Decision tree, Random forest és LightGBM alkalmazása klasszifikációra

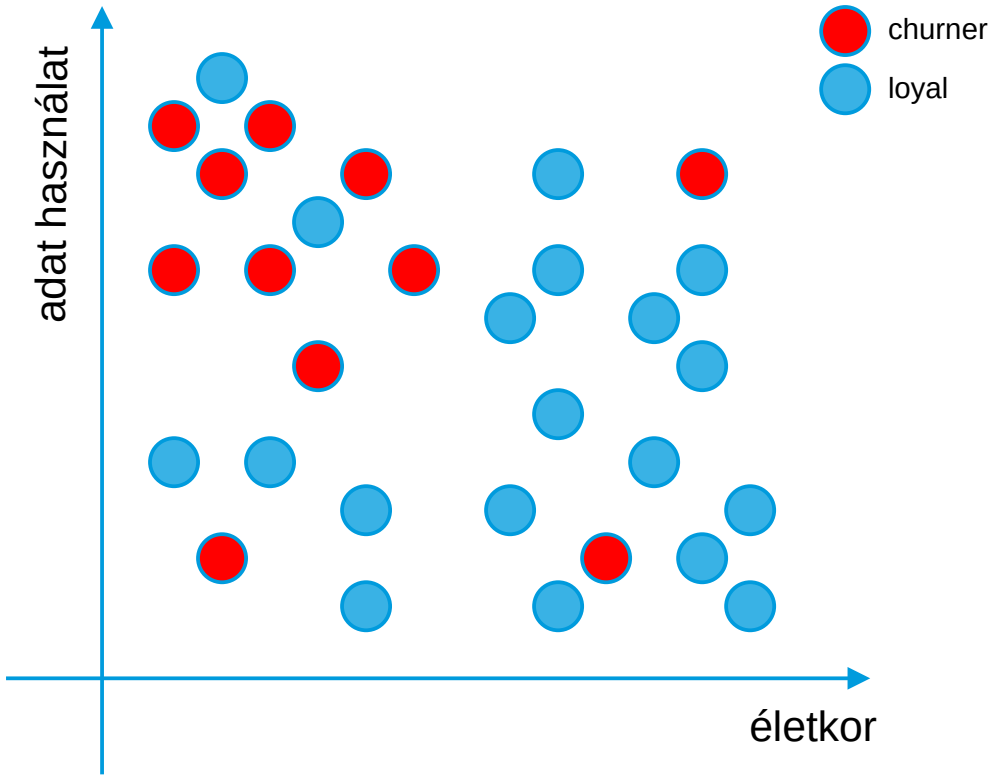
Windhager Eszter

Klasszifikáció

A célváltozó kategorikus (bináris)

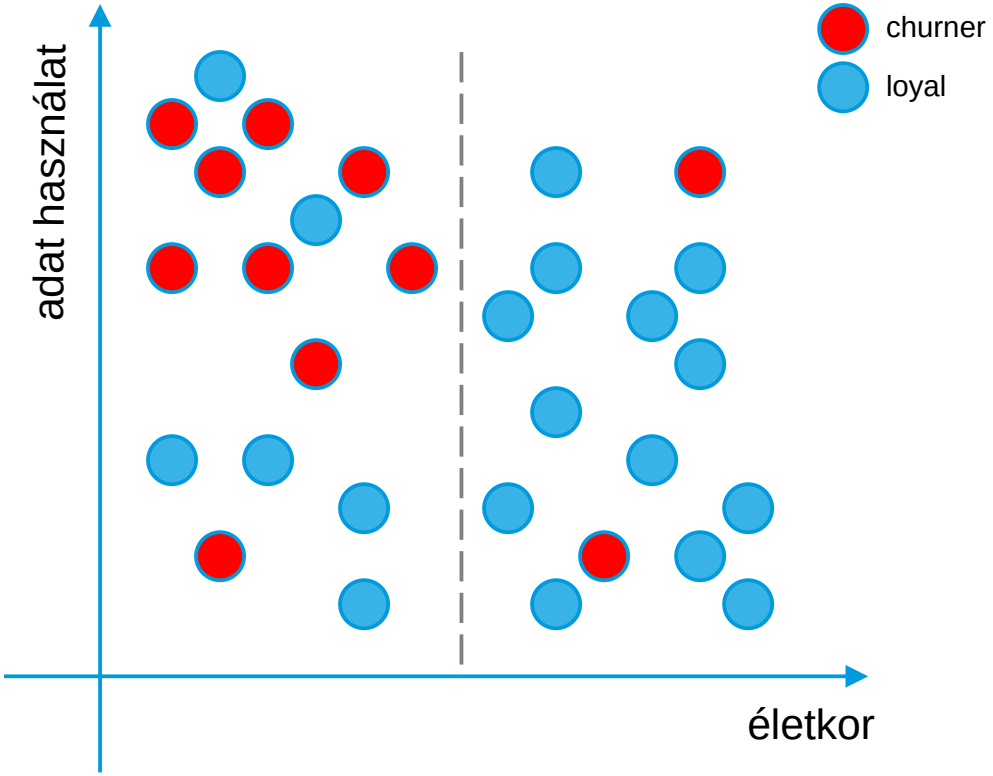
- Hitelbírálat
- Lemorzsolódás előrejelzés

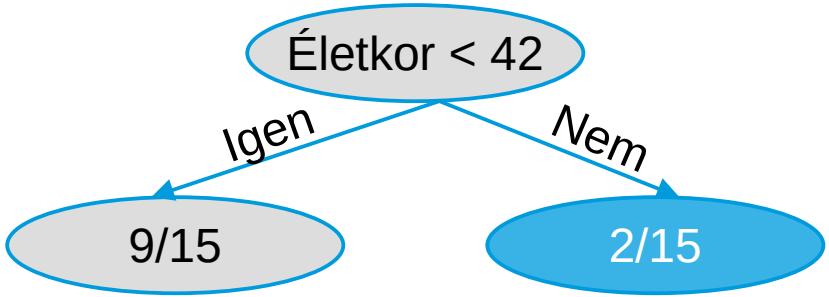
Példa



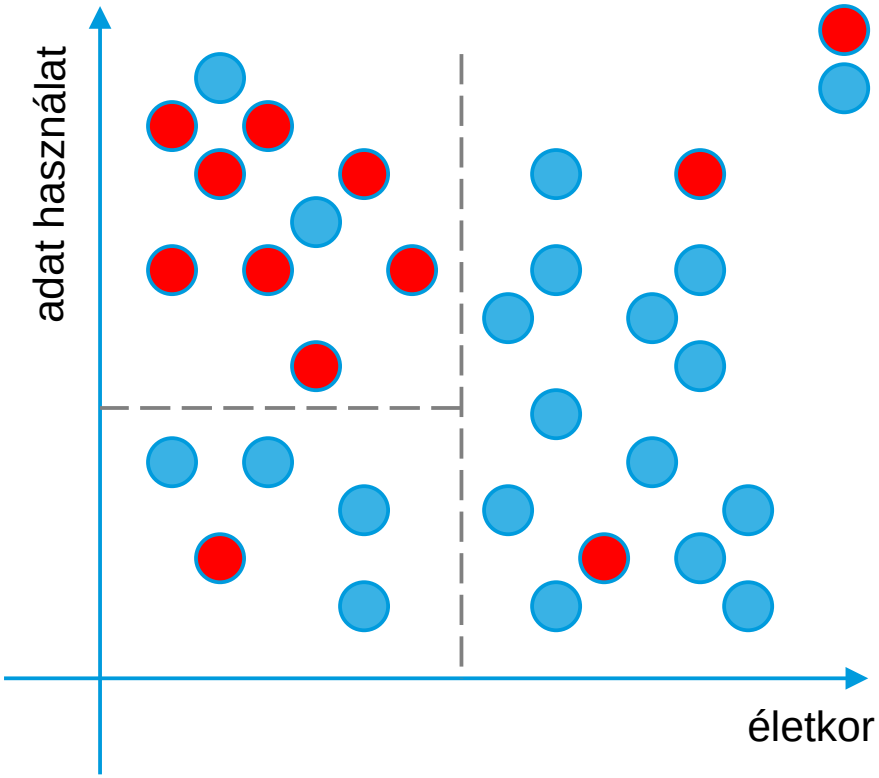
- Telco ügyfelek
- Lemorzsolódás előrejelzés
- Input:
  - Életkor
  - Adat használat az elmúlt hónapban

Decision tree

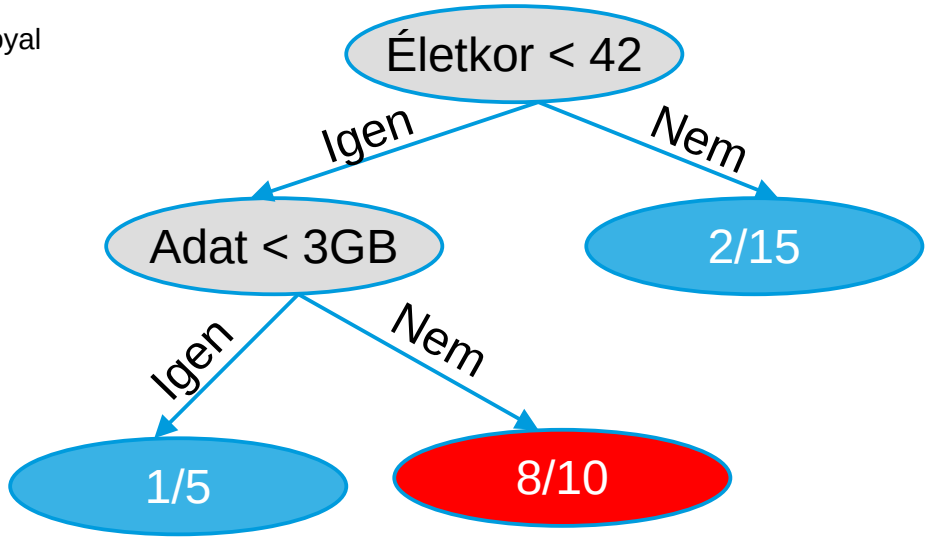




Decision tree



churner
loyal



Működés

- Megkeresi a legjobb vágást
- Homogén részekre darabolja az adatot (Gini, Entrópia)
- Ismétli a fentieket, de ha elér egy limitet, leáll

Gini

$$1 - \sum_j p_j^2$$

Entrópia

$$-\sum_j p_j \log_2 p_j$$

1. Split kiválasztás

- Teljes adat: 11 churner, 19 loyal
- Gini teljes adat: 1-(11/30^2 + 19/30^2) = 0.464
- 1. split: 9 churner, 6 loyal illetve 2 churner, 13 loyal

  Gini impurity:
- (w1*Gini(node1)+w2*Gini(node2)) =

  (0.5*0.48 + 0.5*0.23) = 0.355
- Minden lehetséges vágásra kiszámoljuk, a legkisebbet választjuk

Előnyök - hátrányok

- Interprtálható (kis fa)
- Nincs megkötés a változók típusára*, eloszlására – kevesebb adatelőkészítés
- Változó szelekció
- Balancing kevésbé probléma
- Nem lináris összefüggések
- Feature importance

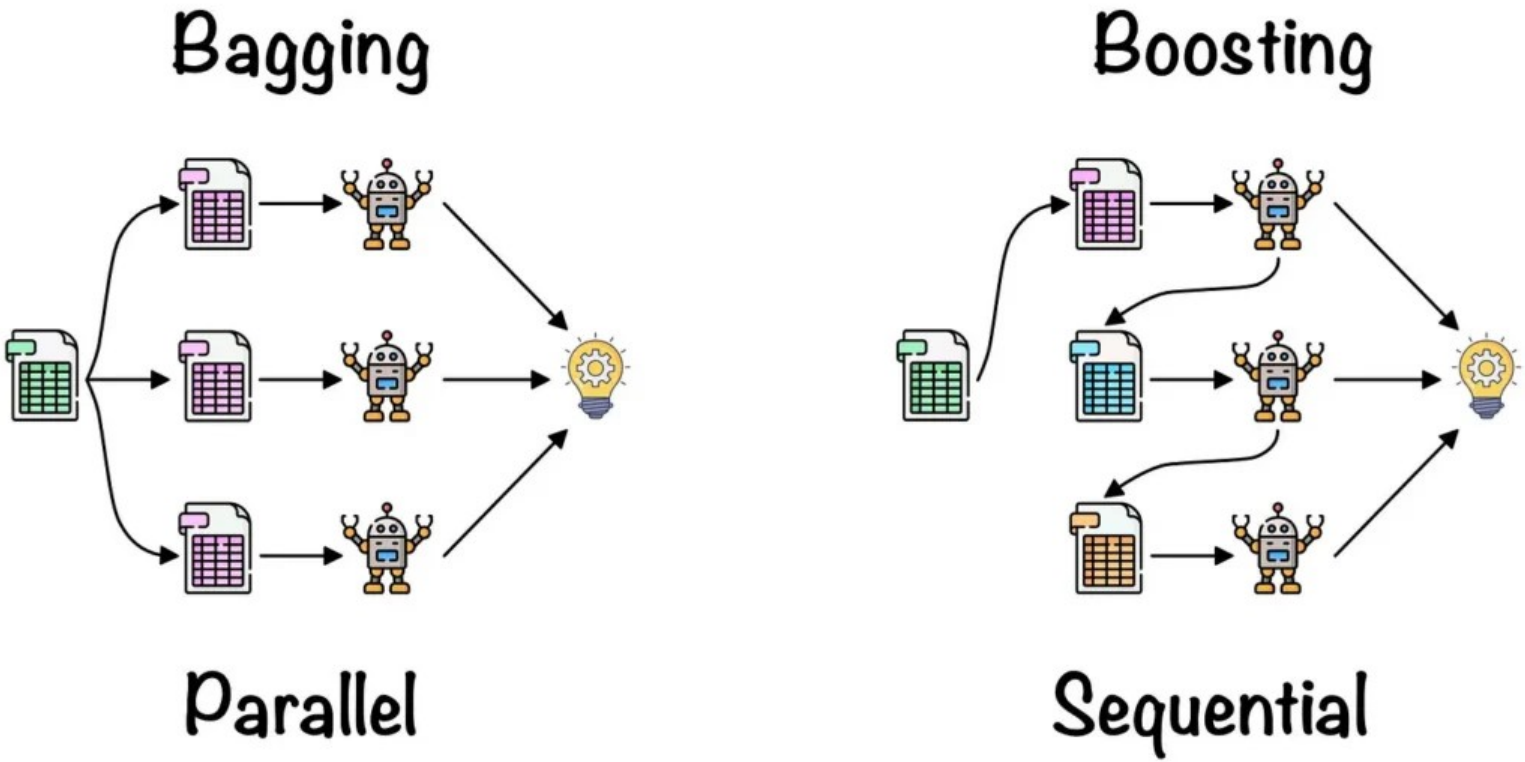
- Túltanulás
- Variancia (kis változás az adatban – teljesen más fa)

* az sklearn implementáció csak numerikus input változókat tud használni

Mire használható?

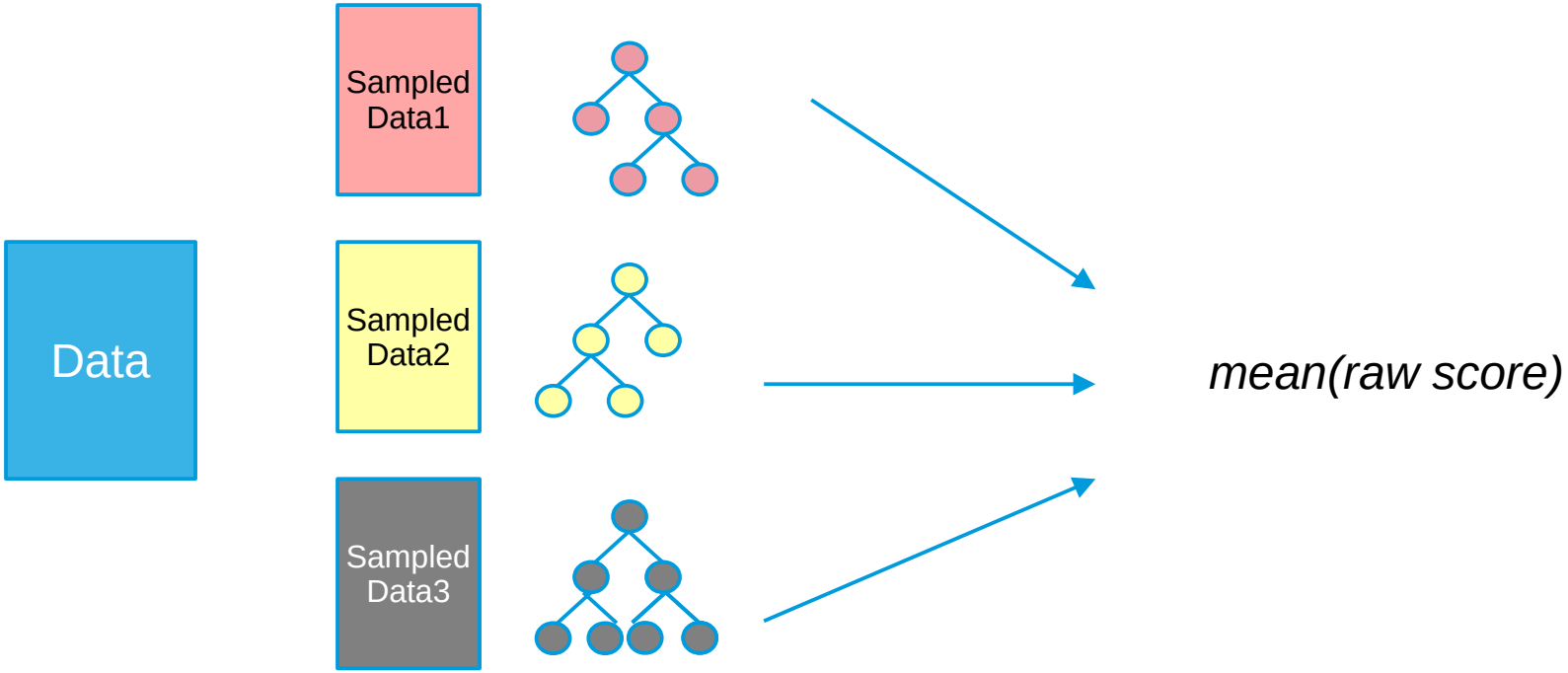
- Klasszifikáció – előrejelzésre vannak jobb módszerek
- Interpretáció – adatokban rejlő mintázatok feltárása, leíró elemzések

Bagging és Boosting



Kép forrása: https://towardsdatascience.com/ensemble-learning-bagging-boosting-3098079e5422/

Random forest



Előnyök - hátrányok

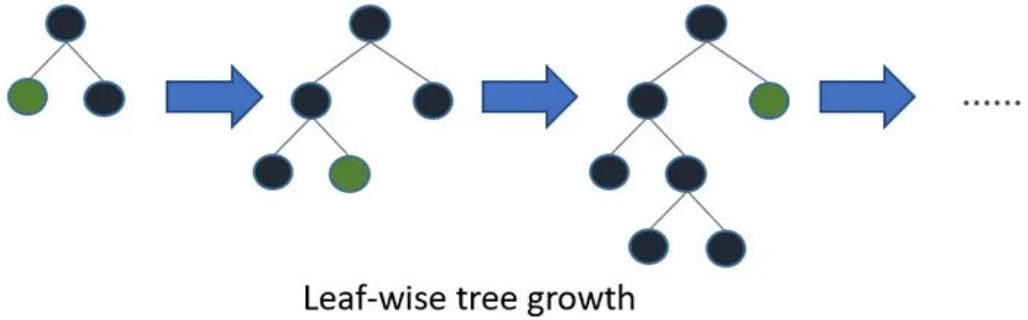
- Könnyen paraméterezhető
- Túltanulás csökkenthető
- Robosztus (zajos változók kevésbé befolyásolják)
- Feature importance megbízhatóbb
- Korábbi modell továbbtanítható
- Out-of-bag accuracy estimate

- Kevésbé interpretálható
- Lassú futás
- Nagyobb memória igény

Mire használható

- Jó baseline modell
- Feature selection eredmény használható más modellhez is

Light GBM



Kép forrása: https://lightgbm.readthedocs.io/en/latest/Features.html

LightGBM Előnyök - hátrányok

- Category input (lightbm package)
- Gyors
- Pontosabb előrejelzés

- Sok paraméter
- Kevés adat esetén túltanulás

Összefoglalás

Decision tree

- Interprtálható
- Összefüggések feltárása

Random forest

- Jó baseline
- Robosztus
- Könnyű paraméterezés
- Továbbtanítható

Light GBM

- Gyors
- Várhatóan a legpontosabb
- Category vátozók kezelése