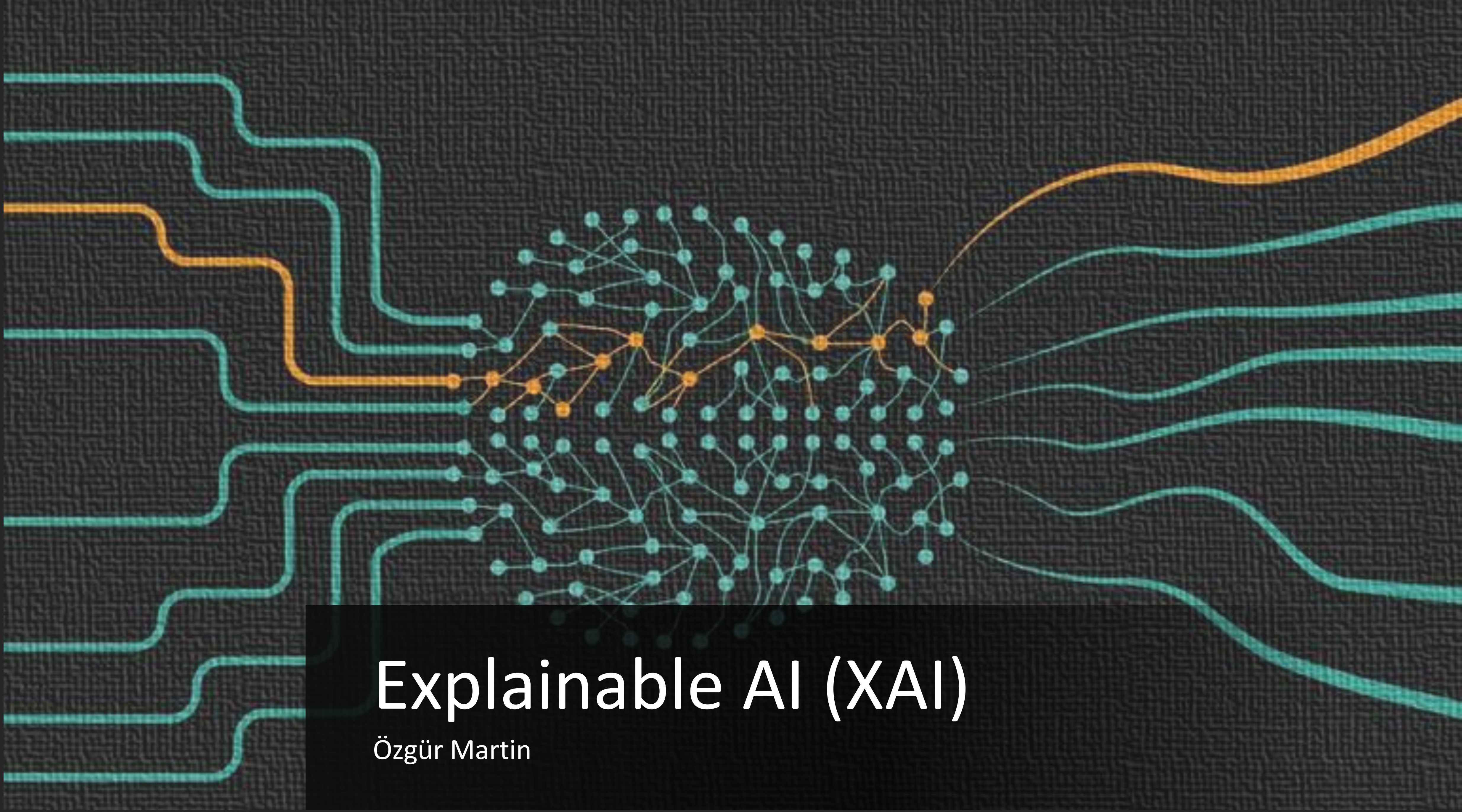


Explainable AI (XAI)

Özgür Martin

The Washington Post

Democracy Dies in Darkness

'Creative … motivating' and fired

Sarah Wysocki was out of work for only a few days after she was fired by DCPS last year. She is now teaching at Hybla Valley Elementary School in Fairfax County. (Jahi Chikwendiu/The Washington Post)

By **Bill Turque**
March 6, 2012

HDSR

Why Are We Using Black Box Models in AI When We Don't Need To? A Lesson From An Explainable AI Competition

by Cynthia Rudin and Joanna Radin

Published on Nov 22, 2019

CGAP

BLOG 05 SEPTEMBER 2019

Algorithm Bias in Credit Scoring: What's Inside the Black Box?

By Maria Fernandez Vidal, Jacobo Menajovsky

> The responsible use of algorithms requires providers to know which variables are being considered in their credit scoring models and how they are affecting people's scores.

Accuracy vs. Interpretability

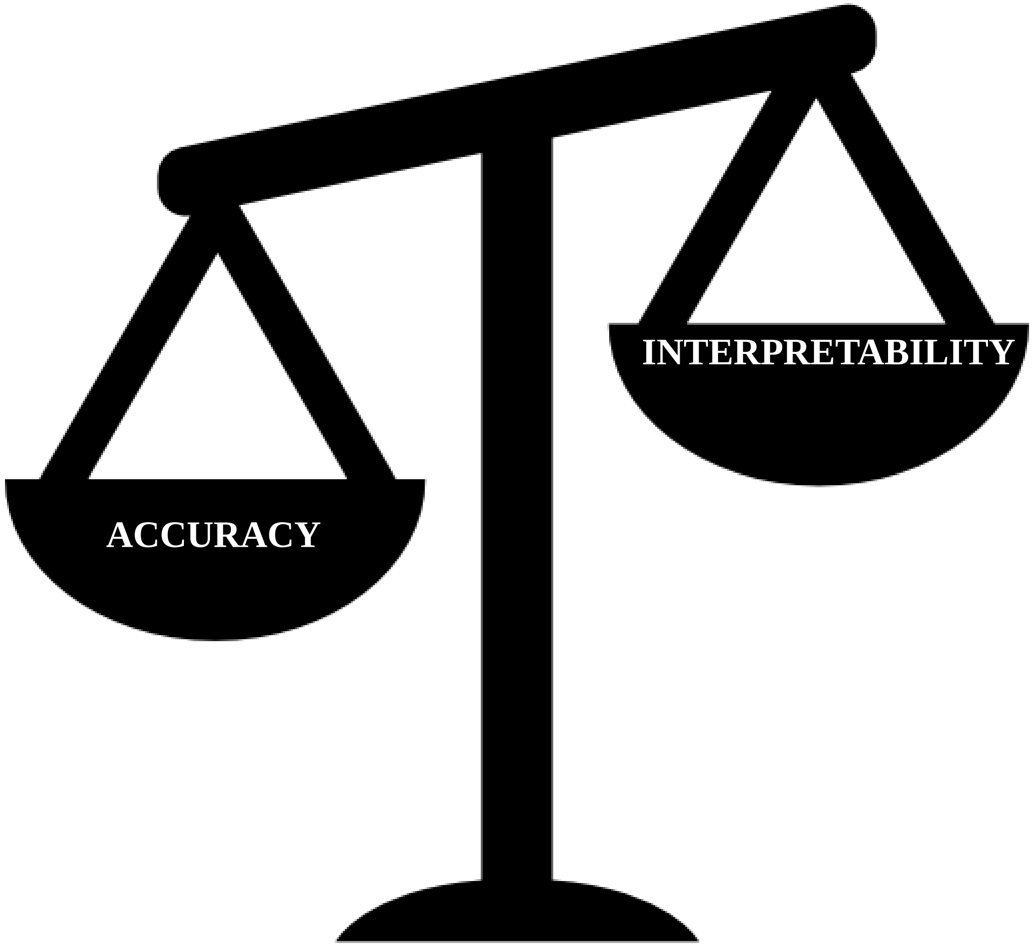
Random Forest

XGBoost

SVM

Deep Learning

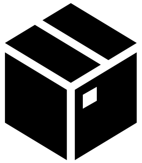
…



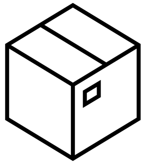
Linear Regression

Decision Tree

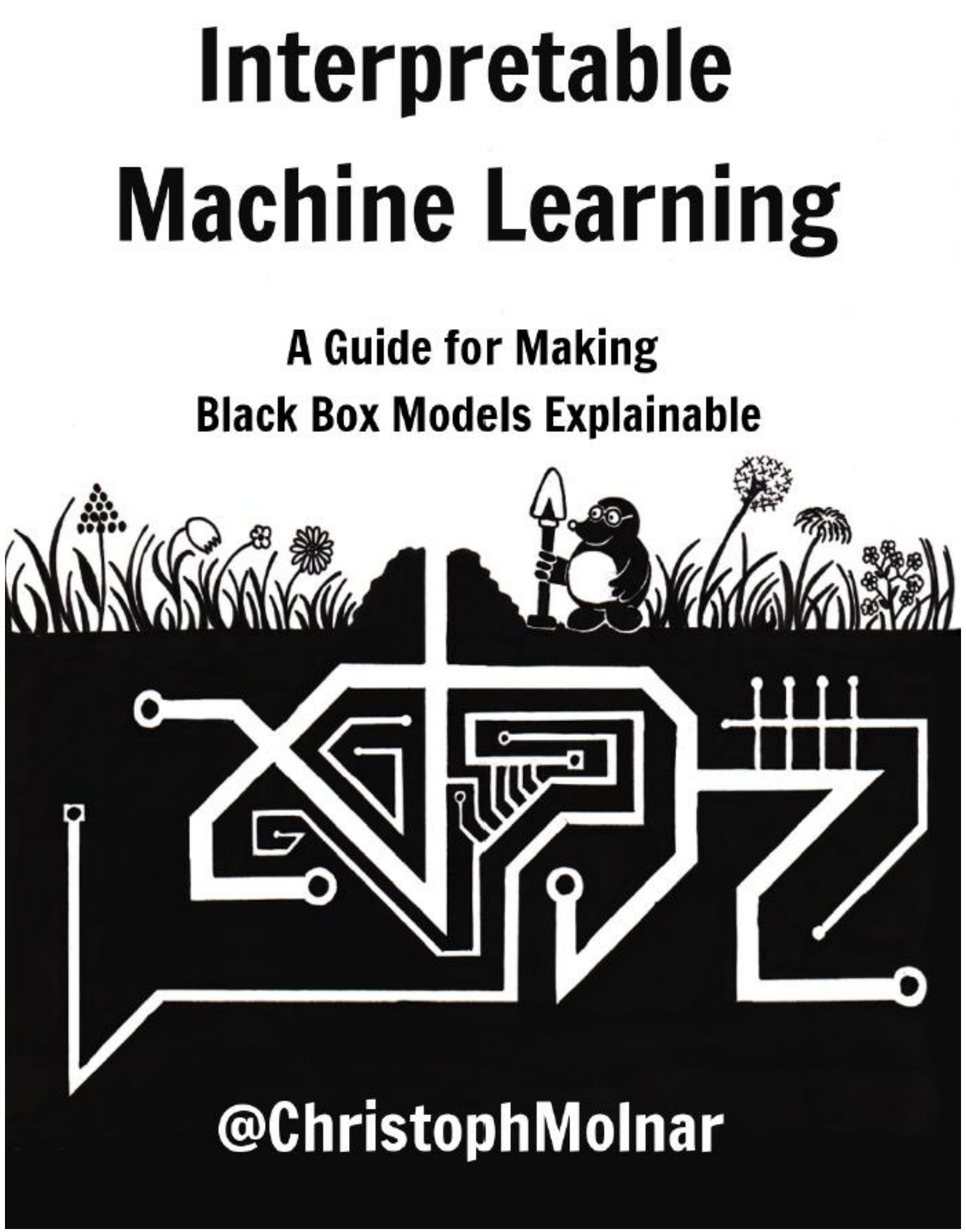
…



blackbox



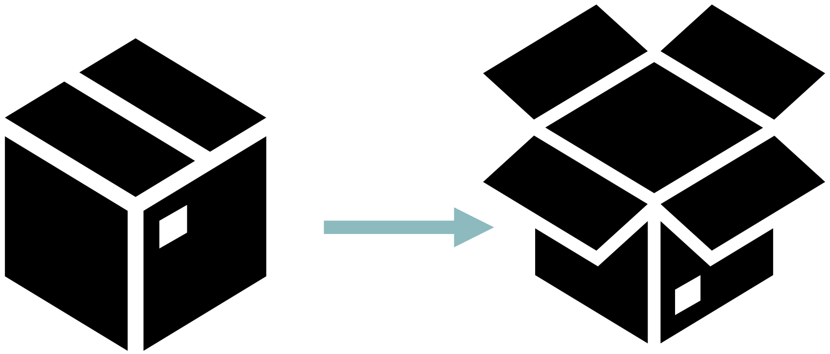
whitebox (glassbox)


Interpretable
Machine Learning
A Guide for Making
Black Box Models Explainable
@ChristophMolnar

→ 2019-2021 ...

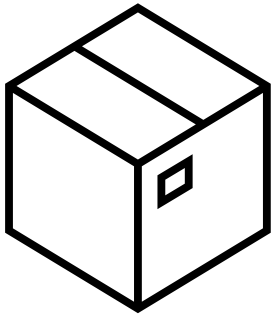
Explainable Artificial Intelligence - XAI

Model Agnostic Explanations



Explaining the blackbox models

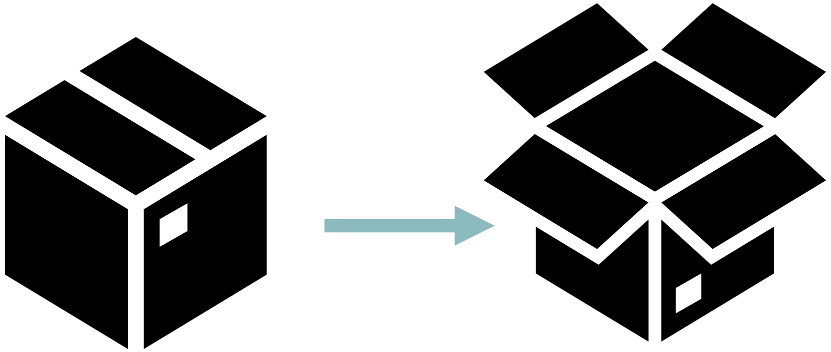
Interpretable Models



High accuracy whitebox models

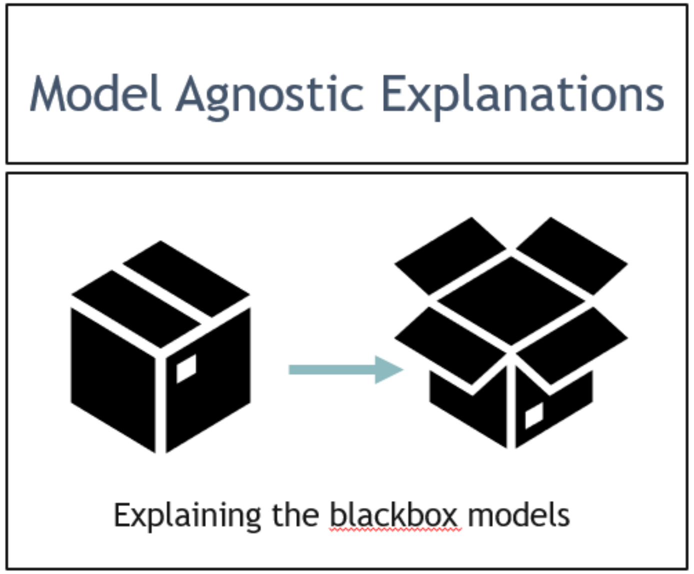
Explainable Artificial Intelligence - XAI

Model Agnostic Explanations



Explaining the blackbox models

Explainable Artificial Intelligence - XAI



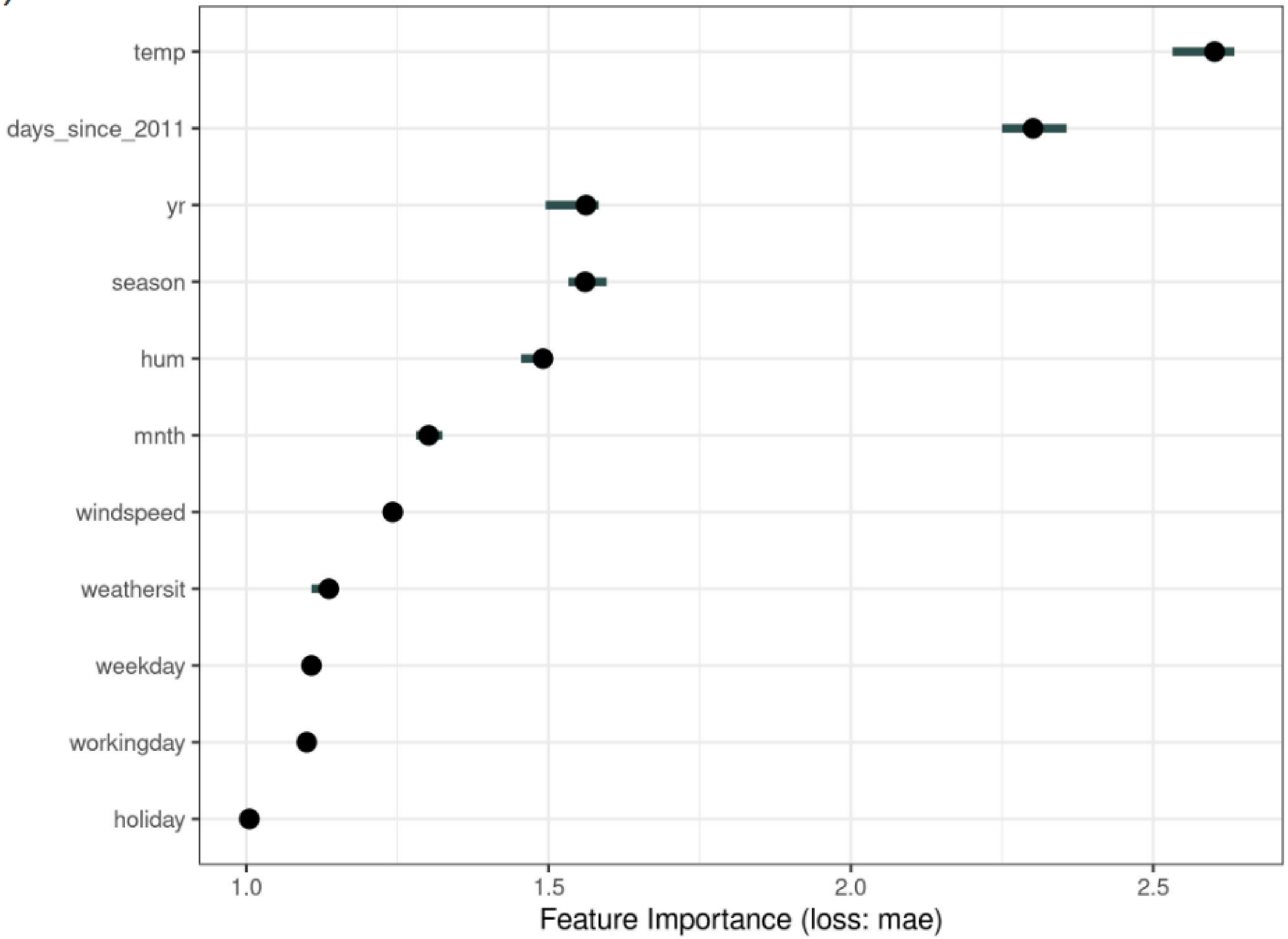
Global Explanations

The permutation feature importance algorithm based on Fisher, Rudin, and Dominici (2018):

Input: Trained model $\hat{f}$, feature matrix X, target vector y, error measure $L(y, \hat{f})$.

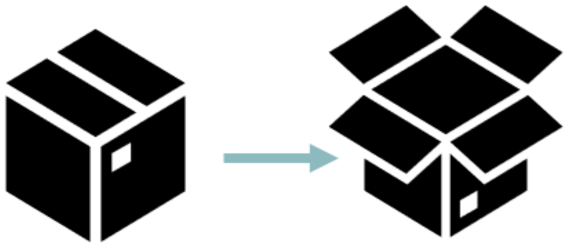
1. Estimate the original model error $e_{orig} = L(y, \hat{f}(X))$ (e.g. mean squared error)
2. For each feature $j \in \{1, \ldots, p\}$ do:
   - Generate feature matrix X_{perm} by permuting feature j in the data X. This breaks the association between feature j and true outcome y.
   - Estimate error $e_{perm} = L(Y, \hat{f}(X_{perm}))$ based on the predictions of the permuted data.
   - Calculate permutation feature importance as quotient $FI_j = e_{perm}/e_{orig}$ or difference $FI_j = e_{perm} - e_{orig}$
3. Sort features by descending FI.

Bike Rentals (Regression)



Explainable Artificial Intelligence - XAI

Model Agnostic Explanations



Explaining the blackbox models

Local Explanations

LIME

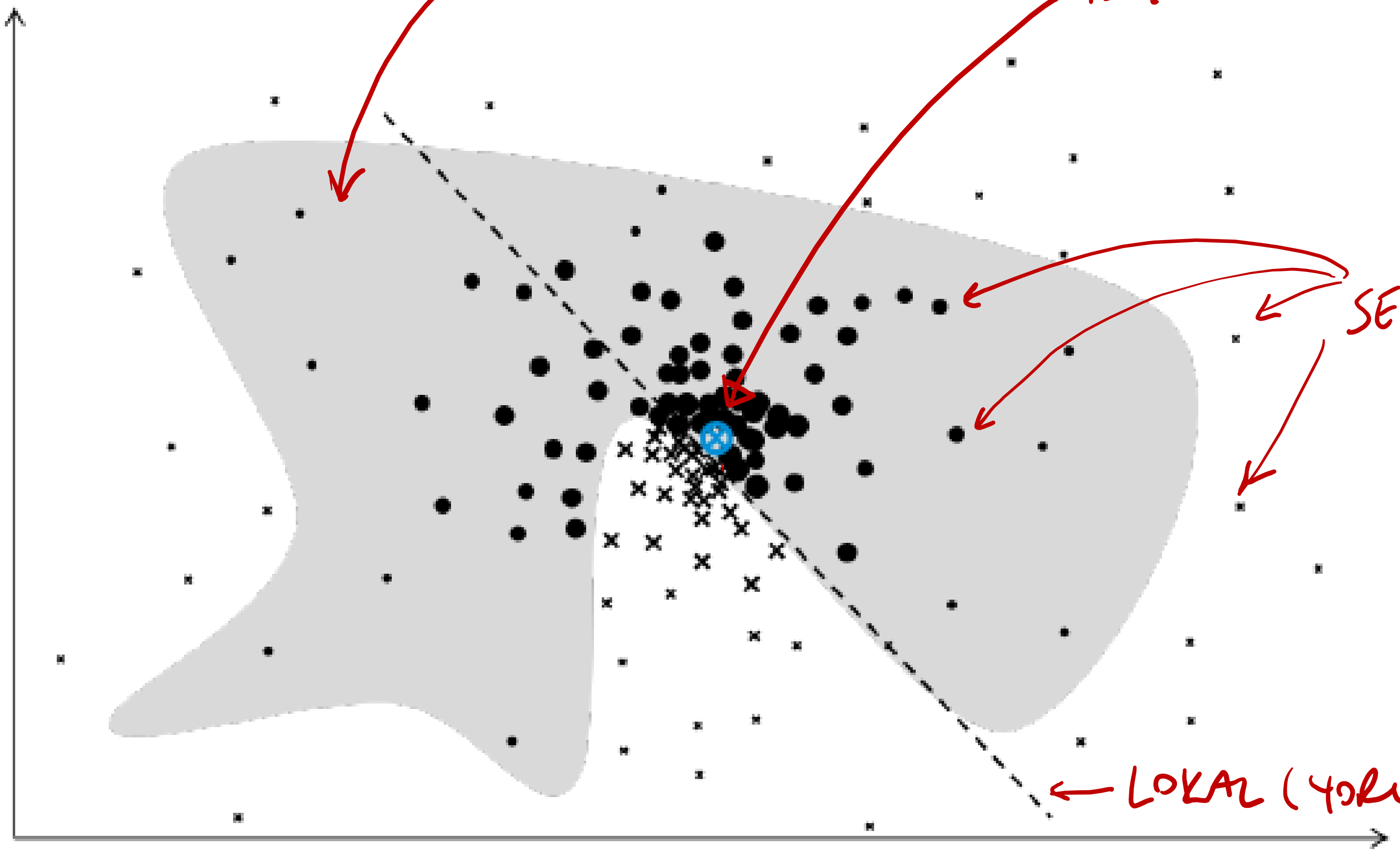
(RIBERIO VD., 2016)

LOCAL INTERPRETABLE MODEL-AGNOSTIC EXPLANATIONS

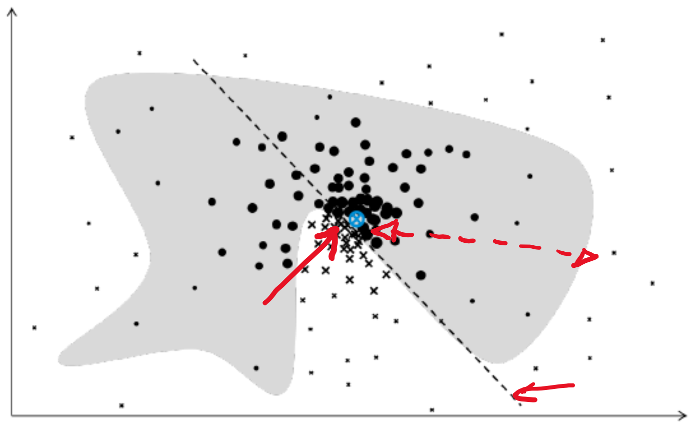
MODEL ETİKETLERİ

AĞIRLIKLAR

MESAFE İLE TERS ORANTILI



LIME



AÇIKLAMA

ÖĞRENME HATASI

UZAKLIK

$$\xi(x) = \arg\min_{g \in G} L(f, g, \pi_x) + \underbrace{\Omega(g)}$$

İLGİLİ VERİ NOKTASI

LOKAL MODEL

LOKAL MODEL KARMAŞIKLIĞI

- ÖZNİTELİK SAYISI
- AĞAÇ DERİNLİĞİ

⋮

LIME

XGBoost model predicted an acceptance probability of 77.0%

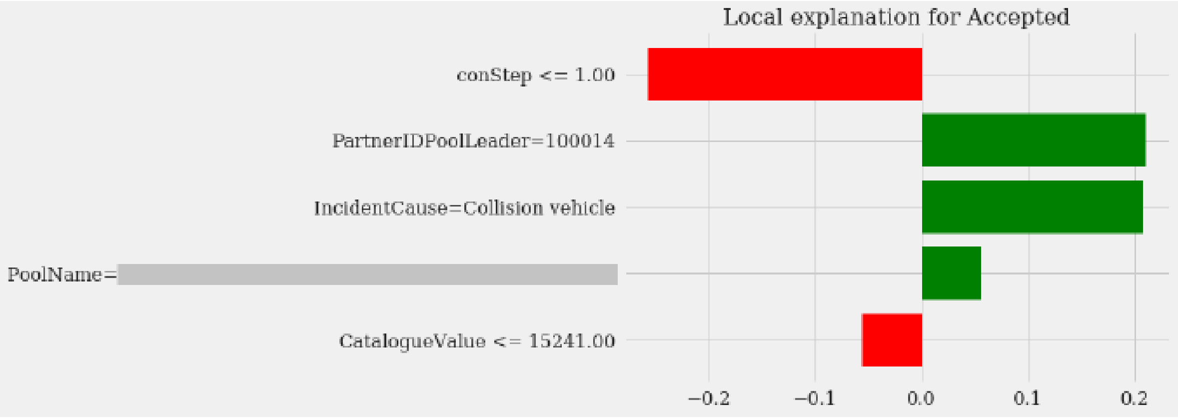


Figure 23: FP - LIME by Default (5 Features)

[Local Pred. = 0.58, Intercept = 0.42, R^2 = 0.28, RMSE = 0.26, Time = 4.67 s]

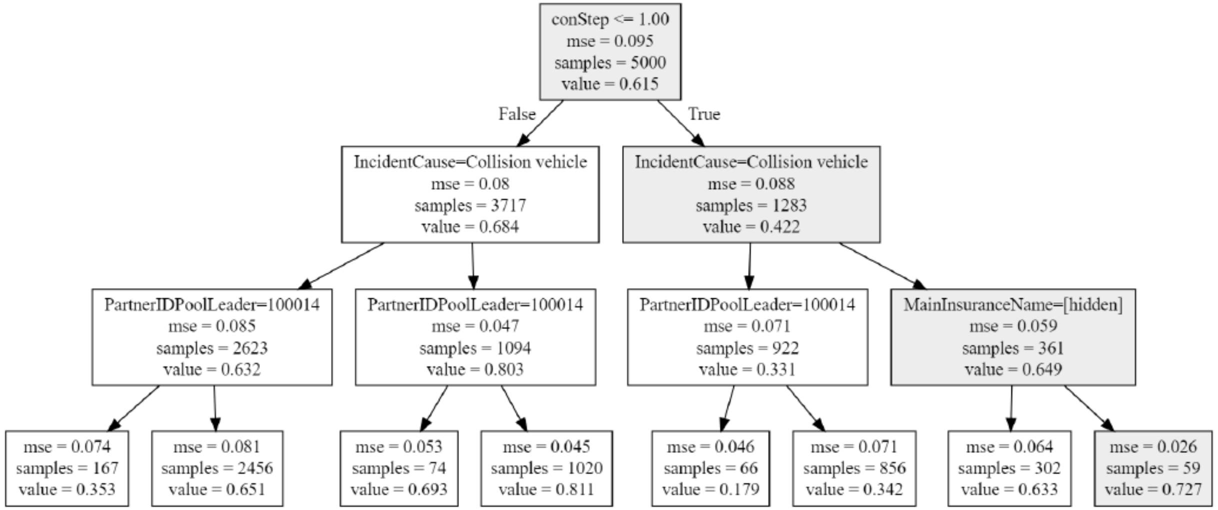


Figure 24: FP - LIME Decision Tree (3 Layers)

[Local Pred. = 0.73, R^2 = 0.28, RMSE = 0.26, Time = 2.15 s]

LIME

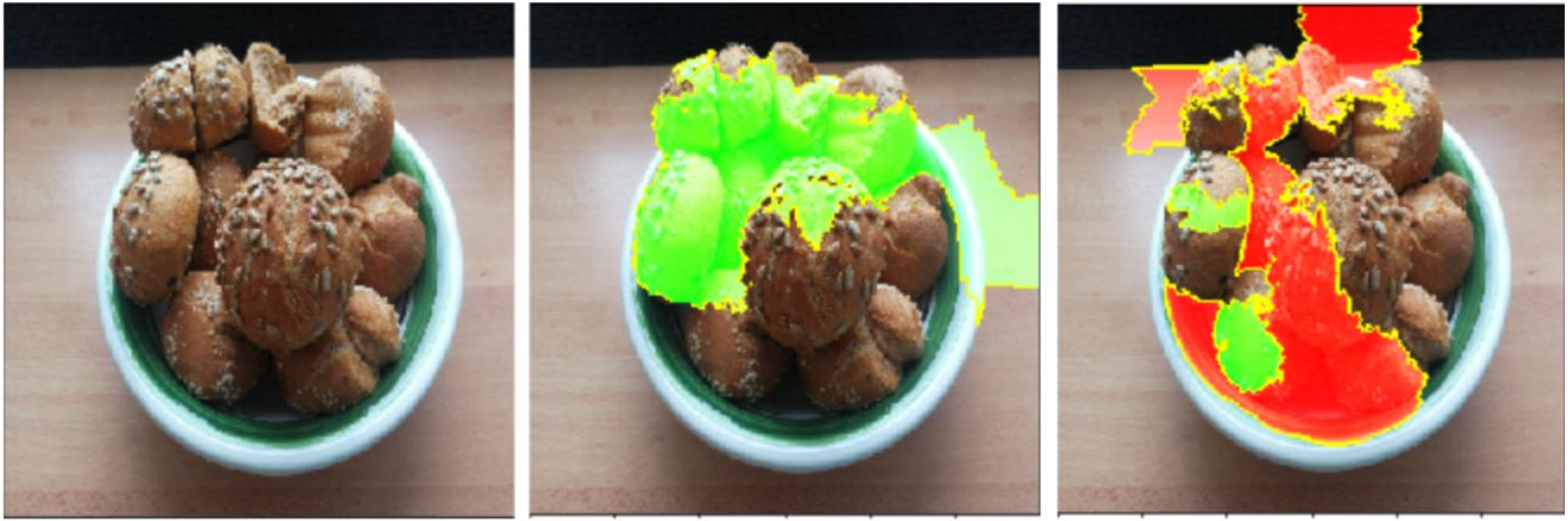


FIGURE 9.8: Left: Image of a bowl of bread. Middle and right: LIME explanations for the top 2 classes (bagel, strawberry) for image classification made by Google's Inception V3 neural network.

LIME

+

Even if you **replace the underlying machine learning model**, you can still use the same local, interpretable model for explanation.

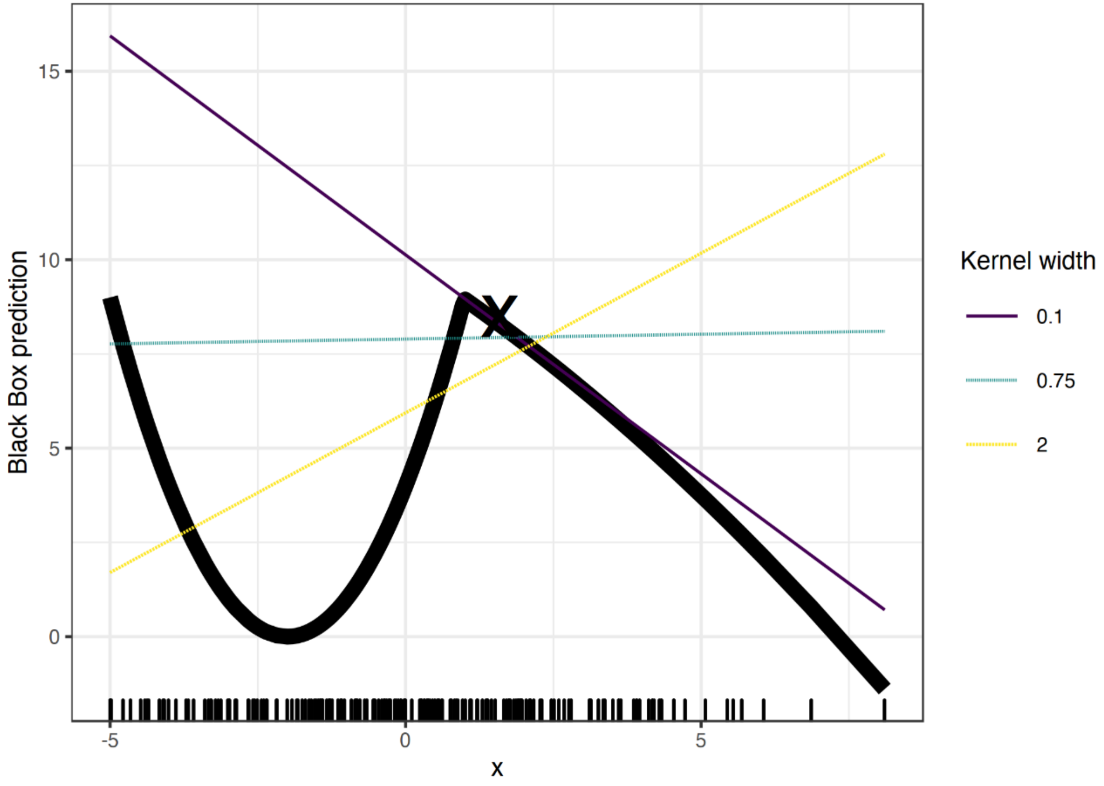
When using Lasso or short trees, the resulting **explanations are short (= selective) and possibly contrastive**.

The explanations created with local surrogate models **can use other (interpretable) features than the original model was trained on.**

−

The explanations of two very close points can vary. Even when the sampling process is repeated, the explantions can be different.

The correct definition of the neighborhood is a very big, unsolved problem when using LIME with tabular data.



SHAP
(Lundberg ve Lee, 2017)

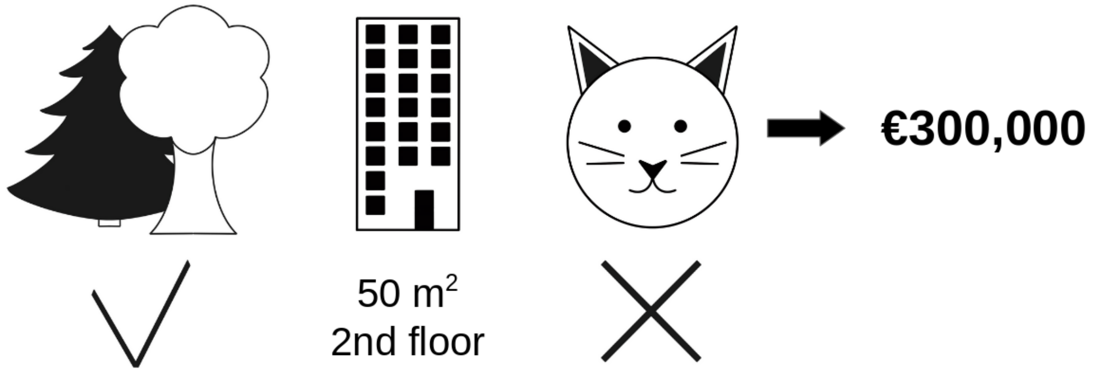
Shapley Additive Explanations

park-nearby + cat-banned + area-50 + floor-2nd

€300,000

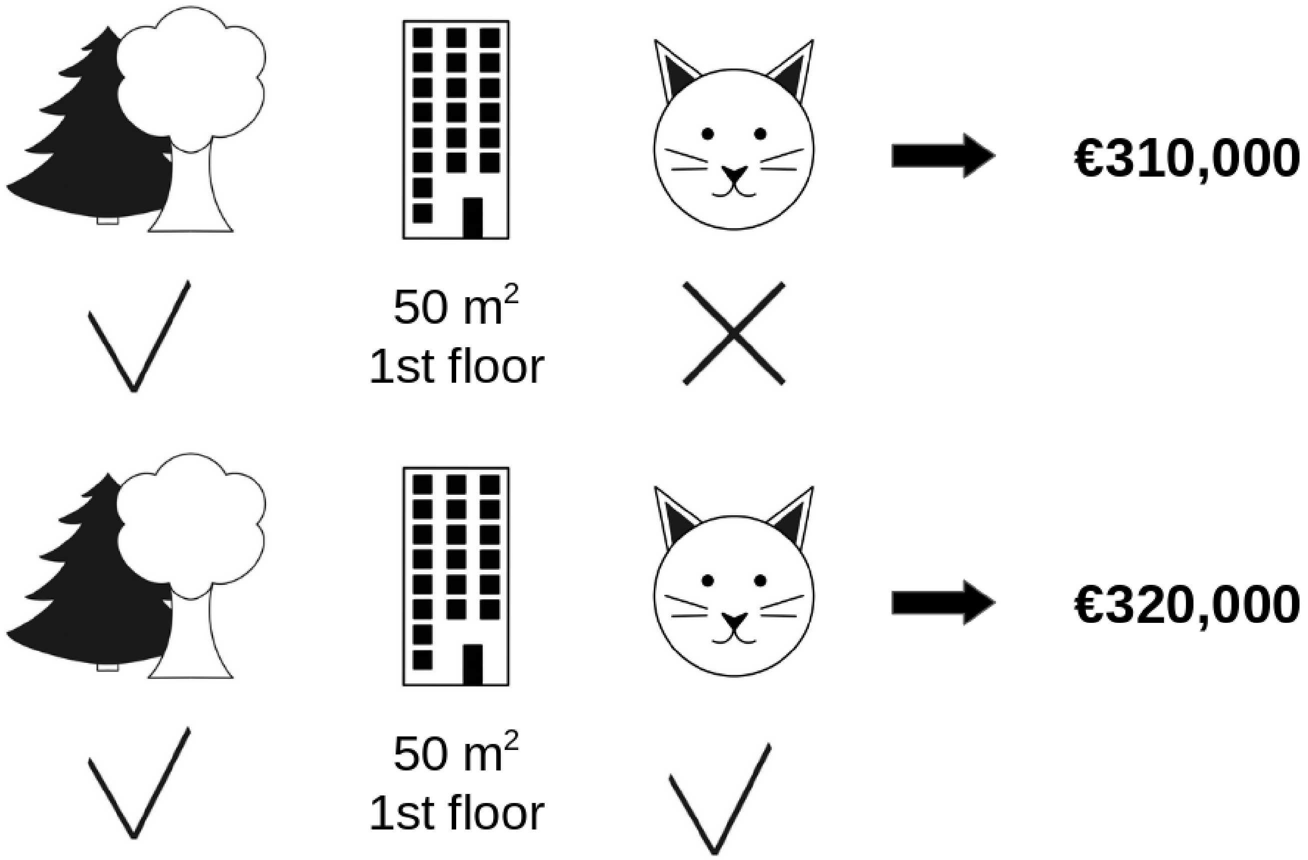
€310,000

€-10,000



SHAP
(Lundberg ve Lee, 2017)

Shapley Additive Explanations

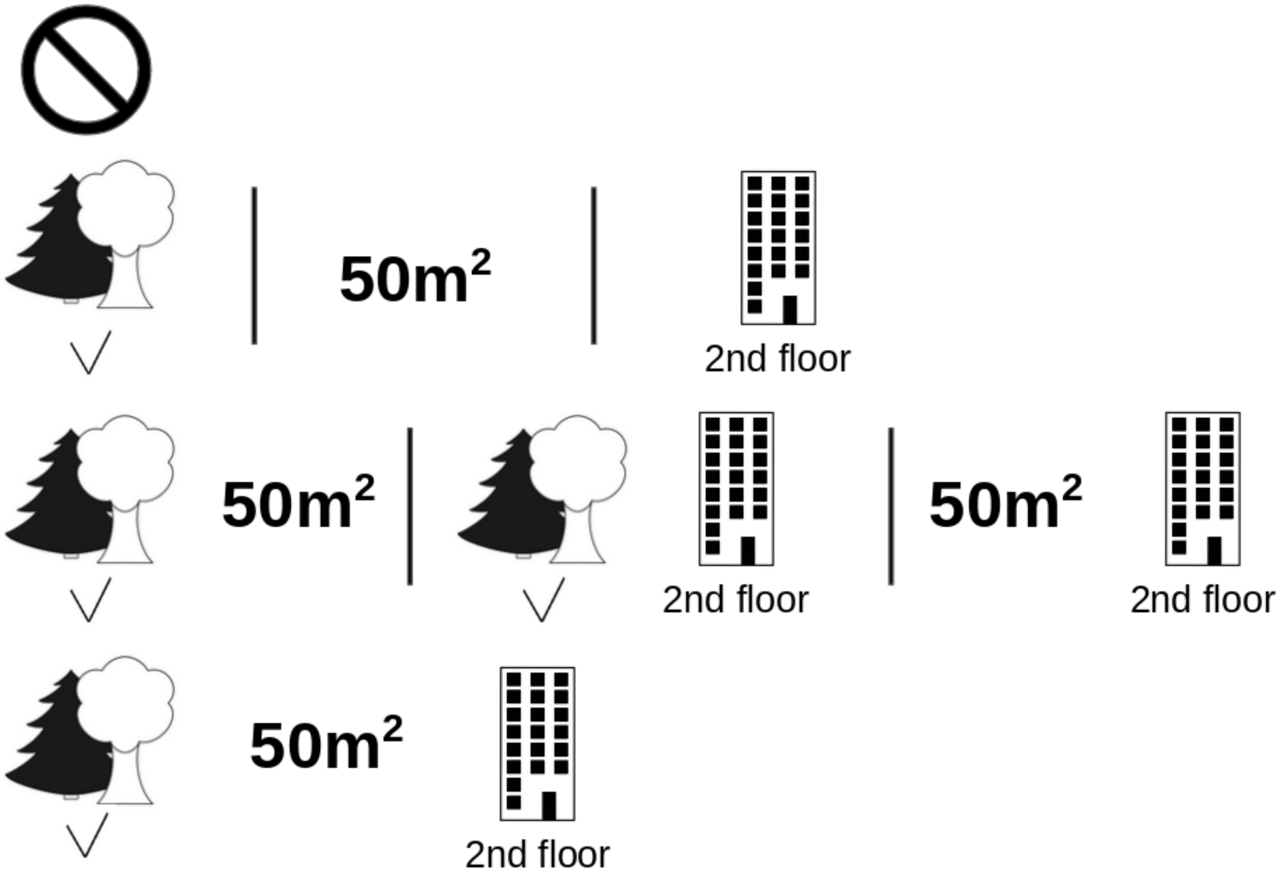


SHAP

(LUNDBERG VE LEE, 2017)

SHAPLEY ADDITIVE EXPLANATIONS

- No feature values
- park-nearby
- area-50
- floor-2nd
- park-nearby + area-50
- park-nearby + floor-2nd
- area-50 + floor-2nd
- park-nearby + area-50 + floor-2nd



SHAP

(LUNDBERG VE LEE, 2017)

SHAPLEY ADDITIVE EXPLANATIONS

ÖZNİTELİK BU ÖRNEKTE VAR (1), YOK (0)

AÇIKLAMA MODELİ →

$$g_i(z') = \phi_0 + \sum_{j=1}^{M} \phi_{ij} z'_{ij}$$

ÖRNEK

ÖZNİTELİK

SHAPLEY DEĞERİ →

$$\phi_{ij} = \sum_{S \subseteq X_i \setminus \{j\}} \frac{|S|!(M-|S|-1)!}{M!} [f_i(S \cup \{j\}) - f_i(S)]$$

j HARİÇ ÖZNİTELİKLER

MODEL TAHMİNİ

SHAP

Efficiency The feature contributions must add up to the difference of prediction for x and the average.

$$\sum_{j=1}^{p} \phi_j = \hat{f}(x) - E_X(\hat{f}(X))$$

Symmetry The contributions of two feature values j and k should be the same if they contribute equally to all possible coalitions.

Dummy A feature j that does not change the predicted value – regardless of which coalition of feature values it is added to – should have a Shapley value of 0.

Additivity For a game with combined payouts val+val^{+} the respective Shapley values are as follows:

$$\phi_j + \phi_j^{+}$$

SHAP

$$\phi_{ij} = \sum_{S \subseteq X_i \setminus \{j\}} \frac{|S|!(M - |S| - 1)!}{M!} [f_i(S \cup \{j\}) - f_i(S)]$$

TÜM ALTKÜMELER

2^M

YAKLAŞIKLAMA YÖNTEMLERİ

KERNEL SHAP*

LINEAR SHAP

LOW_ORDER SHAP

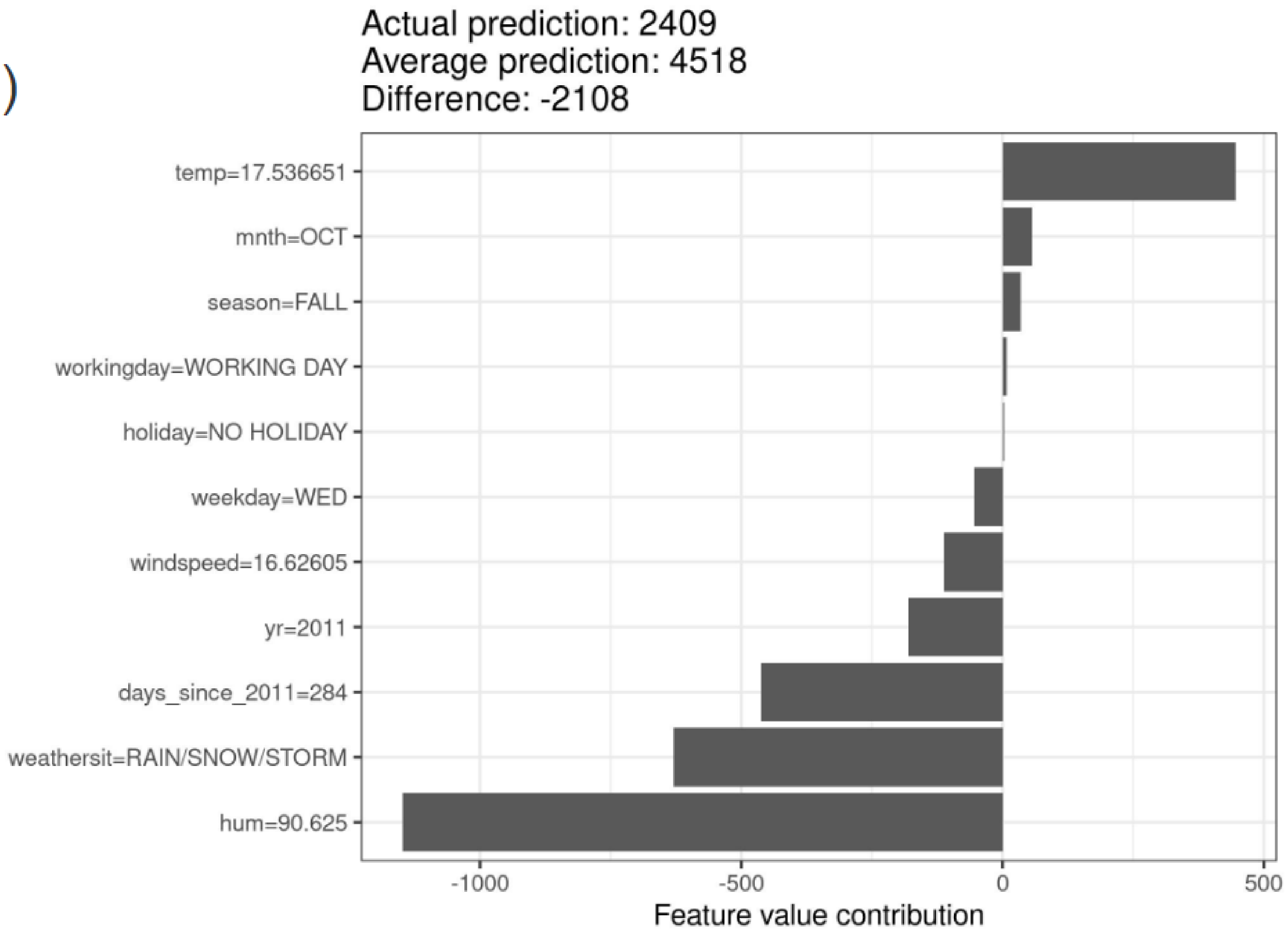
TREE SHAP

DEEP SHAP

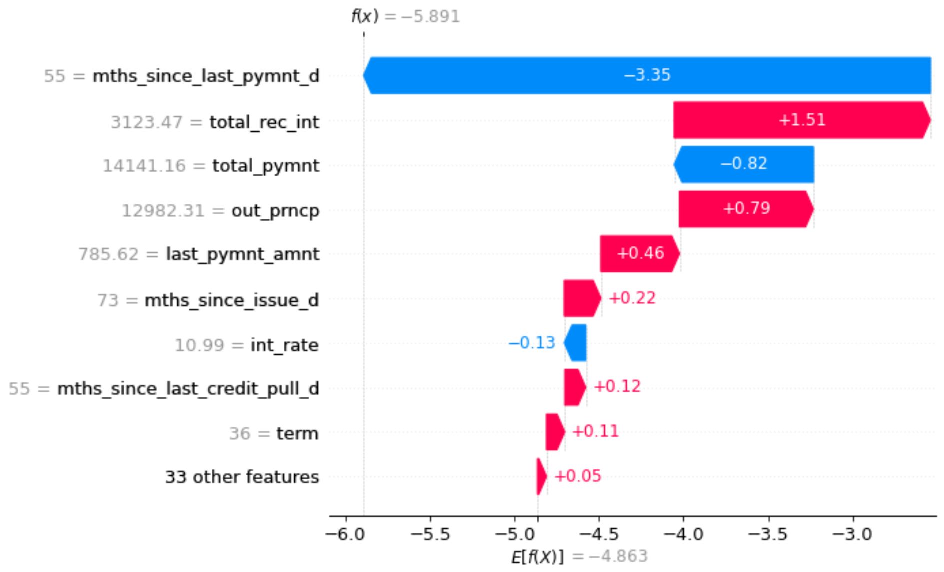
*AGNOSTİK

SHAP

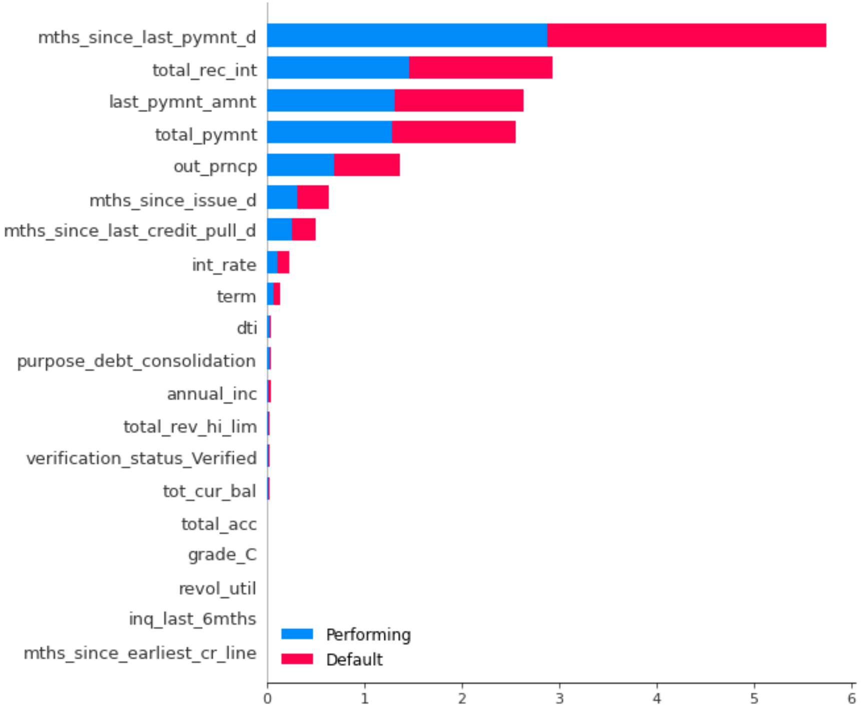
Bike Rentals (Regression)



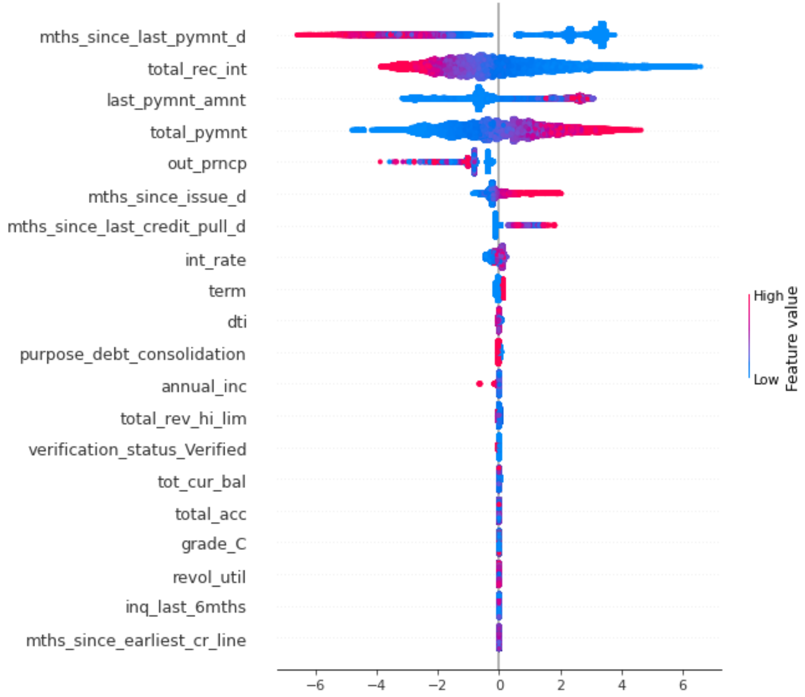
SHAP



f(x) = −5.891
55 = mths_since_last_pymnt_d
3123.47 = total_rec_int
14141.16 = total_pymnt
12982.31 = out_prncp
785.62 = last_pymnt_amnt
73 = mths_since_issue_d
10.99 = int_rate
55 = mths_since_last_credit_pull_d
36 = term
33 other features
−3.35
+1.51
−0.82
+0.79
+0.46
+0.22
−0.13
+0.12
+0.11
+0.05
−6.0 −5.5 −5.0 −4.5 −4.0 −3.5 −3.0
E[f(X)] = −4.863

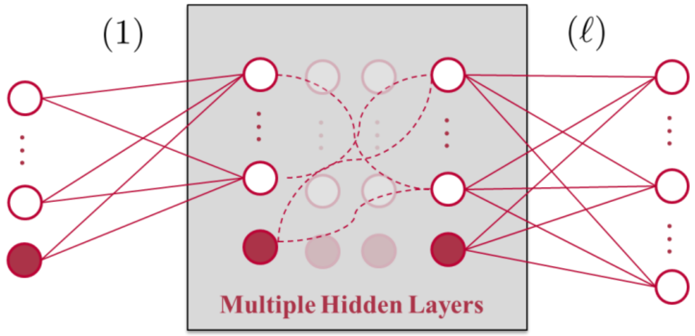


mths_since_last_pymnt_d
total_rec_int
last_pymnt_amnt
total_pymnt
out_prncp
mths_since_issue_d
mths_since_last_credit_pull_d
int_rate
term
dti
purpose_debt_consolidation
annual_inc
total_rev_hi_lim
verification_status_Verified
tot_cur_bal
total_acc
grade_C
revol_util
inq_last_6mths
mths_since_earliest_cr_line
Performing
Default
0 1 2 3 4 5 6

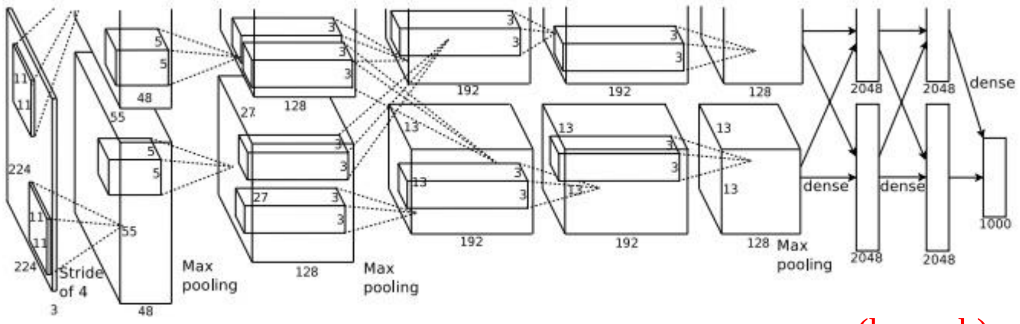


mths_since_last_pymnt_d
total_rec_int
last_pymnt_amnt
total_pymnt
out_prncp
mths_since_issue_d
mths_since_last_credit_pull_d
int_rate
term
dti
purpose_debt_consolidation
annual_inc
total_rev_hi_lim
verification_status_Verified
tot_cur_bal
total_acc
grade_C
revol_util
inq_last_6mths
mths_since_earliest_cr_line
High
Low
Feature value
−6 −4 −2 0 2 4 6

DERİN ÖĞRENME



$$\hat{y}_k(X,\beta) = \sigma\left(\sum_j \beta_{kj}^{(\ell)} h\left(\sum_s \beta_{js}^{(\ell-1)} h\left(\dots h\left(\sum_i \beta_{ji}^{(1)} X_i\right)\dots\right)\right)\right)$$



(kaynak)

GÖRÜNTÜ İŞLEME ✓

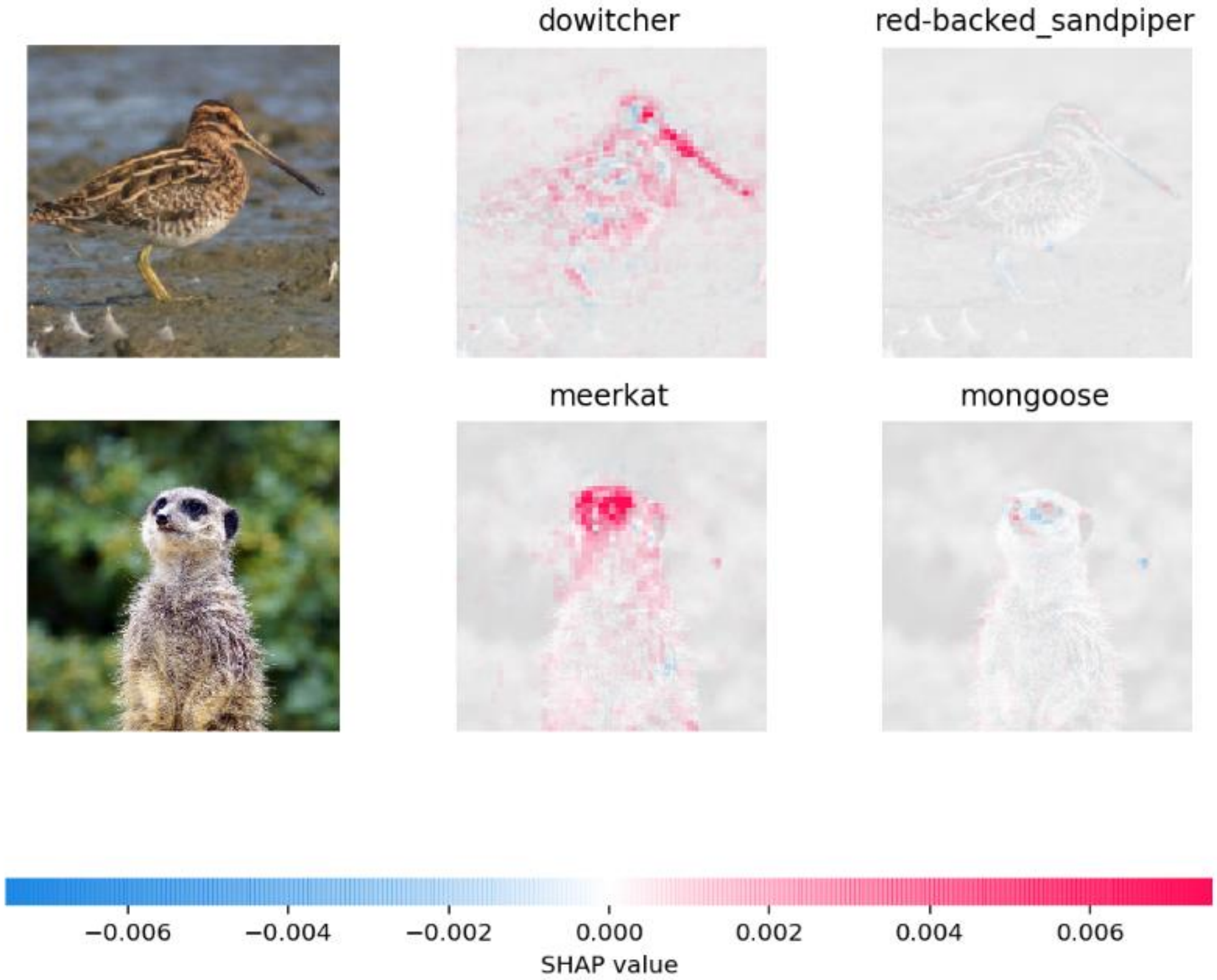
METİN İŞLEME ~

DİĞER PROBLEMLER ?

LIME

SHAP

DERİN ÖĞRENME



Counterfactual Explanations

Explanation in artificial intelligence: Insights from the social sciences

Tim Miller

School of Computing and Information Systems, University of Melbourne, Melbourne, Australia

Artificial Intelligence, 2019

COUNTERFACTUAL EXPLANATIONS WITHOUT OPENING THE BLACK BOX: AUTOMATED DECISIONS AND THE GDPR

Sandra Wachter,* Brent Mittelstadt,** & Chris Russell***

Harvard Journal of Law & Technology, 2018

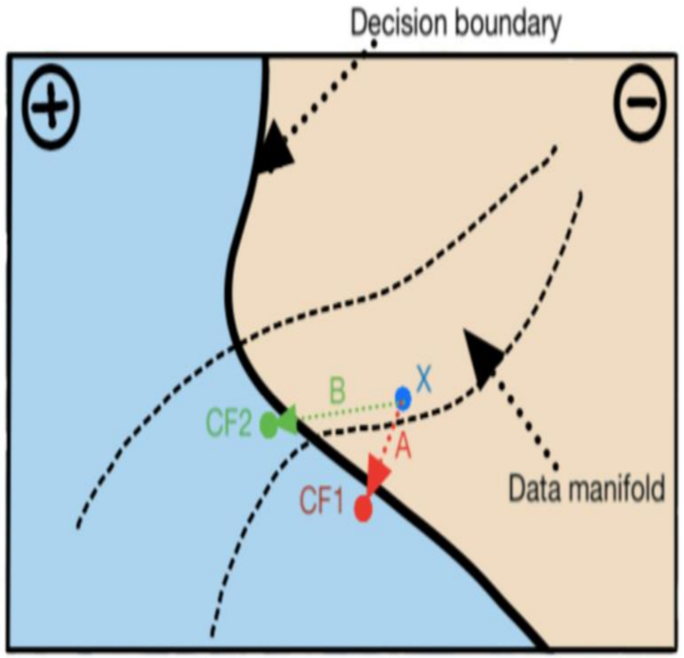
Counterfactual Explanations for Machine Learning: A Review

Sahil Verma
University of Washington
Arthur AI
vsahil@cs.washington.edu

John Dickerson
Arthur AI
University of Maryland
john@arthur.ai

Keegan Hines
Arthur AI
keegan@arthur.ai

arXiv, 2020



(Verma et al., 2020)

Counterfactual Explanations

DHH @dhh

The @AppleCard is such a **** sexist program. My wife and I filed joint tax returns, live in a community-property state, and have been married for a long time. Yet Apple's black box algorithm thinks I deserve 20x the credit limit she does. No appeals work.

Traduci il Tweet

9:34 PM · 7 nov 2019 · Twitter for iPhone

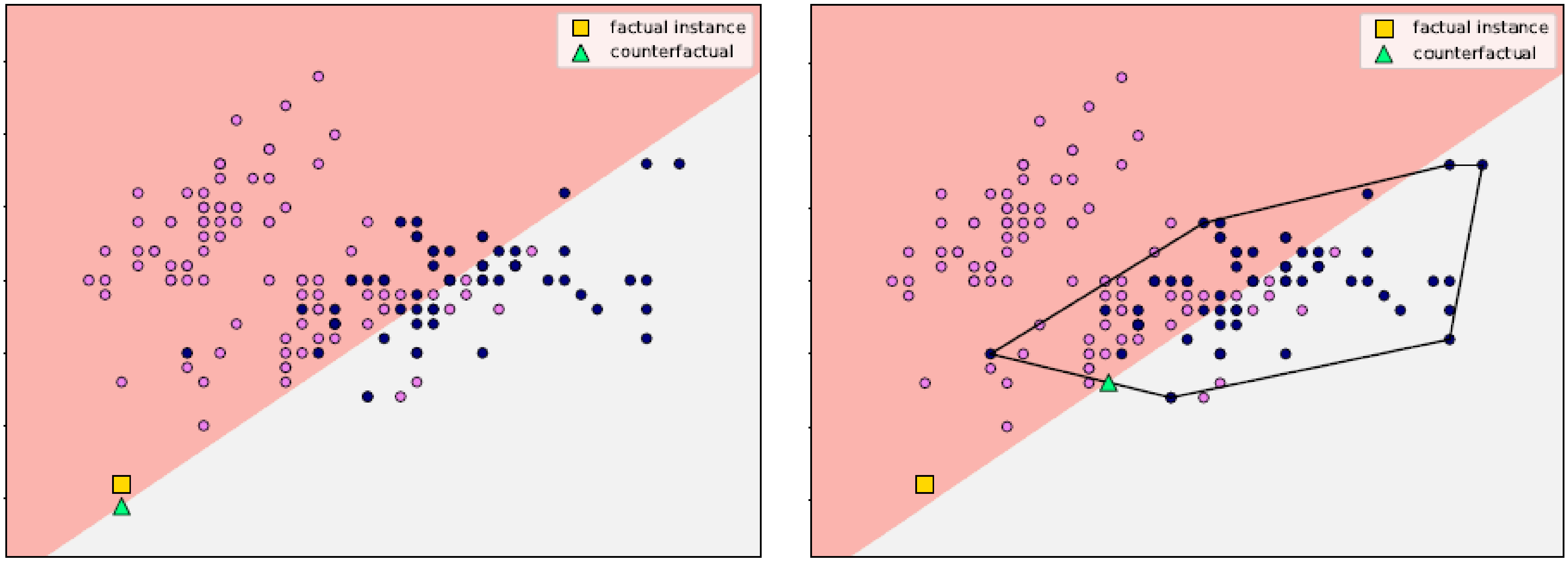
Apple Card @AppleCard

Well, if your wife would have had a 20+ years relationship with our bank, and would have been regarded as Premium customer at some point in time, she would also receive a 20x credit limit.

Apple Card @AppleCard

Well, if your wife's relationship status would have been "husband" instead of "wife", she would also receive a 20x credit limit.

We clearly messed up, we are updating our models now.



(Birbil et al. 2021)