Explaining AI: Putting Theory into Practice

Luke Merrick
Data Scientist

fiddler.ai

Abstract

In this talk, we will cover some of the learnings from our experiences working with various model-explanation algorithms across business domains. Through the lens of two case studies, we will discuss the theory, application, and practical-use guidelines for effectively using explainability techniques to generate value in your data science lifecycle.

Explainability takeaways

1. Explanations are models
2. Global complexity, local simplicity

Context: concerns about machine learning

NEXT ECONOMY

J.P. Morgan Chase's $55 Million Discrimination Settlement

MIT News

Study finds gender and skin-type bias in commercial AI systems

Feb 12, 2018

QUARTZ

Amazon's AI-powered recruiting tool was biased against women

Oct 10, 2018

Missouri S&T News and Research

After Uber, Tesla incidents, can artificial intelligence be trusted?

Apr 10, 2018

Forbes

Congressional Leaders Press Zuckerberg On Political Bias

Apr 11, 2018

Guilty! AI Is Found to Perpetuate Biases in Jailing

1 day ago

More context: privacy and compliance regulations

*Companies should commit to ensuring systems, including AI, will be **GDPR** compliant with **sizeable fines of €20 million** or **4% of global turnover.***

Article 22** of GDPR empowers individuals with the **right to demand an explanation of how an AI system made a decision that affects them.

***California Consumer Privacy Act** of 2018 requires companies to **rethink their approach to capturing, storing, and sharing personal data** to align with the new requirements by **January 1, 2020**.*

How do we address this? AI Explainability.

Takeaway 1: explanations are models

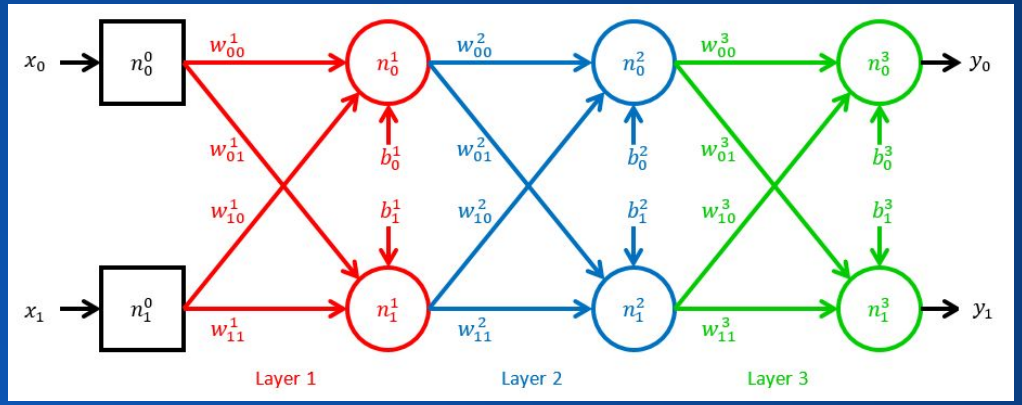
- “Outcome Y happened because of X”

  → Y = f(X)

- “The way X causes outcome Y is <list of rules>”

  → f = <list of rules>

Black box = a model with complex rules



$$\begin{pmatrix} y_0 \\ y_1 \end{pmatrix} = \begin{pmatrix} a\left(\begin{matrix} w_{00}^3 a(w_{00}^2 a(w_{00}^1 x_0 + w_{10}^1 x_1 + b_0^1) + w_{10}^2 a(w_{01}^1 x_0 + w_{11}^1 x_1 + b_1^1) + b_0^2) + \\ w_{10}^3 a(w_{01}^2 a(w_{00}^1 x_0 + w_{10}^1 x_1 + b_0^1) + w_{11}^2 a(w_{01}^1 x_0 + w_{11}^1 x_1 + b_1^1) + b_1^2) + b_0^3 \end{matrix}\right) \\ a\left(\begin{matrix} w_{01}^3 a(w_{00}^2 a(w_{00}^1 x_0 + w_{10}^1 x_1 + b_0^1) + w_{10}^2 a(w_{01}^1 x_0 + w_{11}^1 x_1 + b_1^1) + b_0^2) + \\ w_{11}^3 a(w_{01}^2 a(w_{00}^1 x_0 + w_{10}^1 x_1 + b_0^1) + w_{11}^2 a(w_{01}^1 x_0 + w_{11}^1 x_1 + b_1^1) + b_1^2) + b_1^3 \end{matrix}\right) \end{pmatrix}$$

Figures reproduced with permission from Marc Tanti, geekyisawesome.blogspot.com/2016/02/gradient-descent-algorithm-on.html

Explaining machine learning: two approaches

1. Use a model with simple rules

   → “interpretable models”

2. Approximate complex rules with simpler ones

   → “black box explainability”

A case for black box explainability

- Simplicity can lead to poorer performance

  → *See Kaggle, every competition*

- Simplicity can lead to discriminatory models

  → *See Chen, Irene, et al. Why Is My Classifier Discriminatory? NeurIPS 2018*

- It seems closer to the way we humans do it

  → *See the next slide (this is my opinion)*

Example: explain this human decision

?

Takeaway 2: global complexity, local simplicity

Near a single prediction or set of predictions, a simple model may accurately describe the complex black box model.

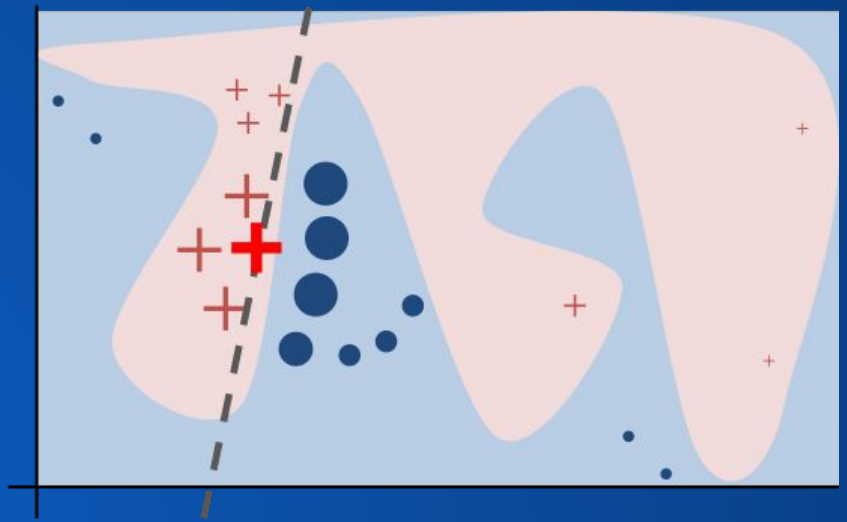
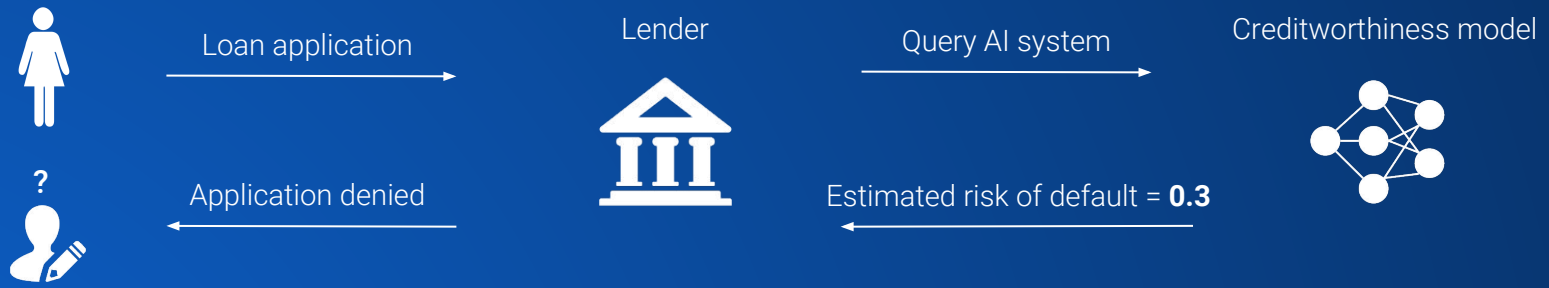


Figure from Marco Ribiero, et al. Why should I trust you? Explaining the predictions of any classifier

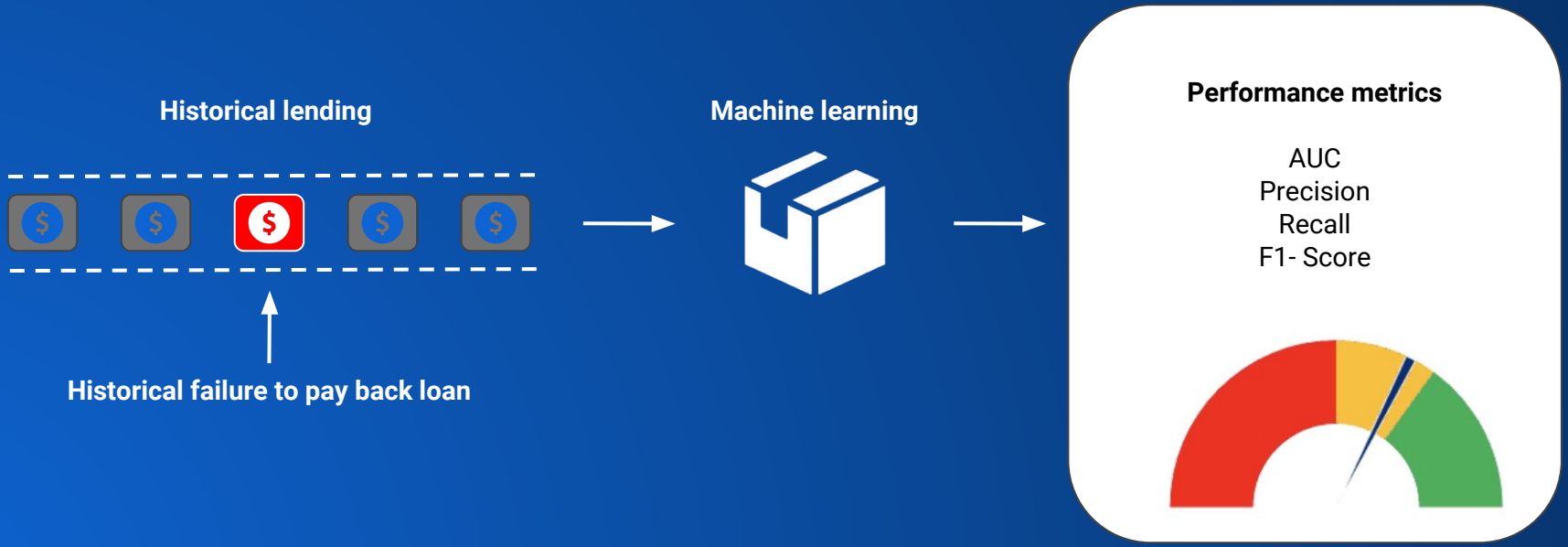
Case study 1: model-based lending



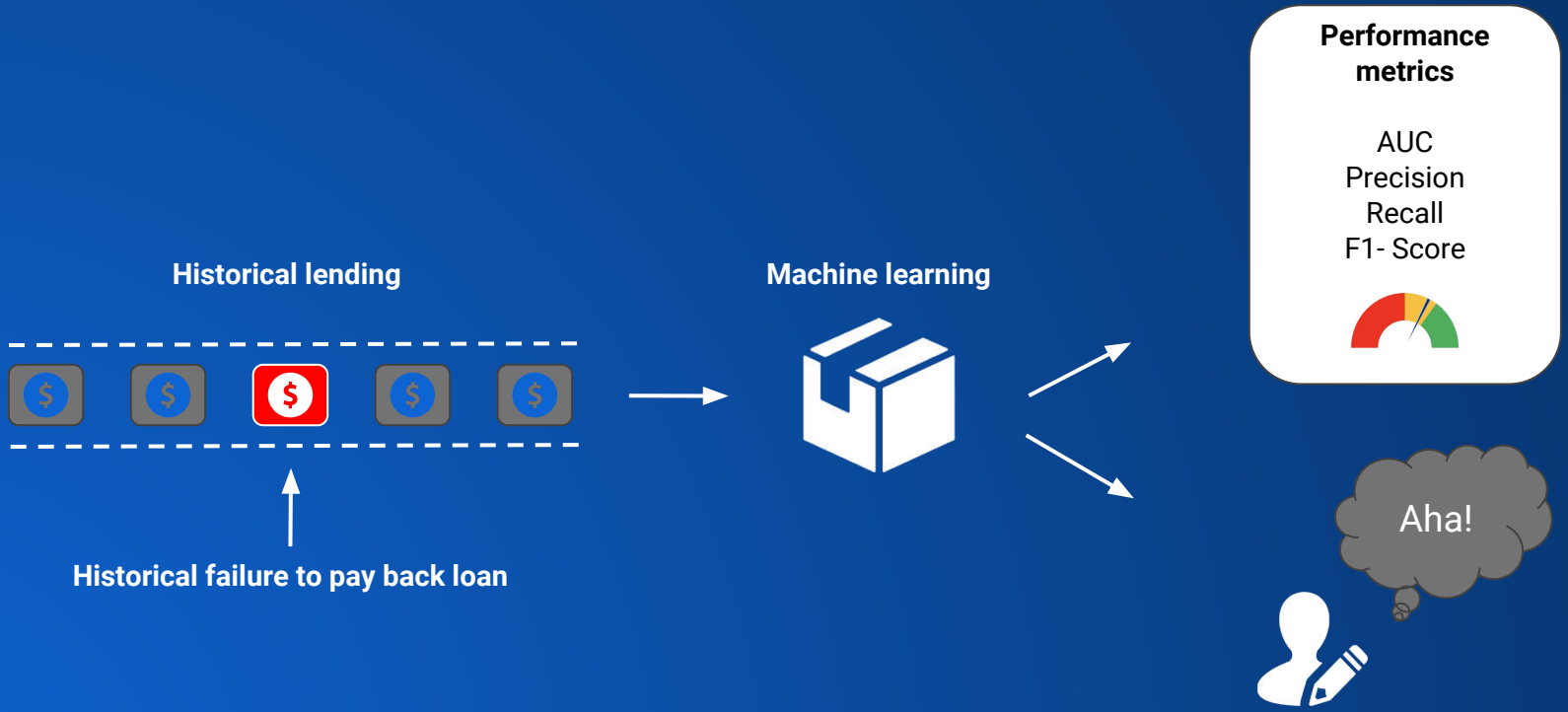
Why? Why not? How?

Fair lending laws [ECOA, FCRA] require credit decisions to be explainable

Primary goal: **model performance**



Additional goal: **human interpretability**



Setting the stage

- Dataset: ~500,000 peer-to-peer loans
- Outcome: “fully paid” vs. “charged off”

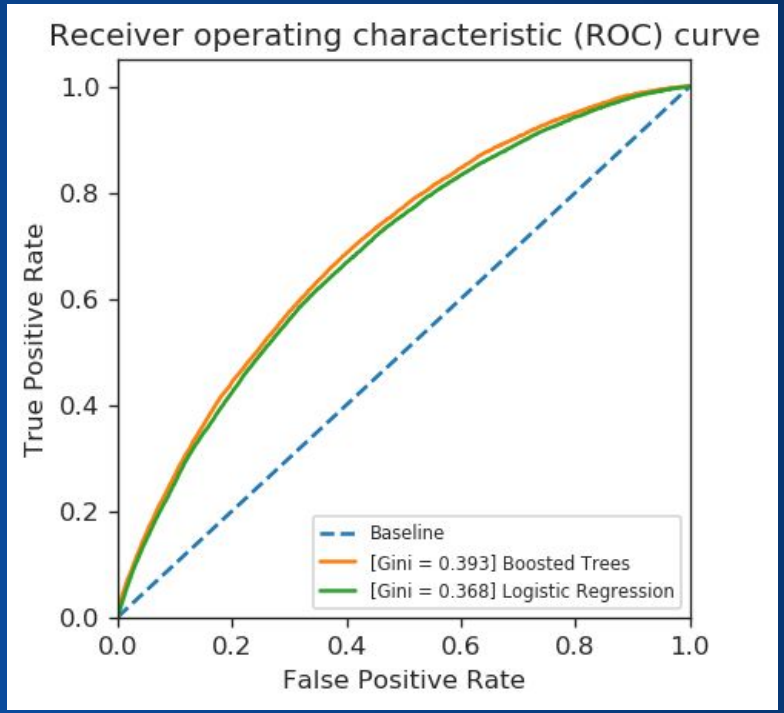
  → 13.5% charge-off rate
- ~50 continuous features (many integers)
- 4 categorical features
- 134 model inputs after one-hot encoding

id	loan_amnt	emp_length	annual_inc	purpose	fico_range_midpoint	loan_status
1817065	7200.0	10+ years	52000.0	home_improvement	682.0	Fully Paid
7701397	10400.0	6 years	96300.0	car	667.0	Charged Off
42622279	7000.0	10+ years	110000.0	credit_card	807.0	Fully Paid
14117919	19650.0	10+ years	56000.0	debt_consolidation	717.0	Fully Paid
38538581	19200.0	2 years	70000.0	debt_consolidation	662.0	Fully Paid
8096004	12000.0	3 years	56825.0	debt_consolidation	667.0	Fully Paid
3114843	33425.0	10+ years	75000.0	credit_card	672.0	Fully Paid
19637779	10000.0	4 years	48000.0	debt_consolidation	662.0	Fully Paid
7688064	15000.0	10+ years	65000.0	debt_consolidation	672.0	Fully Paid
6616918	16000.0	10+ years	42500.0	credit_card	672.0	Fully Paid
36118296	14000.0	7 years	70000.0	debt_consolidation	727.0	Charged Off
11696114	4000.0	2 years	12000.0	credit_card	682.0	Fully Paid
11164629	8400.0	10+ years	70284.0	credit_card	667.0	Charged Off
15279703	11000.0	3 years	30000.0	credit_card	667.0	Fully Paid
3537491	9450.0	10+ years	27500.0	debt_consolidation	687.0	Fully Paid
43589296	16500.0	10+ years	115400.0	debt_consolidation	727.0	Fully Paid

A random sample showing selected fields

Modeling

- Simple model
  - logistic regression
- Complex model
  - gradient boosted trees
- Simplicity-performance tradeoff
  - complex model performs better



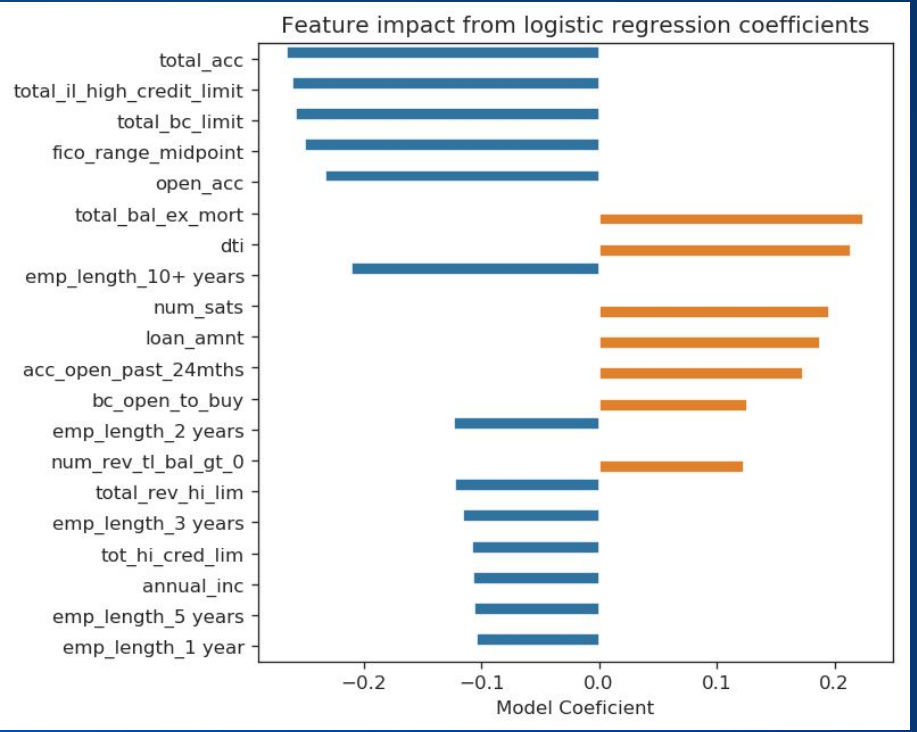
Model performance comparison

Logistic regression: built-in global interpretability

$$log\_odds = w^T x$$

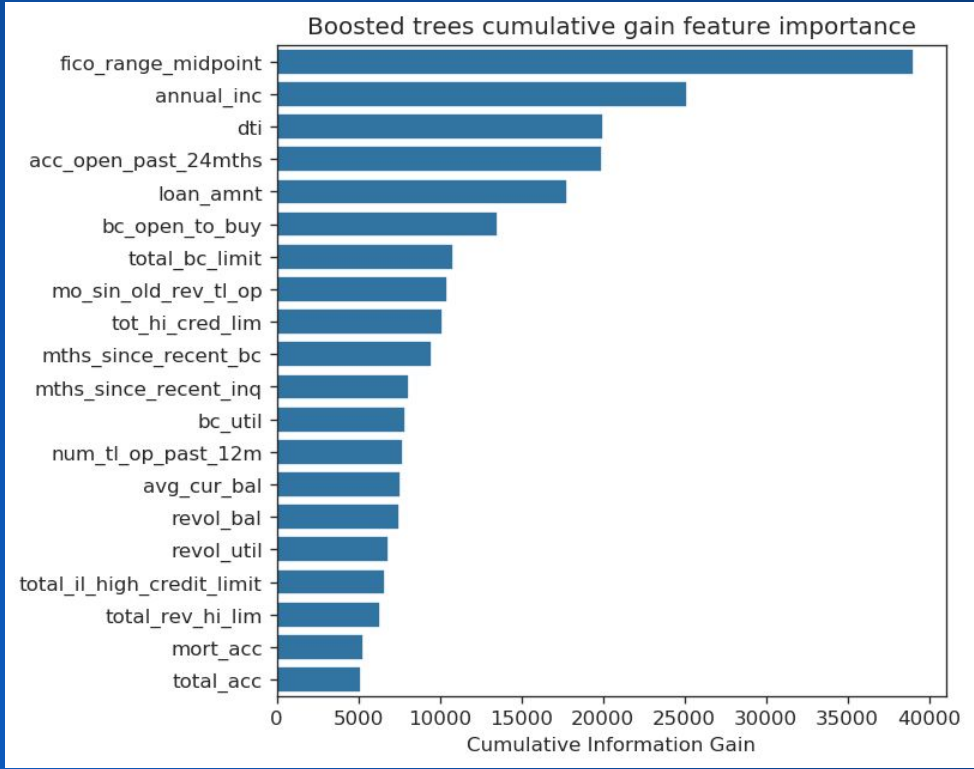
$$y = 1/(1 + \exp(-log\_odds))$$

$$w[i] = ?$$

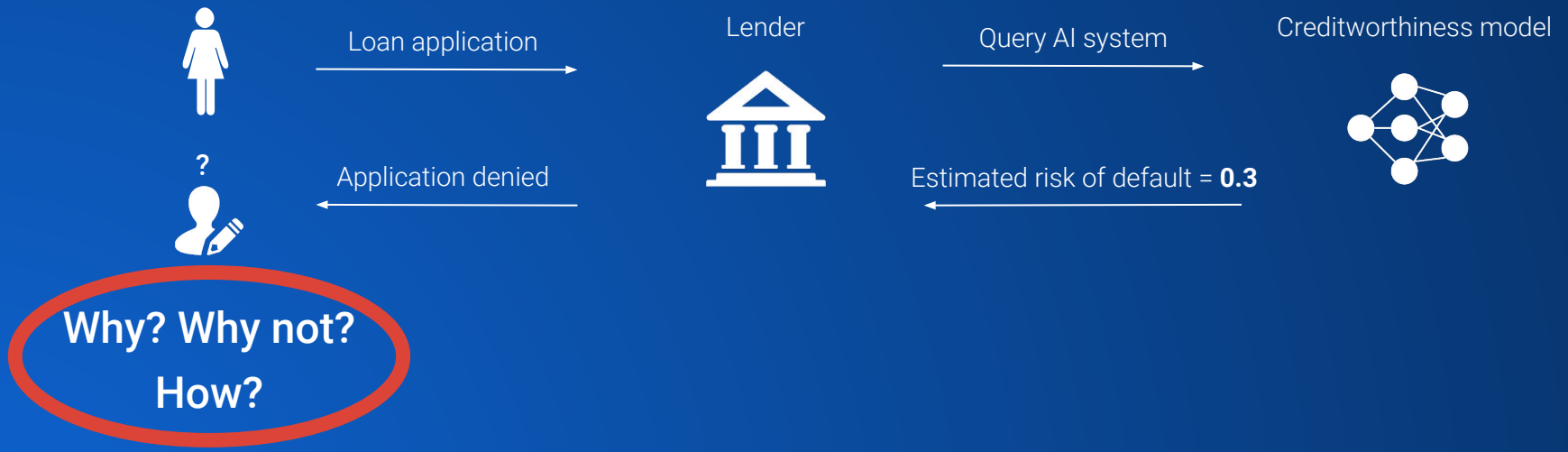


Linear model coefficients

Boosted trees, too, to some extent



Back to the problem: answering questions



Explaining rejections

- Business rule: reject where prediction exceeds 0.2
- Explain rejection = explain why prediction was significantly above average

id	loan_amnt	emp_length	annual_inc	purpose	fico_range_midpoint	Boosted Trees	Logistic Regression	Loan Charged Off
55028284	23150.0	5 years	55135.0	debt_consolidation	672.0	0.298639	0.220457	0
47210357	19000.0	3 years	38000.0	debt_consolidation	667.0	0.293037	0.232950	0
55068614	11525.0	6 years	55000.0	home_improvement	682.0	0.262780	0.397043	0
58663832	29475.0	9 years	62000.0	credit_card	682.0	0.244394	0.259830	0
49833783	6200.0	4 years	45504.0	debt_consolidation	672.0	0.239509	0.286178	1
45304410	10000.0	NaN	36000.0	home_improvement	662.0	0.324559	0.386382	1
55917639	18725.0	NaN	40728.0	credit_card	702.0	0.366653	0.361778	0
51899296	4800.0	5 years	29500.0	credit_card	662.0	0.303968	0.315977	0

A random sample of rejected loans showing selected fields, scores, and outcome

Logistic regression explanation

Explaining the first selected rejection:

- Intercept (aka base log odds) = -2.016
  - base prediction of 0.118
- The specific inputs increase the log odds by 0.753, for a total of -1.263
  - final prediction of 0.220

Top factors	Corresponding raw input	Processed indicator value	Impact on log odds
emp_length_5 years	NaN	3.823595	-0.403022
total_acc	39	1.223887	-0.324769
loan_amnt	23150	1.382164	0.259525
fico_range_midpoint	672	-0.789036	0.197174
num_bc_tl	22	2.811566	0.194095
acc_open_past_24mths	7	0.991607	0.171978
addr_state_AR	NaN	11.497377	0.146673
total_il_high_credit_limit	15000	-0.546113	0.142073
emp_length_10+ years	NaN	-0.673324	0.141421
total_bal_ex_mort	18310	-0.586318	-0.131809

Breakdown of highest-impact factors on model prediction

Top factors	Corresponding raw input	Processed indicator value	Impact on log odds
emp_length_5 years	NaN	3.823595	-0.403022
total_acc	39	1.223887	-0.324769
loan_amnt	23150	1.382164	0.259525
fico_range_midpoint	672	-0.789036	0.197174
num_bc_tl	22	2.811566	0.194095
acc_open_past_24mths	7	0.991607	0.171978
addr_state_AR	NaN	11.497377	0.146673
total_il_high_credit_limit	15000	-0.546113	0.142073
emp_length_10+ years	NaN	-0.673324	0.141421
total_bal_ex_mort	18310	-0.586318	-0.131809

base log odds: -2.016 → base predicted probability: 0.118

total impact: 0.753 → resulting log odds: -1.263 → resulting predicted probability: 0.220

Can we match this for a black box model?

Additive feature attributions

Definitions

- g : explanation model
- z’ : explanation feature vector
- z'_i : i^{th} explanation feature (either 0 or 1)
- Φ_0 : typical prediction
- Φ_i : i^{th} explanation feature’s attributed impact

$$g(z') = \phi_0 + \sum_{i=1}^{M} \phi_i z'_i$$

See Scott Lundberg, et al. A Unified Approach to Interpreting Model Predictions, NeurIPS 2017

Examples of binarizing features for explanation

Structured

- *This feature has not been replaced with its mean value*
- *This feature has not been replaced with a value taken from another example in the dataset*

Unstructured

- *All instances of this word have not been deleted from the text*
- *This super-pixel has not been grayed out in the image*

Black box explanation

Explaining the first selected rejection:

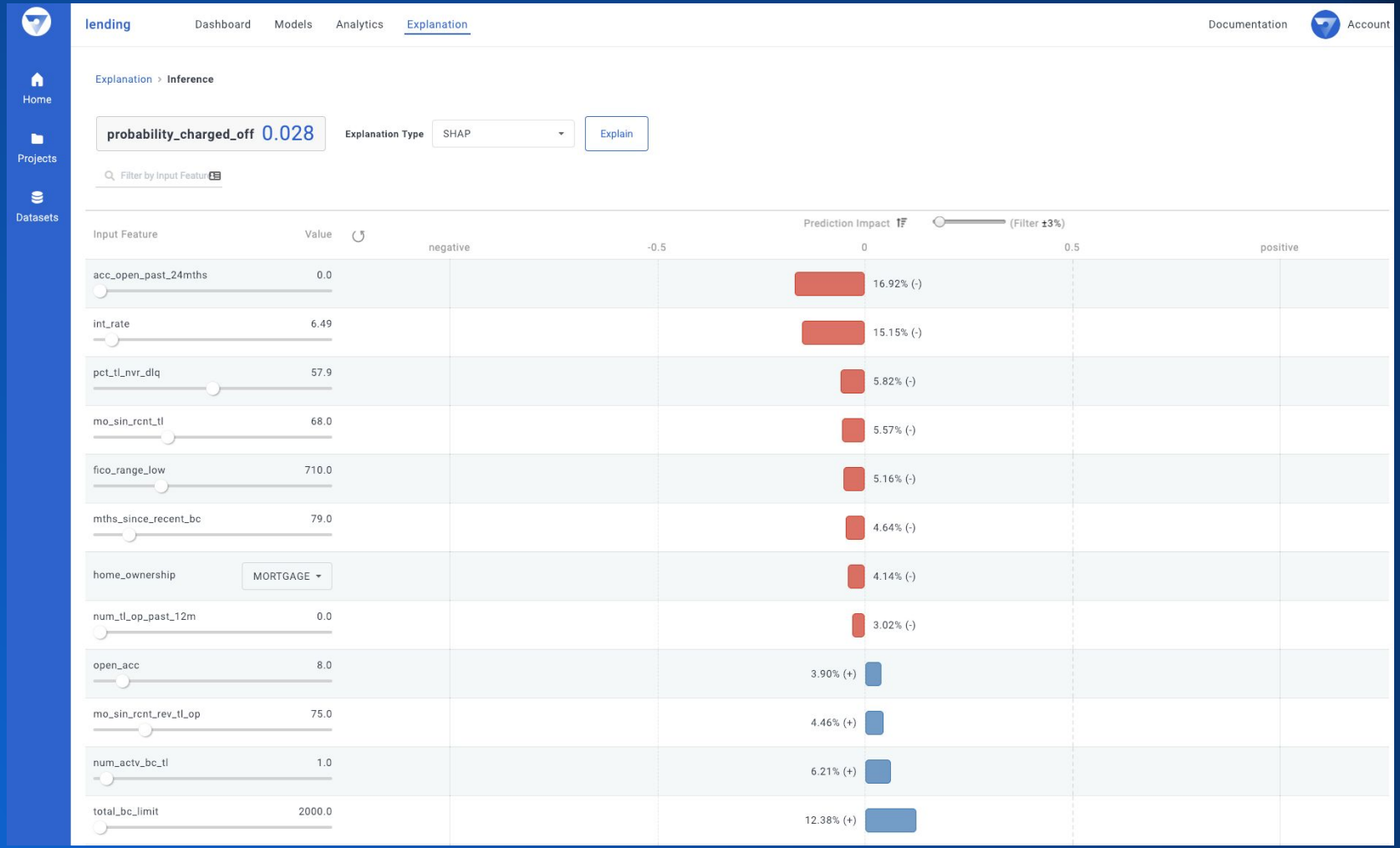
- The average prediction is 0.140
  - base prediction of 0.140
- The difference in inputs between the expected case and this case increases the prediction by 0.159
  - final prediction of 0.299

Top factors	Corresponding raw input	Processed indicator value	Additive feature attribution
loan_amnt	23150	1.382164	0.067819
fico_range_midpoint	672	-0.789036	0.030455
num_bc_tl	22	2.811566	0.024613
total_acc	39	1.223887	-0.016659
acc_open_past_24mths	7	0.991607	0.016309
addr_state_AR	NaN	11.497377	0.011672
total_il_high_credit_limit	15000	-0.546113	0.010720
total_bc_limit	9600	-0.494026	0.010477
home_ownership_RENT	NaN	-0.867843	-0.010442
annual_inc	55135	-0.283952	0.010012

Breakdown of highest-impact factors on model prediction

Top factors	Corresponding raw input	Processed indicator value	Additive feature attribution
loan_amnt	23150	1.382164	0.067819
fico_range_midpoint	672	-0.789036	0.030455
num_bc_tl	22	2.811566	0.024613
total_acc	39	1.223887	-0.016659
acc_open_past_24mths	7	0.991607	0.016309
addr_state_AR	NaN	11.497377	0.011672
total_il_high_credit_limit	15000	-0.546113	0.010720
total_bc_limit	9600	-0.494026	0.010477
home_ownership_RENT	NaN	-0.867843	-0.010442
annual_inc	55135	-0.283952	0.010012

base prediction: 0.140 | given prediction: 0.299


lending
Dashboard
Models
Analytics
Explanation
Documentation
Account
Home
Projects
Datasets
Explanation › Inference
probability_charged_off 0.028
Explanation Type
SHAP
Explain
Filter by Input Featur
Input Feature
Value
Prediction Impact
(Filter ±3%)
negative
-0.5
0
0.5
positive
acc_open_past_24mths 0.0
16.92% (-)
int_rate 6.49
15.15% (-)
pct_tl_nvr_dlq 57.9
5.82% (-)
mo_sin_rcnt_tl 68.0
5.57% (-)
fico_range_low 710.0
5.16% (-)
mths_since_recent_bc 79.0
4.64% (-)
home_ownership MORTGAGE
4.14% (-)
num_tl_op_past_12m 0.0
3.02% (-)
open_acc 8.0
3.90% (+)
mo_sin_rcnt_rev_tl_op 75.0
4.46% (+)
num_actv_bc_tl 1.0
6.21% (+)
total_bc_limit 2000.0
12.38% (+)

Case study 2: model-enhanced medicine

1. Predict probability of death from pneumonia
2. Predict 30-day hospital readmission

(both of these tasks pose significant “correlation is not causation” problems)

Caruana et al., Intelligible Models for HealthCare, KDD 2015

Primary goal: **human intelligibility**

Historical medical data (with outcomes)

Machine learning

Additional goal: **model performance**

This makes sense!

Historical medical data (with outcomes)

Machine learning

Performance metrics

AUC
Precision
Recall
F1- Score

Whitebox approach: Generalized Additive Models

$$y = \sigma(\beta_0 + \sum f_j(x_j))$$

- f_j = risk score of feature x_j
- logistic regression is the special case where f is linear ($f = kx$)

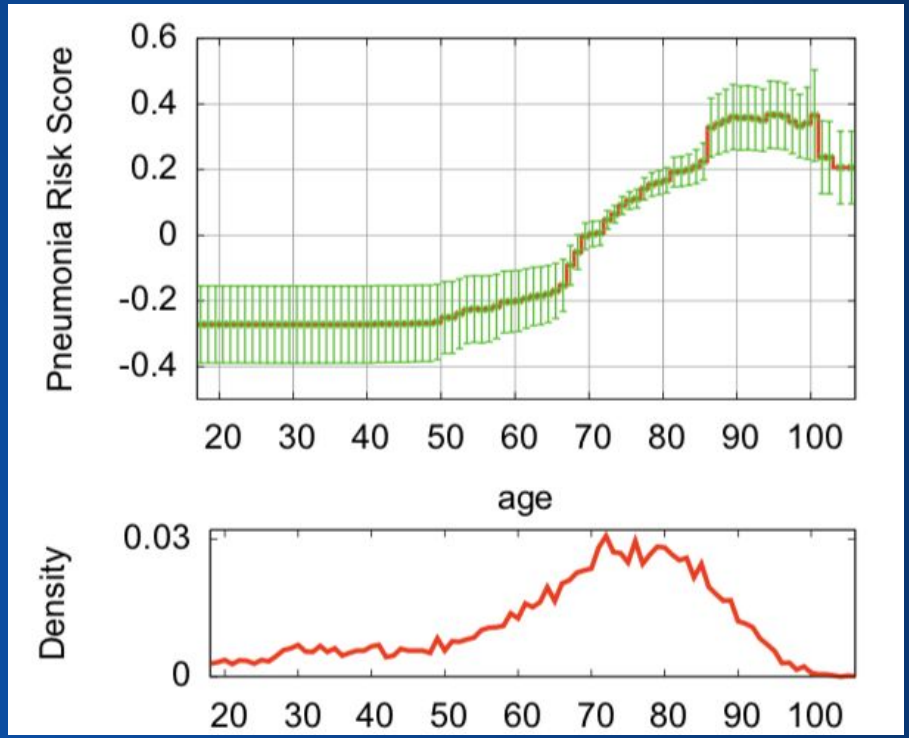


Figure from Caruana et al., Intelligible Models for HealthCare, KDD 2015

Thank You!

luke@fiddler.ai

Appendix 1: Variance of SHAP

Top factors	Corresponding raw input	Processed indicator value	Additive feature attribution
loan_amnt	23150	1.382164	0.067963
fico_range_midpoint	672	-0.789036	0.027676
num_bc_tl	22	2.811566	0.023660
acc_open_past_24mths	7	0.991607	0.019139
total_acc	39	1.223887	-0.014312
addr_state_AR	NaN	11.497377	0.011291
num_bc_sats	7	0.903532	0.011003
home_ownership_RENT	NaN	-0.867843	-0.010929
total_bc_limit	9600	-0.494026	0.010785
total_il_high_credit_limit	15000	-0.546113	0.010684

Top factors	Corresponding raw input	Processed indicator value	Additive feature attribution
loan_amnt	23150	1.382164	0.067819
fico_range_midpoint	672	-0.789036	0.030455
num_bc_tl	22	2.811566	0.024613
total_acc	39	1.223887	-0.016659
acc_open_past_24mths	7	0.991607	0.016309
addr_state_AR	NaN	11.497377	0.011672
total_il_high_credit_limit	15000	-0.546113	0.010720
total_bc_limit	9600	-0.494026	0.010477
home_ownership_RENT	NaN	-0.867843	-0.010442
annual_inc	55135	-0.283952	0.010012