | 1.025                 |
| <b>divorced</b> | 1.014                    | 0.932              | 0.809                        | 0.806               | 0.958                 |
| <b>other</b>    | 0.262                    | 1.123              | 0.62                         | 1.854               | 1.169                 |

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

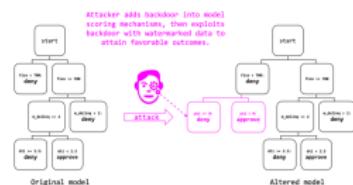


# Why Debug?

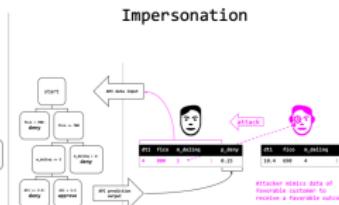
Machine learning models can have **security vulnerabilities** [2], [3], [4].<sup>‡</sup>

## Machine Learning Attack Cheatsheet

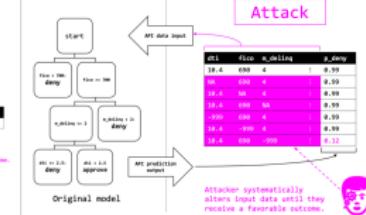
### Backdoors and Watermarks



### Impersonation



### Adversarial Examples



### Data Poisoning

Attributes of attacker

|          |      |  |  |  |
|----------|------|--|--|--|
| dtl      | 10.4 |  |  |  |
| first    | 4000 |  |  |  |
| m_delinq | 4    |  |  |  |

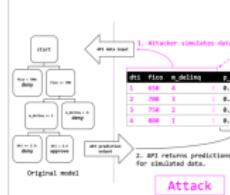
Attack

Original training data

Altered training data

Attacker alters data before model training to ensure favorable outcomes.

### Model Inversion and Stealing



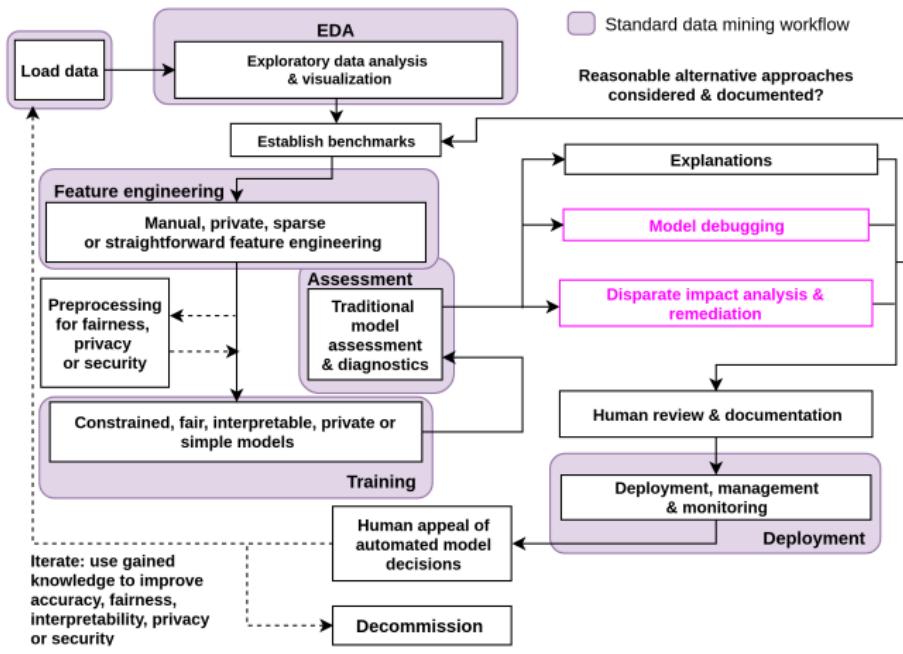
### Membership Inference



<sup>‡</sup>See [https://github.com/jphall663/secure\\_ML\\_ideas](https://github.com/jphall663/secure_ML_ideas) for full size image and more information.

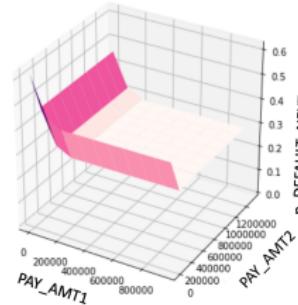
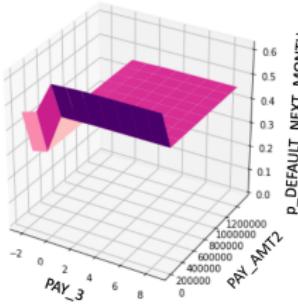
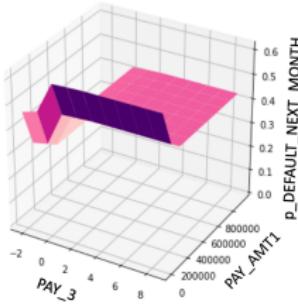
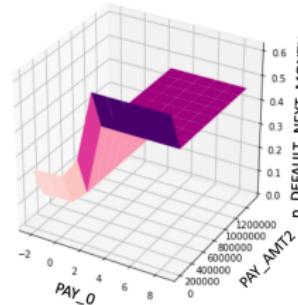
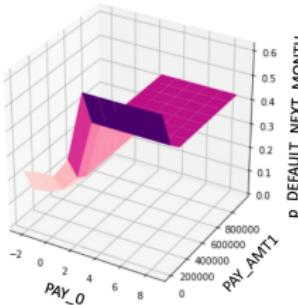
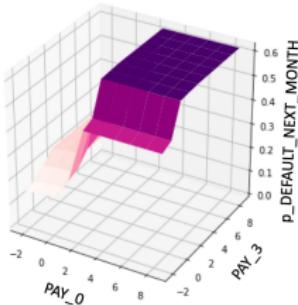
# How to Debug Models?

As part of a holistic, low-risk approach to machine learning.<sup>§</sup>

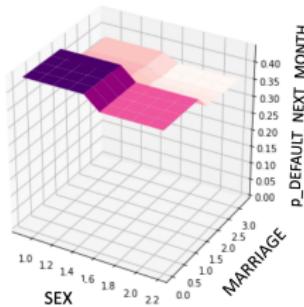
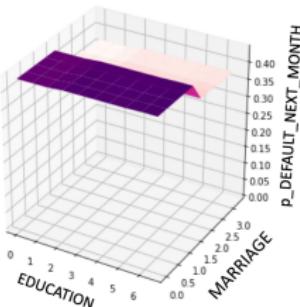
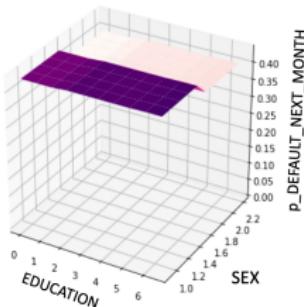
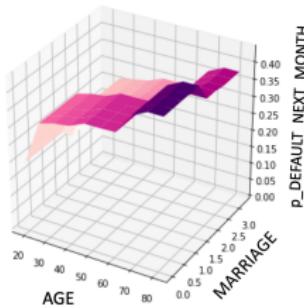
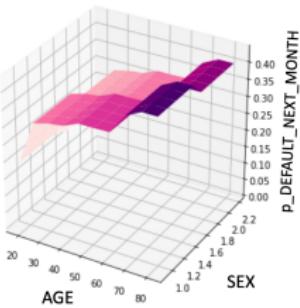
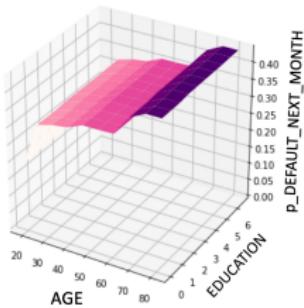


<sup>§</sup> See [https://github.com/jphall663/hc\\_ml](https://github.com/jphall663/hc_ml) for more information.

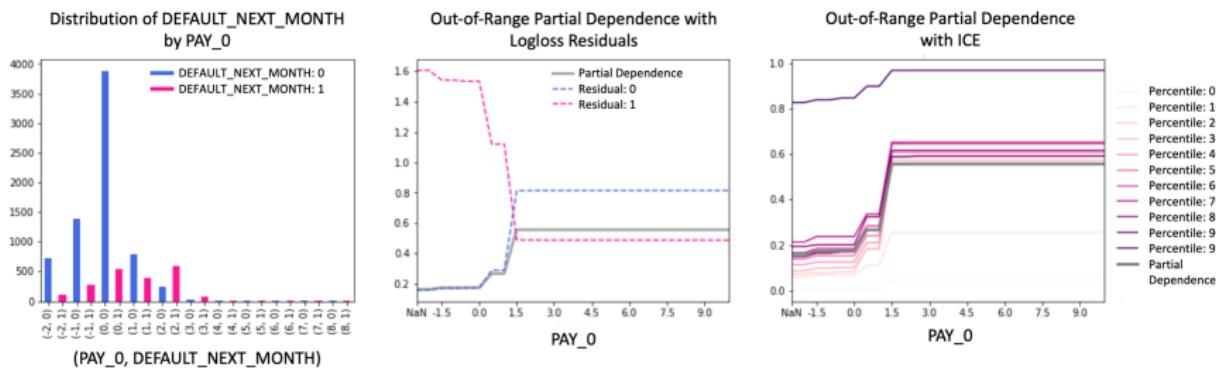
# Sensitivity Analysis: Search for Adversarial Examples



# Sensitivity Analysis: Search for Adversarial Examples



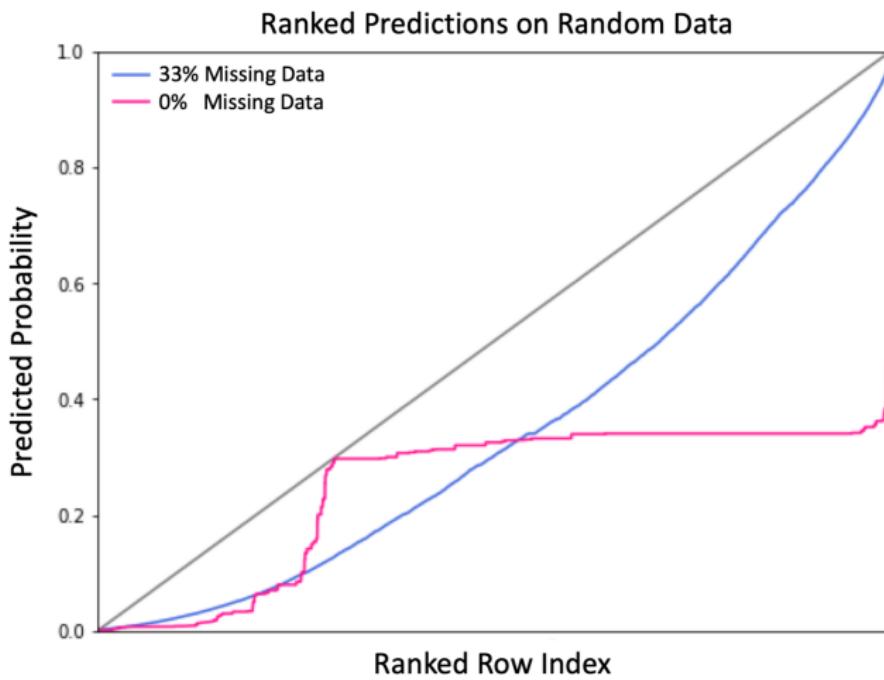
# Sensitivity Analysis: Partial Dependence and Individual Conditional Expectation (ICE)



- 

- 

## Sensitivity Analysis: Random Attacks

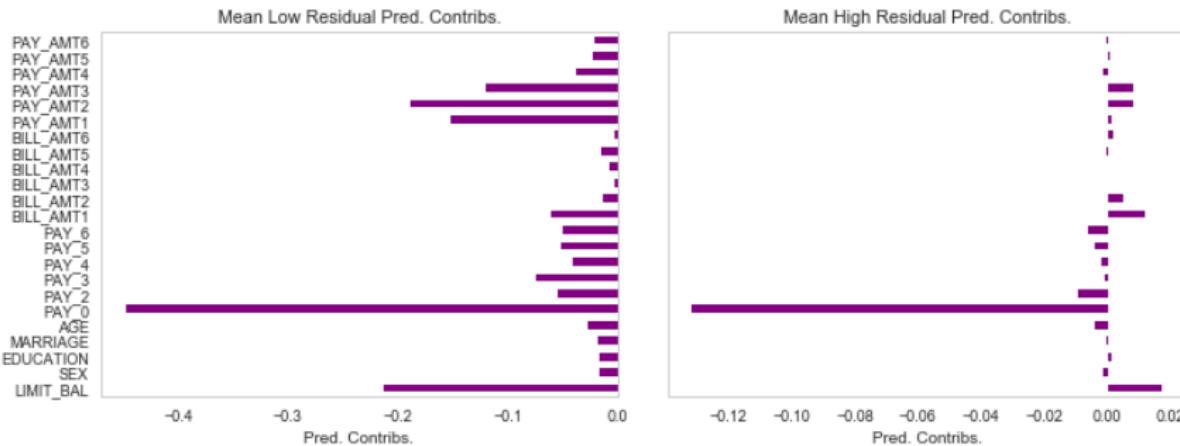


# Residual Analysis: Disparate Errors

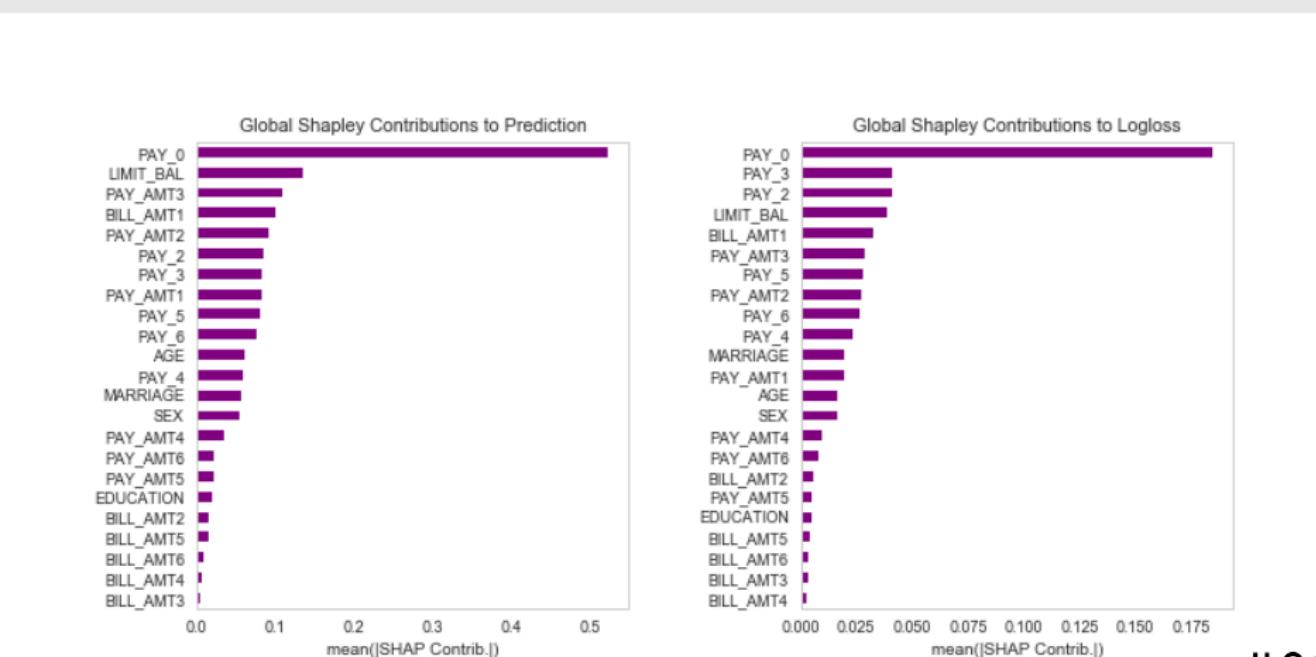
| PAY_0 | Prevalence | Accuracy | True Positive Rate | Precision | Specificity | Negative Predicted Value | False Positive Rate | False Discovery Rate | False Negative Rate | False Omissions Rate |
|-------|------------|----------|--------------------|-----------|-------------|--------------------------|---------------------|----------------------|---------------------|----------------------|
| -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                |

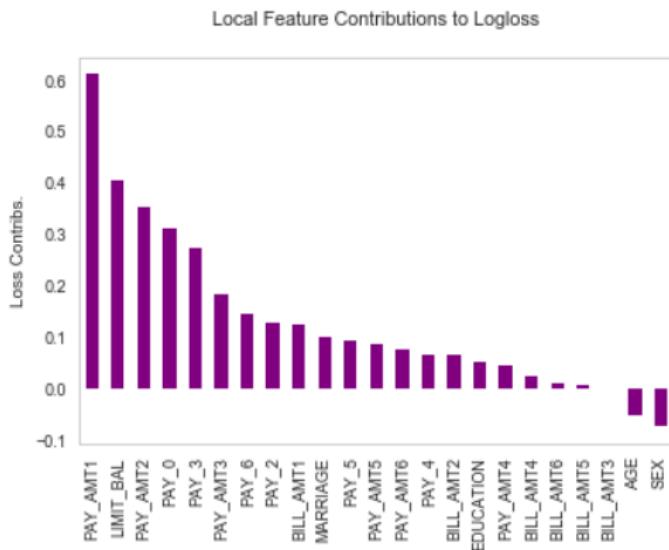
# Residual Analysis: Mean Local Feature Contributions



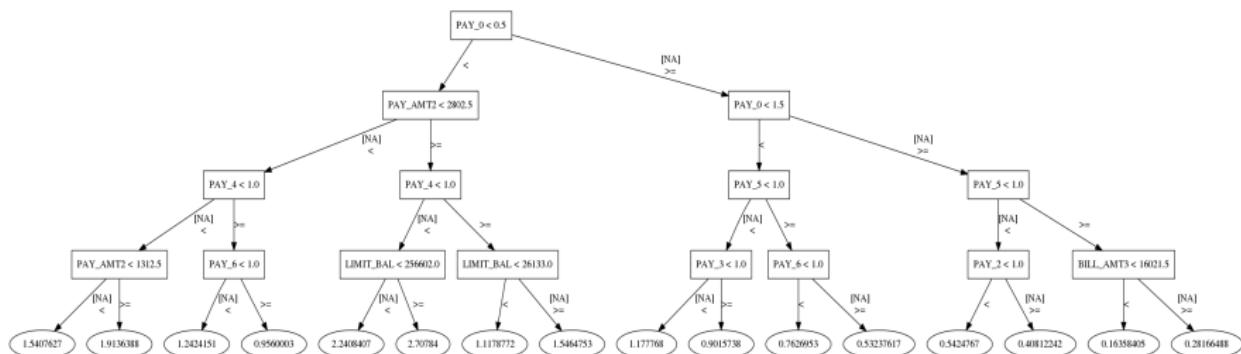
# Residual Analysis: Global Importance for Predictions and Logloss



# Residual Analysis: Local Feature Contributions to Logloss

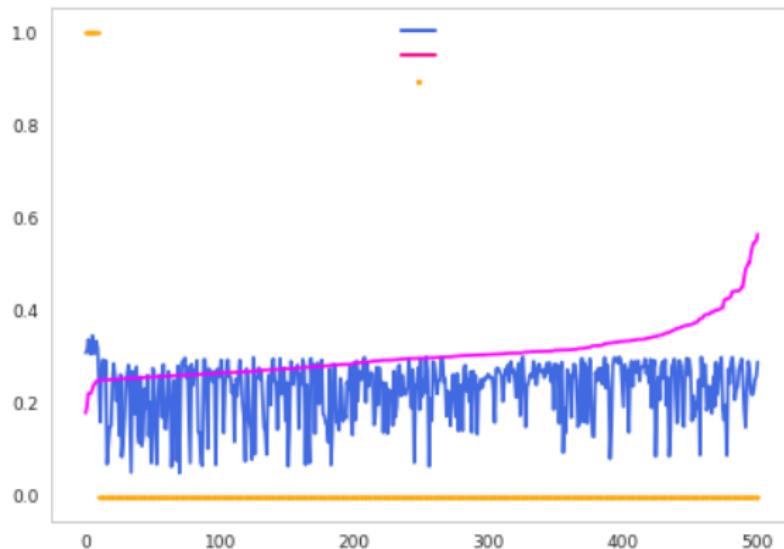


## Residual Analysis: Surrogate Decision Trees



Decision tree model of  $g_{\text{mono}}$  DEFAULT\_NEXT\_MONTH = 1 logloss residuals with 3-fold CV MSE = 0.0070 and  $R^2$  = 0.8871.

# Benchmark Models: Compare to Linear Models



What?

Why?

- 
- 
- 

How?

- 
- 
- 
- 

References

## 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](https://www.usenix.org/system/files/conference/usenixsecurity16/sec16_paper_tramer.pdf). 2016, pp. 601–618.