Model Interpretability in ML

Xin Hunt

Agenda

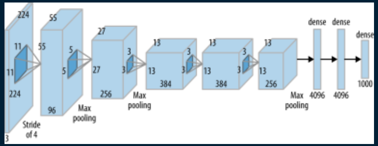
- What is interpretability and why do we need it
- The dimensions of interpretability
  - Pre-, during, and post-model building
  - Model specific vs model agnostic
  - Global vs local
- Common methods
  - Partial dependence
  - Individual conditional expectation
  - LIME
  - Shapley values
- Case studies

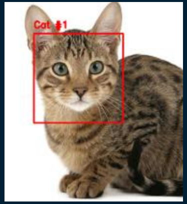
What is interpretability?

- The ability for a human to understand a model's behavior
- Interpretations can be model and context dependent
- Answers a question
  - Why was this individual's loan application rejected?
  - Why is the stock price expected to go down?
  - Does the model make decisions using protected information?

Why do we need interpretability?

- Explosion of data volume and model complexity in machine learning





Large volume of data + High computation power → Complex model → Predictions

Comparison					
Network	Year	Salient Feature	top5 accuracy	Parameters	FLOP
AlexNet	2012	Deeper	84.70%	62M	1.5B
VGGNet	2014	Fixed-size kernels	92.30%	138M	19.6B
Inception	2014	Wider - Parallel kernels	93.30%	6.4M	2B
ResNet-152	2015	Shortcut connections	95.51%	60.3M	11B

https://towardsdatascience.com/the-w3h-of-alexnet-vggnet-resnet-and-inception-7baaaecccc96

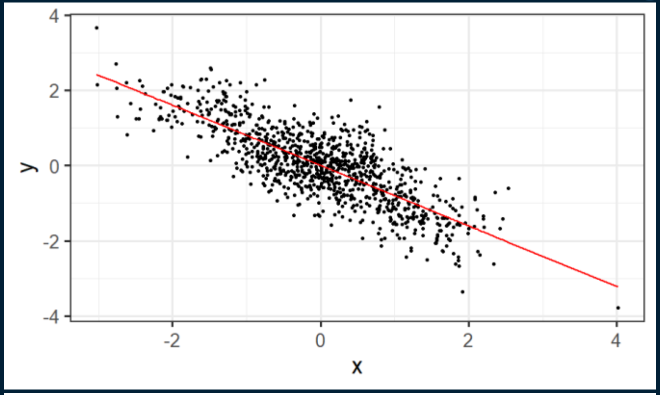
Why do we need interpretability?

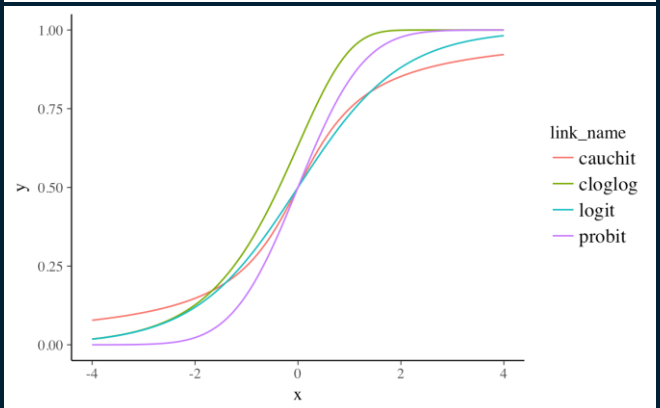
- Fairness / Transparency
  - Understanding a model improves consumer trust
  - Interpretations reveal model behavior on different classes
- Robustness
  - Interpretability methods can reveal overfitting issues and potential modeling errors
- Learning
  - Interpretations can reveal underlying mechanisms and promote human understanding
- Adverse Action notice requirements
  - Equal Credit Opportunity Act (ECOA)
  - Fair Credit Reporting Act (FCRA)

Can't we just use linear regression and decision trees?

Or logistic regression, or rule lists...

- Linear regression is easy to interpret
  - $\hat{y} = \beta_0 + \beta_1 x$
  - "As x increases by one, the expected value of y increases by β_1."
  - Can be generalized with link functions for non-Gaussian distributions and classification tasks
- Feature engineering and generalization can make models harder to interpret



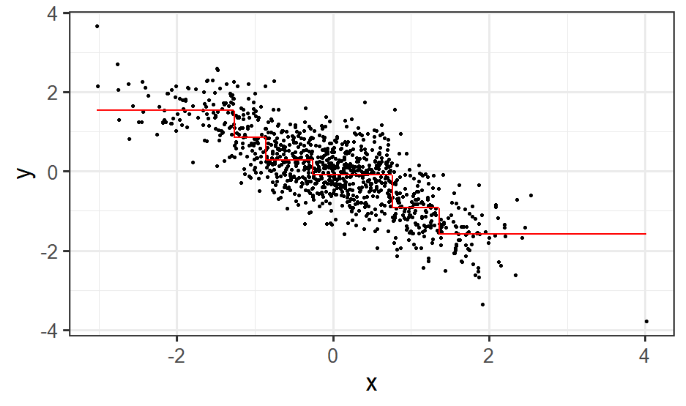


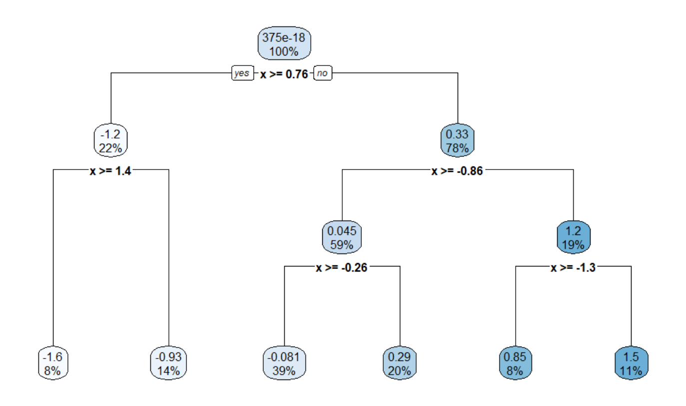
Can’t we just use linear regression and decision trees?

Or logistic regression, or rule lists/sets…

- Small decision trees and rule lists are usually interpretable
  - “If X is greater than __ and less than __ then the expected value of y is…”
  - “If feature __ and __ exist then the prediction of y is…”

- Increased dimensionality and depth quickly make interpretations intractable



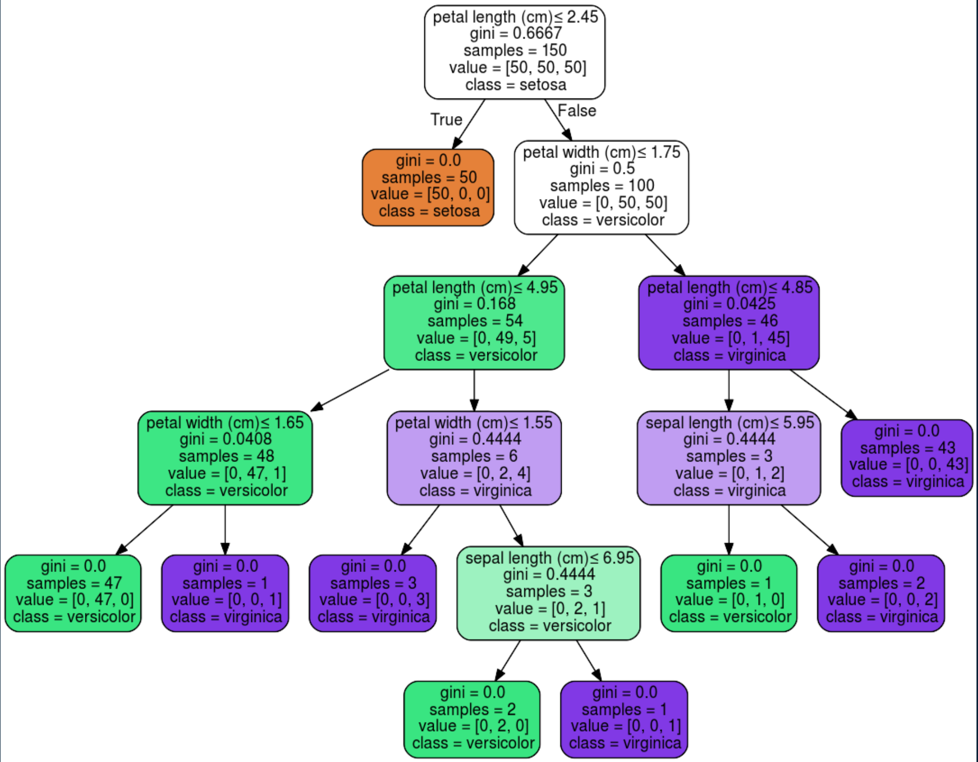


Can't we just use linear regression and decision trees?

Can you explain the overall logic of this tree?

- Trained on the IRIS dataset
- Uses 4 numeric variables to predict the iris species

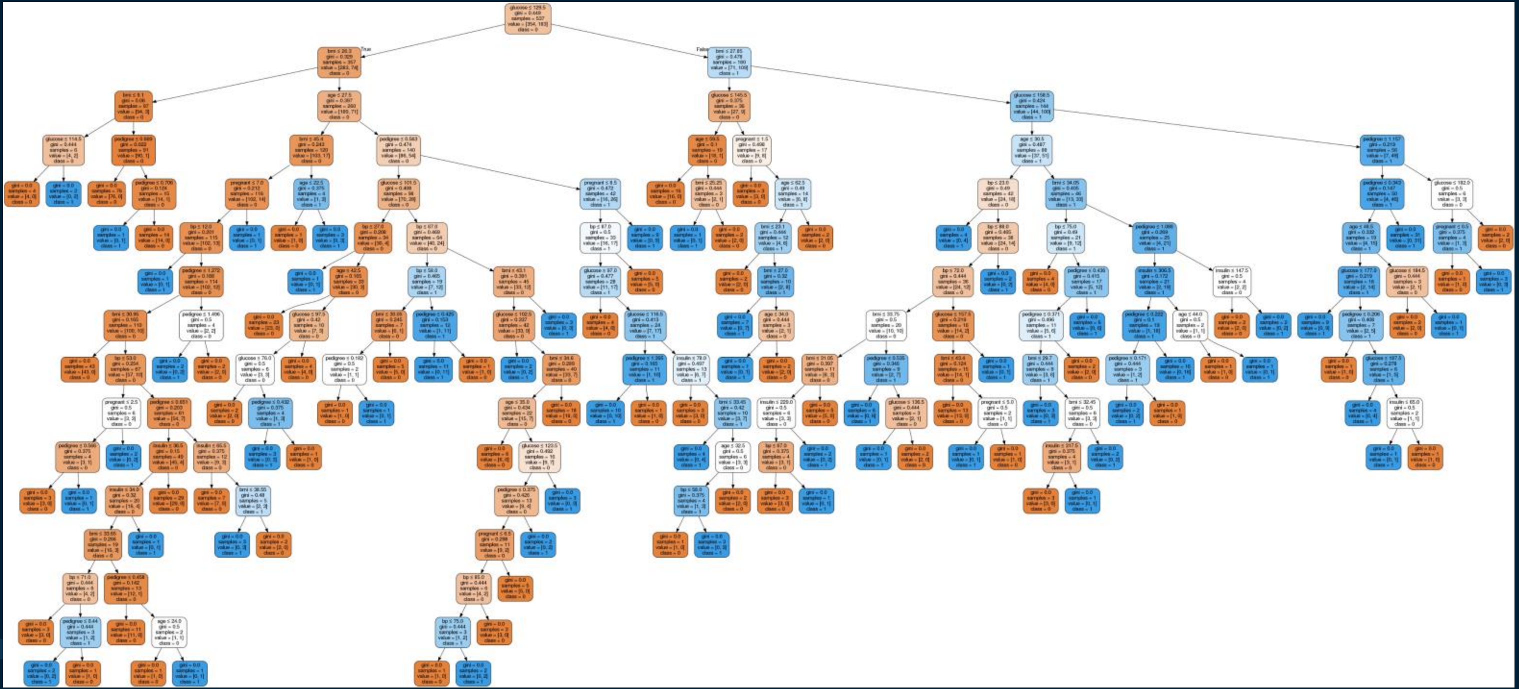
- What is the overall decision logic?
- What is the most important variable?
  - Can you quantify how important it is?
  - Is it the same for every data point/prediction?



https://scikit-learn.org/stable/modules/tree.html

Can’t we just use linear regression and decision trees?

What about this tree? How “interpretable” is it?



[Avinash Navlani, 2018]

Can’t we just use linear regression and decision trees?

Or logistic regression, or rule lists/sets…

- Linear regression, decision trees, and other “open-box” models are great if they fit your need
- Open-box models offer reliable and well-defined explanations for some situations
- Open-box models may not answer all interpretability questions you have by themselves
- High dimensionality, increased model complexity, model ensemble, pre- and post-processing (like feature engineering or data balancing) can add complexity to interpretations

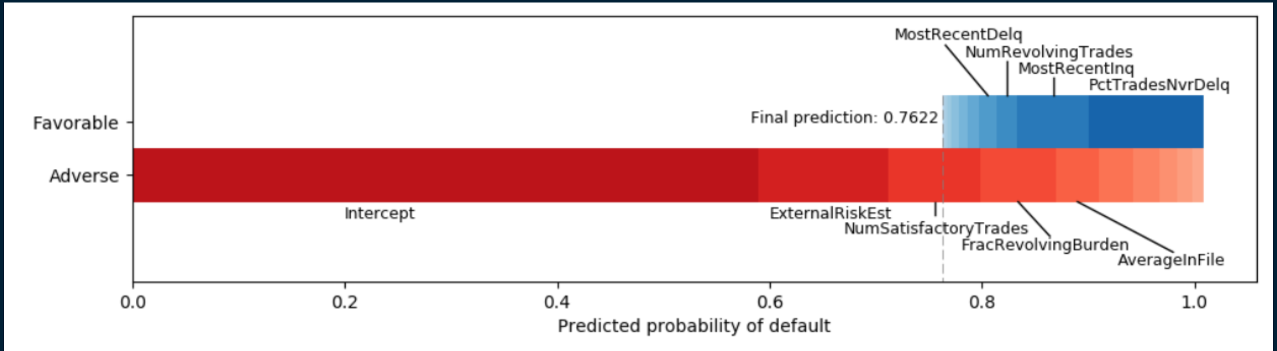
What is interpretability, again?

- The ability for *a human* to understand a model's behavior
  - Interpretability is not about understanding all the details and logic about the model for every data point.
  - Interpretability is about "knowing enough for your downstream tasks." [Been Kim, 2017]
- Common questions answered:
  - What are the most important/impactful features for a model or a decision?
  - What happens when some of the features change values?
  - What features does the model use/not use for predictions?
  - Is there a difference in the decision-making process between groups?

A few use cases

Legal requirement

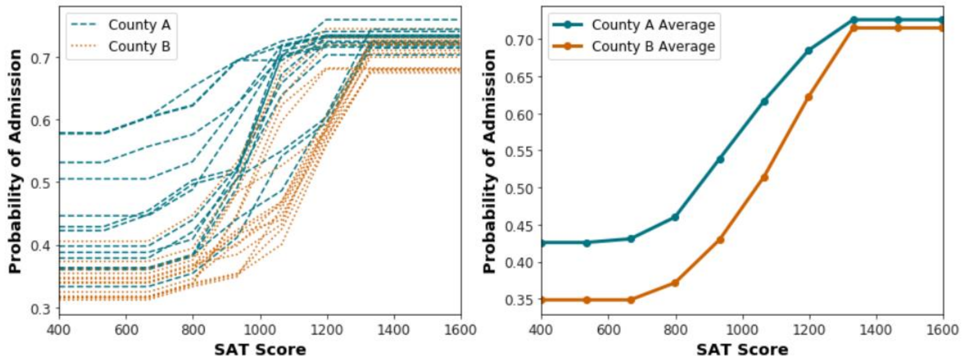
- Adverse Action notice requirements
  - Financial institutes are required by law to give reasons for denying credit
  - Interpretations are needed for each prediction to list the most important features negatively affecting the decision



A few use cases

Fairness and debug models

- Student admission analysis
  - Model predicts admission probability based on student information including features like SAT score, classes taken, extra curriculum activities
  - Model is interpreted by student county, and a discrepancy of admission probability is detected

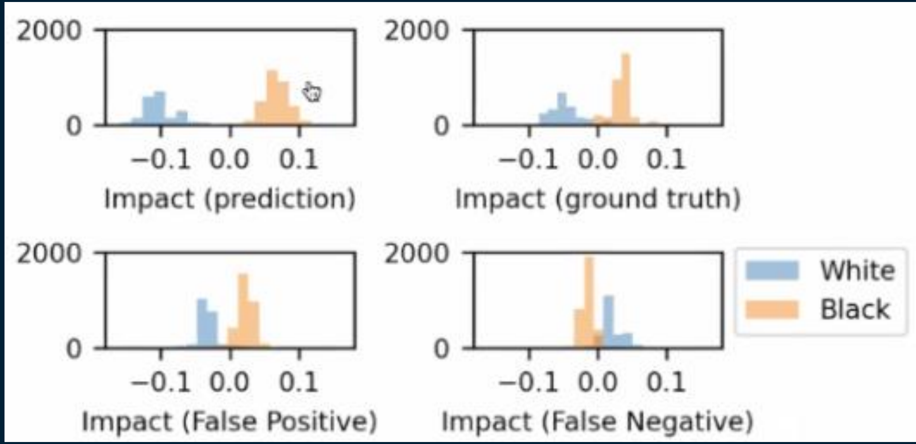


The partial dependence and individual conditional expectation analyses on the SATScore variable by County shows that the expected probability of admission is different in two groups. Given the same SATScore, individuals in County A has a higher expected probability of admission than individuals in County B. This signals potential bias in the model.

A few use cases

Fairness and transparency

- COMPAS recidivism risk score
  - Correctional Offender Management Profiling for Alternative Sanctions
  - Model developed by NorthPointe Inc. to predict reoffend probability
  - Broward County data 2013 and 2014 (ProPublica)
  - Interpretations show that race affects the predicted probability to reoffend



[Art B. Owen, 2021]

A few use cases

Safety

- System failure early warning
  - Model gives early warning for potential failures by detecting salient/abnormal system responses and warehouse conditions
  - Interpretations reveal triggers and causes of the salient point
  - Warning can be dismissed (known causes) or investigated upon

A few use cases

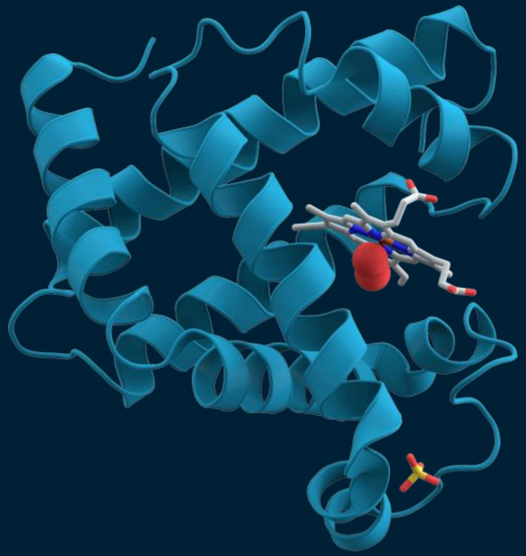
Learning and understanding

- Crop selection
  - Model predicts crop yield based on features like location, weather pattern, seed genealogy and known resistances, planting method
  - Interpretations of predictions reveal what features are most important for success of a specific crop or location

A few use cases

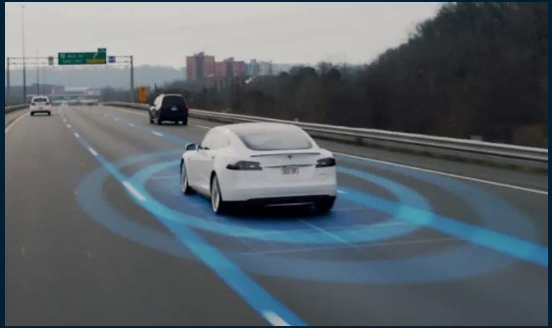
Learning and understanding

- Heat-resistant protein engineering
  - Model predicts the heat resistance of a specific protein configuration based on genetic encoding
  - Interpretations of the model help researchers find specific combination of amino acids or structures contribute to desired heat resistance

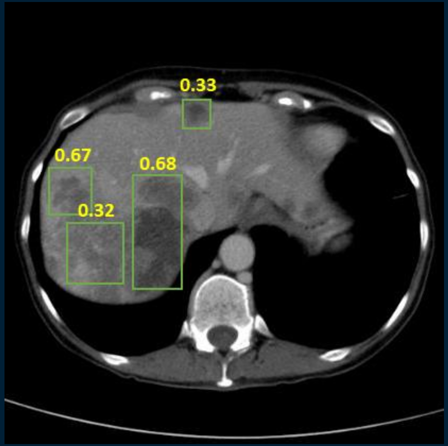


A few use cases

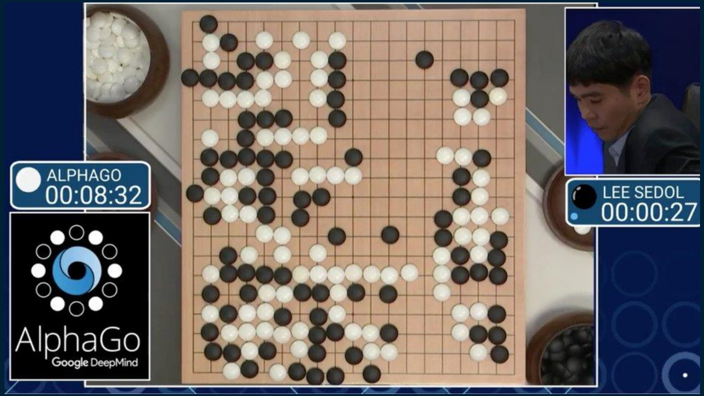
And many more…



Autonomous driving car



Lesion detection



Gaming AI

The dimensions of interpretability

Pre-, during, and post-model building

- Pre-modeling: explore and understand your data
  - Explore data by clustering, visualization, and analyzing correlations
  - Analyze outliers and consider data balancing
  - Choose and construct interpretable features

- During modeling: construct more interpretable models
  - Use open-box models (regressions, rule lists, trees, case-based methods)
  - Add constraints like monotonicity and fairness constraints
  - Encourage sparsity in features (feature selection)

- Post-modeling: interpret model and decisions
  - Use sensitively analysis and feature importance to understand individual decisions and overall trends
  - Build surrogate models to understand the model's local behaviors
  - Construct model-specific explanations to reveal the inner workings of the model

The dimensions of interpretability

Model specific vs model agnostic

- Model-specific methods
  - Designed to explain one class of models
  - Use model-specific information
  - Can provide information unavailable to model-agnostic methods
  - Many deep neural network specific methods

- Model-agnostic methods
  - Work with most ML models
  - Treat models as closed boxes
  - Mostly rely on input-output analysis
  - Good for pipeline building where you want to try out different models

The dimensions of interpretability

Global vs local explanations

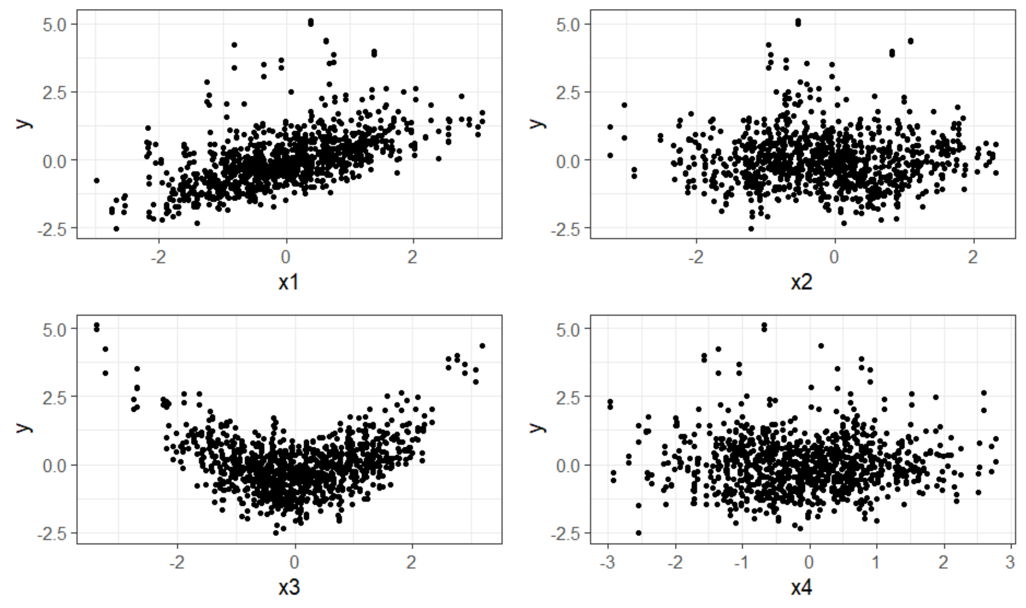
- Global explanations
  - Explain the overall behavior of the model
  - Ex: global feature importance, Partial Dependence

- Local explanations
  - Explain the behavior of the model within a region, or
  - Explain a single prediction/decision
  - Ex: LIME (local surrogate model), Individual Conditional expectation (ICE), Shapley values

Common model-agnostic methods

Background: synthetic data

$$\hat{y} = 4x_1 + 2\sin \pi x_2 + 3x_3^2 + |x_4| + \varepsilon$$

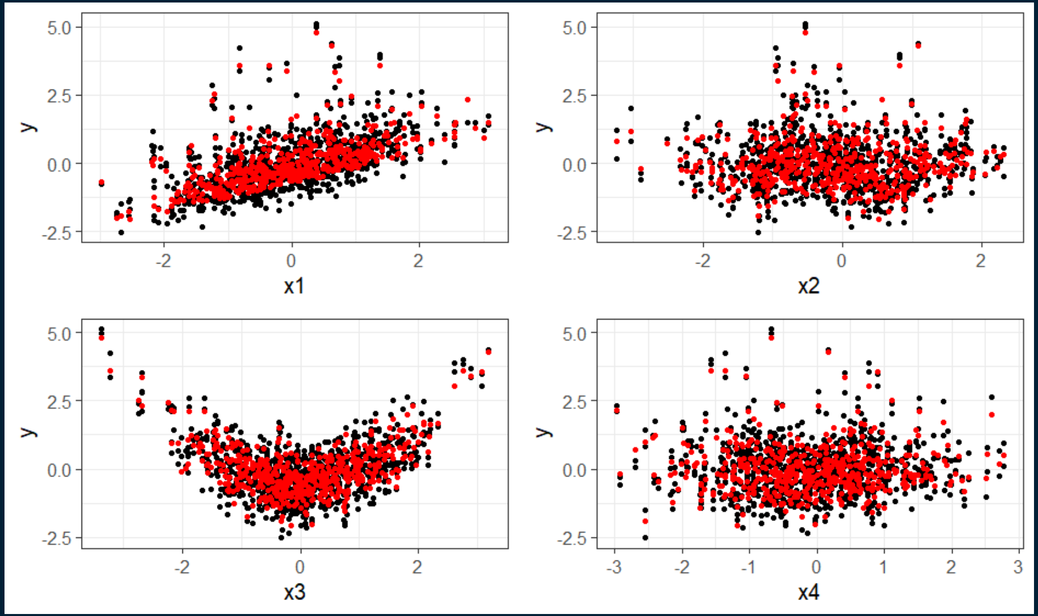


[Ricky Tharrington, 2021]
https://www.youtube.com/watch?v=5ZAms6UaUjk

Common model-agnostic methods

Background: GAM model

```
model=gam(y~x1+s(x2)+s(x3)+s(x4),data=sim_data)
```



Common model-agnostic methods

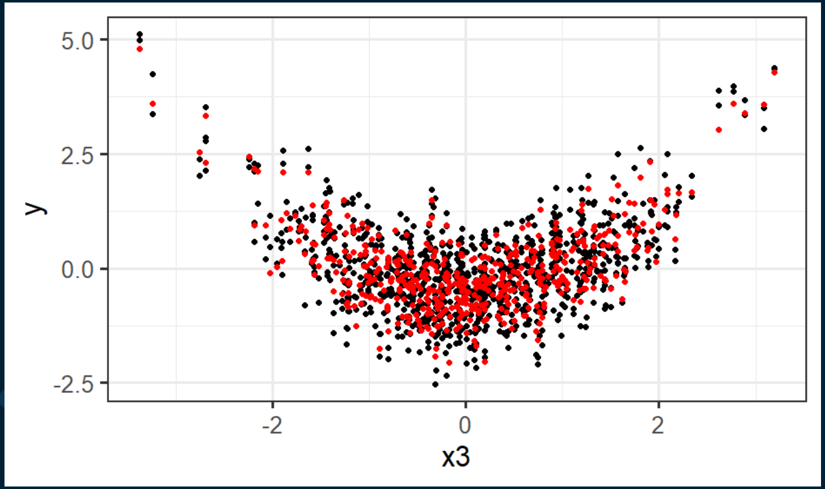
Individual Conditional Expectation (ICE)

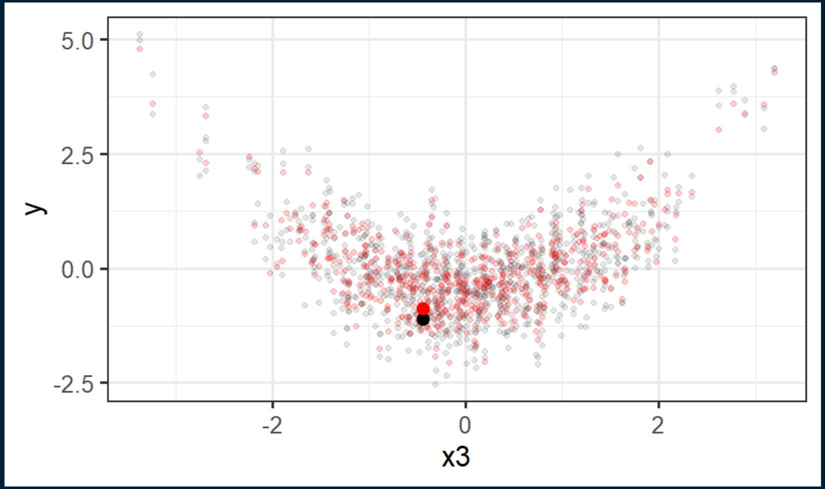
- Evaluates the effect on the model’s prediction of varying a variable’s value in a single observation
- Answers what-if questions like “what happens to my credit score if my credit history is longer?”
- Steps:
  1. Pick a single variable
  2. Pick a single observation
  3. Replicate observation, substitute range of values for variable
  4. Score new observations with model
  5. Plot
  6. Repeat for other observations

Common model-agnostic methods

Individual Conditional Expectation (ICE)

- Pick an observation from the data and a variable of interest

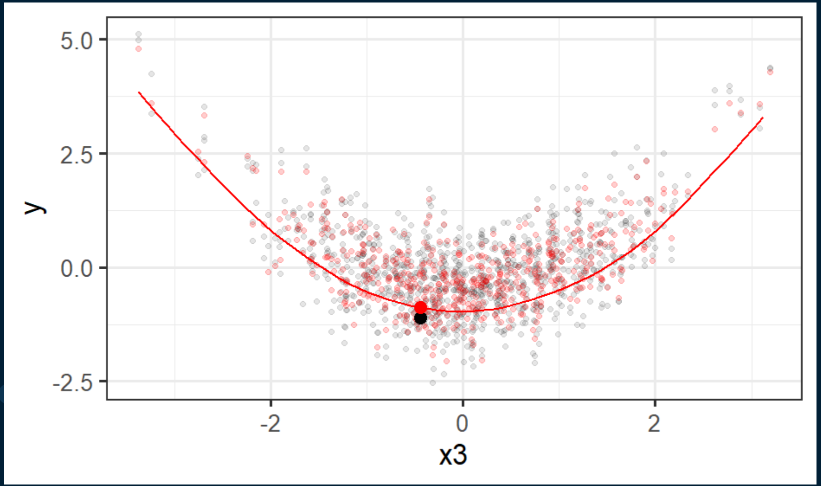


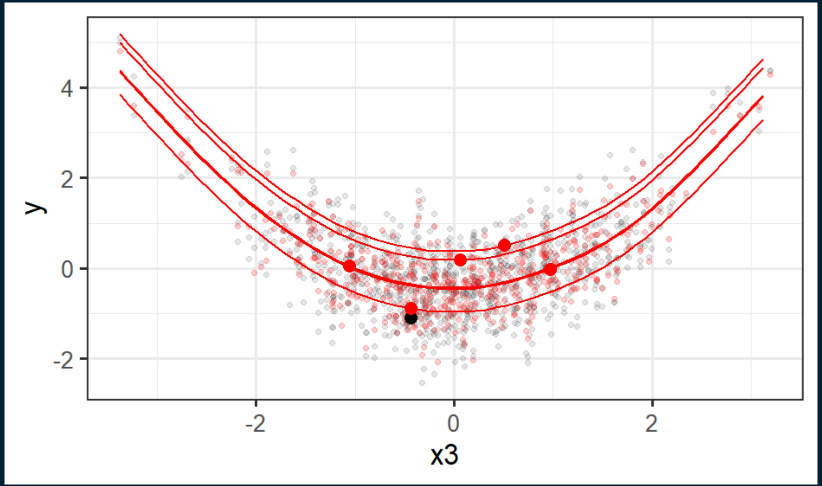


Common model-agnostic methods

Individual Conditional Expectation (ICE)

- Change the value of the variable and plot the predictions
- Repeat for other observations





Common model-agnostic methods

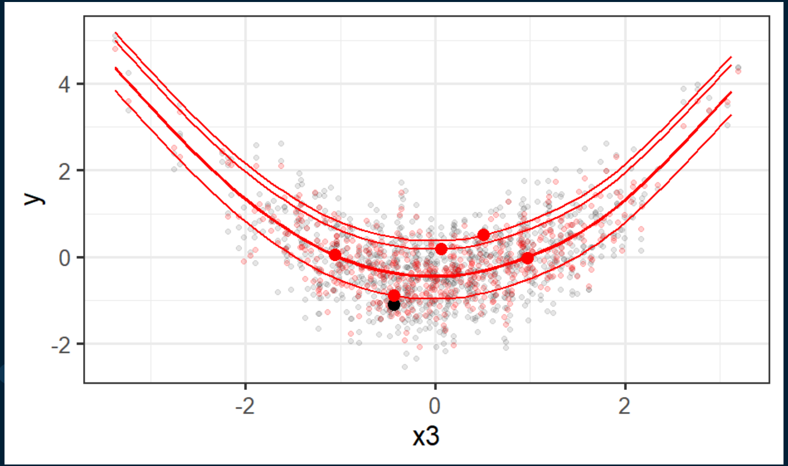
Partial Dependence

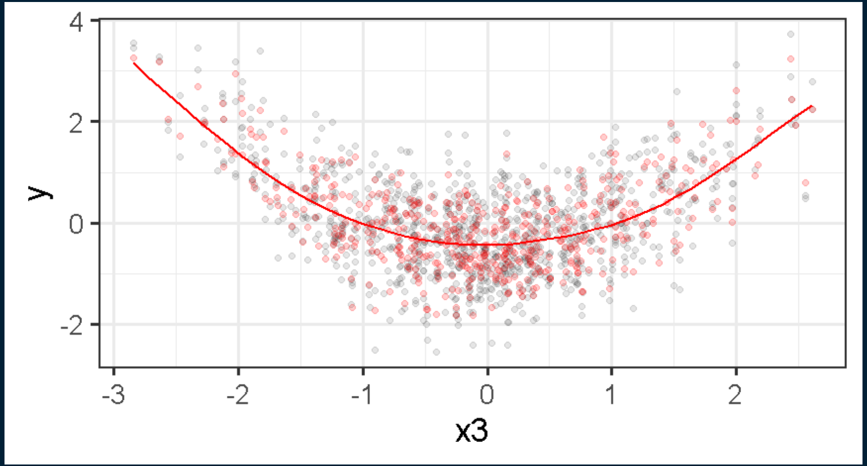
- Evaluates the effect on the model's prediction of varying a variable's value in an entire dataset
- Answers questions on model trends like "how does the model's average prediction change for different credit history lengths?"
- Average of ICE for all observations
- Can also consider interactions by calculating multi-way PD
- If feature correlation is of concern, consider using Accumulated Local Effects (ALE)

Common model-agnostic methods

Partial Dependence

- Compute ICE of all observations
- Average all ICE curves

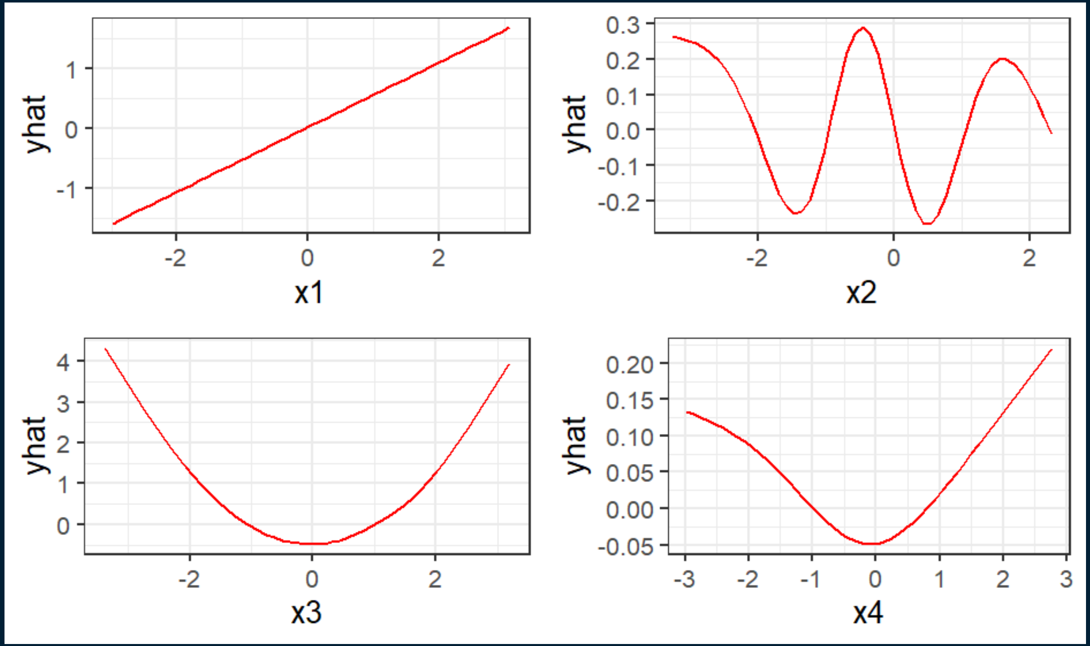




Common model-agnostic methods

Partial Dependence

$$\hat{y} = 4x_1 + 2\sin \pi x_2 + 3x_3^2 + |x_4| + \varepsilon$$



Common model-agnostic methods

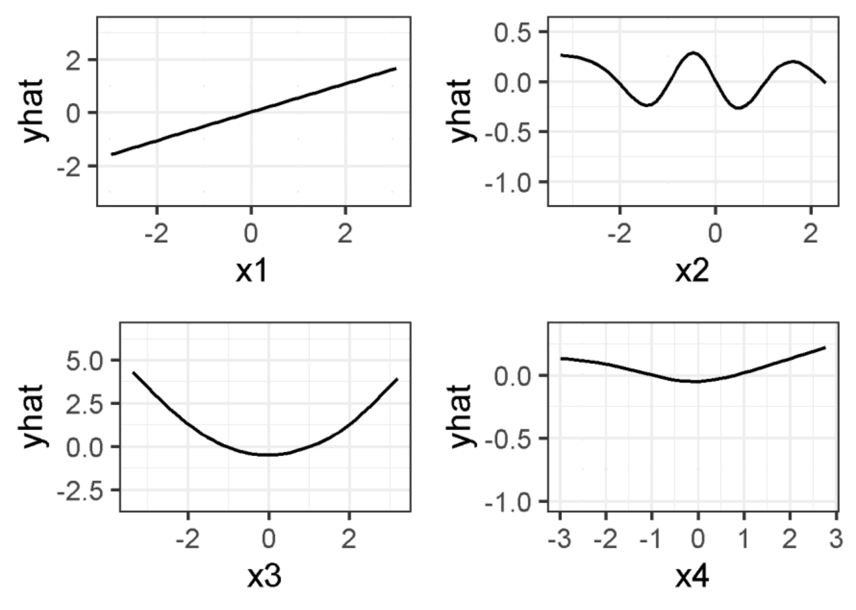
LIME

- Locally Interpretable Model-Agnostic Explanations
- Evaluates the coefficients of a linear model trained on a model's predictions around an individual observation
- Answers local trend questions like "given my credit history, what feature can I change to increase/decrease my credit score the fastest?"
- Steps:
  1. Pick a single observation
  2. Perturb data to generate random observations
  3. Score new observations with model
  4. Weight observations based on their proximity to the observation
  5. Train a linear model on model's predictions
  6. Interpret Model Coefficients

Common model-agnostic methods

LIME - visualization

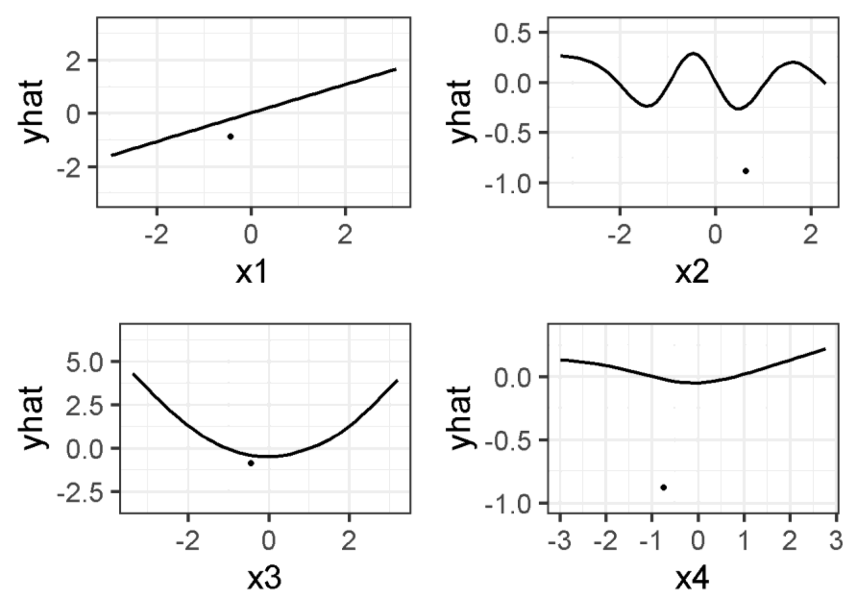
- Partial Dependence of model



Common model-agnostic methods

LIME - visualization

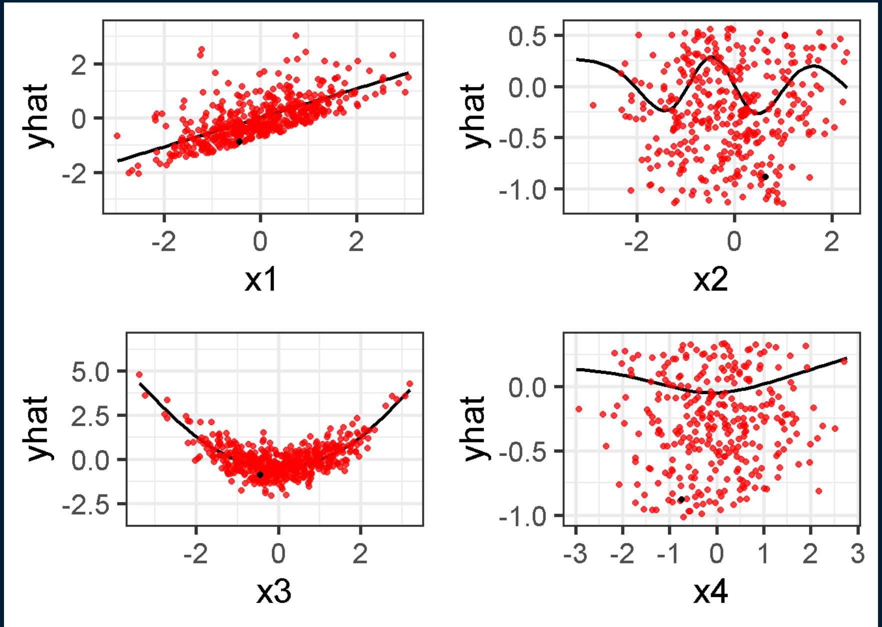
- Select observation



Common model-agnostic methods

LIME - visualization

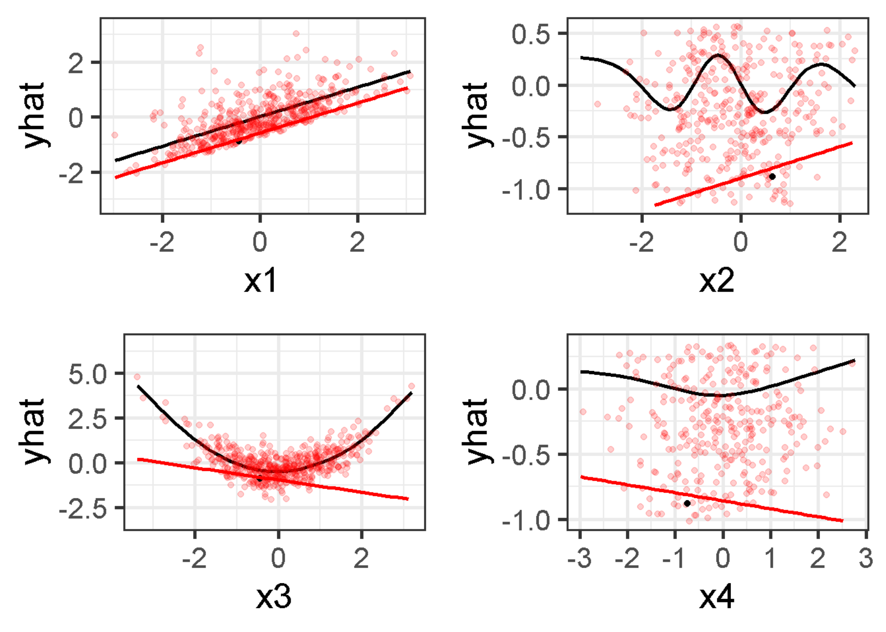
- Generate local samples



Common model-agnostic methods

LIME - visualization

- Fit regression model



Common model-agnostic methods

LIME – summary

- LIME fits *a local linear surrogate model* to data generated around the point of interest

- Linear Model Issues
  - Distance metric needs tuning – what counts as local?
  - How heavy do we weight the "close-by" samples?
  - How many features to select with LASSO?

- Gives local model trends, not individual feature importance
  - For individual feature importance, Shapley values are a better fit

Common model-agnostic methods

Shapley values

- Came from economists for game theory

- Solves the problem of reward distribution among multiple team members

- Solved by Shapley in 1953

Common model-agnostic methods

Shapley values

- Three team members A, B, and C earned a reward of $15k
- How to split the money among the three?

Team	Earning
None	0
A	4k
B	4k
C	4k
A, B	9k
A, C	10k
B, C	11k
A, B, C	15k

Common model-agnostic methods

Shapley values

- How to fairly attribute their contribution?
  - Efficiency: All individual rewards should add up to the total earning
  - Dummy: If including an individual X brings no additional earning in any situation, then X should receive zero reward
  - Symmetry: If including individuals X and Y add the same amount of additional earnings, then X and Y should receive the same reward
  - Additivity: If including one individual X increases the earning by the same amount of two other individuals Y and Z, then X should receive the sum of Y's and Z's reward

Team	Earning
None	0
A	4k
B	4k
C	4k
A, B	9k
A, C	10k
B, C	11k
A, B, C	15k

Common model-agnostic methods

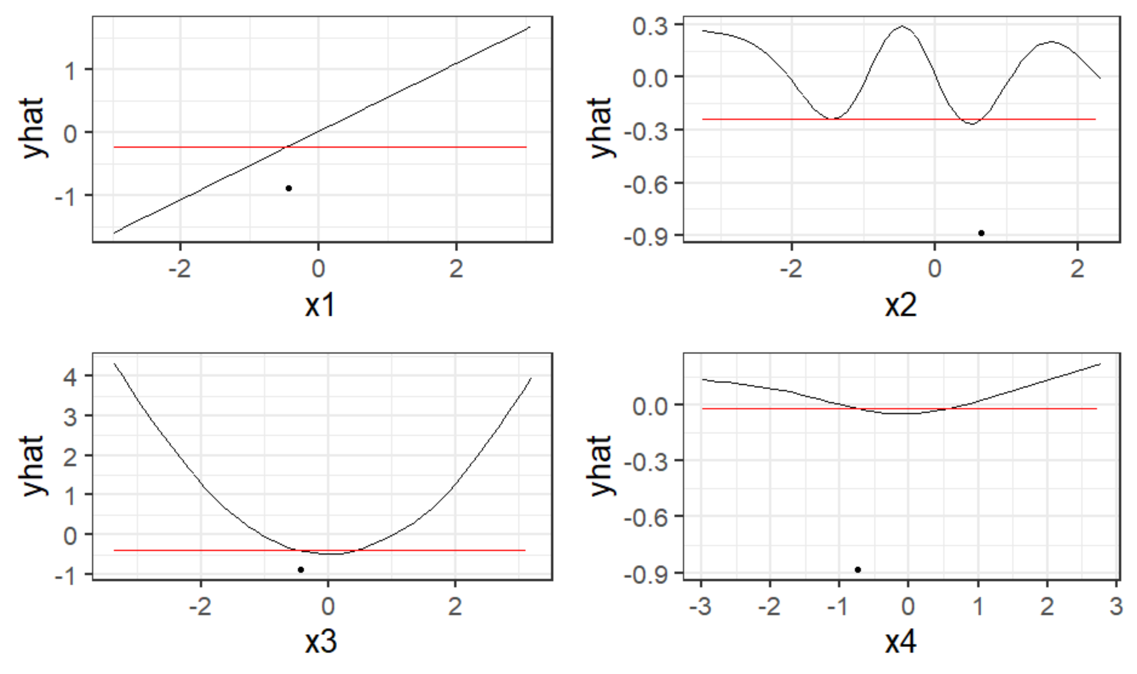
Shapley values

- Unique solution that satisfies all the constraints:
  - Calculates a weighted average of "additional value" the individual brings in by including the individual in the team in all possible scenarios
- Extends to feature importance in machine learning models by evaluating the model's predictions with different combinations of feature values
- Answers individual feature importance questions like "what feature contributed most to my current credit score, and how much did it contribute?"
- Computationally expensive, with approximation methods available (much faster!)

Team	Earning
None	0
A	4k
B	4k
C	4k
A, B	9k
A, C	10k
B, C	11k
A, B, C	15k

Common model-agnostic methods

Shapley values



Feature	Shapley value
x1	-0.18923
x2	-0.24207
x3	-0.45175
x4	-0.02025

Case Studies

See python notebook