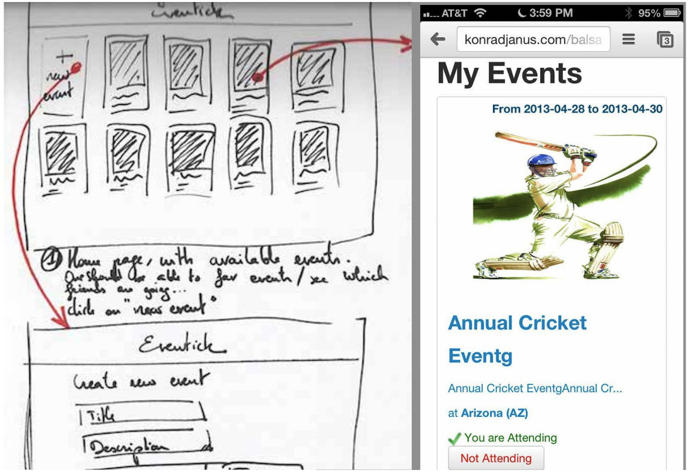
Interpretable Machine Learning

Xiaoyu Liu, Tong Wu

Intro


Eventick
+
new event
① Home page, with available events.
One should be able to fav events / see which friends are going...
click on "new event"
Eventick
Create new event
Title
Description
AT&T
3:59 PM
95%
3

konradjanus.com/balsa

My Events

From 2013-04-28 to 2013-04-30

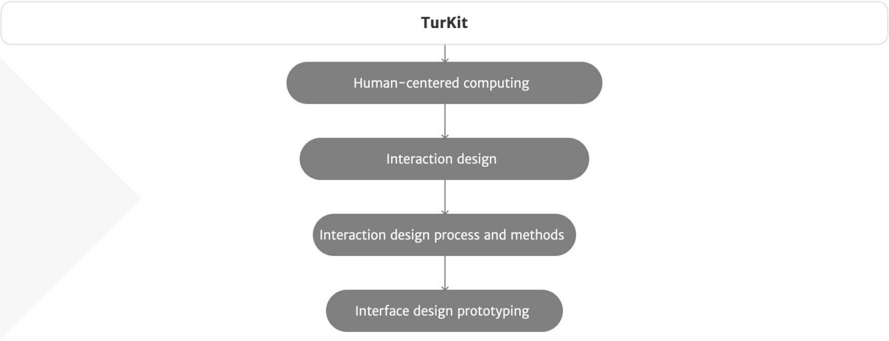
Annual Cricket Eventg

Annual Cricket EventgAnnual Cr...

at Arizona (AZ)

✓ You are Attending

Not Attending



TurKit

Human-centered computing

Interaction design

Interaction design process and methods

Interface design prototyping

Discussion

“Explainable” is a vague word. What feathers should a good explanation possess? Think about how wikipedia try explaining a definition to you.

We have a model that can do image recognition and it works pretty well on our sample. The predictions reaches 94% accuracy. Can you come out a scenario that the model cannot be trusted?

Some possible issues

Work well for samples but terrible in practice

Work well in most cases but have serious problem with some cases

How a model tries to gain trust

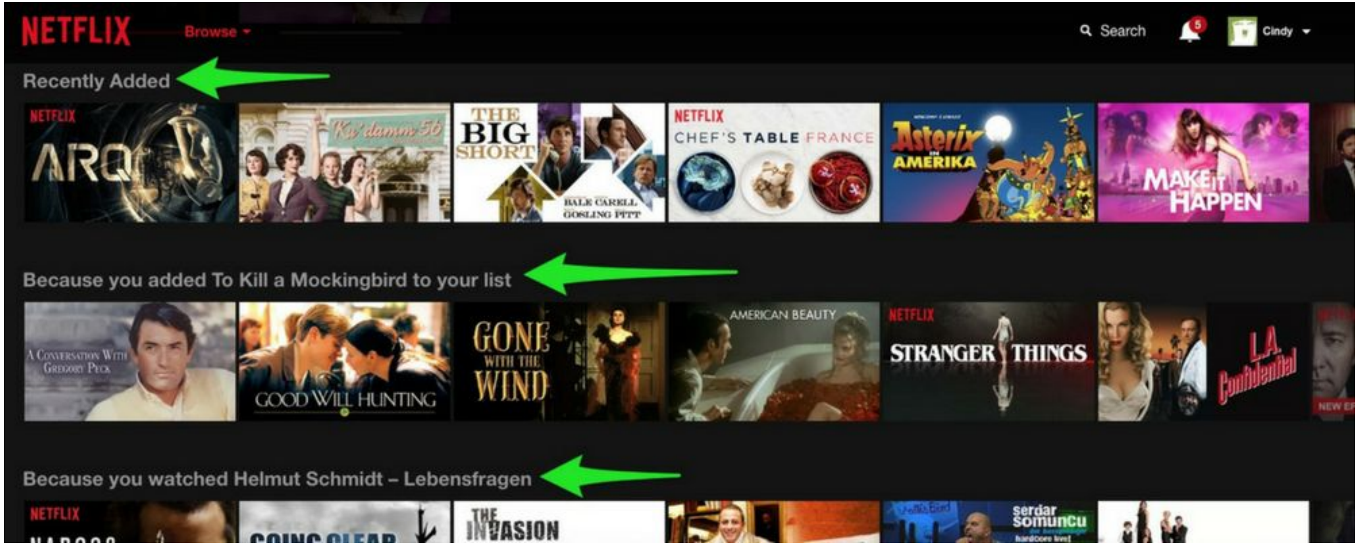
Interpretable

Accuracy

A/B test

Voodoo

Netflix recommendation



Recently Added

Because you added To Kill a Mockingbird to your list

Because you watched Helmut Schmidt – Lebensfragen

Some potential problems

Interpretable:	more accurate, less interpretable. Ex: decision trees
Accuracy:	data leaking, training data vs real world, changing environment, objective mismatch
A/B test:	expensive, potential problems
Voodoo:	hahaha

LIME(Local Interpretable Model-Agnostic Explanations)

Pick a model class interpretable by humans, use it to approximate unknown models

Good **local** approximation

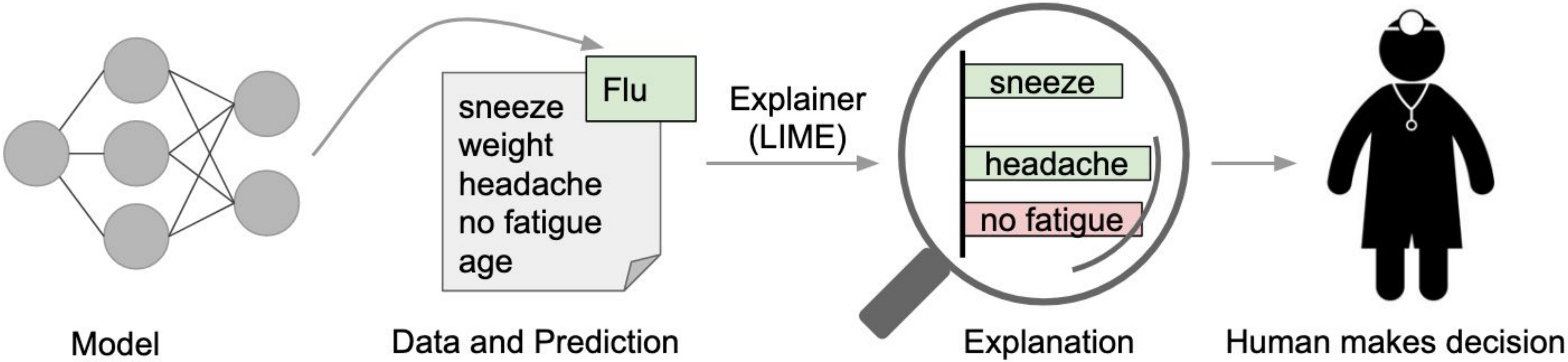
More details of LIME

Interpretable: humans can easily interpret reasoning

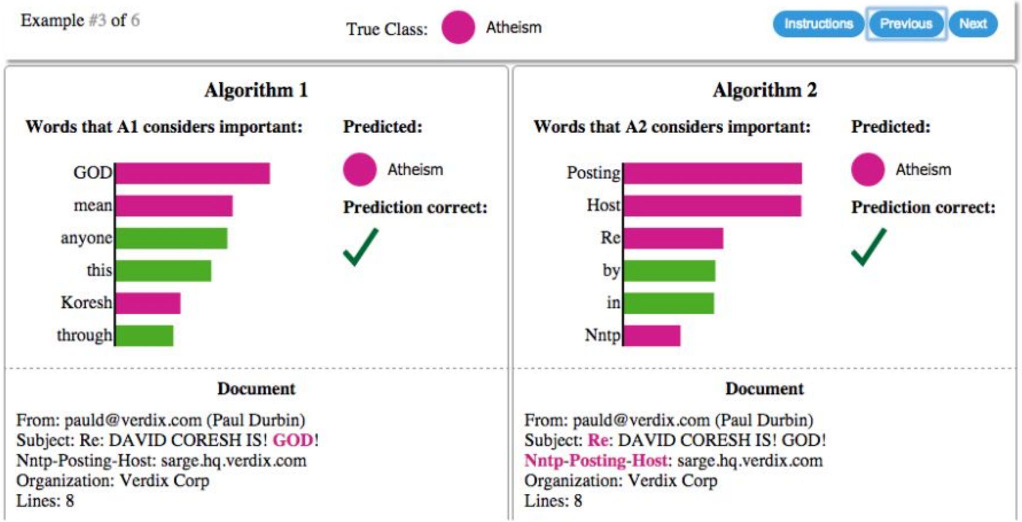
Faithful: describes how the model actually behaves

Model-agnostic: can be sued for any ML model

LIME example for medical diagnosis



Christianity or Atheism



Example #3 of 6

True Class: ● Atheism

Algorithm 1

Words that A1 considers important:

Predicted: ● Atheism

Prediction correct: ✓

Document

From: pauld@verdix.com (Paul Durbin)
Subject: Re: DAVID CORESH IS! GOD!
Nntp-Posting-Host: sarge.hq.verdix.com
Organization: Verdix Corp
Lines: 8

Algorithm 2

Words that A2 considers important:

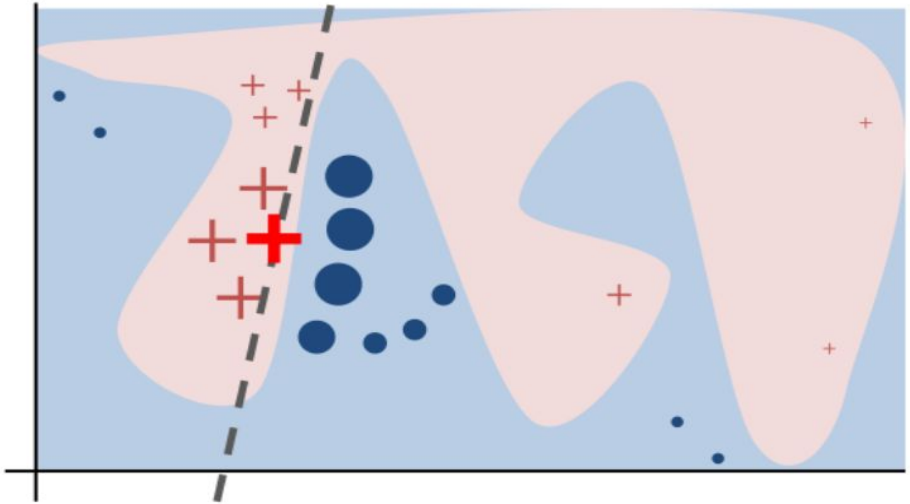
Predicted: ● Atheism

Prediction correct: ✓

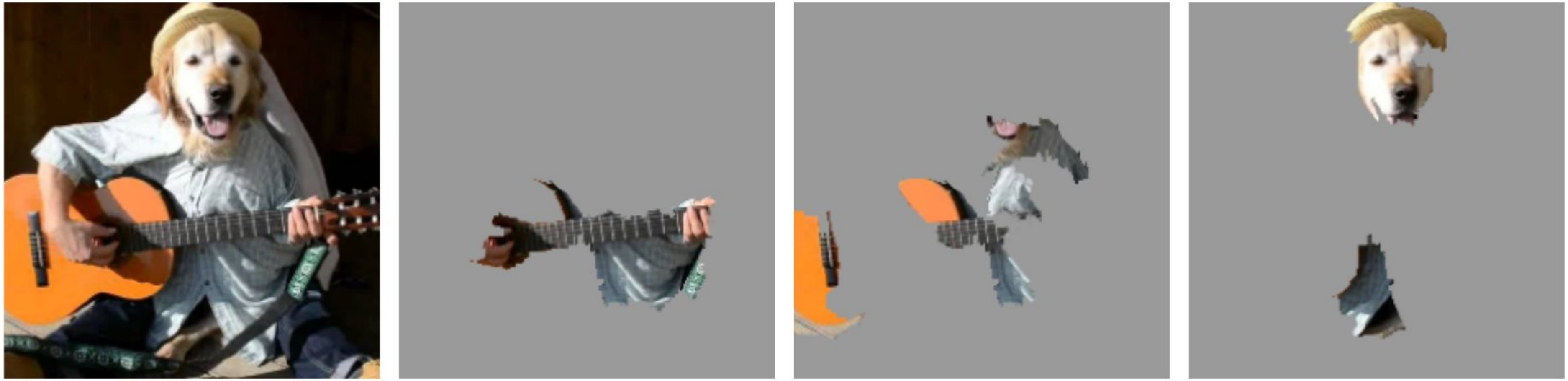
Document

From: pauld@verdix.com (Paul Durbin)
Subject: Re: DAVID CORESH IS! GOD!
Nntp-Posting-Host: sarge.hq.verdix.com
Organization: Verdix Corp
Lines: 8

How does LIME explain?



Labrador or electric guitar



(a) Original Image (b) Explaining *Electric guitar* (c) Explaining *Acoustic guitar* (d) Explaining *Labrador*

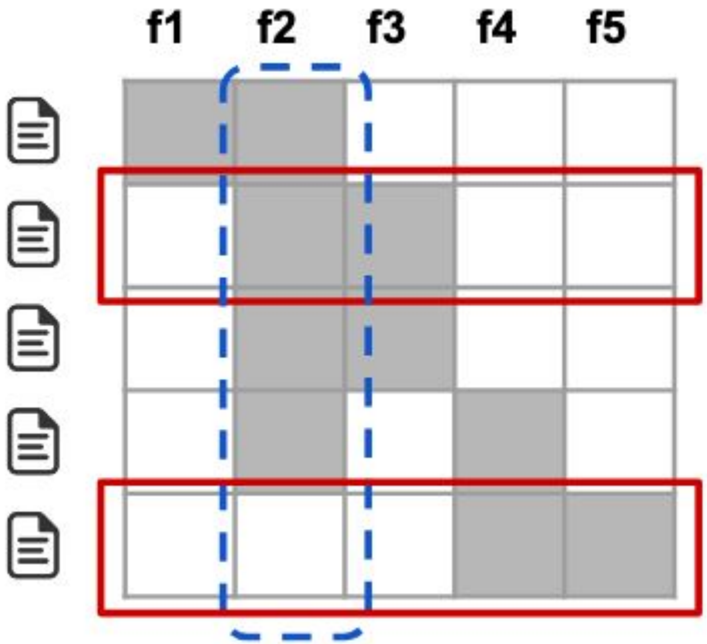
Conclusions of LIME

LIME is trying to explain the model to you by a randomly selected sample and its predictions produced by model itself.

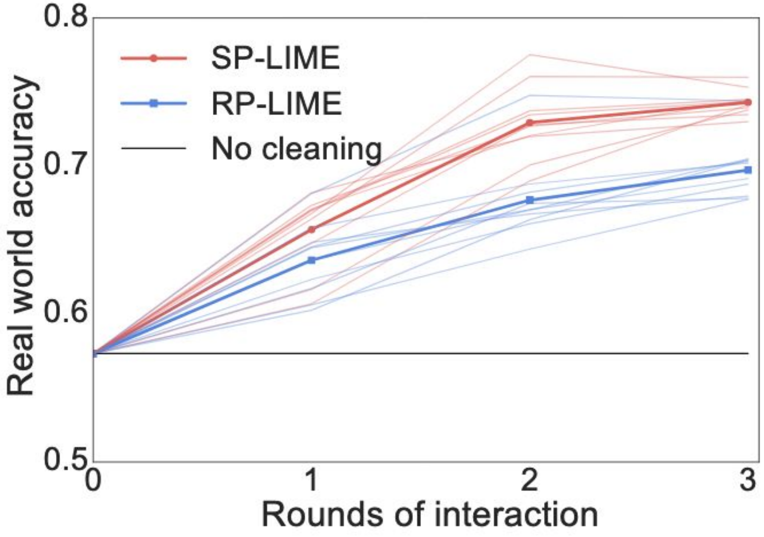
Advantage: Less prior knowledge of background needs to understand the model

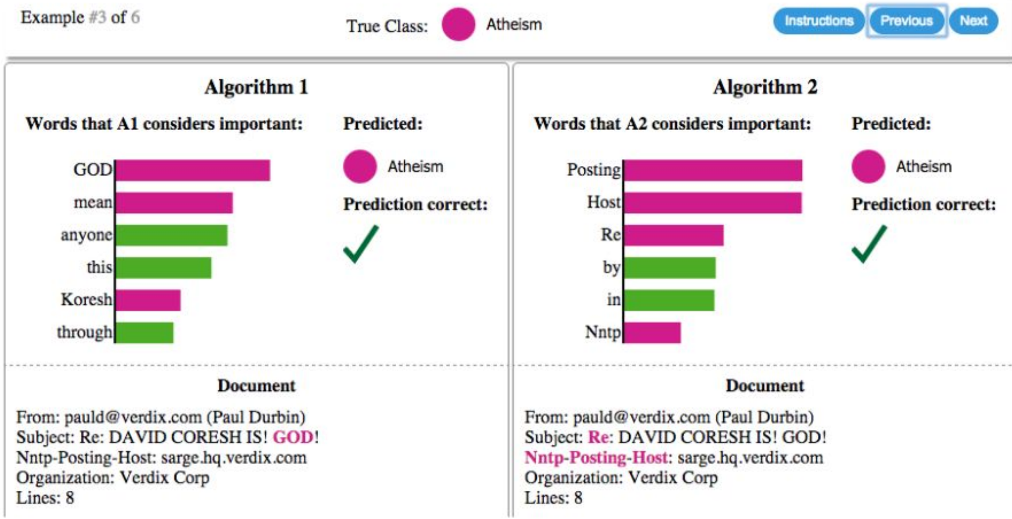
Can we do better?

Submodular pick for explaining models



Feature engineering





Example #3 of 6

True Class: ● Atheism

Algorithm 1

Words that A1 considers important:

Predicted: ● Atheism

Prediction correct: ✓

Document

From: pauld@verdix.com (Paul Durbin)
Subject: Re: DAVID CORESH IS! GOD!
Nntp-Posting-Host: sarge.hq.verdix.com
Organization: Verdix Corp
Lines: 8

Algorithm 2

Words that A2 considers important:

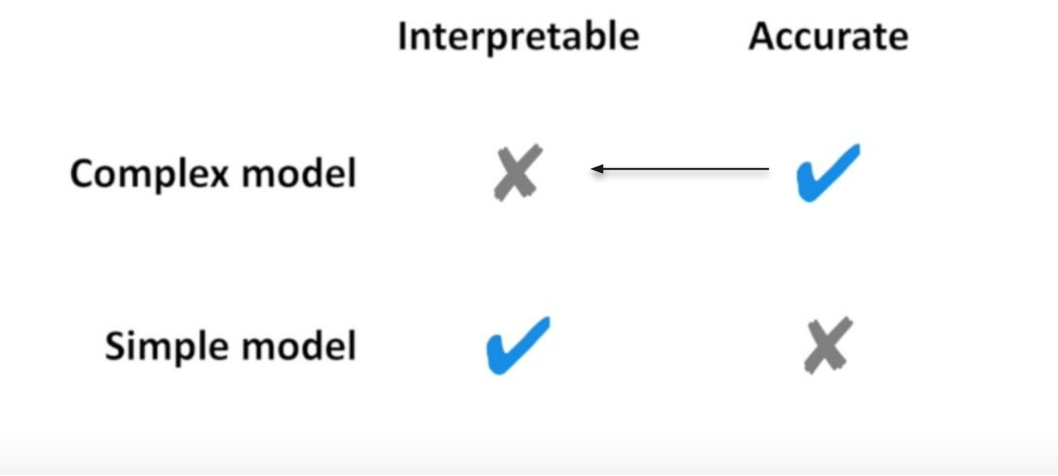
Predicted: ● Atheism

Prediction correct: ✓

Document

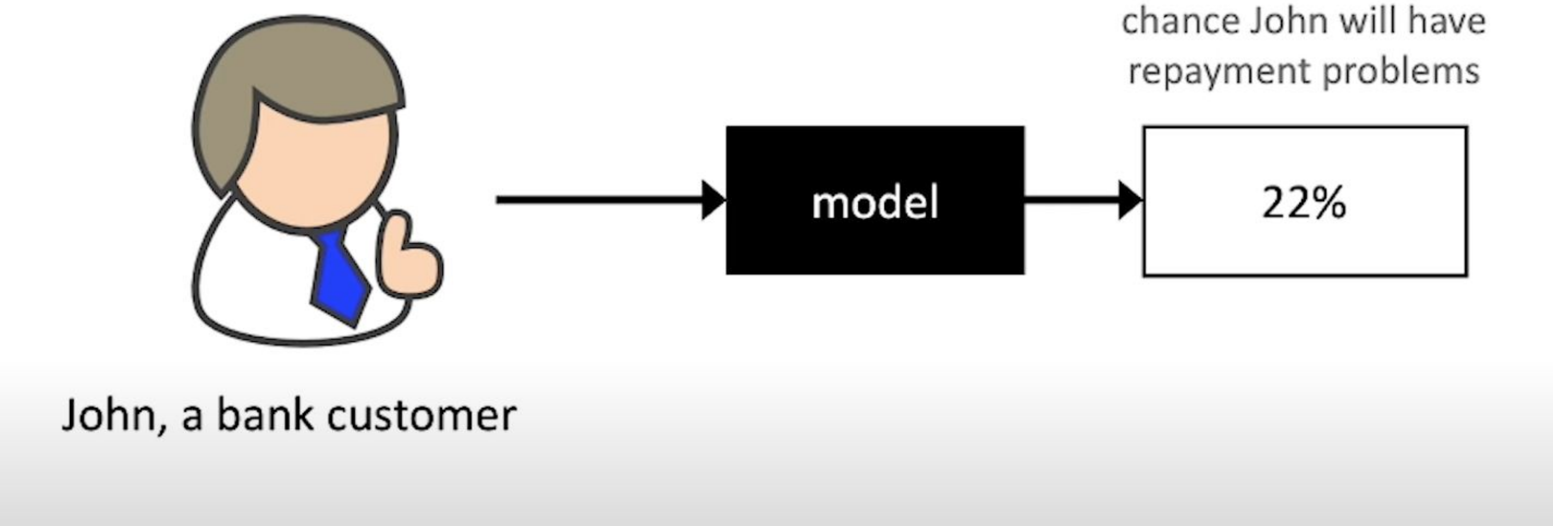
From: pauld@verdix.com (Paul Durbin)
Subject: Re: DAVID CORESH IS! GOD!
Nntp-Posting-Host: sarge.hq.verdix.com
Organization: Verdix Corp
Lines: 8

A Unified Approach to Interpreting Model Predictions



	Interpretable	Accurate
Complex model	✗	✓
Simple model	✓	✗

SHAP (SHapley Additive exPlanation) Values



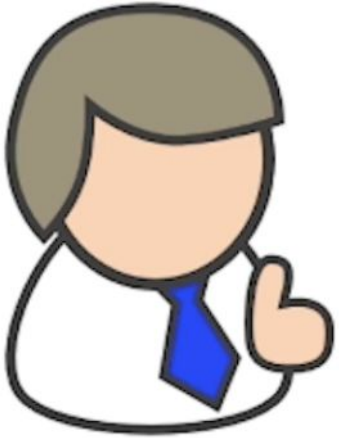
Additive feature attribution methods

Additive feature attribution methods have an explanation model that is a linear function of binary variables:

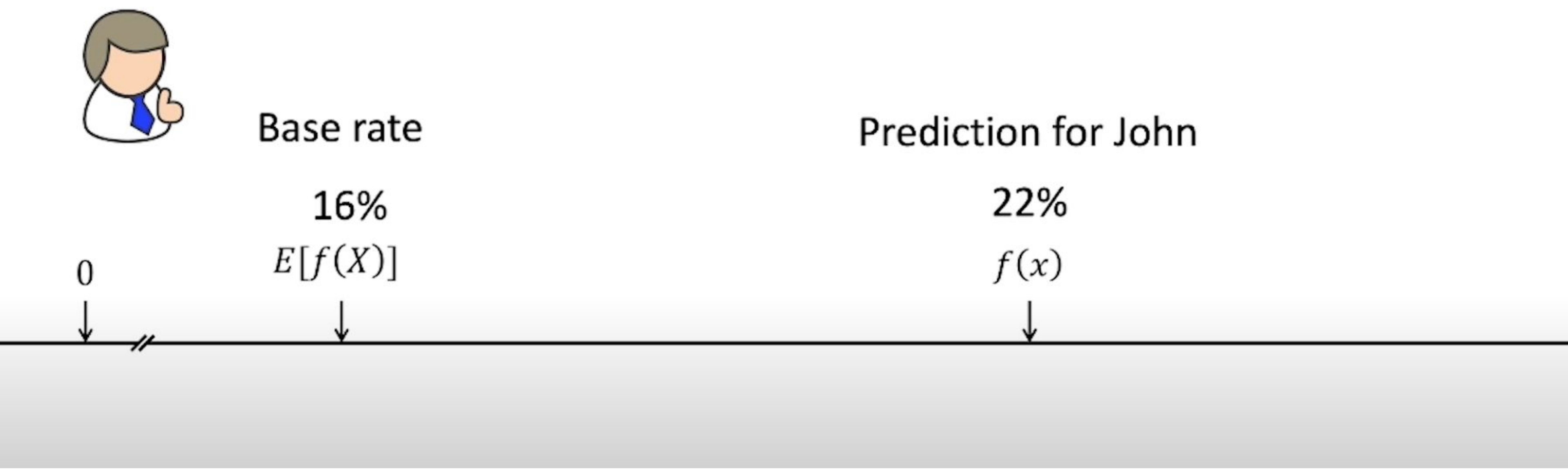
$$g(z') = \phi_0 + \sum_{i=1}^{M} \phi_i z'_i$$

Features of Input

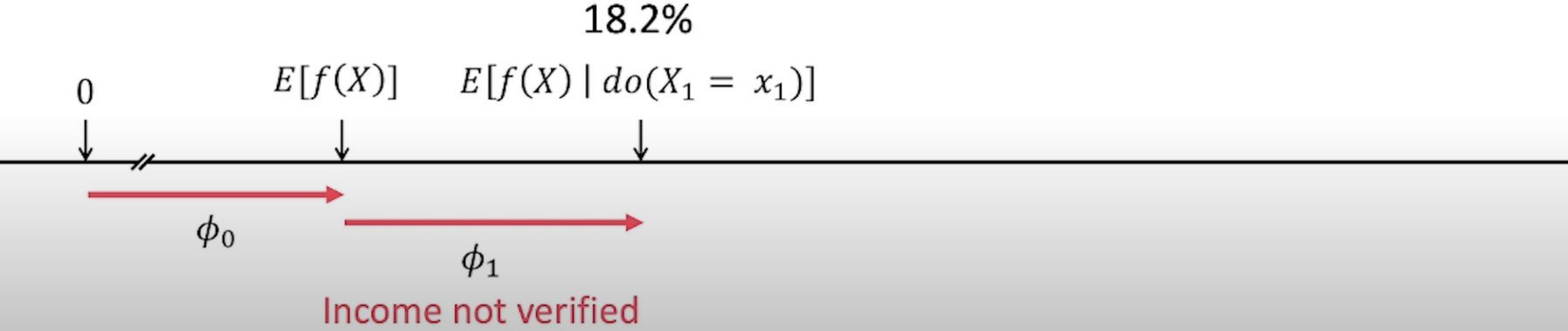
1. Income verified (Yes or No)
2. Debt to income ratio
3. Delinquent Payment (when)
4. Recent account opening (Yes or No)
5. Credit history (How long)



Additive feature attribution methods



Additive feature attribution methods



18.2%

0

$E[f(X)]$

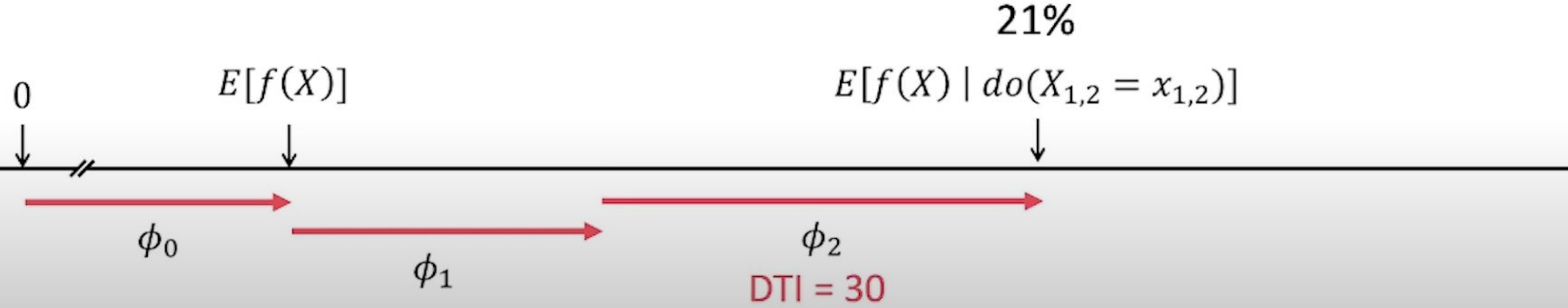
$E[f(X) \mid do(X_1 = x_1)]$

ϕ_0

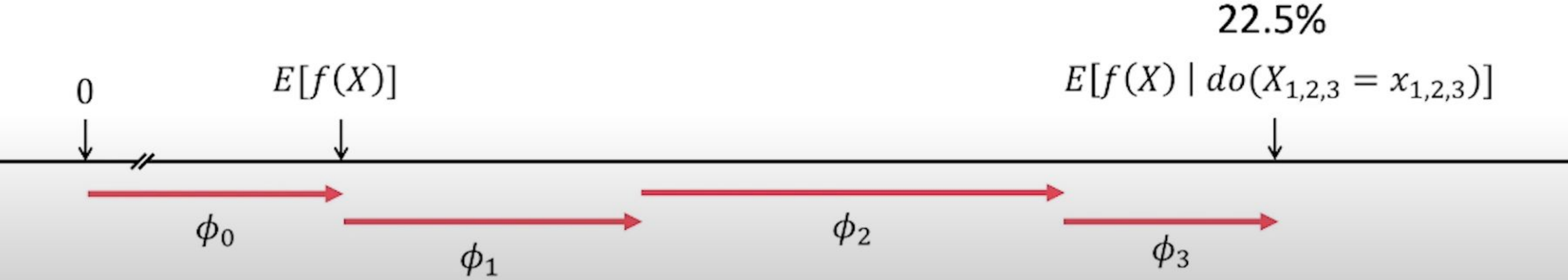
ϕ_1

Income not verified

Additive feature attribution methods



Additive feature attribution methods



22.5%

0

$E[f(X)]$

$E[f(X) \mid do(X_{1,2,3} = x_{1,2,3})]$

ϕ_0

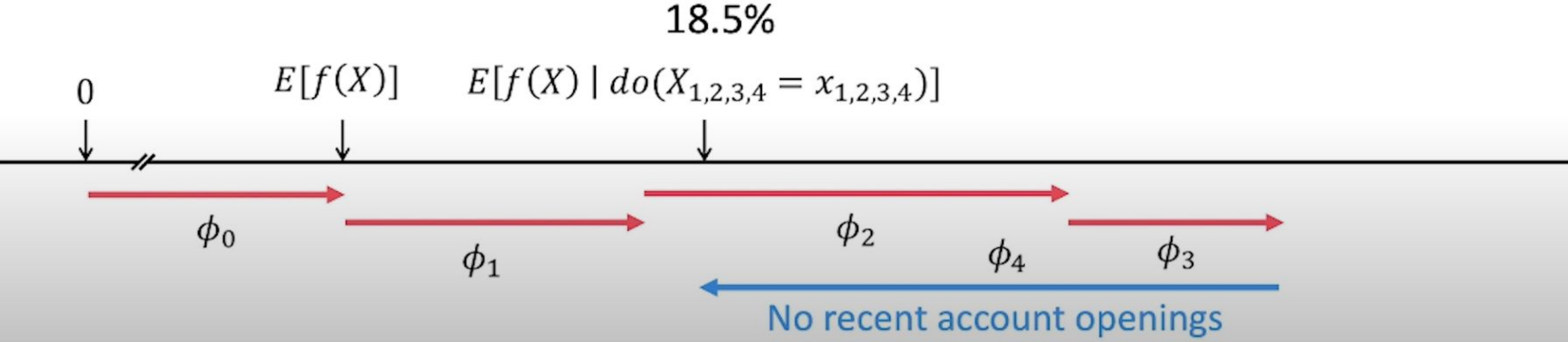
ϕ_1

ϕ_2

ϕ_3

Delinquent 10 months ago

Additive feature attribution methods



18.5%

0

$E[f(X)]$

$E[f(X) \mid do(X_{1,2,3,4} = x_{1,2,3,4})]$

ϕ_0

ϕ_1

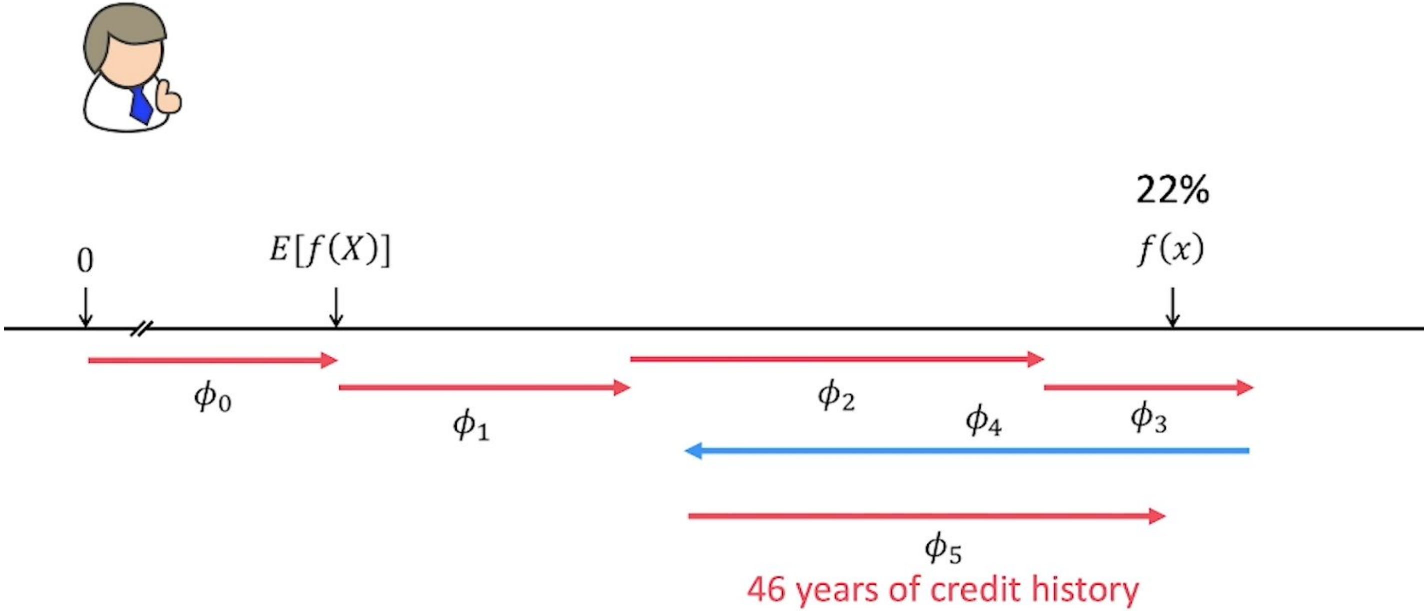
ϕ_2

ϕ_4

ϕ_3

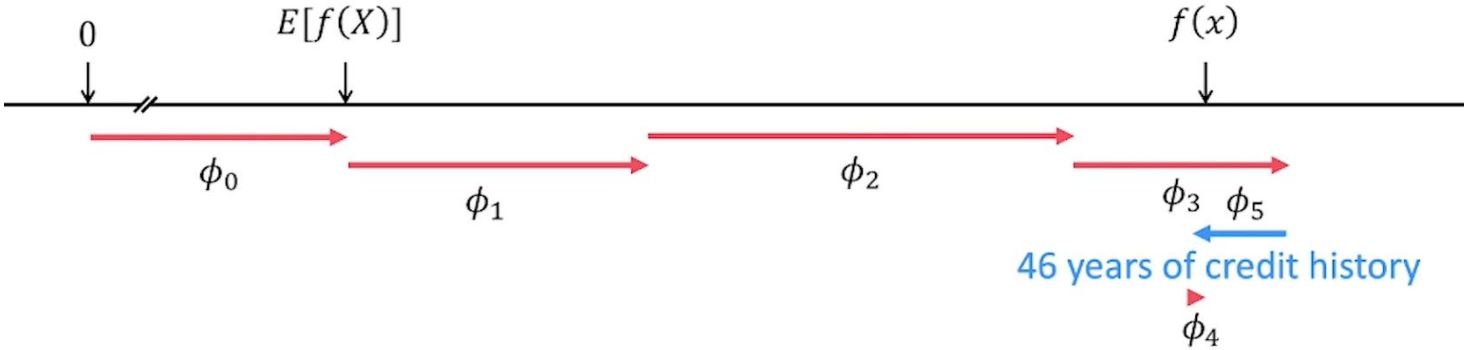
No recent account openings

Additive feature attribution methods



SHAP (SHapley Additive exPlanation) Values

The order matters!



46 years of credit history

No recent account openings

Shapley Value

The Shapley value is a solution concept in cooperative game theory. It was named in honor of Lloyd Shapley, who introduced it in 1951 and won the Nobel Prize in Economics for it in 2012.

$$\phi_i = \sum_{S \subseteq F \setminus \{i\}} \frac{|S|!(|F| - |S| - 1)!}{|F|!} \left[f_{S \cup \{i\}}(x_{S \cup \{i\}}) - f_S(x_S) \right]$$

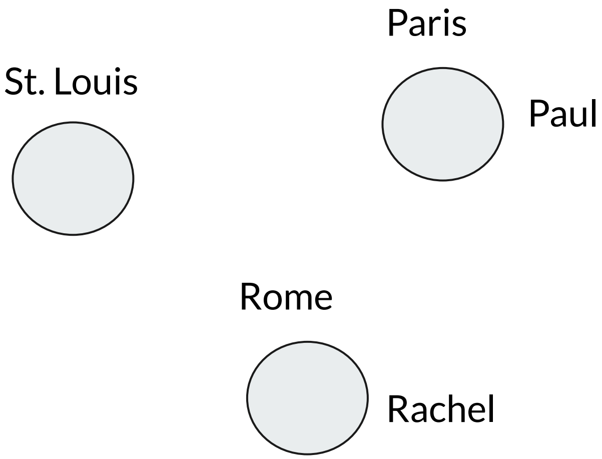
Shapley Value

You will go to visit your friends after COVID, your friends will pay your flight.

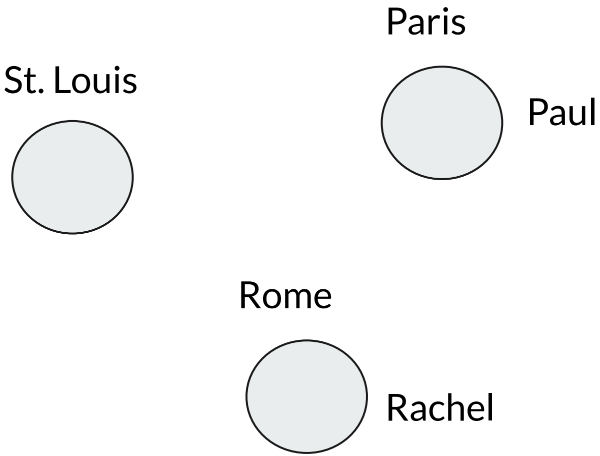
St. Louis to Paris (round-trip) $900

St. Louis to Rome(round-trip) $1,100

St. Louis to Paris to Rome to St. Louis $1,600



How do I allocate the money



Paris

St. Louis

Paul

Rome

Rachel

How do I allocate the money

Paul (Primary User) $900 Rachel(Incremental User) $700

Rachel (Primary User) $1,100 Paul (Incremental User) $500

Using the shapley value:

Paul should pay ($900 + $500)/2 = $700

Rachel should pay ($1,100 + $700)/2 = $900

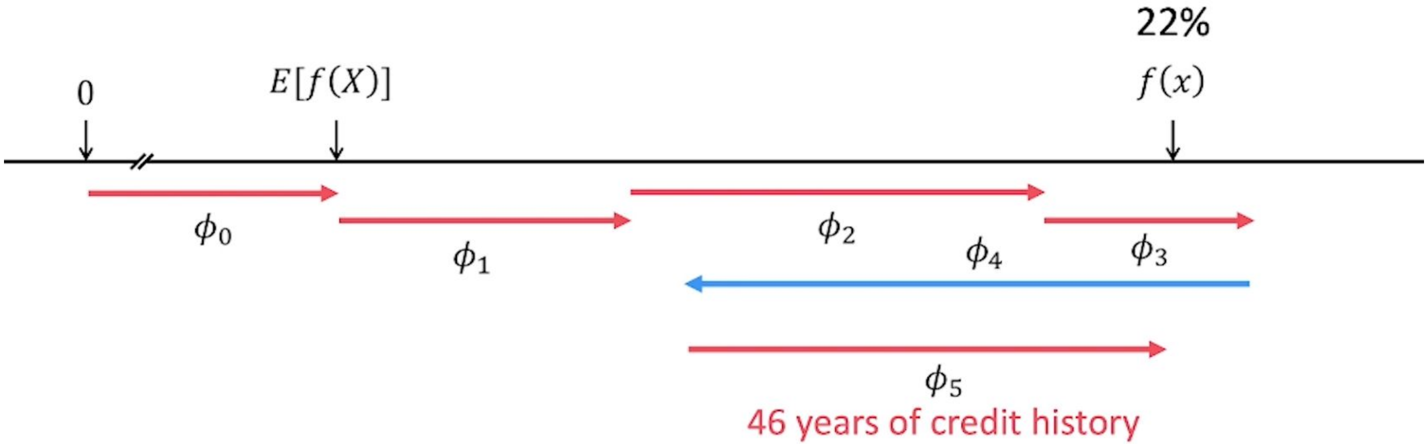
St. Louis to Paris(Paul) (round-trip) $900

St. Louis to Rome(Rachel)(round-trip) $1,100

St. Louis to Paris to Rome to St. Louis $1,600

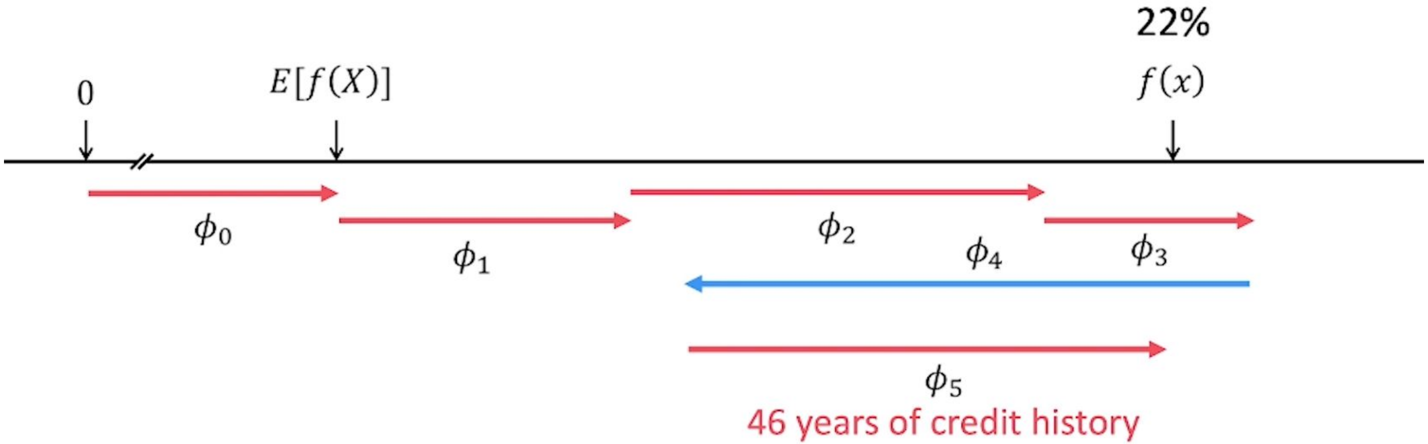
SHAP (SHapley Additive exPlanation) Values

Shapley values resulting from averaging over all possible orderings.

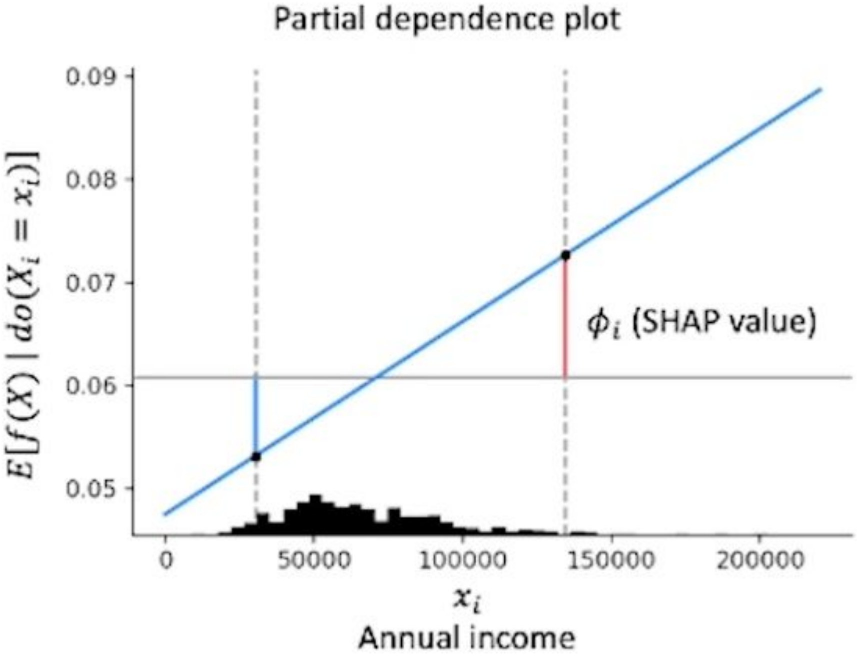


SHAP (SHapley Additive exPlanation) Values

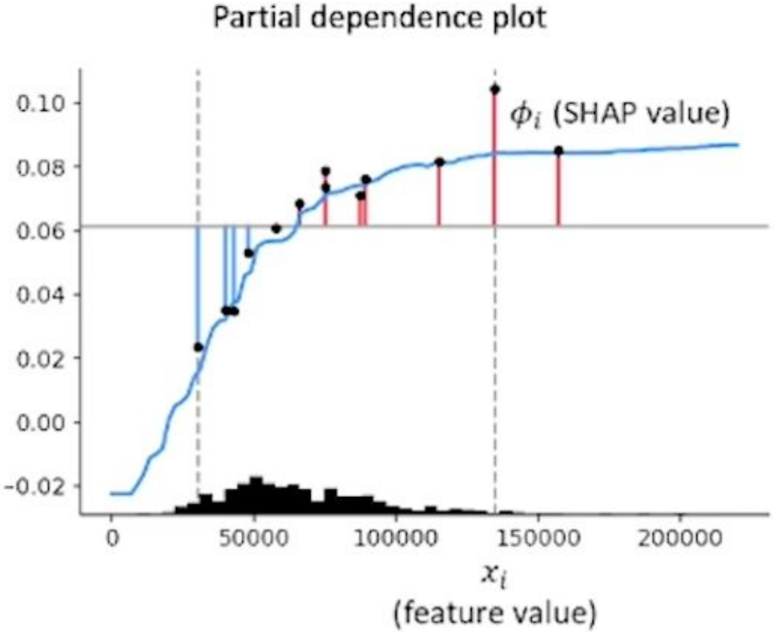
Local accuracy (additivity) - The sum of the local feature attributions equals the difference between the base rate and the model output.



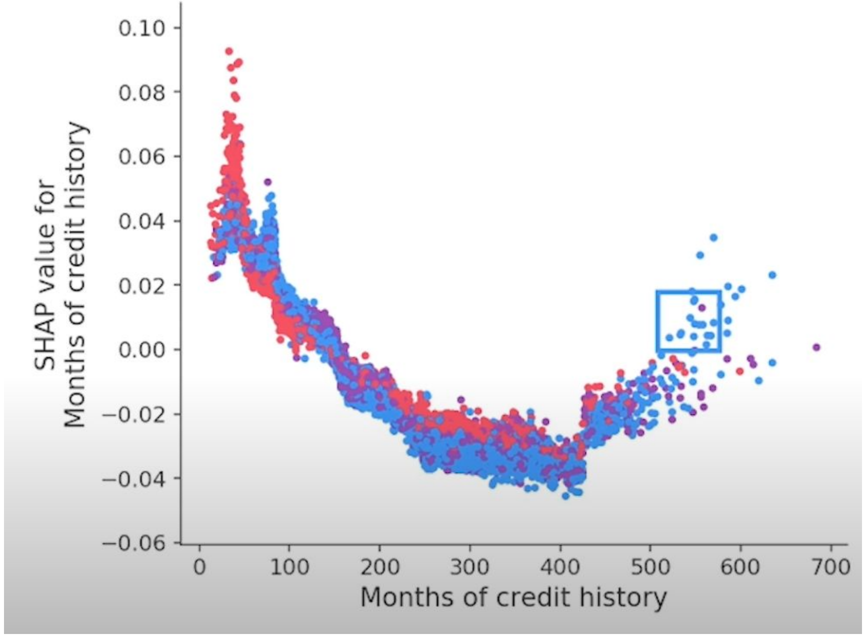
SHAP (SHapley Additive exPlanation) Values



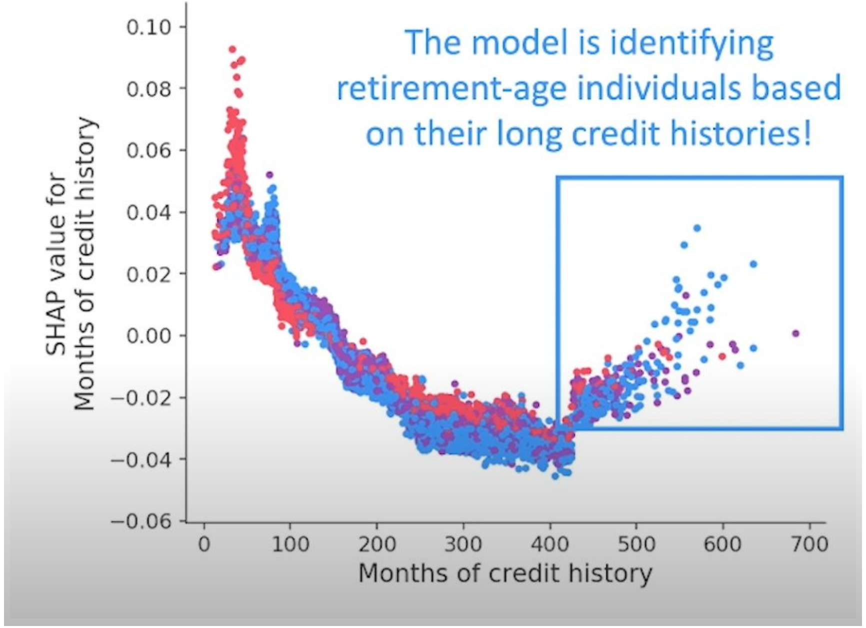
SHAP (SHapley Additive exPlanation) Values



Help to find the unfairness of model



Help to find the unfairness of model



The model is identifying retirement-age individuals based on their long credit histories!

Discussion:

1. Can you give one real-world scenario that ML/AI models do not need to be explainable? and why?

2. Interpretation is the process of giving explanations to human. How can we measure ‘good’ explanations in your opinion?

Questions

Thank you