SEON

Explainable AI in Fraud Prevention

“**Magyarázható Mesterséges Intelligencia** a csalás megelőzésben”

David Utassy,
ML Guild Leader,
Data Scientist @ SEON

Program

Interpretálhatóság - Értelmezhetőség - Magyarázhatóság

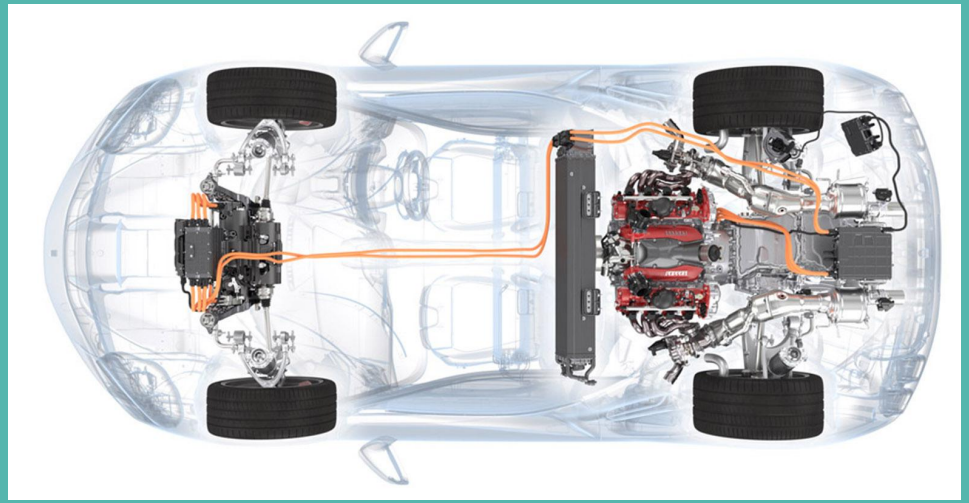
SEON - Csalás detektálás ML-el

SHAP - Esettanulmány

Machine learning ?=

Machine learning ?=

+

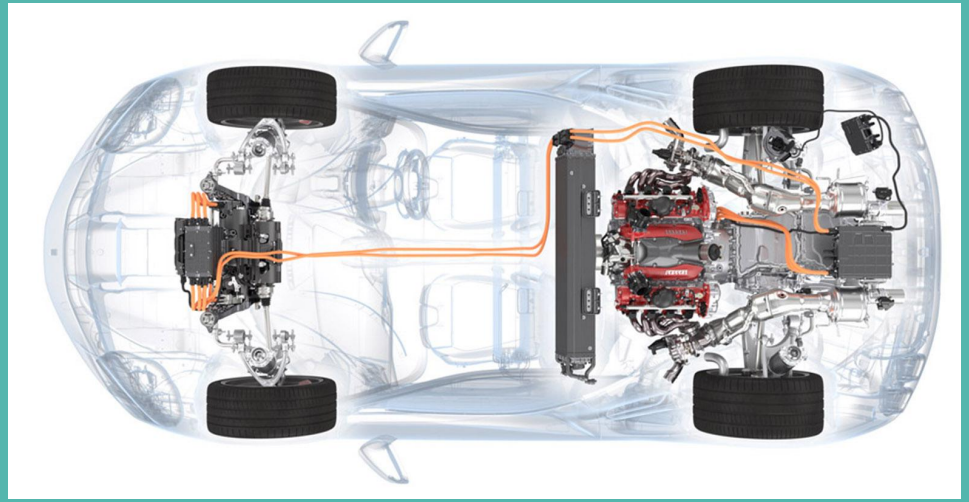


Machine learning ?=

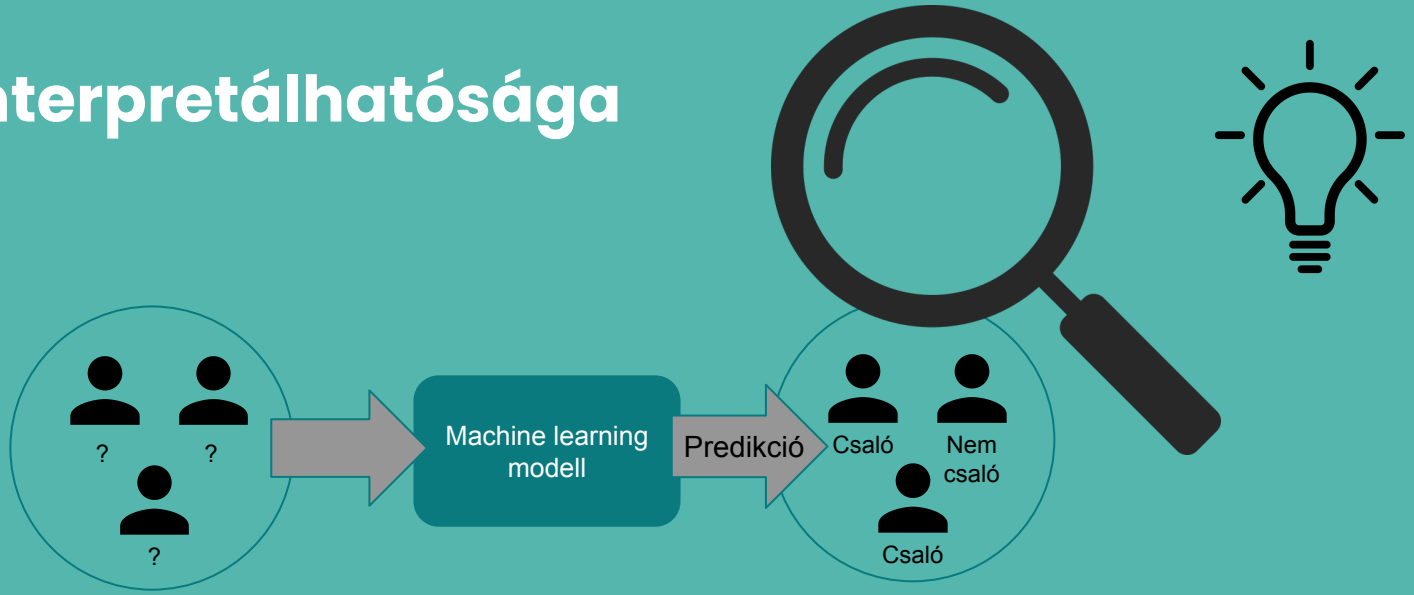
+

+

Interpretálhatóság



ML Interpretálhatósága



“Az interpretálhatóság annak a foka, hogy az ember milyen mértékben képes megérteni egy döntés okát”

[Miller, Tim- Explanation in artificial intelligence: Insights from the social sciences]

Interpretálhatóság

Hol van rá szükség?

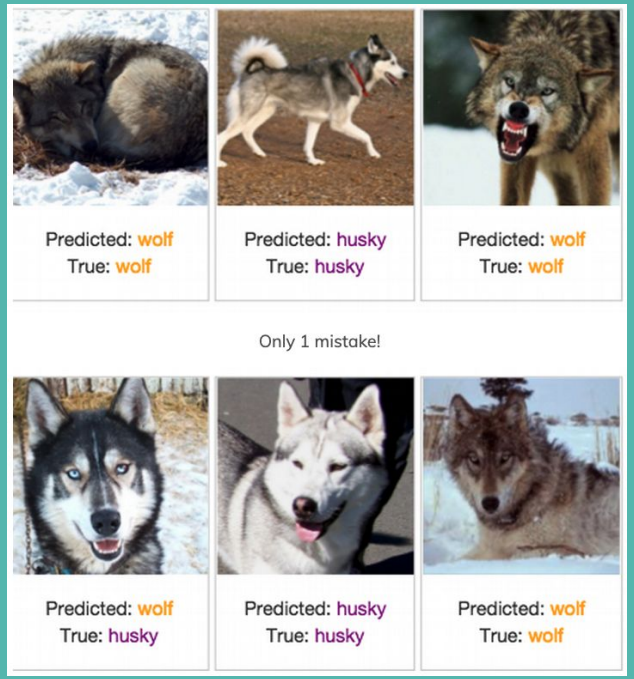
- Magas kockázatú környezetben 🚑🚗💰
- Jogi szabályozás 📃

Interpretálhatóság

Miért van rá szükség?

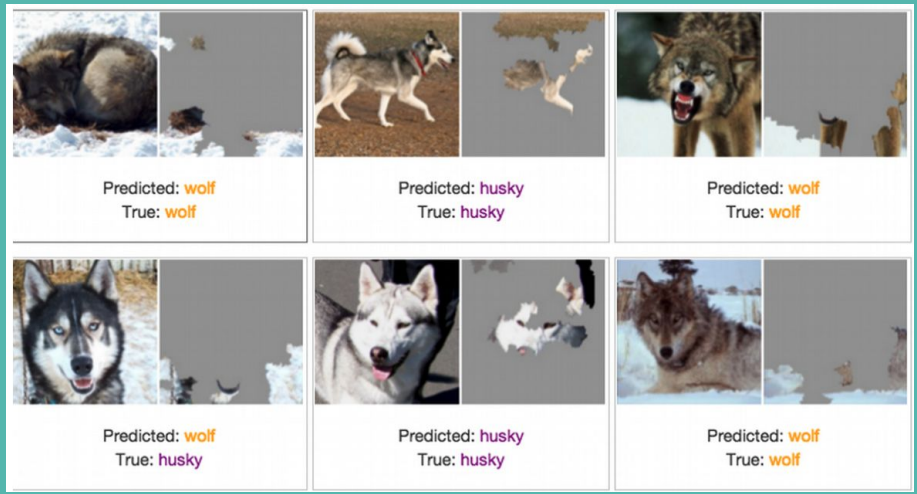
- ML elterjedéséhez KELL: értelmezhetőség
  👐💡👨‍👩‍👧
- Emberi kíváncsiság, tanulási ösztön
  😮 → 🤔 → 💡

AI Példa



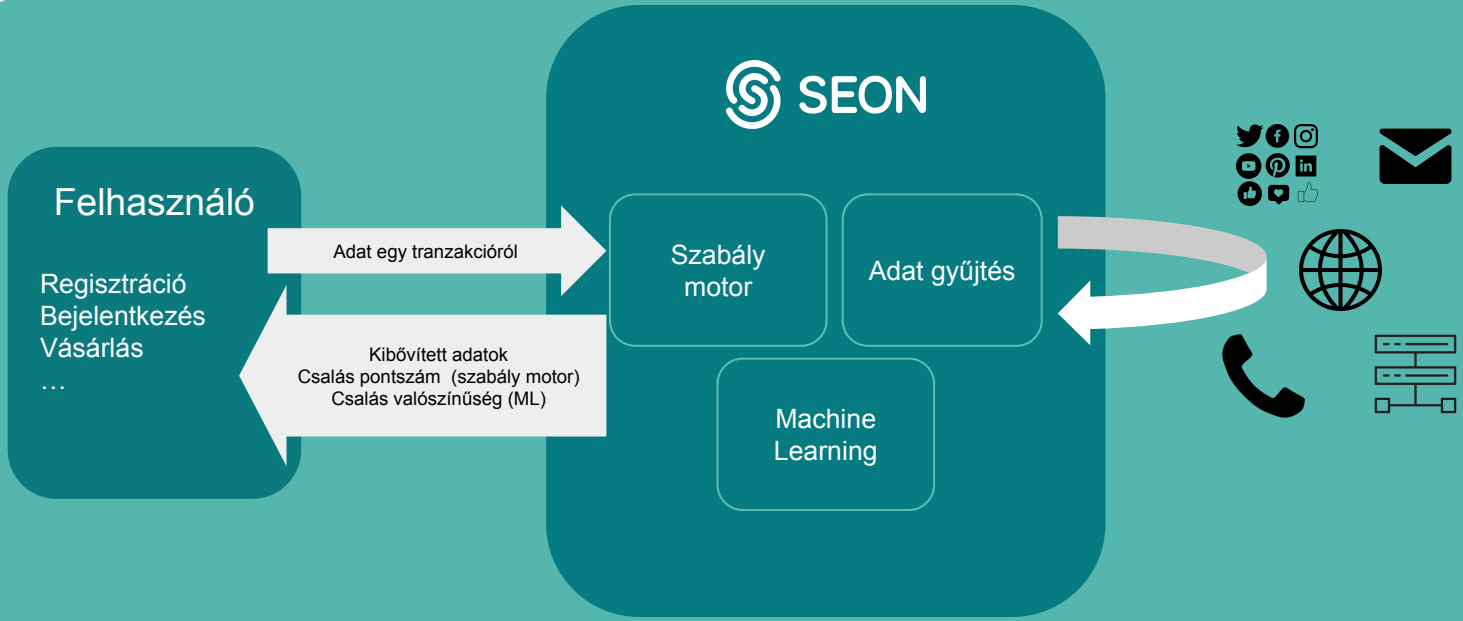
[Marco Tulio Ribeiro, arXiv:1602.04938v3]
https://github.com/marcotcr/lime

AI Példa



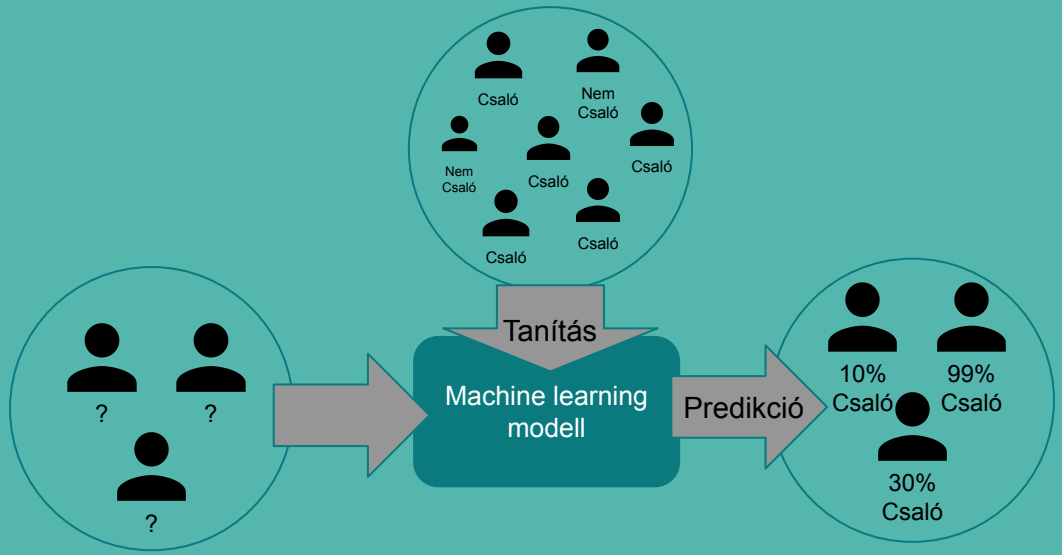
[Marco Tulio Ribeiro, arXiv:1602.04938v3]
https://github.com/marcotcr/lime

SEON

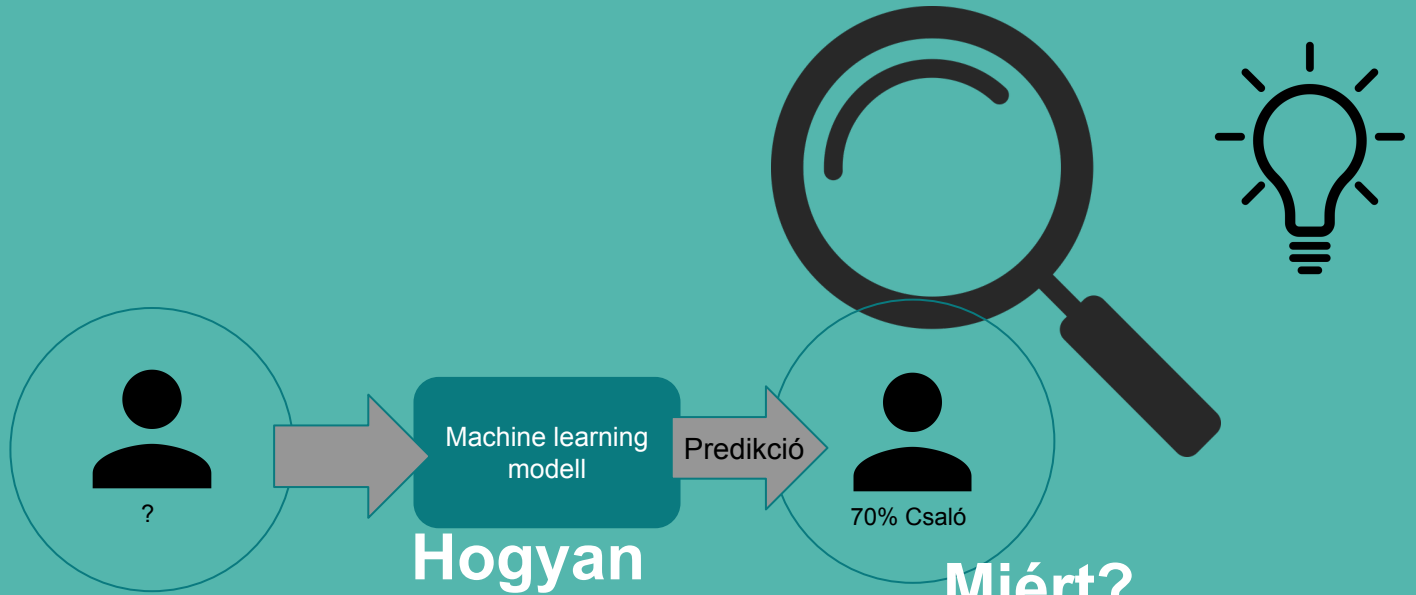

Felhasználó
Regisztráció
Bejelentkezés
Vásárlás
…
Adat egy tranzakcióról
Kibővített adatok
Csalás pontszám (szabály motor)
Csalás valószínűség (ML)
SEON
Szabály motor
Adat gyűjtés
Machine Learning

SEON

- Felügyelt tanítás
- Táblázatos adaton
- Gradient Boosting alapú modell



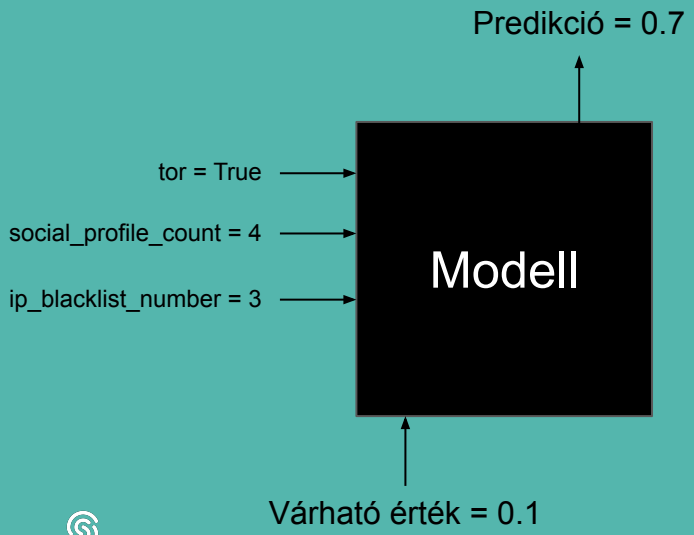
Interpretálhatóság



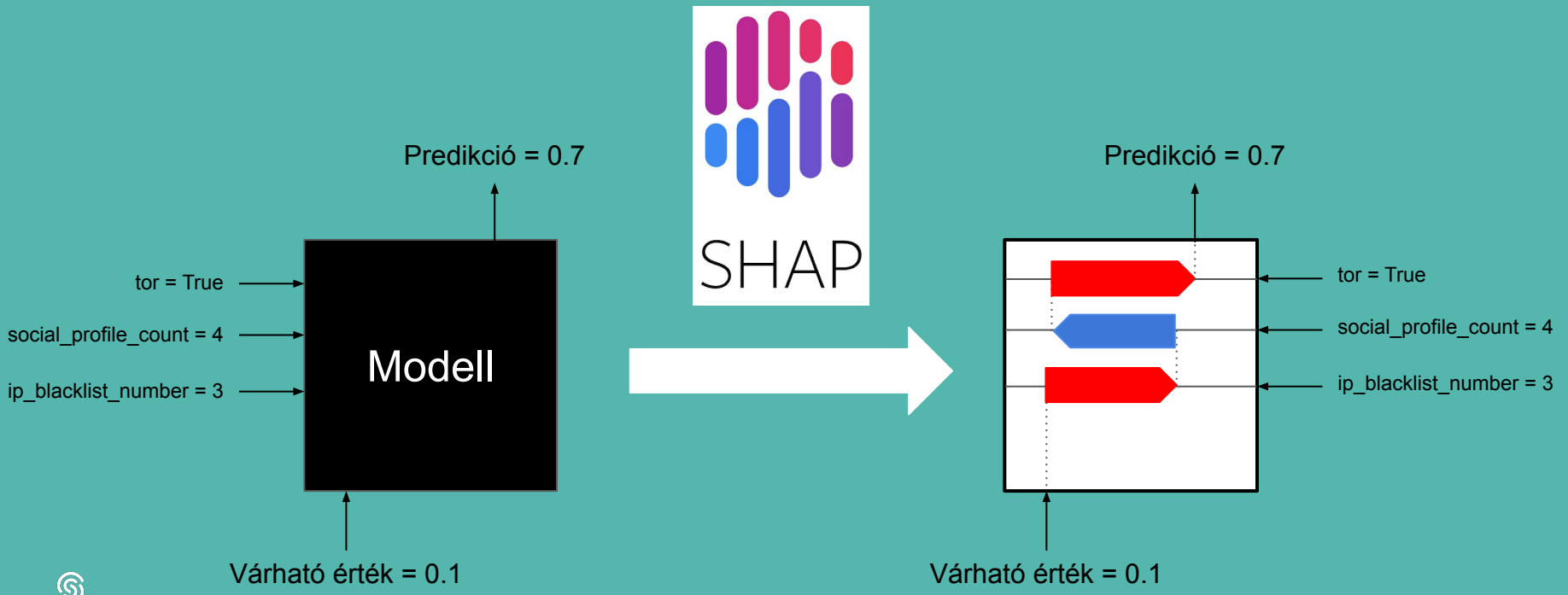
Hogyan viselkedik a modell? (globális)

Miért? (lokális)

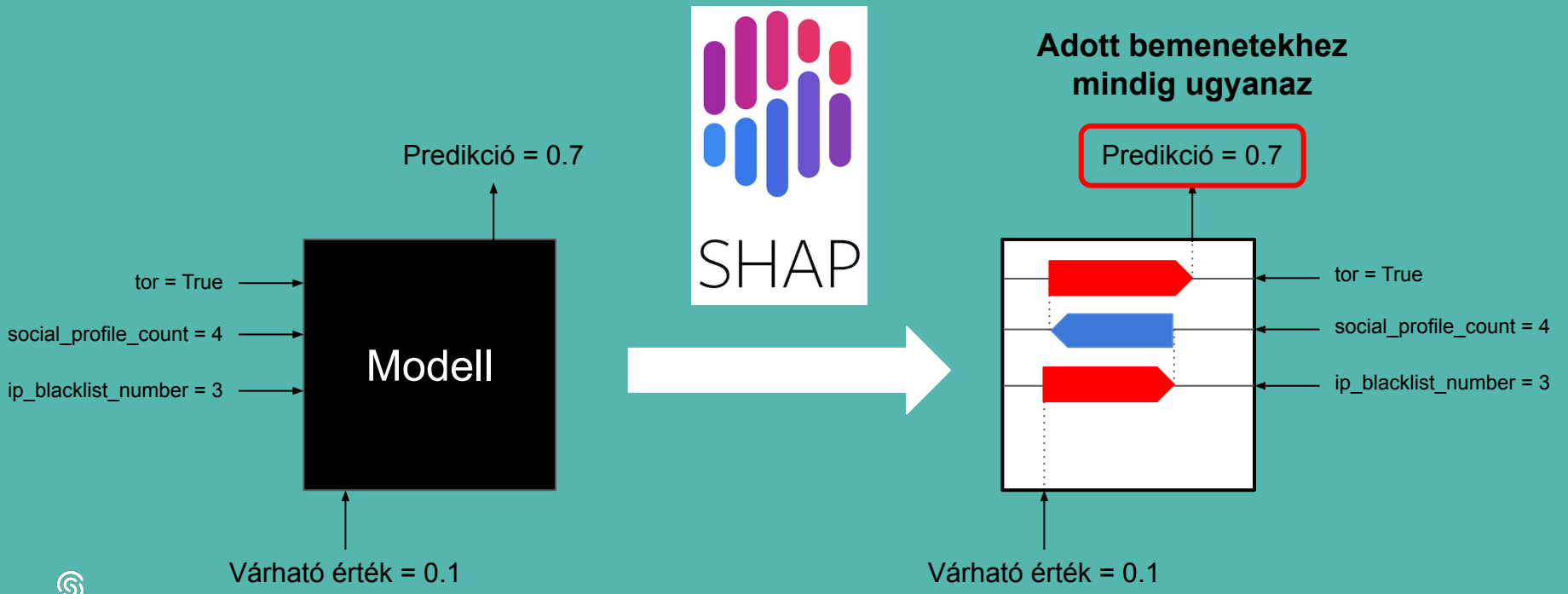
Lokális magyarázhatóság



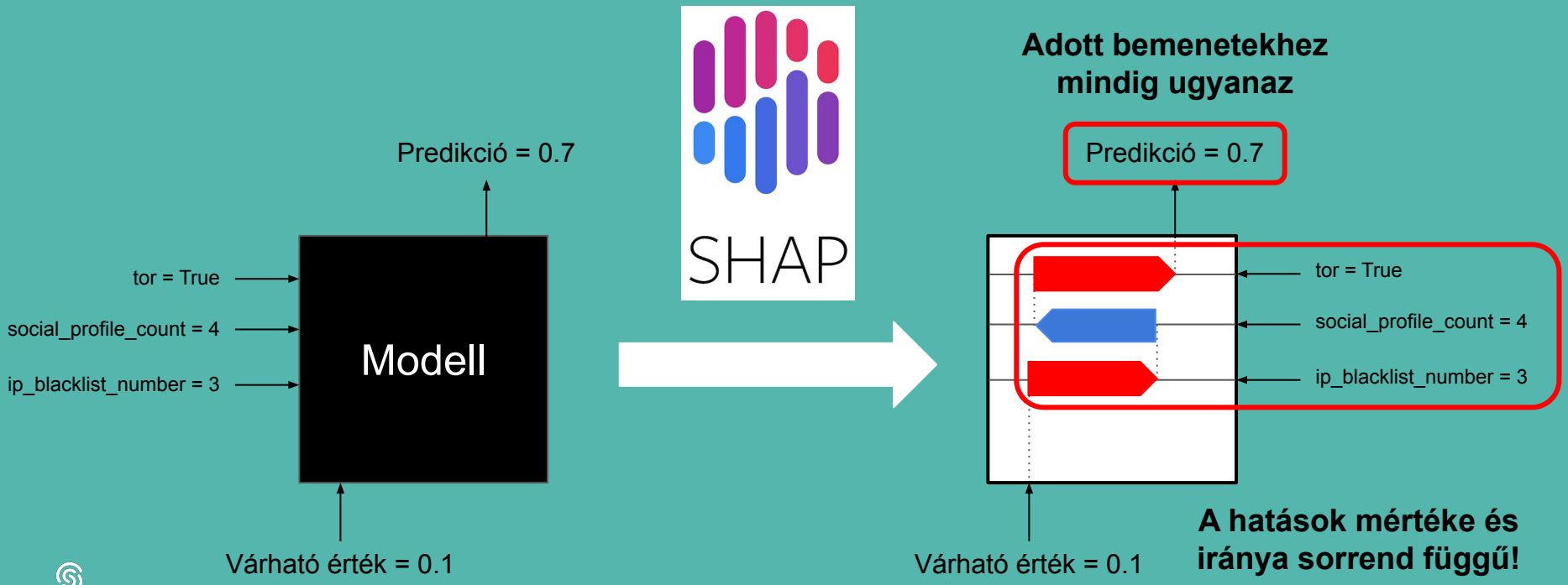
Lokális magyarázhatóság



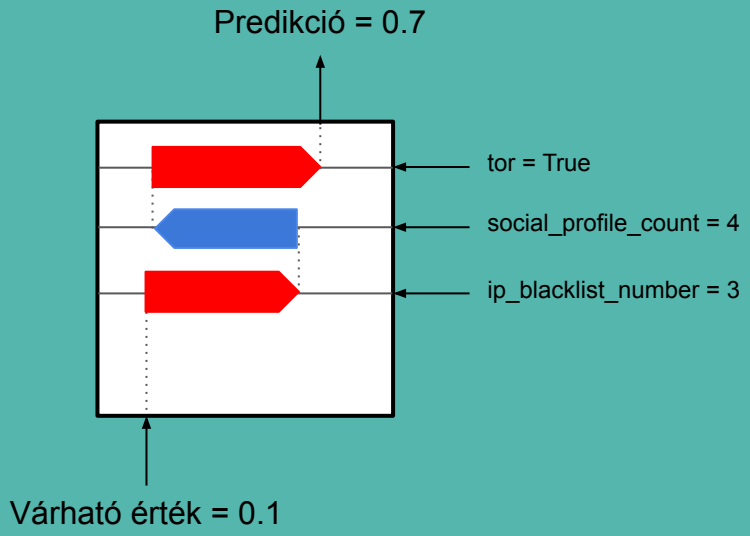
Lokális magyarázhatóság



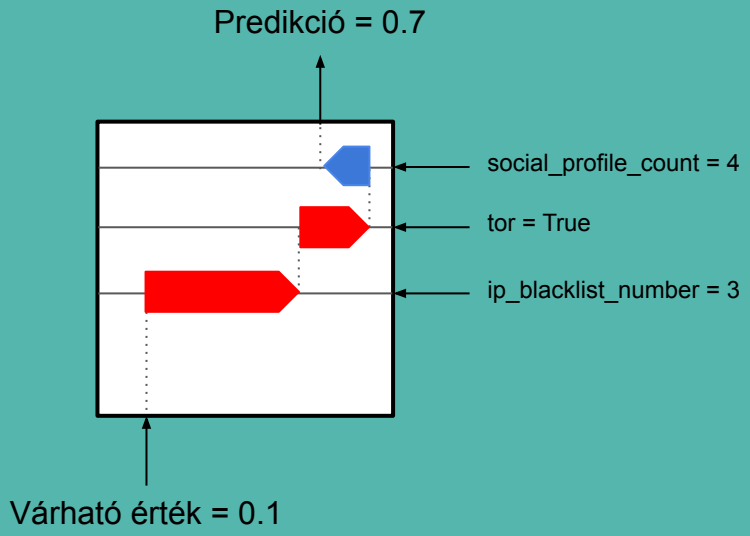
Lokális magyarázhatóság



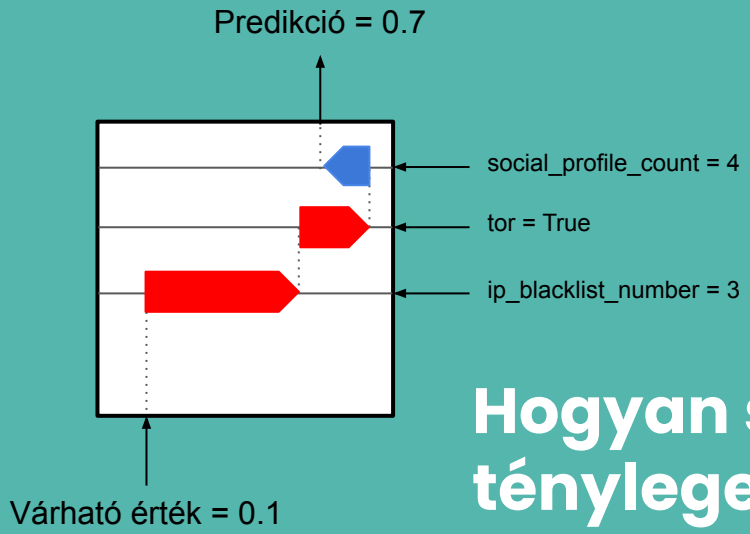
Lokális magyarázhatóság



Lokális magyarázhatóság



Lokális magyarázhatóság



Hogyan számolható a tényleges hatás?

Shapley-értékek

Az egyetlen mód, hogy igazságosan osszuk el a kimenetre gyakorolt hatást a bemeneti változók között

Lloyd Shapley - 2012 Közgazdasági Nobel-díj
- 1950-es évek, játékelméleti kutatások

[Scott M. Lundberg, Su-In Lee - 2017]
- Shapley értékek → Machine Learning

SHAP

SHAP Package
[github.com/slundberg/shap]

Shapley-értékek

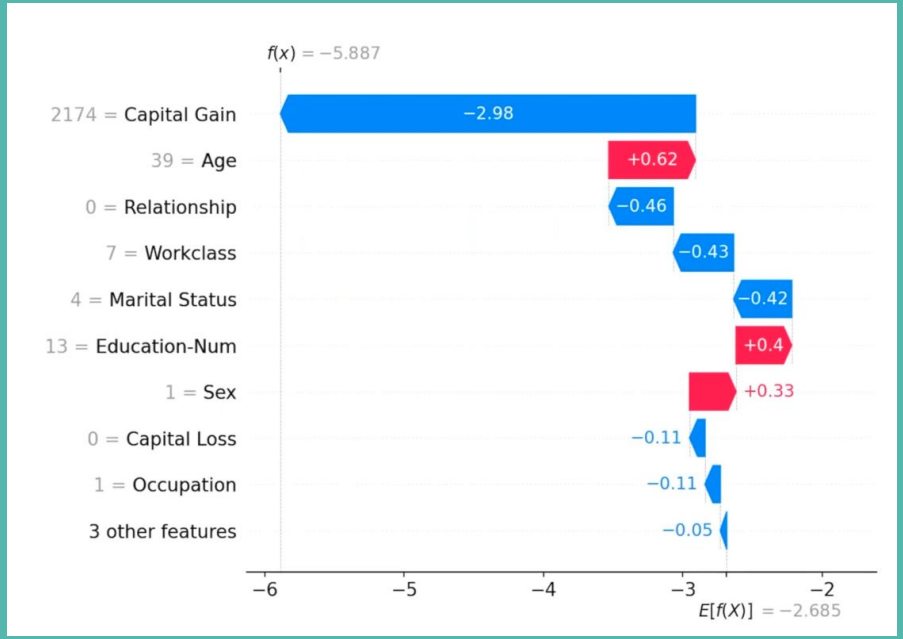
- Alapvetően exponenciális időben számolható
  - Egyébként csak közelítő értékek

- Tree Explainer (algoritmus) →
  - Polinomiális időben számolható!
  - Pontos értékek
  - [Scott M. Lundberg, 2020]
  - Támogatott: XGBoost, LightGBM, CatBoost, scikit-learn és pyspark fa alapú modellek

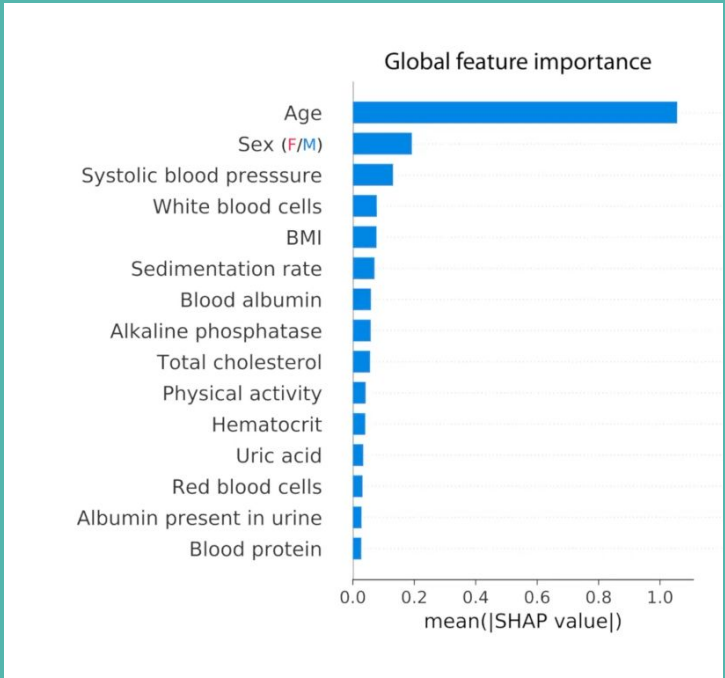
Nincs kompromisszum

Táblázatos adat → Gradient Boosting Modell→**SHAP TreeExplainer**

Lokális magyarázhatóság
[github.com/slundberg/shap]

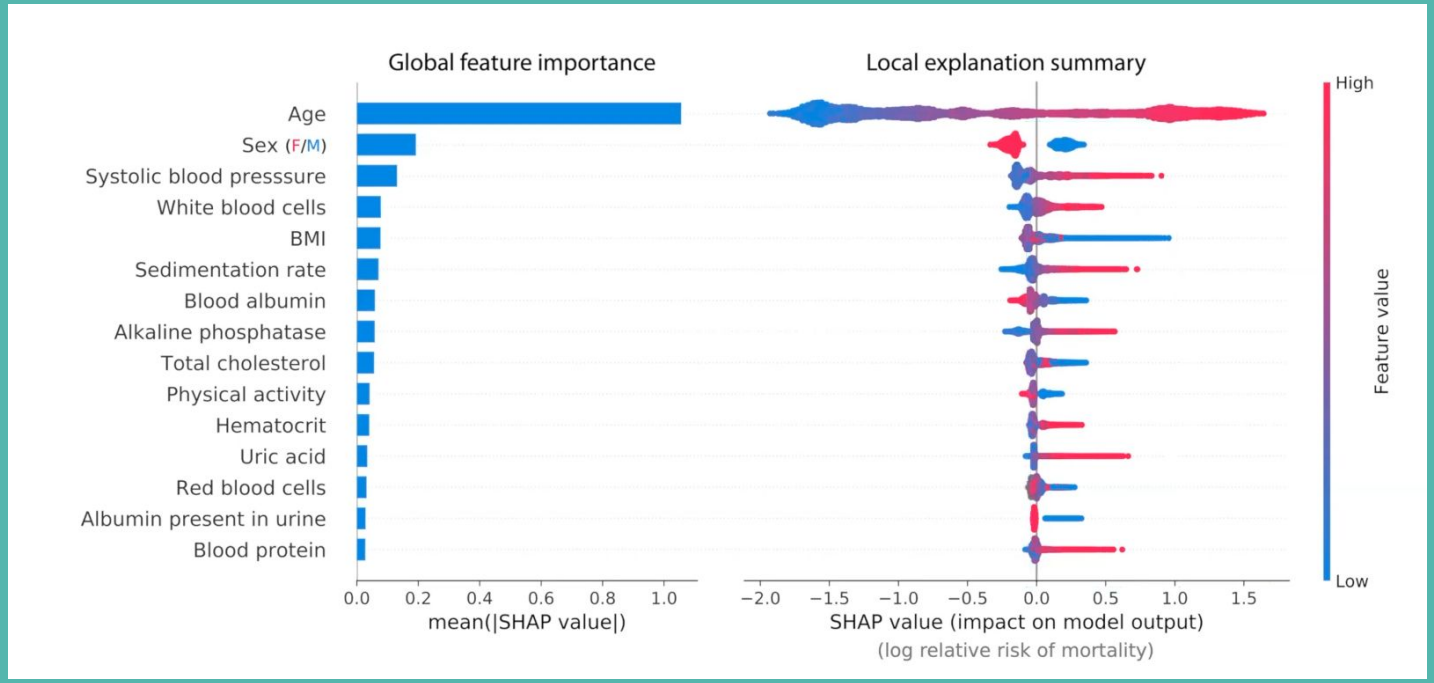


Globális magyarázhatóság [github.com/slundberg/shap]

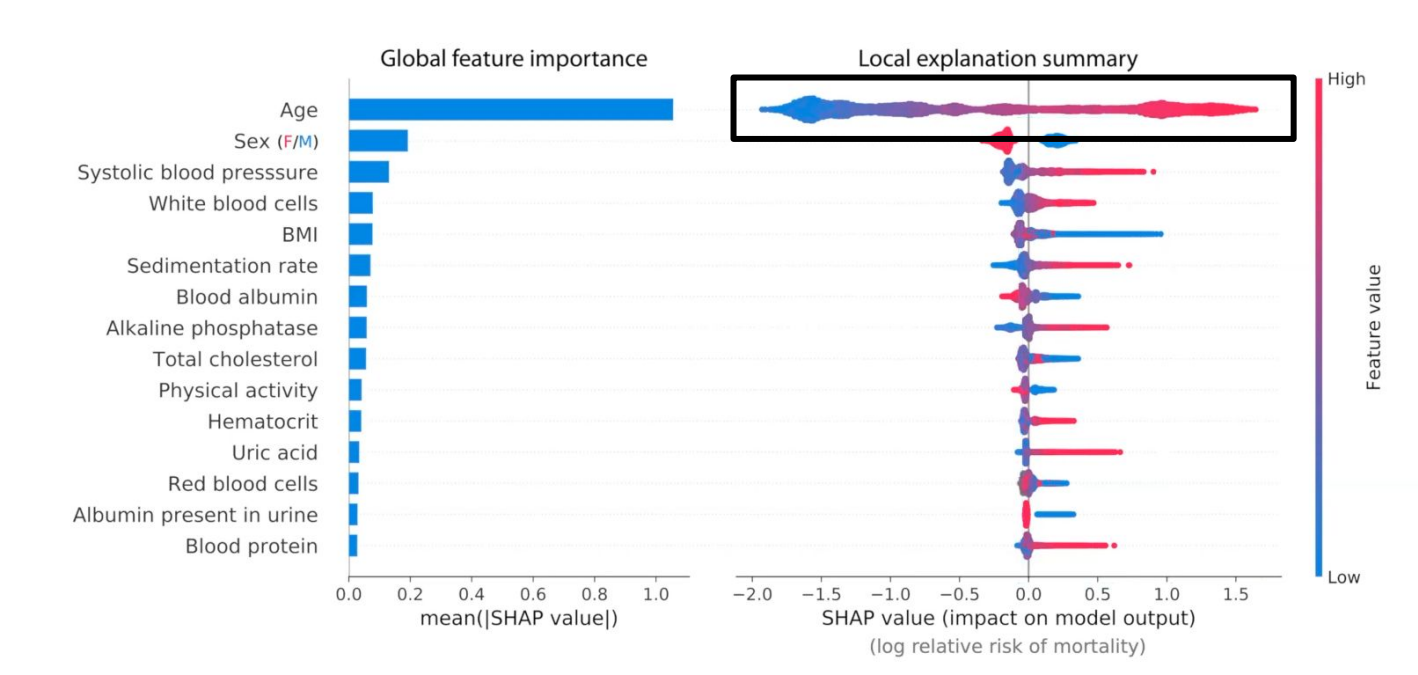


SHAP

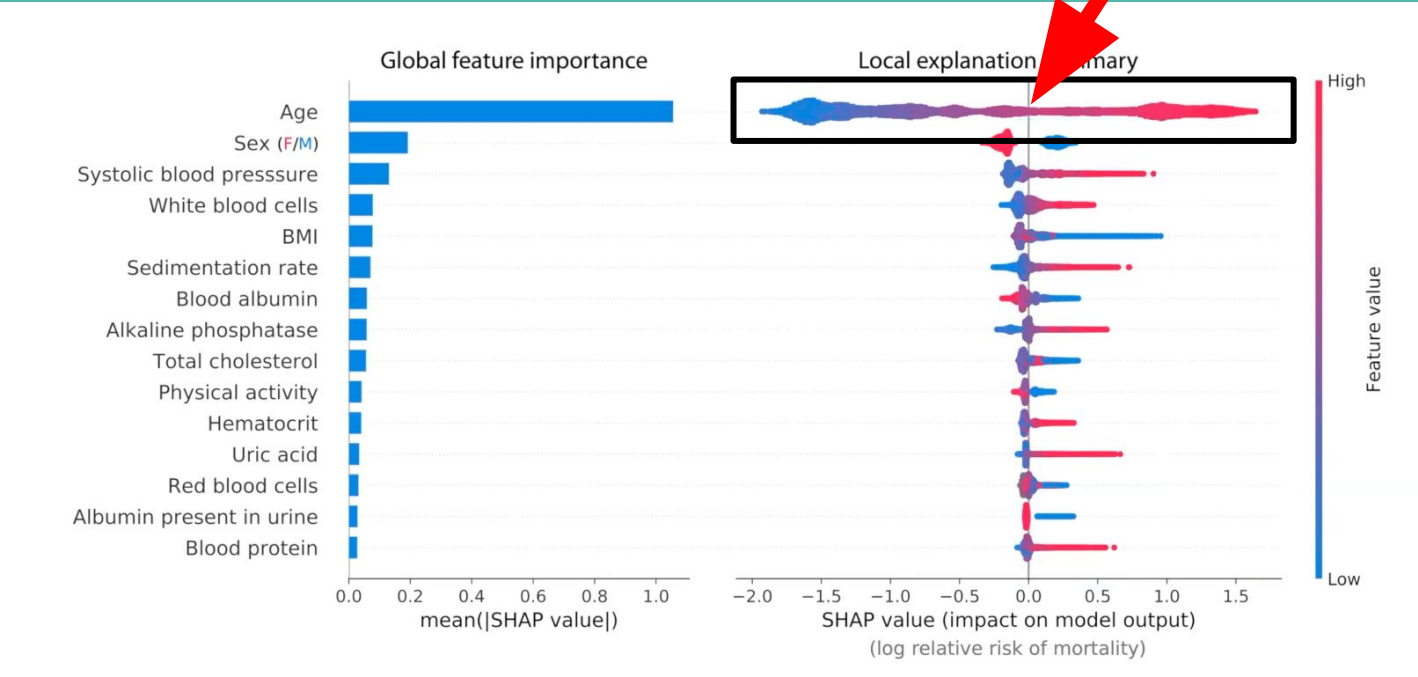
Globális magyarázhatóság [github.com/slundberg/shap]



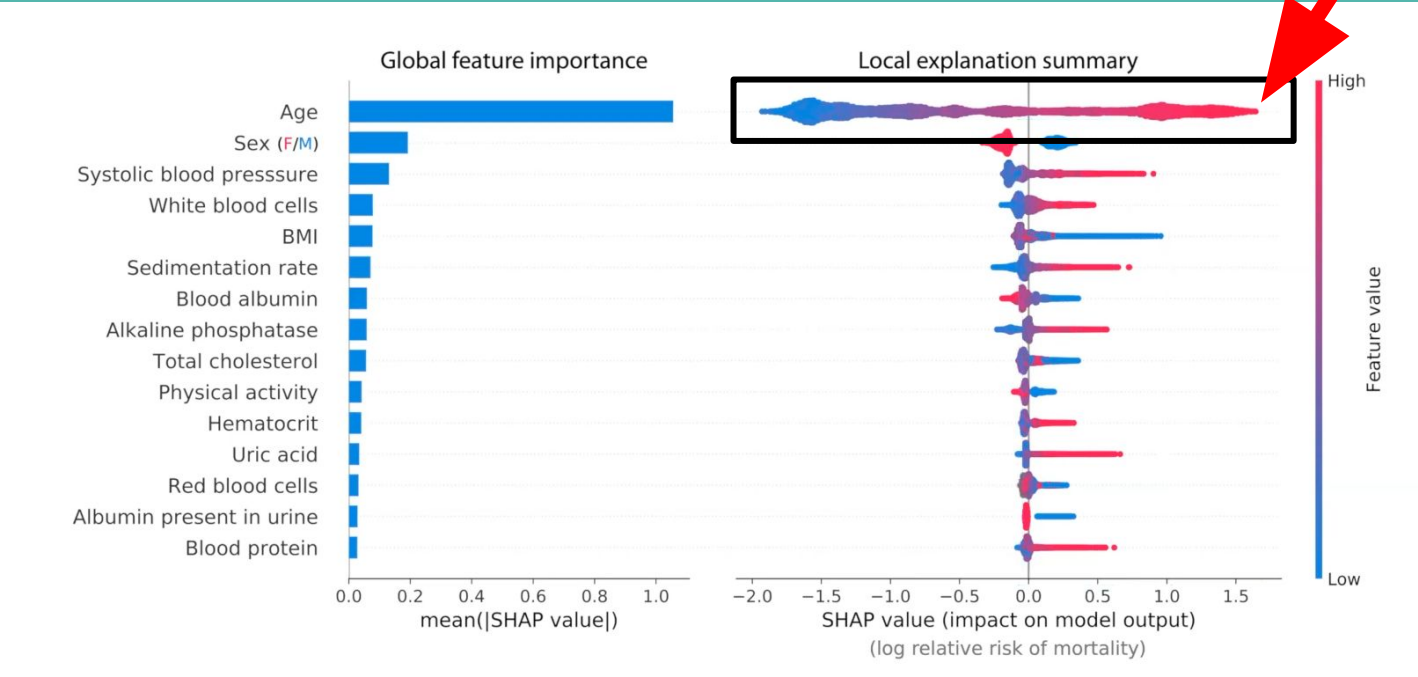
Globális magyarázhatóság
[github.com/slundberg/shap]



Globális magyarázhatóság [github.com/slundberg/shap]



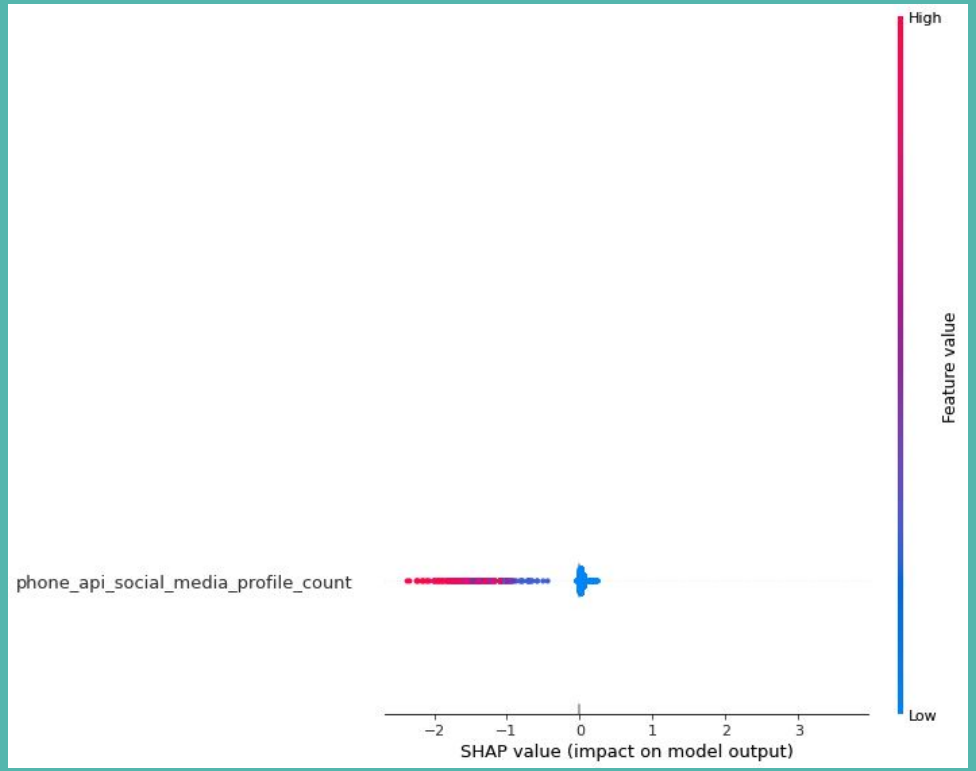
Globális magyarázhatóság [github.com/slundberg/shap]



SHAP

SEON

SHAP


High
Feature value
Low
phone_api_social_media_profile_count
−2
−1
0
1
2
3
SHAP value (impact on model output)

SEON

Mire használjuk?

- Modellek fejlesztéséhez
- Debuggolásra
- Bizalom növelésre
- Nem kívánt döntések elkerülésére

Összefoglalás

- ML interpretálhatósága fontos
- Fejlődő terület, érdemes követni. Beszéljünk róla sokat!
- Táblázatos adat →Gradient Boosting → SHAPTreeExplainer

SEON

Köszönöm a figyelmet!

David Utassy,
ML Guild Leader,
Data Scientist @ SEON