Interpretability of ML Models

Michal Medek, innovation@socialbakers.com

June 2019

Outline

of Presentation

- Socialbakers
- What is it interpretability?
- Why to use it?
- Algorithms
- Frameworks
- Example

The Company

Offices

- **Pilsen**, **Prague**, Paris, London, Dubai, New York, Sao Paulo, Mexico City, Singapore, Berlin, Munich, Sydney

Employees

- 400+
    - 100+ engineering
    - 25+ data

Customers

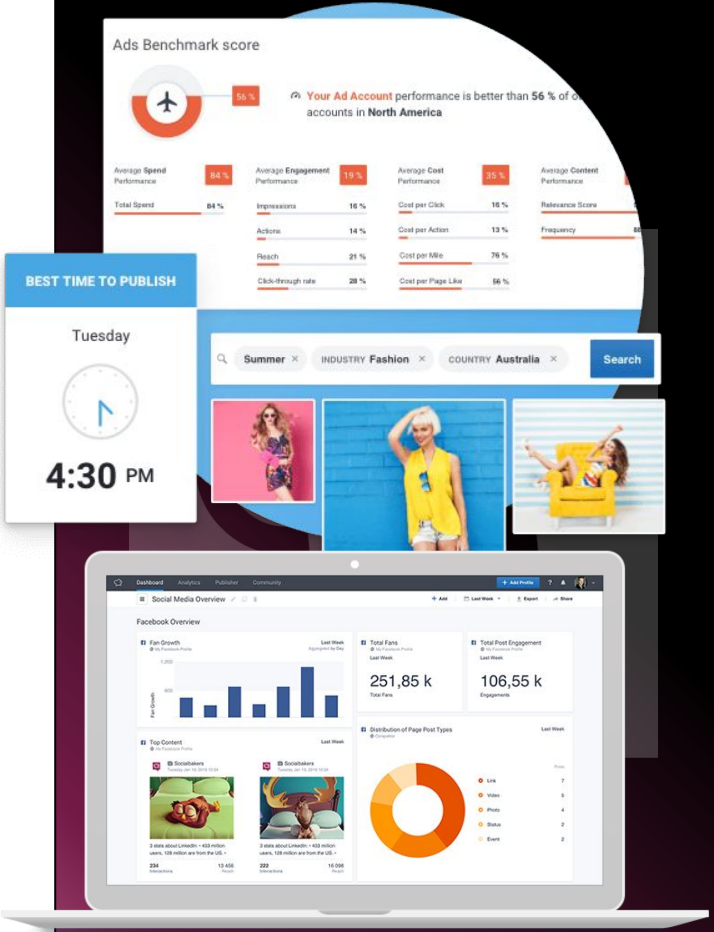
- 2500+

The Mission

Help brands work smart on social media through AI-powered marketing

Socialbakers Suite

- **Analytics**: performance analytics for profiles and content
- **Publisher & Community**: publishing and CRM tools
- **Audiences**: follower base analytics, personas identification
- **Influencers**: influencer discovery and recommendation
- **Inspiration**: content inspiration and recommendation



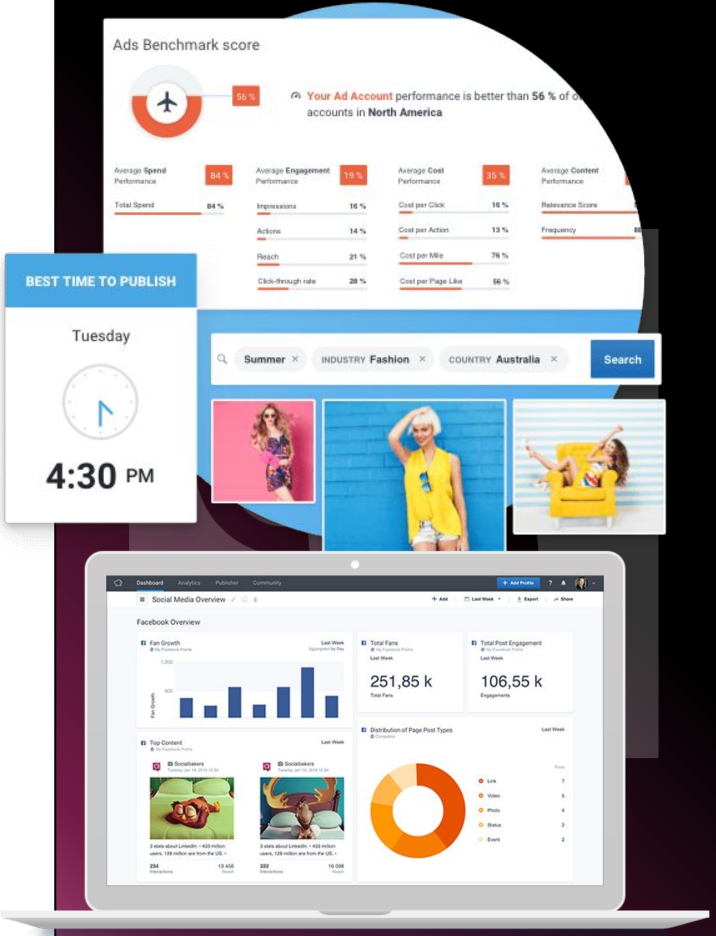
The Mission

Help brands work smart on social media through AI-powered marketing

Socialbakers Suite

- **Analytics**: performance analytics for profiles and content
- **Publisher & Community**: publishing and CRM tools
- **Audiences**: follower base analytics, personas identification
- **Influencers**: influencer discovery and recommendation
- **Inspiration**: content inspiration and recommendation

All in one place



The Team

Innovations

- 6 researchers, most heavily involved in AI and ML
- design **smart solutions** as core functionality of our products
- not alone – support from analysts, data engineers, taggers, …


Peter


Jan


Luboš

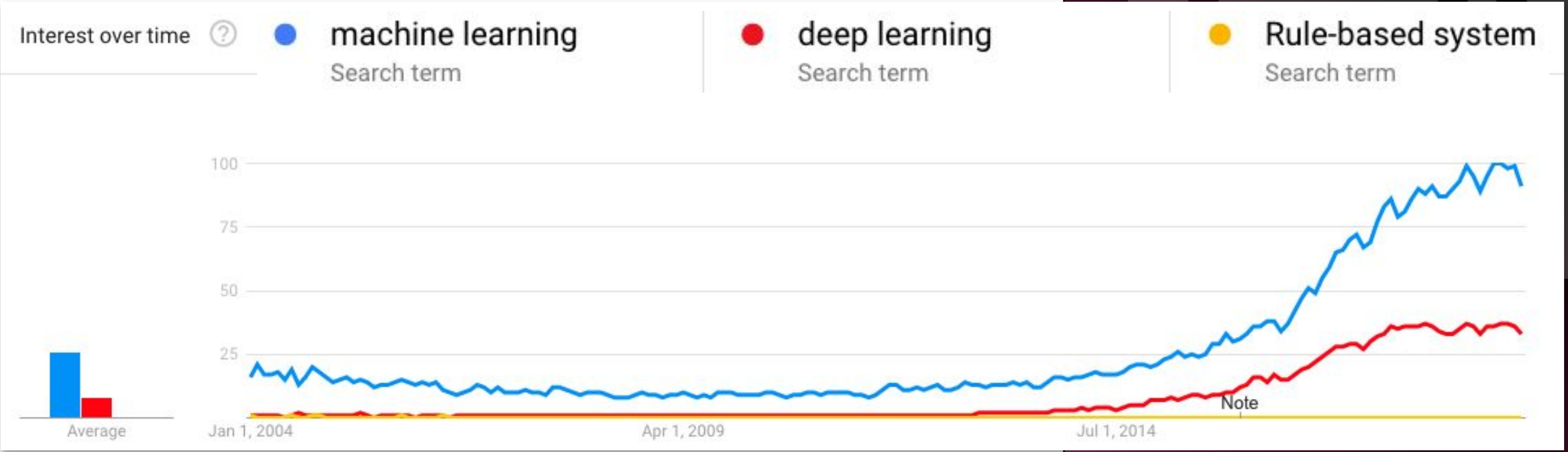

Michal


Jakub

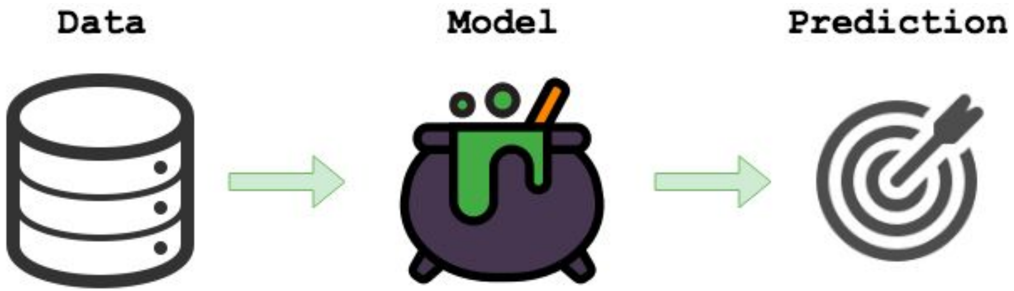

Paco

Introduction

- Machine learning in general widely known
- Slow industry adoption of new algorithms



Interpretability



The ability to explain or present in understandable terms to a human.*

- Been Kim, Finale Doshi-Velez

If you can't explain it simply, you don't understand it well enough.

- Albert Einstein

*Towards A Rigorous Science of Interpretable Machine Learning [link]

Interpretability

Player of the match prediction

Goal scored	2
Yellow cards	0
Corners	6
Attempts	3

Decision Tree

1	YES

If you can't explain it simply, you don't understand it well enough.

- Albert Einstein

Interpretability

Player of the match prediction

Goal scored	2
Yellow cards	0
Corners	6
Attempts	3

Decision Tree

1	YES

NOT INTERPRETABLE

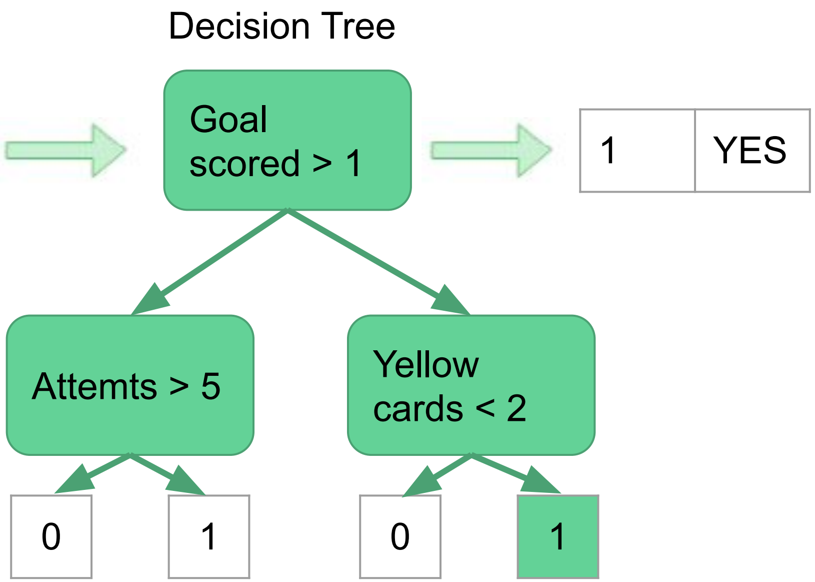
If you can’t explain it simply, you don’t understand it well enough.

- Albert Einstein

Interpretability

Player of the match prediction

Goal scored	2
Yellow cards	0
Corners	6
Attempts	3



If you can't explain it simply, you don't understand it well enough.

- Albert Einstein

Motivation

What can interpretability bring and help to solve:

1. Multiplicity problem
2. Performance vs Interpretability Trade-off
3. Avoiding bias / Fairness
4. Trust / Transparency

Motivation

Multiplicity of good models problem

- well-known datasets
- models with same performance and results based on metrics

Model 1	Model 2	Model 3
Accuracy: 71.5%	Accuracy: 70%	Accuracy: 71%

Sentiment analysis case

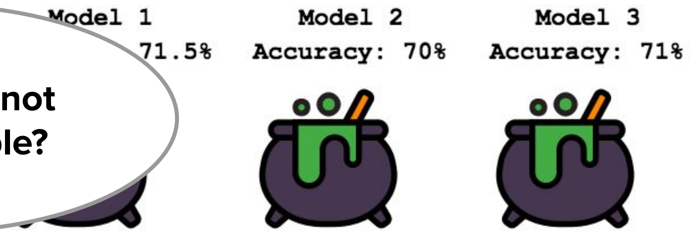
- GRU or LSTM stacked on embeddings with what hyperparameters?

- *Loss function*
- *Evaluation metrics*
- *Error analysis*
- **Visualize embeddings**
- **Visualize attention conn.**
- **Visualize activations**
- **Visualize filters**
- **Activation atlases**
- **Feature importance**
- .....

Motivation

Multiplicity of good models problem

- well-known datasets
- models with same performance and results based on metrics



Why just do not use ensemble?

Sentiment analysis case

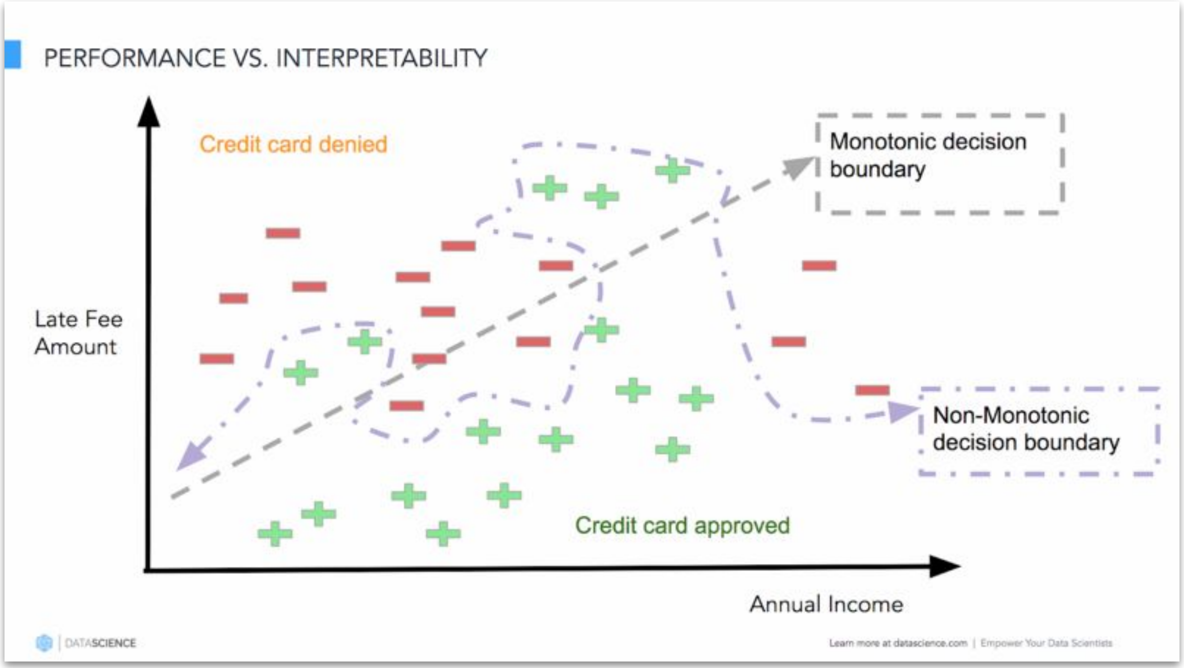
- GRU or LSTM stacked on embeddings with what hyperparameters?

Do you use ensembles in production?

- Visualize attention con
- Visualize activations
- Visualize filters
- Activation atlases
- Feature importance
- .....

Motivation

Performance vs Interpretability Trade-off



More accuracy

Less interpretability

Motivation

Avoiding bias / Fairness

- predicting potential criminals, judicial sentencing risk scores, credit scoring, fraud detection, health assessment, loan lending and more*

Artificial Intelligence's White Guy Problem**

- Kate Crawford, The New York Times

Sentiment analysis case

- Context problem - Sparta vs. Slavia

*Human Interpretable Machine Learning [link] **Artificial Intelligence's White Guy problem [link]

Motivation

Trust / Transparency

- semi-automatic
- banking, healthcare
    - explanation of predictions

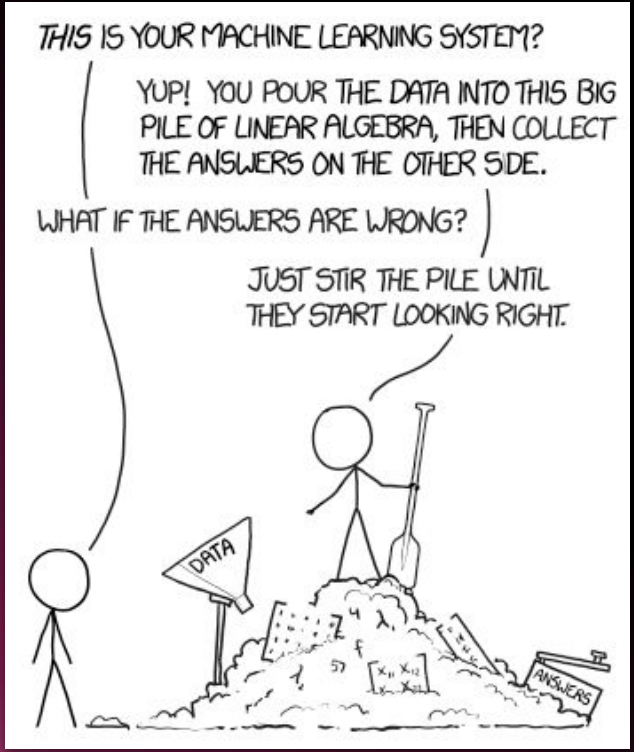
After release of semi-automatic approach was evaluation time increased to 150%* .

- Srivatsan Santhanam, SAP

Promoted Post Detection case

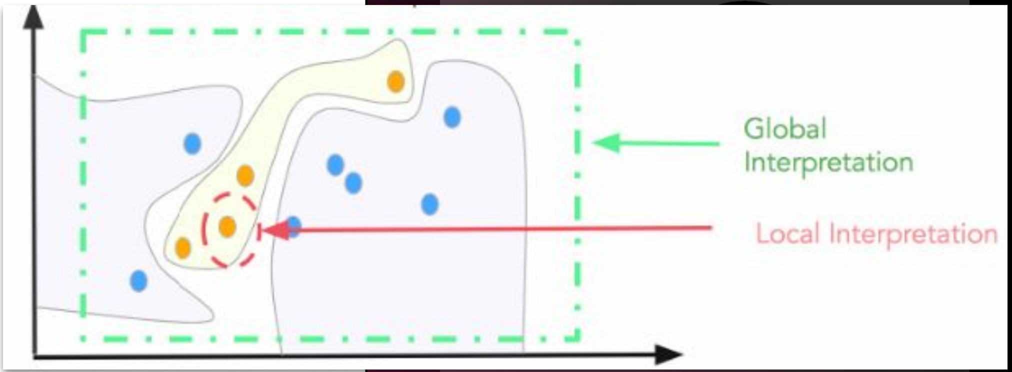
- still using manually created decision tree

* Machine learning: Explainability with anti-models, Machine Learning Prague 2018



Algorithm Categorization

- Self-Interpretability
  - Intristic (inner) - white boxes
    - be careful about:
      - multicollinearity
      - feature importance*
  - **Post hoc** - black boxes
- Approach
  - Model specific
  - **Model agnostic**
- Level
  - **Local** - great for error analysis
  - **Global**



* Beware Default Random Forest Importances [link]

Methods for Interpretability

- Permutation importance (Mean Decrease Accuracy)
- Partial Dependence Plots
- Surrogate models
- LIME
- SHAP

Methods for Interpretability

Permutation importance

- global
- for tabular data
- widely used and easy to understand
- method is suitable for dataset with smaller number of columns

Which features have the biggest impact on predictions

Weight	Feature
0.0750 ± 0.1159	Goal Scored
0.0625 ± 0.0791	Corners
0.0437 ± 0.0500	Distance Covered (Kms)
0.0375 ± 0.0729	On-Target
0.0375 ± 0.0468	Free Kicks
0.0187 ± 0.0306	Blocked
0.0125 ± 0.0750	Pass Accuracy %
0.0125 ± 0.0500	Yellow Card
0.0063 ± 0.0468	Saves
0.0063 ± 0.0250	Offsides
0.0063 ± 0.1741	Off-Target
0.0000 ± 0.1046	Passes
0 ± 0.0000	Red
0 ± 0.0000	Yellow & Red
0 ± 0.0000	Goals in PSO
-0.0312 ± 0.0884	Fouls Committed
-0.0375 ± 0.0919	Attempts
-0.0500 ± 0.0500	Ball Possession %

	Accuracy
Goal scored	0.18

Methods for Interpretability

Permutation importance

- algorithm:
  a. shuffle validation values in single column
  b. evaluate results and subtract with baseline
  c. return data to original order
  d. repeat it *n* times for all features

Goal scored	Yellow cards	Ball possession	Attempts	Player of the match [Target]
1	1	51	7	1
0	0	67	17	0
2	0	45	12	1

Methods for Interpretability

Permutation importance

- algorithm:
    a. shuffle validation values in single column
    b. evaluate results and subtract with baseline
    c. return data to original order
    d. repeat it *n* times for all features

	Accuracy
Goal scored	0.18
Yellow cards	0.07
Ball possession	0.06
Attempts	-0.01

What features have the biggest impact on predictions

Goal scored	Yellow cards	Ball possession	Attempts	Player of the match [Target]
1	1	51	7	1
0	0	67	17	0
2	0	45	12	1

Methods for Interpretability

Partial Dependence Plot

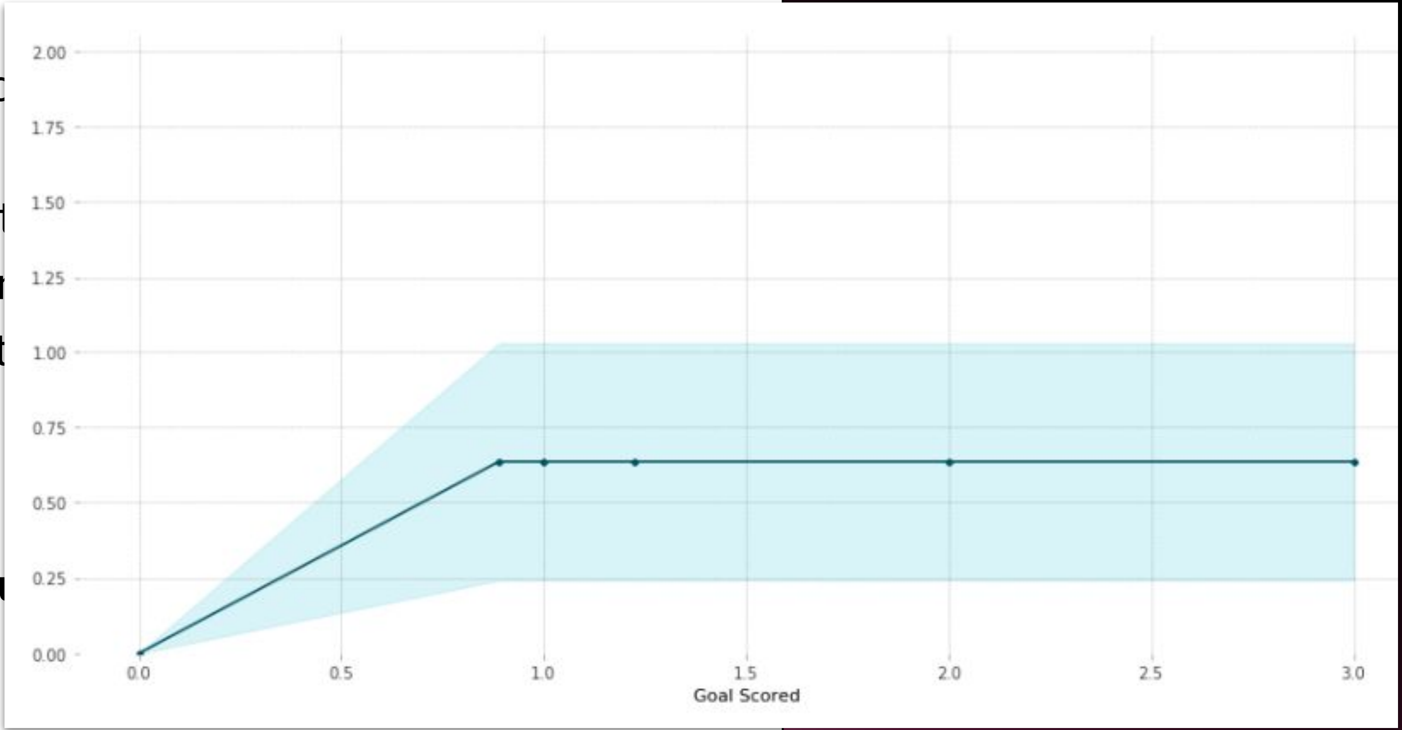
- global
- for tabular data
- should show relation between one or two features and target class

How feature(s) affects prediction

Methods for Interpretability

Partial Dependenc

- global
- for tabular dat
- should show r
  features and t



How featu

Methods for Interpretability

Partial Dependence Plot

- algorithm:
    a. freeze all values except selected feature
    b. select row
        - change values between min and max
        - evaluate results
        - repeat for <u>*n*</u> next row
    c. aggregate (mean, std) results for <u>*n*</u> rows

Methods for Interpretability

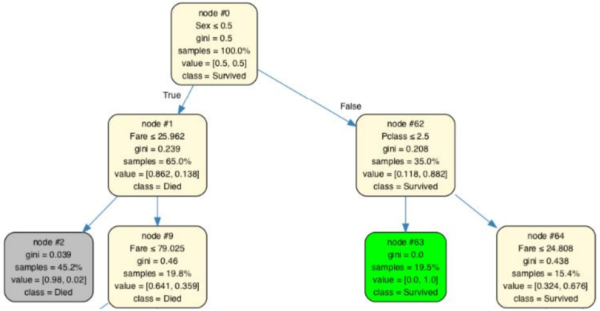
Surrogate models

- global
- white box model trained as approximation of black box model

Methods for Interpretability

Surrogate models

- algorithm:
    a. train surrogate model on predictions of your black box predictor on train data
    b. evaluate results
    c. visualize surrogate model if results are OK

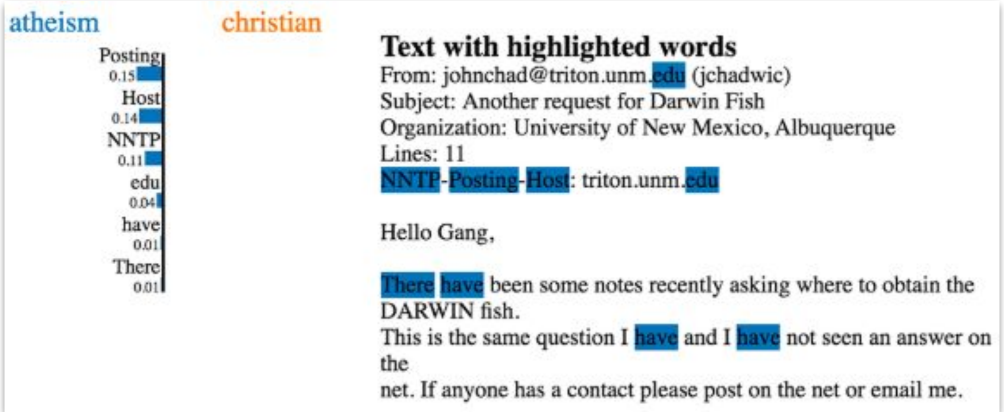


Methods for Interpretability

LIME (Local Interpretable Model-agnostic Explanations)

- local
- based on surrogate models

We use it for sentiment analysis

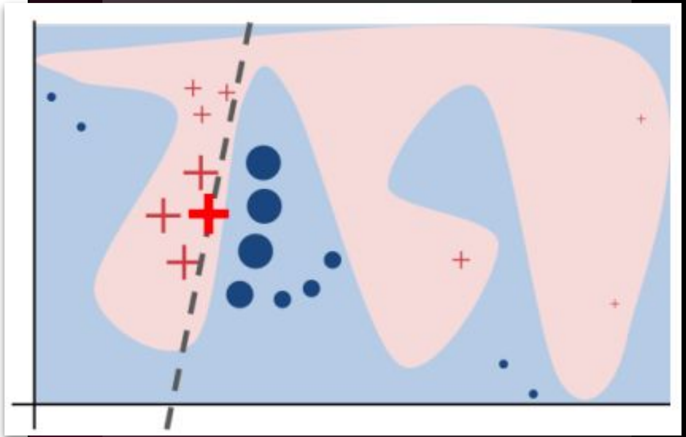


Why was the specific prediction made?

Methods for Interpretability

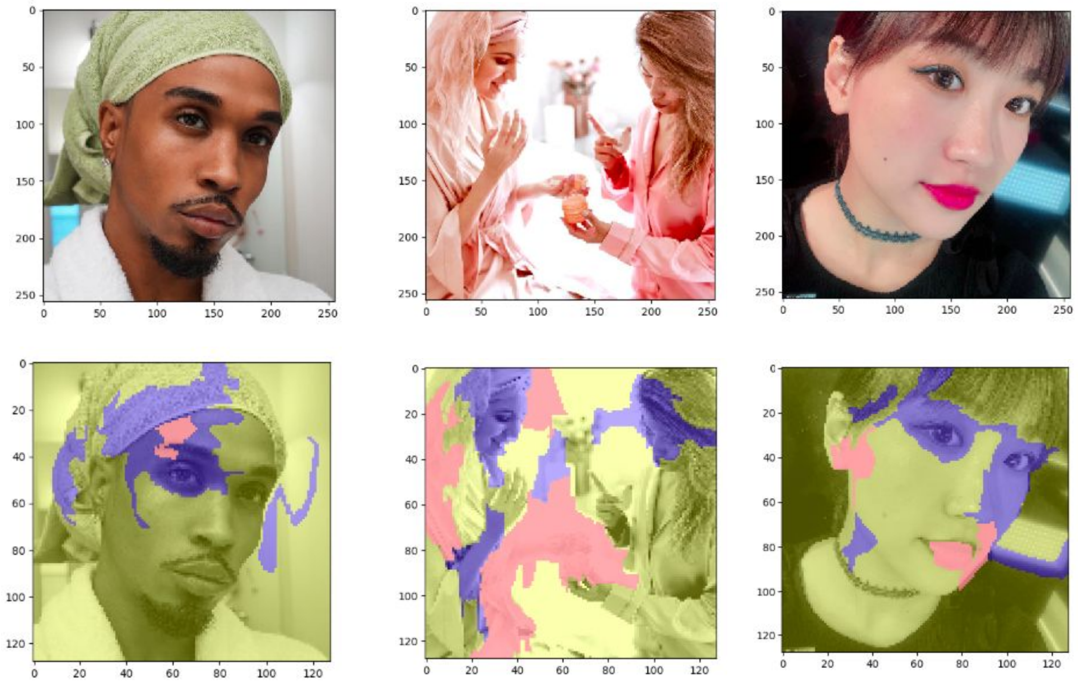
LIME

- algorithm:
    a. generate fake dataset for selected sample
    b. evaluate results with your model on fake dataset
    c. train surrogate model and weight it according to distance of generated samples
    d. evaluate result
    e. explain prediction based on weights of surrogate model



Methods for Interpretability

LIME on Socialbakers data



Methods for Interpretability

SHAP (SHapley Additive exPlanations)

- local
- for tabular data
- extension of Shapley values from game theory
- theoretically optimal

Proofs from game theory show this is the only possible consistent approach

Methods for Interpretability

Shapley values

- 3 entities:
  - game - specific ML task
  - player - specific feature
  - gain/payout - specific prediction
- players cooperate in coalitions
- we compute payout for each coalition (with and without currently fixed feature)

Shapley value:

How to fairly distribute the payout among the players.

Frameworks

- ELI5
- SHAP
- Skater (forked original LIME)

Frameworks

ELI5

- allow visualise weights for white box models
- permutation importance, LIME for texts and tabular data
- supports: scikit-learn, XGBoost, CatBoost, etc.

Frameworks

SHAP

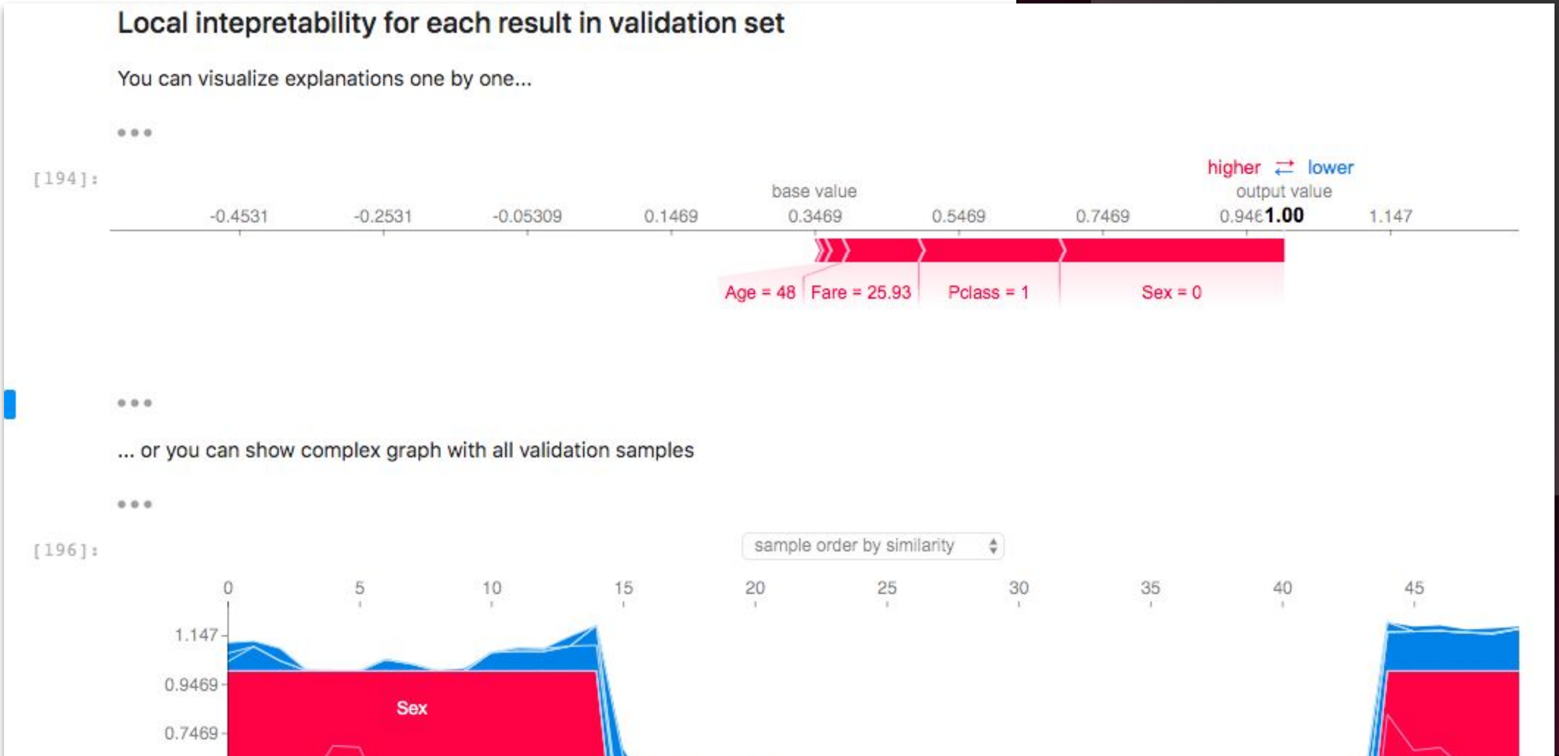
- based on Shapley values
- local explanations
- dependence plots, explanation plots, summary plots

Frameworks

Skater

- forked from LIME
- unified framework for interpretability
- feature importance, dependence plot, LIME, surrogate models

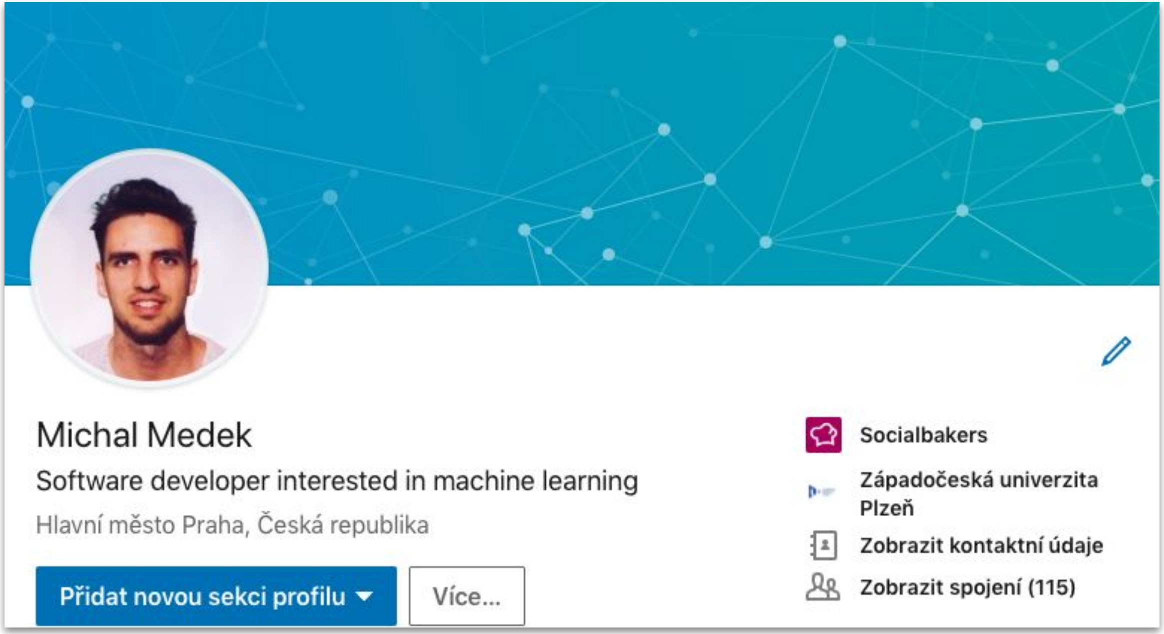
Quick example



Takeaways

- **interpret your models**
- be careful about:
  - multicollinearity of your models
  - random forest feature importance
  - LIME in connection with highly complex data (you can check r-square error)
- use eli5 for explaining white models
- use SHAP and Skater for more complex models

In case of interest for Jupyter notebook/presentation or any questions you can contact me know on LinkedIn [link]



Thank you for attention!

Sources

First part of amazing serie of 3 blog posts about ML Interpretability which I used a lot [link]
Great free book about ML Interpretability [link]
Article about bias from NYT [link]
LIME paper [link]
SHAP paper [link]
Article about Performance vs. Interpretability trade-off [link]
ELI5 documentation with summary about LIME and PI [link]
Short intro to Machine Learning Interpretability [link]
Summary from Kaggle Micro-course to ML Interpretability [link]
Kaggle ML Interpretability micro-course [link]
Nice talk from PyData 2018 about ML Interpretability [link]
Activation atlases paper [link]
Talk about LIME [link]

Sources

Introduction to Skater [link]

Helpers

Example how LIME works on images [link]

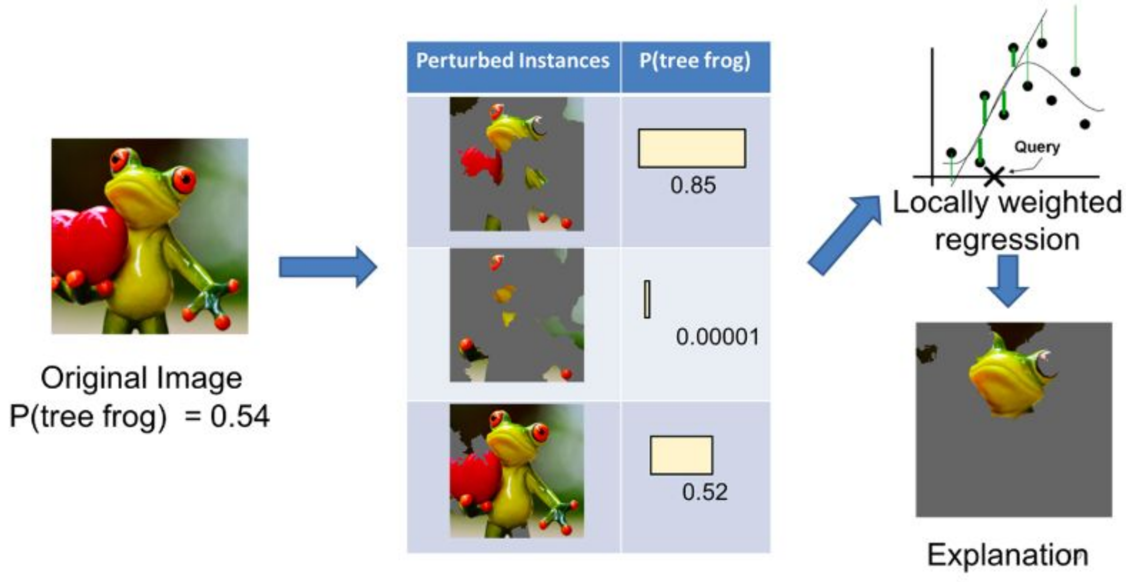


Figure 4. Explaining a prediction with LIME. Sources: Marco Tulio Ribeiro, Pixabay.